Lecture Notes in Electrical Engineering 113

Yuhang Yang
Maode Ma *Editors*

Green Communications and Networks

Proceedings of the International Conference on Green Communications and Networks (GCN 2011)

Part 1

Springer

Lecture Notes in Electrical Engineering

Volume 113

For further volumes:
http://www.springer.com/series/7818

Yuhang Yang · Maode Ma
Editors

Green Communications and Networks

Proceedings of the International Conference on Green Communications and Networks (GCN 2011)

Springer

Yuhang Yang
Department of Electronic Engineering
Shanghai Jiao Tong University
Dongchuan Road 800
Shanghai 200240
People's Republic of China
e-mail: yhyangsjtu@gmail.com

Maode Ma
Electrical and Electronic Engineering
Nanyang Technological University
Nanyang Avenue
Singapore 639798
Singapore
e-mail: emdma@ntu.edu.sg

ISSN 1876-1100
ISBN 978-94-007-2168-5
DOI 10.1007/978-94-007-2169-2
Springer Dordrecht Heidelberg London New York

e-ISSN 1876-1119
e-ISBN 978-94-007-2169-2

Library of Congress Control Number: 2011940205

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Welcome to the proceedings of the International Conference on Green Communications and Networks (GCN 2011), which was held in Chongqing, China, July 15–17, 2011.

GCN 2011 will be a venue for leading academic and industrial researchers to exchange their views, ideas and research results on innovative technologies and sustainable solutions leading to greener communications and networks. The conference will feature keynote speakers, a panel discussion and paper presentations.

The objective of GCN 2011 is to facilitate an exchange of information on best practices for the latest research advances in the area of green communications and networks, which mainly include the intelligent control, or efficient management, or optimal design of access network infrastructures, home networks, terminal equipment, and etc. GCN 2011 will provide a forum for engineers and scientists in academia, industry, and government to address the most innovative research and development including technical challenges, social and economic issues, and to present and discuss their ideas, results, work in progress and experience on all aspects of advanced green communications and networks engineering.

The GCN 2011 conference provided a forum for engineers and scientists in academia, industry, and government to address the most innovative research and development including technical challenges and social, legal, political, and economic issues, and to present and discuss their ideas, results, work in progress and experience on all aspects of information computing and applications.

There were a very large number of paper submissions (956), representing 21 countries and regions, not only from Asia and the Pacific, but also from Europe, and North and South America. All submissions were reviewed by at least three Program or Technical Committee members or external reviewers. It was extremely difficult to select the presentations for the conference because there were so many excellent and interesting submissions. In order to allocate as many papers as possible and keep the high quality of the conference, we finally decidedto accept 190 papers for presentations, reflecting a 19.9% acceptance rate. We believe that all of these papers and topics not only provide novel ideas, new results, work in progress and state-of-the-art techniques in this field, but also stimulated the future research activities in the area of information computing and applications.

The exciting program for this conference was the result of the hard and excellent work of many others, such as Program and Technical Committee members, external reviewers and Publication Chairs under a very tight schedule. We are also grateful to the members of the Local Organizing Committee for supporting us in handling so many organizational tasks, and to the keynote speakers for accepting to come to the conference with enthusiasm. Last but not least, we hope you will enjoy the conference program, and the beautiful attractions of Chongqing, China.

China, June 2011

Maode Ma
Yuhang Yang

Contents

Part I Communication Systems

Part II Data Management and Database System

Part III Digital Image Processing

Part IV Education and Informatics

Part V Enabling Technologies

Part VI Forensics, Recognition Technologies and Applications

Part VII Fuzzy System and Control

Part VIII Green Computing

Part IX Graphics and Visualizing

Part X Internet Growth Modelling and Virtualized Networks

Part XI Network Components and Application

Part XII Network Design Methodolog

Part XIII Semantic Web and Intelligent Search

Part XIV Social and Economical Systems

Part XV Web Service

Part XVI Wireless Network

Organization

GCN 2011 was organized by Chongqing Normal University, BeiHang University , Peking University , and sponsored by the National Science Foundation of China, Shanghai Jiao Tong University, Nanyang Technological University. It was held in cooperation with Lecture Notes in Electrical Engineering (LNEE) of Springer.

Executive Committee

General Chairs	Maode Ma Nanyang Technological University, Singapore Yuhang Yang, Shanghai Jiao Tong University, China
Program Chairs	Yuhang Yang, Shanghai Jiao Tong University, China Qi Jing,Peking University, China Hongsong Chen, University of Science and Technology Beijing, China
Local Arrangement Chairs	Xilong Qu, Hunan Institute of Engineering, China Pan Deng, BeiHang University, China Wenjiang Du, Chongqing Normal University, China
Steering Committee	Qun Lin, Chinese Academy of Sciences, China Maode Ma, Nanyang Technological University, Singapore Nadia Nedjah, State University of Rio de Janeiro, Brazil Lorna Uden, Staffordshire University, UK Yiming Chen, Yanshan University, China Aimin Yang, Hebei united University, China Chunying Zhang, Hebei United University, China

	Dechang Chen, Uniformed Services University of the Health Sciences, USA Mei-Ching Chen, Tatung University, Taiwan Rong-Chang Chen, National Taichung Institite of Technology, Taiwan Chi-Cheng Cheng, National Sun Yat-Sen University, Taiwan Donald C. Wunsch, University of Missouri Rolla, USA
Publicity Chairs	Aimin Yang, Hebei United University, China Xilong Qu, Hunan Institute of Engineering, China
Publication Chairs	Yuhang Yang, Shanghai Jiao Tong University, China
Financial Chair	Wenjiang Du, Chongqing Normal University, China
Local Arrangement Committee	Defang Luo, Chongqing Normal University, China Linyan Chen, Chongqing Normal University, China Pan Deng, BeiHang University, China Yuhuan Cui, Hebei Polytechnic University, China
Secretaries	DuWu Cui, Xian University of Technology, China Jinai Qu, Defense Security Command, Korea Yaang Yang, Shanghai University, China Lichao Feng, Defense Security Command, Korea

Program/Technical Committee

Yuan Lin	Norwegian University of Science and Technology, Norwegian
Yajun Li	Shanghai Jiao Tong University, China
Yanliang Jin	Shanghai University, China
Mingyi Gao	National Institute of AIST, Japan
Yajun Guo	Huazhong Normal University, China
Haibing Yin	Peking University, China
Jianxin Chen	University of Vigo, Spain
Miche Rossi	University of Padova, Italy
Ven Prasad	Delft University of Technology, Netherlands
Mina Gui	Texas State University, USA
Nils Asc	University of Bonn, Germany
Ragip Kur	Nokia Research, USA

On Altintas	Toyota InfoTechnology Center, Japan
Suresh Subra	George Washington University, USA
Xiyin Wang	Hebei Polytechnic University, China
Dianxuan Gong	Hebei Polytechnic University, China
Chunxiao Yu	Yanshan University, China
Yanbin Sun	Beijing University of Posts and Telecommunications, China
Guofu Gui	CMC Corporation, China
Haiyong Bao	NTT Co., Ltd., Japan
Xiwen Hu	Wuhan University of Technology, China
Mengze Liao	Cisco China R&D Center, China
Yangwen Zou	Apple China Co., Ltd., China
Liang Zhou	ENSTA-ParisTech, France
Zhanguo Wei	Beijing Forestry University, China
Hao Chen	Hu'nan University, China
Lilei Wang	Beijing University of Posts and Telecommunications, China
Xilong Qu	Hunan Institute of Engineering, China
Duolin Liu	ShenYang Ligong University, China
Xiaozhu Liu	Wuhan University, China
Yanbing Sun	Beijing University of Posts and Telecommunications, China
Yiming Chen	Yanshan University, China
Hui Wang	University of Evry in France, France
Shuang Cong	University of Science and technology of China, China
Haining Wang	College of William and Marry, USA
Zengqiang Chen	Nankai University, China
Dumisa Wellington Ngwenya	Illinois State University, USA
Hu Changhua	Xi'an Research Insti. of Hi-Tech, China
Juntao Fei	Hohai University, China
Zhao-Hui Jiang	Hiroshima Institute of Technology, Japan
Michael Watts	Lincoln University, New Zealand
Tai-hon Kim	Defense Security Command, Korea
Muhammad Khan	Southwest Jiaotong University, China
Seong Kong	The University of Tennessee USA
Worap Kreesuradej	King Mongkuts Institute of Technology Ladkrabang, Thailand
Uwe Kuger	Queen's University of Belfast, UK
Xiao Li	CINVESTAV-IPN, Mexico
Stefa Lindstaedt	Division Manager Knowledge Management, Austria

Paolo Li	Polytechnic of Bari, Italy
Tashi Kuremoto	Yamaguchi University, Japan
Chun Lee	Howon University, Korea
Zheng Liu	Nagasaki Institute of Applied Science, Japan
Michiharu Kurume	National College of Technology, Japan
Sean McLoo	National University of Ireland, Ireland
R. McMenemy	Queens University Belfast, UK
Xiang Mei	The University of Leeds, UK
Cheol Moon	Gwangju University, Korea
Veli Mumcu	Technical University of Yildiz, Turkey
Nin Pang	Auckland University of Technology, New Zealand
Jian-Xin Peng	Queens University of Belfast, UK
Lui Piroddi	Technical University of Milan, Italy
Girij Prasad	University of Ulster, UK
Cent Leung	Victoria University of Technology, Australia
Jams Li	University of Birmingham, UK
Liang Li	University of Sheffield, UK
Hai Qi	University of Tennessee, USA
Wi Richert	University of Paderborn, Germany
Meh shaffiei	Dalhousie University, Canada
Sa Sharma	University of Plymouth, UK
Dong Yue	Huazhong University of Science and Technology, China
YongSheng Ding	Donghua University, China
Yuezhi Zhou	Tsinghua University, China
Yongning Tang	Illinois State University, USA
Jun Cai	University of Manitoba, Canada
Sunil Maharaj Sentech	University of Pretoria, South Africa
Mei Yu	Simula Research Laboratory, Norway
Gui-Rong Xue	Shanghai Jiao Tong University, China
Zhichun Li	Northwestern University, China
Lisong Xu	University of Nebraska-Lincoln, USA
Wang Bin	Chinese Academy of Sciences, China
Yan Zhang	Simula Research Laboratory and University of Oslo, Norway
Ruichun Tang	Ocean University of China, China
Wenbin Jiang	Huazhong University of Science and Technology, China
Xingang Zhang	Nanyang Normal University, China
Qishi Wu	University of Memphis, USA
Jalel Ben-Othman	University of Versailles, France

Part I
Communication Systems

Chapter 1
Model of Cyber-Physical Systems for Underground Coal Mine

Yanjing Sun, Man Yu, Yanjun He and Xiaohui Ding

Abstract Aiming at problems in work safety and accident rescue occurred in coal mine, cyber-physical systems (CPS) is introduced into underground coal mine. According to environment parameters and considering special applications, a model of CPS for coal mine, which has the ability of sensing surroundings, auto warning, intelligent control and personnel location, is constructed. Based on relative theoretical researches, three key technologies for CPS under complicated environment are presented, which are information perceiving, computing and collaborative control. The implementation of mine-oriented CPS could improve the safety in production and accelerate the development of coal mine.

Keywords Coal mine · Cyber-physical systems · Information acquisition · Collaborative control

Y. Sun (✉) · M. Yu · Y. He · X. Ding
School of Information and Electrical Engineer,
China University of Mining and Technology,
Xuzhou 221116, China
e-mail: yanjingsun_cn@163.com

M. Yu
e-mail: ymcumt@163.com

Y. He
e-mail: yjhcumt@163.com

X. Ding
e-mail: xhding@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_1,

1.1 Introduction

In recent years, the industrial structure of coal mine has been optimized vigorously to enhance safety assurance in China. Unfortunately, the overall level of coal industry and mine safety is not high enough, so that the amount of accidents is still great and grave accidents also happen from time to time, all what have seriously threatened the miners' life safety. Therefore, it is imperative to design and establish an intelligent system, which is targeted for disaster prevention and accident rescue, with advanced technologies such as monitoring, data communication, information processing and automatic control.

Cyber-physical systems (CPS) have the ability of communication and computation, which integrate computation and physical processes reliably and efficiently in runtime by monitoring or controlling the physical world [1]. Communication, computation and control are the cores of CPS, making such engineering systems highly integrated and controllable. So, CPS achieve the inter-infiltration of virtual (cyber) world and real (physical) world, what compared with traditional embedded systems and wireless sensor networks (WSN) are more intelligent and complicated. In [2], it is reported that CPS involve ubiquitous environment perception, embedded computing, network communication and networked control in the future, all that enable physical system possess five functions of computation, communication, accuracy control, remote collaboration and autonomy management.

While, CPS for coal mine is created in mines underground, taking surroundings as plants to monitor and control through wireless sensors and embedded devices, by means of distributed information processing, data fusion and cooperative control, etc. On one hand, this can obtain a lot of accurate and reliable information at any time and any place, realizing a concept of computing anywhere indeed. On the other hand, it makes computation processes interact with physical environments precisely, dependably and rapidly. Above all, it can be anticipated that setting up a CPS for coal mine will not only eliminate potential safety hazards efficiently and reduce accident rate, but also provide guidance for accident rescue scientifically and decrease casualties.

1.2 Background

The influence of CPS in economy and society is more than recognized before. Developed countries leading with the US and EU have been aware of this potential and devoted great resources into related fields research. In 2006, NSF took CPS as a key point of scientific research and made great efforts to propel its development and application. The EU invested 5.4 billion EUR in Artemis in order to become the world leader of intelligent electronic till 2016.

In China, it is pointed out in The National Basic Research Program (973 Program) that the program is seeking proposals that address research challenges in CPS themes of methodologies and models. Furthermore, the CPS program has been put forward from gestation to practice study.

Currently, three main grand challenge problem areas were developed more fully: distributed energy systems, transportation systems and healthcare systems [3–7]. Professor Edward A. Lee and his group in UC Berkeley presented a programing model based on time-centric named PTIDES (Programming Temporally Integrated Distributed Embedded Systems) [8], which focus on the correlations of computers on a network and how they interact with and through physical processes via sensors and actuators. Besides, University of Pennsylvania has carried out research projects on transportation systems and medical devices aiming at addressing real-time and concurrent computation problem. In China, T. John Koo commits to hybrid systems and control systems related to CPS. While, researchers in Dalian University of Technology mainly focus on QoS and network optimization of CPS combining WSN.

To our knowledge there has not been a study of CPS directing at coal mine underground. There are two reasons for this: (1) CPS is a promising approach but still in its infancy, and (2) the working environment of coal mine is so particular that people rarely pay attention to this field. In our view, it is feasible and credible to design and set up a CPS for coal mine. In addition, we can benefit from other application domain on common problems which have been solved and consider parameter characteristic of mines to construct a high efficient cooperative control system for coal mine.

1.3 Model of CPS for Coal Mine

Compared with traditional embedded systems, standard CPS are designed to be networked, where the components cooperate with each other but not independently [9]. As information science and engineering technology are developing, the relationship between computation and physical process will be improved, what is more, the adaptation, independence, function, reliability, security and availability of CPS will be promoted greatly.

Demands of coal mine are illustrated in Fig. 1.1, including environment monitoring, auto warning and control, rescue guidance, transmission of wireless multimedia, instrument reading, locomotive localization and navigation together with personnel location and tracking, we put forward a model of CPS for coal mine and analyze information perceiving and collaborative control technologies in complicated environment.

Model-based design is the main method to develop CPS. The model of CPS for coal mine can be divided into a lot of models that have been built by computing science, but also need to establish some new models of subsystems to abstract time, location, energy loss and potential sensing data. These models should define physical systems exactly and then blend computation and physical process well. In consideration of actual applicative requirements, designing such model for coal mine involves control or computation algorithm and abstract of feedback loop. Moreover, the model will be able to estimate and deal with boundary condition, so as to operate on specified controller.

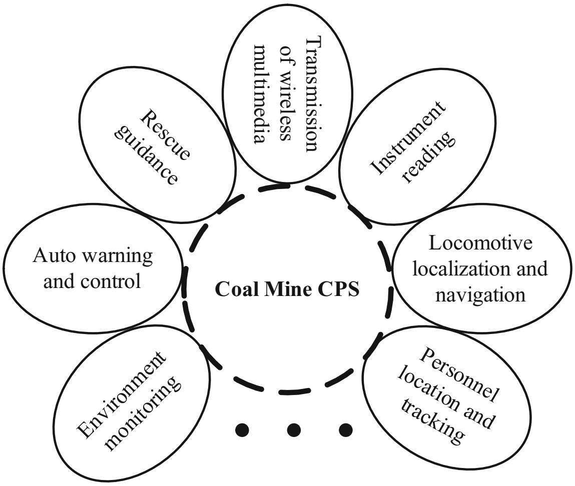

Transmission of wireless multimedia
Rescue guidance
Instrument reading
Auto warning and control
Coal Mine CPS
Locomotive localization and navigation
Environment monitoring
Personnel location and tracking

Fig. 1.1 Functional demands of CPS for coal mine

Most of time, CPS are made up of series of network agents, like sensors, actuators, controllers together with communication devices. Between cyber system and physical system, there is a feedback loop to keep the whole system stable and reliable. Data from sensors could be sent to controller directly and then controllers give commands and instructions to control or influence physical objects. In this way, the feedback loop capable of self-organizing transmission and intelligent control is formed.

Considering the work environment of coal mine and application requirements, a model of CPS for coal mine is proposed as shown in Fig. 1.2.

In fact, CPS is a hybrid system which consist of a large amount of components that differ from each other on location and coverage. Theoretically, exchange and feedback may occur in both cyber and physical world, when in physical world, the feedback control model will change all plants' states. To hybrid systems for now, discrete-time models and continuous-time models do not work together, and so do event-driven models and time-driven models, while which kind of models to choose for designing CPS is significant. This is why the hybrid model suitable for CPS needs to be developed.

In coal mine CPS, modeling should be carried on both cyber and physical resources because states of plants including voltage and distance correspond to elements in cyber field such as computer memory, CPU and communications capability. Factors of sensing, wireless communication, noisy and mobility always influence the abstraction of physical system and result in uncertainty. Other issues, for example, error and packet loss also should be taken into account. The solution satisfying robustness and real-time is significant because of the random, uncertainty and delay attribute of physical process. Since the physical world is not prescient, what leads CPS may run out of control, so it brings out the requirement that CPS have to be robust enough to deal with unexpected situations and be tolerant of fault in subsystems.

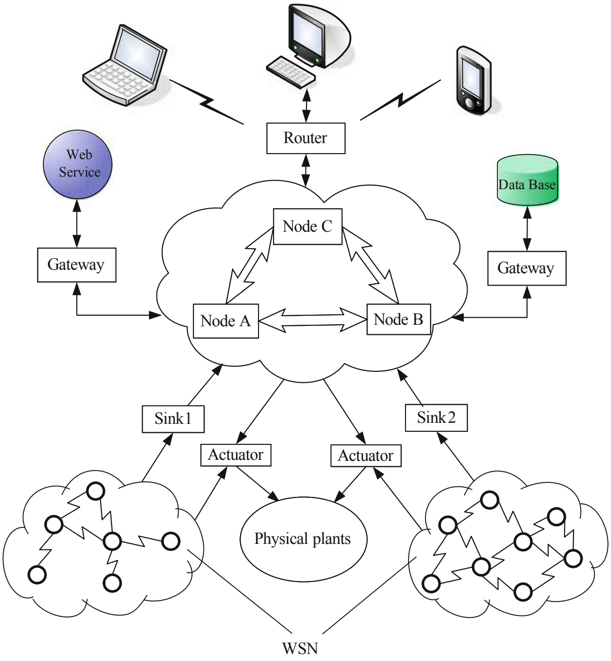


Fig. 1.2 A Model of CPS for coal mine

1.4 Key Technologies

Such domains as embedded system, hybrid system, dynamical system, distributed system, wireless network, microcontroller, sensor and actuator are to be based via control theory and stochastic process to implement CPS. However, present distributed control, sensing and computing technology yet cannot be used to real-time and secure CPS. The main reason is that they are not able to collect enough information and take precise measures in the distributed environment. Consequently, it is not appropriate to apply traditional distributed model, sensing model and communication model to time accurate and high safety CPS, most important, the information perceiving, computing and collaborative control technology are to be addressed.

1.4.1 Information Perceiving

1.4.1.1 The Shortage of Traditional Models

Traditional sensing, computing, decision-making and actuating models are based on low-potential environment, where data is the latest and actuators also act immediately. On the contrary, CPS are networked so that potential factors can initiate decision process uncertainty and lead to delay actuation that result in serious problems of system.

1.4.1.2 Solutions

Firstly, the coordination pattern of reality and virtual should be addressed to fix on the mapping relationship between physical process and cyber process. Secondly, we need to describe sensors' attribute in detail into perception approach, including acquisition standard, context information and some other uncertain factors. After that, the information have been gathered is fed back to the physical layer, in this way, we can regulate sampling rate or change state of sensors to increase the amount of information. Apart from this, there is a request to build new models for the whole processing sectors which transform physical entity to data and then to credible information.

1.4.1.3 Technology Challenge

Compared to static data, the acquisition of real-time data flow is different not only on throughput but also in essence. In order to scan and look up data fleetly, we need to save and dispose them preferably. In some case, information collected may be alike to sample data, while others require to be abstracted. At the same time, perceiving information and historical data memory area should be reserved to recover and search important information. Beyond that, the adaptability of acquisition techniques nowadays also has some defects because they face certain but not all layers of system. If problems raised above not be solved well, the opening feature of CPS for coal mine will be limited.

1.4.2 Computing

1.4.2.1 Semantic Integration

Sensing devices together with embedded control software make up a coal mine CPS. This kind of complex system depends largely on fault tolerance, security and distributed control, therefore it raises a higher requirement for computing power. For that reason, CPS needs new abstract model and computation logic. Semantic integration is the process of interrelating information from diverse sources, both on cyber and physical side. A major way is that redefine computation method and instruction to map actions of physical plant to computing processes through detailed attribute of observed object. By semantic integration, operation on massive data and command execution can be prompt and veracious to ensure real-time and trustworthy system. Crucially, semantic integration can solve the problem of modeling software or hardware elements in isolation, without ignoring important interactions between them and thus risks are restricted and avoided to a certain extent.

1.4.2.2 Software

Usually, it is the software that affects system complexity and even cost. Multi-core and multi-thread have tremendously improved the performance and capability of servers, leading the convergence of computing progresses to the server. And as a result, middleware which can provide usability and high capacity of parallel computing, together with OS and DBMS should be developed further. On the other side, owing to networked property of CPS, we can develop production software tools with only an appropriate framework, so open source is inevitable. Meanwhile, cloud computing makes server cluster possess more powerful computational ability at server side, yet it has a good application potential in CPS. Apart from this, the development of software also needs to improve validation and certification to assure functional and interconnected embedded system and prevent system from unpredictable influence of external factors at runtime.

1.4.2.3 Hardware

Cooperation among embedded devices is implemented through high-speed data flow in CPS. When they communicate via wireless connection, not only the energy of individual nodes will consume fast, but also wide bandwidth is needed. To this point, current infrastructure for CPS is deficient. The design and development of embedded systems will be cared about more on its unpredictability and robustness, but not high efficiency ever. To adapt the development of CPS, embedded systems need to adopt multi-core architecture and expand various kinds of network communication interfaces. Besides, with formidable hardware development kit and software package, resource-constraint devices could accomplish complicated business function reliably.

1.4.3 Collaborative Control

1.4.3.1 Theoretical Approaches

The development of CPS needs new theory to support, such as information integration, cascading failures and self-adaptation, because traditional uncertain decision-making model is not fit for control loop of CPS. The research of collaborative control refers to multi-layer collaboration, proprietary/share backup path design, service fast recovery, fault-tolerant management and control of active networks, renewable networks model, survivability of any cast and multicast, self-recovery technology, all what can make CPS available and dependable.

1.4.3.2 Modeling Analysis

Generally, discrete mathematics is used to describe computing model, instead, differential equation and description of system behaviors are for cybernetic model. Thus, how to combine scatter and continuation together is a key point when building models for complicated system. From another point of view, a lot of control theories are event-driven strategy, while most process function of computing system are asynchronous. To these various kinds of models, we need to fuse them respectively, otherwise, the physical devices may not able to compute and communicate.

1.4.3.3 Technical Realization

Abstraction method can simplify information by removing redundant data so as to improve the capability of message processing, and traditional computing abstraction is based on data transfer structure and never interrelated to ever-changing characteristics of physical object. Nonetheless, it cannot be applied straight for CPS on account of time accuracy between cyber and physical processes. Therefore, design of continuous dynamic feedback loop and management of real-time control loop involve mode conversion, fault detection and real-time interaction to put into practice. Meanwhile, factors, for example, stability, transient response and parameter variation of system, remain to be taken into consideration.

1.5 Conclusion

Safety in coal mine is very important which concerns people’s life safety and state property. For the sake of decreasing accident rate and satisfying the demand of emergent accident rescue to ensure production safety, we suggest applying CPS in coal mine underground, establishing a Coal Mine CPS capable of hierarchical control. On the basic of introducing the concept of CPS, we indicate three main application fields and summarize researches of CPS at home and abroad. Subsequently, a model of CPS for underground coal mine is proposed, and problems which should be paid attention to while building the model are also pointed out. At last, we particularly analyze information perceiving and collaborative control as key technique using coal mine CPS. This paper will enrich the models of CPS. Furthermore, it can play a certain role in the development of next generation network technology and speedup information construction of coal mine.

Acknowledgments Financial support for this work, sponsored by Qing Lan Project, National Science Foundation of China under grant No.50904070, the China Postdoctoral Science Foundation No. 20100471412, the Key Project of the National High Technology Research and Development Program of China (863 Program) under grant No. 2008AA062200, the National High Technology Research and Development Program of China (863 Program) under grant No. 2007AA01Z180, the Key Project of Jiangsu under grant No. BG2007012, and the Science Foundation of China University of Mining and Technology under Grant No. OC080303, are gratefully acknowledged.

References

1. Lee E (2006) Cyber-physical systems: are computing foundations adequate? Position paper, NSF workshop on Cyber-Physical systems: research motivation techniques and roadmap, London
2. He J (2010) Cyber-physical systems. In: Communications of the CCF Beijing, 47:25–29
3. Morris T, Srivastava A, Reaves B (2009) Engineering future Cyber-physical energy systems: challenges, research needs and roadmap. In: 41st North American power symposium
4. Work D, Bayen A (2008) Impacts of the mobile internet on transportation cyberphysical systems: traffic monitoring using smartphones. National workshop for research on high-confidence transportation cyber-physical systems, Washington, pp 18–20
5. Tang H, McMillin B (2008) Analysis of the security of information flow in the advanced electric power grid using flexible alternating current transmission system (FACTS). In: 1st Annual IFIP international conference on critical infrastructure protection, Springer, pp 43–56
6. Work D, Bayen A, Jacobson Q (2008) Automotive cyber physical systems in the context of human mobility. National workshop on high-confidence automotive cyber-physical systems. Troy, MI
7. David A, Fischmeister S, Goldman J, Lee I, Robert T (2009) Plug-and-play for medical devices experiences from a case study. Biomed Instrum Technol 43:313–317
8. Lee A, Matic S, Seshia S, Zou J (2009) The case for timing-centric distributed software. 29th IEEE international conference on distributed computing systems, Canada, pp 57–64
9. Lee A (2008) Cyber physical systems: design challenges. In: 11th IEEE international symposium on object/component/service-oriented real-time distributed computing (ISORC ‘08), pp 363–369

Chapter 2
Design Signal Detection Project of MIMO Communication Systems Based on Improved Grover Algorithm

Lu Xin-Bo

Abstract The signal detection scheme of multiple inputs multiple output (MIMO) communication system is introduced, and the existing problems of Grover algorithm is analyzed and improved. The signal detection scheme of MIMO system based on Grover algorithm and improved Grover algorithm is developed, and Grover algorithm is applied to find the minimum value in order to decide the sending sequence. In order to test the efficiency and reliability of this algorithm, the analysis and comparison between Grover algorithm, improved Grover algorithm and other traditional algorithms are made by MATLAB simulation. Experimental results show that it can reduce the complexity, while achieving the same performance of the traditional optimum detection algorithms.

Keywords Grover quantum search algorithm · Quantum parallel computation · MIMO detection algorithm · Quantum register

Multiple-input multiple-output (MIMO) wireless communication systems refer to transmitters and receivers that are equipped with multiple antenna units for data transferring process in the wireless communication system. Compared with single-input single-output (SISO) wireless communication systems, additional degrees of freedom (space resource) can be created in MIMO system without increasing neither the bandwidth nor transmit power, which can be exploited for significant improvement of system capacity and enhancement of transmission reliability. As the key technology of the fourth generation mobile communication system (4G), MIMO signal detection has been extensively studied and put forward many signal detection algorithm, such as zero-forcing (ZF) algorithm and minimum

L. Xin-Bo (✉)
Hebei University of Science and Technology,
Institute of Information Science and Engineering,
Shijiazhuang 050026, Hebei Province, China
e-mail: newave616@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_2,

mean square error (MMSE) algorithm. However, these algorithms generally are based on flat fading MIMO channel. For frequency selective broadband MIMO communication system, its receiving signals exist serious inter-symbol interference (ISI). MIMO detection algorithm cannot be overcome. We must use that can confront the frequency selective decline of technology. It can not only overcome the frequency channel choice function but also make further improvement on the frequency utilization that OFDM technology is introduced into MIMO. Grover algorithm which is based on the quantum parallel computation, can find the desired value precisely with $O\left(\sqrt{N}\right)$ iterations in a large unsorted database which contains N elements. However, to find the desired value in the database, any classical algorithm would need at least $O(N)$ steps. So quantum algorithm, for the system signal detection, can slash algorithm complexity [1].

In this chapter a kind of signal detection scheme of communication systems based on Grover algorithm is developed. It can not only effectively reduce the algorithm complexity but also reach basically the same performance of classic best receiving algorithm.

2.1 MIMO Communication System

The block diagram of MIMO system signal detection based on spatial multiplexing without coding is shown in Fig. 2.1 [2].

Assume the sender has M root antenna, the receiver has N root antenna. The channel, between the M ($M = 1, 2 \ldots$) sender root antenna and the N ($N = 1, 2\ldots$) receiving root antenna, is the multipath fading channel which obey Rayleigh distribution. OFDM subcarrier number is K. In the sender, input bit data stream converted into M parallel data stream after serial-to-parallel conversion, so as to realize the multiple antennas output. For each path flow, we should first signal mapping then IFFT transform. IFFT transform realize the OFDM modulation function here. In other words, the function is that low-speed of multi-channel parallel data flow are modulated to mutually orthogonal K sub-carrier at the same time. In order to reduce ISI, protect interval that usually use circle prefix form is joined in among symbols after IFFT transformation. At last, the data stream is transmitted after parallel-to-serial conversion.

In the receiver, each antenna received signals that were sent by M transmitting antennas and MIMO channel linear superposition. The system first desterilizes and removes circulation prefix from every data stream then according to receiving antenna FFT transform, from the time domain transform to frequency domain. At last, parallel data flow is delivered to modem after detector processing and get recovery information bit stream by serializer.

To MIMO system which have K sub-carriers, we can assume the receiver completely know channel state information and channel characteristic is changeless in a OFDM symbol duration and cyclic prefix is greater than channel delay

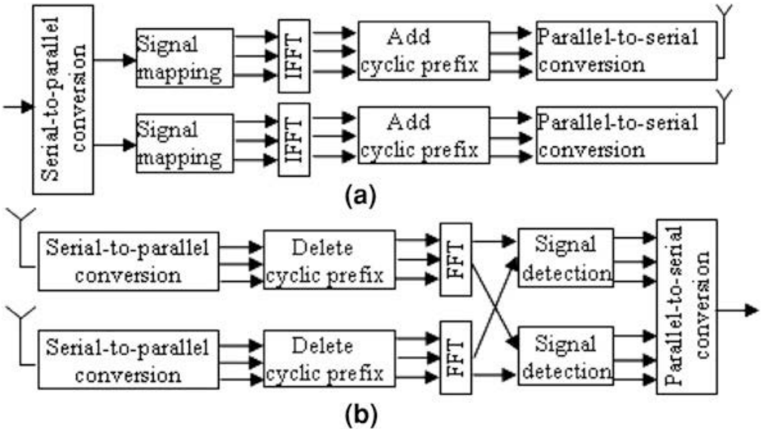

Serial-to-parallel conversion
Signal mapping
IFFT
Add cyclic prefix
Parallel-to-serial conversion
Signal mapping
IFFT
Add cyclic prefix
Parallel-to-serial conversion
(a)
Serial-to-parallel conversion
Delete cyclic prefix
FFT
Signal detection
Serial-to-parallel conversion
Delete cyclic prefix
FFT
Signal detection
Parallel-to-serial conversion
(b)

Fig. 2.1 The block diagram of MIMO system signal detection. **a** The sending end. **b** The receiving end

spread and the system does not exist ISI. Therefore, the multipath channel between any pair antennas can be expressed as k parallel frequency sub channels in a OFDM symbol duration [3].

2.2 Grover Algorithm

Grover algorithm intent is increasing probability amplitude of target quantum states by the unitary transformation of initial equal amplitude superposition state while simultaneously reducing probability amplitude of other off-target quantum states. In the end, the more probability amplitude of target quantum states the bigger probability the right target is searched (namely measured).

Grover algorithm described as follows [4]:

(1) Initialization

$$\left(\overbrace{\frac{1}{\sqrt{N}},\frac{1}{\sqrt{N}},\ldots,\frac{1}{\sqrt{N}}}\right)$$

Hypotheses n are quantum bits, $N = 2^n$. In other words, all basic vectors have initially equal amplitudes. This can be achieved by Walsh–Hadamard transform acts on quantum states $|0000\ldots0\rangle n$. Total need $O(\log N)$ steps.

(2) Grover iteration. Repeat step (a) and (b) $O\left(\sqrt{N}\right)$ times. According to the Ref. [5], best iteration number is $j \approx \left[\pi/4\sqrt{N/m}\right]$, lowercase m letter is the correct solution number, [] show integer.

(a) Selective rotation transformation U_f (corresponds to solution set marked). Hypotheses S is a basic vector of input vectors. Vector S rotate 180° when $C(S) = 1$, unchanged when $C(S) = 0$.

(b) D transforms matrix acts on every input component. D defined as follows:

$$D = \begin{cases} \dfrac{2}{N}, & i = j \\ -1 + \dfrac{2}{N}, & i = j \end{cases} \qquad D = WRW \tag{2.1}$$

W is W–H transform matrix, R is conditions transfer matrix. R matrix reverse probability amplitudes of quantum state $|S\rangle$which content with $C(|S\rangle) - 1$

$$\begin{aligned} R_N &= 0 \ (i \neq j \,\text{and}\, i \neq 0) \\ R_N &= 1 \ (i = 0) \\ R_N &= -1 \ (i = j \,\text{and}\, i \neq 0) \end{aligned} \tag{2.2}$$

(3) Measure input, observation is $|\varphi_r\rangle$, we can get results if $C(|\varphi_r\rangle) = 1$, or we should start algorithm again.

Grover algorithm can effectively search the database. But it is invalid under the following circumstances:

(a) $m = N/2, \theta = \pi/4$, no matter how many times iteration occurs, the effect of iteration and no iteration is same and at this time algorithm is invalid.

(b) $m > N/4$, to ensure the algorithm has greater success probability, iteration number does not meet $\mathrm{O}\left(\sqrt{N/m}\right)$.

In order to solve the above problems, Grover algorithm was improved in Ref. [6]. In the following, we will study and compare Grover algorithm and improved algorithm effect on MIMO signal detection.

2.3 Signal Detection Project Based on Improved Grover Algorithm

Firstly, design two databases. The first database which has 2^{n_τ} registers save all possible sending sequence. As necessary judgment information, the correlation matrix RS and receiving signal stored in the computer. We can get 2^{n_τ} decision values by each may be sending sequence and correlation matrix and channel matrix H are calculated according to $\|y - Hx\|^2$. These judgment values put in the second database. Message sending sequences in the first database and judgment values in the second database constitute one-to-one relationships. Message sequences which correspond to minimum decision values are calculated according to $\|y - Hx\|^2$, and are the sending sequences which are detected by the best detection scheme. Therefore, the following work is looking for the smallest judgment value and judging sending sequence.

Grover Algorithm can solve this problem.

The basic thought: Hypothesis in the second database there is one group of judgment value $(x_1, x_2, x_3, \ldots, x_n, \quad N = 2^n)$, which correspond to N quantum ground-states $(\varphi_1, \varphi_2, \varphi_3, \ldots \varphi_N, \quad N_{N-1} = 2^n)$. N quantum ground-states placed in a quantum register. So there is $x_i = \min\limits_{i=0}^{N-1}(x_i)$ in the second database, and we need to make its position in the database. The judgment value and quantum state is corresponding, so we can think, once finding x_i corresponding quantum state position in quantum registers, also determined x_i position in the second database and then finding x_i corresponding sending sequence, and then the sequence is thought of as the message sequence by judgment. Solved classic problems by quantum algorithms become possible.

Specific algorithm described as follows:

(1) Construct $N = 2$ bit quantum registers, contains quantum ground-state $(|\varphi_1\rangle, |\varphi_2\rangle \ldots |\varphi_N\rangle)$ which correspond with decision value $(x_1, x_2, x_3, \ldots, x_N)$. They can be shown by $T(|\varphi_i\rangle) = x_i \quad (i = 1, 2, \ldots, N)$.

(2) Initialization quantum registers by Walsh–Hadamard transform. At this time the quantum states in quantum register is shown by

$$\left|\varphi\left(k_i^{(0)}, l_i^{(0)}\right)\right\rangle = \sum_{k=1}^{N} \frac{1}{\sqrt{N}} |\varphi_1\rangle \quad k_i^{(0)}, l_i^{(0)} = \frac{1}{\sqrt{N}} \tag{2.3}$$

(3) In quantum register random took a quantum ground-state, its corresponding judgment value as a threshold. Here we used judgment value x_1 which corresponds with the first quantum ground-state $|\varphi_1\rangle$in quantum register quantum as threshold. Then through Grover algorithm, in quantum register we find quantum ground-state which correspond with judgment value which is less or equal to x_i Use rotation operation R to rotate its probability amplitude, otherwise remain unchanged. R defines for:

$$1 \le m \le \frac{N}{3} \quad R_{pq} = \begin{cases} 0 & p \ne q \\ 1 & p = q, T\left(\left|\varphi_p\right\rangle\right) > T(|\varphi_1\rangle) \\ -1 & p = q, T\left(\left|\varphi_p\right\rangle\right) \le T(|\varphi_1\rangle) \end{cases} \tag{2.4}$$

$$m > \frac{N}{3} \quad R_{pq} = \begin{cases} 0 & p \ne q \\ 1 & p = q, T\left(\left|\varphi_p\right\rangle\right) > T(|\varphi_1\rangle) \\ i & p = q, T\left(\left|\varphi_p\right\rangle\right) \le T(|\varphi_1\rangle) \end{cases} \tag{2.5}$$

(4) Vector of quantum ground-state probability amplitude unitary transform by matrix D, amplify the probability amplitude of quantum ground-state which is searching. Matrix D defines for:

$$1 \leq m \leq \frac{N}{3} \quad D_{pq} = \begin{cases} \dfrac{2}{N} & p \neq q \\ -1 + \dfrac{2}{N} & p = q \end{cases} \tag{2.6}$$

$$m > \frac{N}{3} \quad D_{pq} = \begin{cases} \dfrac{1}{N} + \dfrac{i}{N} & p \neq q \\ \dfrac{1}{N} - i\left(\dfrac{N-1}{N}\right) & p = q \end{cases} \tag{2.7}$$

(5) After iteration function of R operator and D operator, all the quantum ground-state have been discovered which correspond judgment value less or equal to x_1 and are in quantum register, seeking scope of target quantum states dramatically narrow. We repeat (3) (4) operation for all searched quantum state and search the quantum ground-state which correspond judgment value less than or equal to threshold by Grover algorithm until the only quantum ground-state remains.

(6) At that time, quantum state is $\left|\varphi\left(k_i^{(0)}, l_i^{(0)}\right)\right\rangle$. By measured quantum registers, we can obtain petitions solution.

2.4 Simulation Results and Analysis

In order to analyze and compare performance of MIMO detection method based on Grover algorithm, this chapter does following hypothesis: (1) $K = 16$, K is subcarrier number of OFDM, each carrier send symbols number is 128, (2) Channel matrix H known and unchanged in 128 symbols cycle, then independently changed, (3) The receiver know the exact channel state information, the sender use no coding QPSK modulation, sending power is 1, (4) Noise is white Gauss noise, (5) 4×4 single diameter MIMO-OFDM systems. On the basis of hypothesis, this thesis gives Grover algorithm detection (GD) and improves Grover algorithm detection (IGD) apply to MIMO-OFDM signal detection and compares with the traditional maximum likelihood (ML) algorithm, ZF algorithm and MMSE algorithm detection. If the difference of detected signals in receiver and the original signals in sender is in a certain range, the receiver judges no error and correct answer. This range defines for search errors.

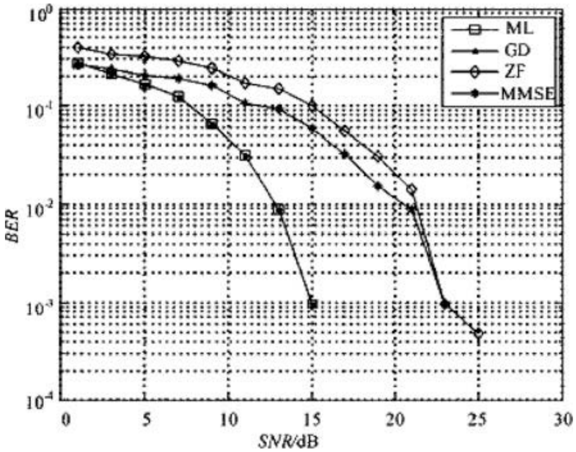


Fig. 2.2 GD and the traditional algorithm performance comparison when search error is 0.001

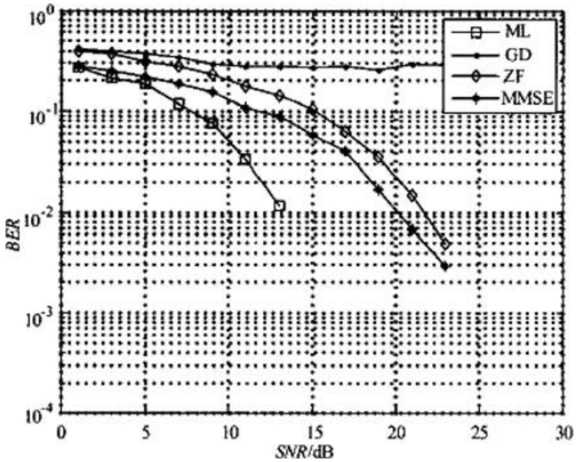


Fig. 2.3 GD and the traditional algorithm performance comparison when search error is 0.0001

Simulation results analysis: It can be seen from the Fig. 2.2 that GD performance is better in MMSE and ZF detection performance and close to ML detection performance even the same when search error is 0.001. But the original GD search error narrowed to 0.0001 even smaller, its detection effect is very poor and search failure, this can be seen from the Fig. 2.3. However, when search error is 0.0001 IGD performance is better in MMSE and ZF detection performance and GD performance in Fig. 2.2. This can be seen from Fig. 2.4. IGD effect is very good when taking smaller search error, since the selection of threshold is random at the beginning. The different selection of search error size not only relation computational complexity but also affects the algorithm's accuracy.

IGD search results are more accurate and valid than the original GD, when GD algorithm has many solutions its maximum iterating times appear indeterminacy. However IGD can assure solutions with great probability by iteration when the number of solutions is greater than $N/3$. All these show the improved Grover

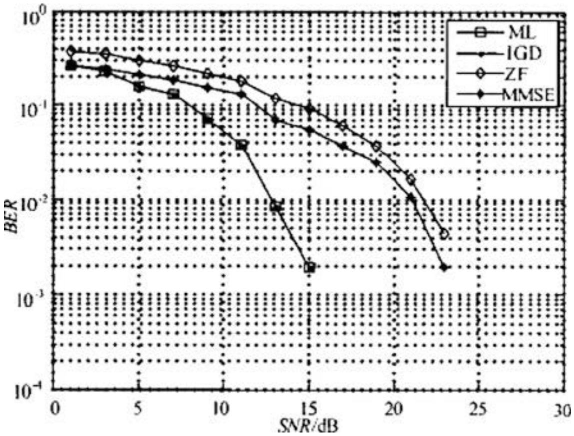


Fig. 2.4 IGD and the traditional algorithm performance comparison when search error is 0.0001

algorithm can clearly improve performance of traditional MIMO detection algorithm and achieve the performance which is very near to ML detection algorithm during effectively reducing the algorithm complexity.

2.5 Conclusions

The complexity of ML detection algorithm is exponential relation with antenna numbers. When antenna number and modulation order number is bigger, this all-over search process is difficult to realize for real-time or cannot come true in actual system for computing complexity, so it applies only to the theoretical analysis. ZF and MMSE receiver greatly reduces the computing complexity, but the two algorithms have to sacrifice property for complexity reduction because their performance was decreased obviously in high signal-to-noise ratio. For Grover algorithm, search steps reduce from classic N to $\sqrt{N}$, its time complexity is $O\left(\sqrt{N}\right)$, thus showing quantum acceleration. Time complexity of IGD is $O\left(\sqrt{N/t}\right)$ that has been proved by Ref. [6] and the algorithm is more effective and fast. The detection scheme that has been presented by this text can effectively reduce the complexity of the algorithm, and its effect is consistent with classic optimum detecting scheme in error code performance.

References

1. Nielsen MA, Chuang IL (2000) Quantum computation and quantum information [J]. Cambridge University Press, Cambridge, pp 1–12
2. Berenguer I, Adeane J (2002) Lattice-reduction-aided receivers for MIMO-OFDM in spatial multiplexing systems[Z]. Lab for communication engineering. University of Cambridge, Cambridge

3. Foschini J, Gan MJ (1998) On limits of wireless communications in a fading environment when using multiple antennas. J Wirel Pers Commun 6(3):311–335
4. Grover L (1996) A fast quantum mechanical algorithm for database search. In: Proceedings of the 28th annual ACM symposium on the theory of computing. ACM Press, Philadelphia, pp 212–2l9
5. Deutsch D (1989) Quantum computational networks. J Proc Roy Soc Lond A 425:73–90
6. Hui S (2002) An improved quantum searching algorithm (in Chinese). J Comput Eng Sci 24:4–8

Chapter 3
High-Performance Speed Control of Induction Motor Using Combined LSSVM Inverse System

Yi Zhang, Guohai Liu, Haifeng Wei and Wenxiang Zhao

Abstract In this paper, a new speed control method based on combined least squares support vector machines (LSSVM) inverse system for induction motor is proposed. It is characterized by that feedback of rotor flux and inverse model building are combined through LSSVM training and fitting. Firstly, LSSVM is used to build the inverse model of induction motor from the input–output data, and the inverse mode is served as the basis for the inverse controller design. The combined LSSVM inverse is composed of a LSSVM to approximate nonlinear mapping, an integrator and a differentiator. Cascading the LSSVM inverse with induction motor, induction motor system is transformed to a pseudo-linear system. Finally, simulation of the control method is performed to validate its feasibility. The results show that presented method has clear dynamic structure, which is effective for induction motor control.

Keywords Induction motor · Speed control · Rotor flux · Stator current · Least squares support vector machines inverse

3.1 Introduction

Induction motor with simple structure, high reliability and low maintenance cost advantages has been widely used in industry production [1]. Induction motor is a nonlinear, multivariable, strong coupling of time-varying system. It is a

Y. Zhang (✉) · G. Liu · W. Zhao
School of Electrical and Information Engineering, Jiangsu University,
Zhenjiang 212013, China
e-mail: zyi82@126.com

H. Wei
School of Electrical and Information, Jiangsu University of Science
and Technology, Zhenjiang 212003, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_3,

challenging job to achieve high-performance speed control. By transforming the physical current into a rotational vector using Clarke and Park transforms, vector control allows induction motor control with conventional techniques, as with a DC motor. Vector control is really approximate decoupling control under static state, since torque and flux can be decoupled completely only when flux is invariable [2]. Nonlinear control strategies for induction motor such as differential geometry [3], direct feedback linearization [4] and inverse system [5] are introduced to improve the decoupling control performance. LSSVM inverse is a modified inverse system control method for induction motor's decoupling control [6]. It is necessary to measure and feedback rotor flux accurately in the implementation of LSSVM inverse control to decouple torque and flux dynamically. Numerous studies have existed on rotor flux identification, including: (1) Direct calculation [7], this method depends on lots of motor parameters and it is open loop without compensation; (2) Method based on miscellaneous observers [8], when the system state is estimated, the observer adopts different gain matrix at different speed for system stability; (3) Integral method based on back electromotive force [9], the low-speed estimation of this method is not accurate enough.

The objective of this paper is to propose a new LSSVM inverse control method. Identification and feedback of rotor flux is implemented through a function that consists of the flux-producing component of stator current and its first-order derivative in the proposed control scheme. In Sect. 3.2, mathematical model is presented for induction motor working in vector control mode, and its invertibility is proved to confirm its feasibility. In Sect. 3.3, the relationship between rotor flux and the flux-producing component of stator current is analyzed, and the combined LSSVM inverse system is built. In Sect. 3.4, simulation results are presented to verify its effects of the proposed method. Last, conclusions are drawn in Sect. 3.5.

3.2 Mathematical Model and its Invertibility

An induction motor can be described by the following state equations in a rotor flux oriented reference frame:

$$\begin{cases} \frac{\mathrm{di_{sm}}}{\mathrm{dt}} = -\left(\frac{L_m^2 R_r + L_r^2 R_s}{\sigma L_s L_r^2}\right) i_{sm} + \left(\omega_r + \frac{L_m R_r}{L_r \psi_r} i_{st}\right) i_{st} + \frac{L_m R_r}{\sigma L_s L_r^2} \psi_r + = \frac{u_{sm}}{\sigma L_s} \\ \frac{di_{\mathrm{st}}}{\mathrm{dt}} = -\left(\frac{L_m^2 R_r + L_r^2 R_s}{\sigma L_s L_r^2}\right) i_{st} - \left(\omega_r + \frac{L_m R_r}{L_r \psi_r} i_{st}\right) i_{sm} - \frac{L_m \omega_r}{\sigma L_s L_r} \psi_r + \frac{u_{st}}{\sigma L_s} \\ \frac{d\psi_r}{\mathrm{dt}} = -\frac{R_r}{L_r} \psi_r + \frac{L_m R_r}{L_r} i_{sm} \\ \frac{d\omega_r}{dt} = (\omega_1 - \omega_r) \frac{n_p^2}{J R_r} \psi_r^2 - \frac{n_p}{J} T_L \end{cases} \tag{3.1}$$

where ψ_r is rotor flux; R_s is stator resistance; R_r is rotor resistance; L_s is stator self inductance; L_r is rotor self inductance; L_m is mutual inductance; $\sigma = 1 - L_m^2 / L_s L_r$ is flux leakage coefficient; ω_r is electrical angular velocity; ω is synchronous

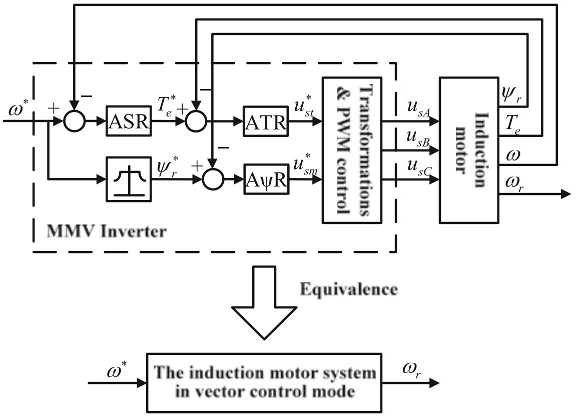


Fig. 3.1 The induction motor system in vector control mode

angular velocity; n_p is number of pole-pairs; J is inertia moment of rotor; and T_L is load torque.

In vector control mode, an induction motor with inverter system can be simplified as a single input–single output (SISO) system whose input is ω and output is ω_r. The induction motor system is shown in Fig. 3.1.

For the control system, state variable is $x = [x_1, x_2, x_3, x_4]^T = [\mathrm{i_{sm}}, \mathrm{i_{st}}, \psi_\mathrm{r}, \omega_\mathrm{r}]^T$, and control variable is $u = \omega_1$. The equations above are described as

$$\begin{cases} \dot{x} = f(x, u) \\ y = x_4 \end{cases} \tag{3.2}$$

where

$$f(x,u) = \begin{cases} -\left(\frac{L_m^2 R_r + L_r^2 R_s}{\sigma L_s L_r^2}\right) x_1 + \left(x_4 + \frac{L_m R_r}{L_r x_3} x_2\right) x_2 + \frac{L_m R_r}{\sigma L_s L_r^2} x_3 + \frac{u_{\mathrm{sm}}}{\sigma L_s} \\ -\left(\frac{L_m^2 R_r + L_r^2 R_s}{\sigma L_s L_r^2}\right) x_2 - \left(x_4 + \frac{L_m R_r}{L_r x_3} x_2\right) x_1 - \frac{L_m}{\sigma L_s L_r} x_3 x_4 + \frac{u_{\mathrm{st}}}{\sigma L_s} \\ -\frac{R_r}{L_r} x_3 + \frac{L_m R_r}{L_r} x_1 \\ (u - x_4) \frac{n_p^2}{J R_r} x_3^2 - \frac{n_p}{J} T_L \end{cases}$$

To analyze its invertibility, derivative of the output can be calculated firstly:

$$\dot{y} = (u - x_4) \frac{n_p^2}{J R_r} x_3^2 - \frac{n_p}{J} T_L \tag{3.3}$$

So Jacobin matrix is

$$A(x,u) = \left[\frac{\partial \dot{y}}{\partial u}\right] = \left[\frac{n_p^2}{JR_r} x_3^2\right] \tag{3.4}$$

When $x \in \Omega = \left\{x \in \mathrm{R}^4 : x_3^2 \neq 0\right\}$, the Jacobin matrix $A(x,u)$ is nonsingular. Thus, invertibility of the motor working in vector control mode is existent. The inverse system can be achieved as:

$$u = \bar{\phi}(x_1, x_2, x_3, x_4, \dot{y}) = \bar{\phi}(x_1, x_2, x_3, y, \dot{y}) \tag{3.5}$$

3.3 Implementation of the Combined LSSVM Inverse System

In the inverse system expressed as Eq. (3.5) , rotor flux ψ_r is involved. For an induction motor, the flux-producing component of stator current i_{sm} is easier to measure compared with ψ_r. Therefore, ψ_r is considered to be expressed by the function $\tilde{\phi}$ that consists of i_{sm} and its lower order derivatives.

It is essential to verify the existence of $\tilde{\phi}$. The corresponding Jacobin matrix is:

$$\frac{\partial [x_1, \dot{x}_1]^T}{\partial x_3} = \left[0, -\frac{L_m R_r}{L_r x_3^2} x_2^2 + \frac{L_m R_r}{\sigma L_s L_r^2}\right] = \left[0, \frac{L_m R_r}{L_r x_3^2} \left(\frac{x_3^2}{\sigma L_s L_r} - x_2^2\right)\right] \tag{3.6}$$

As known, flux leakage coefficient $\sigma \leq 1$, so $\psi_r^2 / (\sigma L_s L_r) > i_{\mathrm{st}}^2$. Therefore, $\mathrm{rank}\left(\partial [x_1, \dot{x}_1]^T / \partial x_3\right) = 1$. The function $\tilde{\phi}$ is existent. The following equation

$$x_3 = \tilde{\phi}(x_1, \dot{x}_1) \tag{3.7}$$

holds according to the inverse system theory.

On the basis of the analysis above, a new combined inverse control system is obtained by combining the function $\bar{\phi}$ and function $\tilde{\phi}$:

$$u == \bar{\phi}(x_1, x_2, x_3, y, \dot{y}) = \bar{\phi}\left(x_1, x_2, \tilde{\phi}(x_1, \dot{x}_1), y, \dot{y}\right) = \phi(x_1, \dot{x}_1, x_2, y, \dot{y}) \tag{3.8}$$

Cascading the combined inverse system with the induction motor, a pseudo-linear system is completed as shown in Fig. 3.2, where a LSSVM is used to approximate the nonlinear function ϕ. The motivation for choosing LSSVM as approximation tool is its higher generalization capability, as well as the achievement of an almost global solution in a reasonably short period of training time [10].

Fig. 3.2 Structure of the pseudo-linear system

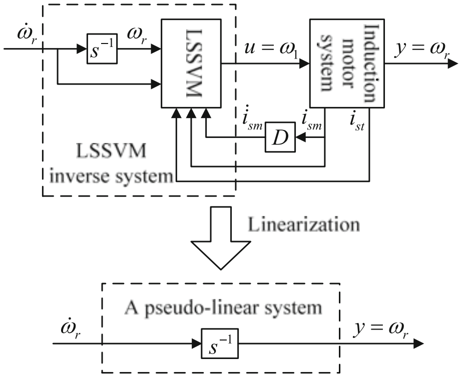
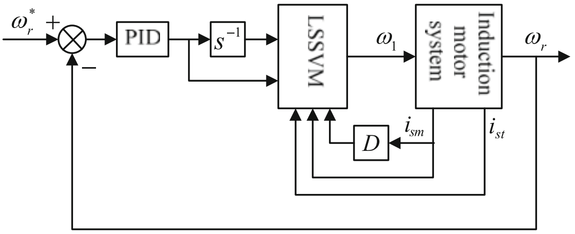


Fig. 3.3 Block diagram of the combined LSSVM inverse control



PID controller can be designed based on the pseudo-linear system to get high control performance. The combined LSSVM inverse control is implemented, and the block diagram is shown in Fig. 3.3.

The new control method is implemented as the following steps:

A. Data Acquisition

The data set used in training is obtained from the induction motor operating stably in vector control mode, when the system is excited by random square supply. Sample field data which include i_{sm}, i_{st}, ω_r and ω. Meanwhile the first-order derivative of i_{sm} are obtained by using 5-point numerical derivative method to guarantee high accuracy.

B. LSSVM's Training

The radial basis function (RBF) $K(x, x_i) = \exp\left(-\|x - x_i\|^2 \big/ 2\sigma^2\right)$ is used as kernel function of LSSVM, and then two key parameters of γ and σ are involved. γ is a regularization parameter which determines penalties to estimation errors, and σ represents the width of RBF kernel. In this paper, the tuning of γ and σ parameters is performed via cross validation.

C. Control Implementation

Fig. 3.4 Fit error curve of the combined LSSVM inverse

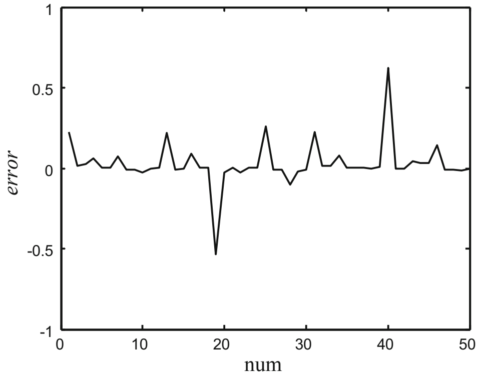
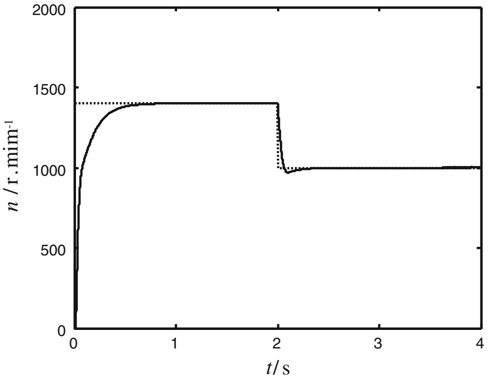


Fig. 3.5 Simulation result when given speed is shift



After training, LSSVM inverse model with rotor flux feedback is built. Cascading the combined LSSVM inverse system with induction motor system, and designing PID controller, the close-loop LSSVM inverse control is completed finally.

3.4 Simulation Results

In order to evaluate the performance of the proposed control method, simulation model of the whole control system is set up through MATLAB. The induction motor parameters; $P_e = 1.1\,\mathrm{KW}$, $R_s = 5.9\,\Omega$, $R_r = 5.6\,\Omega$, $L_s = 0.574\,\mathrm{H}$, $L_r = 0.58\,\mathrm{H}$, $L_m = 0.55\,\mathrm{H}$, $J = 0.0021\,\mathrm{kg} \cdot m^2$, rated load $T_L = 7.5\,\mathrm{N} \cdot m$, rated speed $n = 1400\,\mathrm{r/min}$. Figure. 3.4 is fit error curve of the combined LSSVM inverse system, which shows that the proposed method preserves important values such as high accuracy.

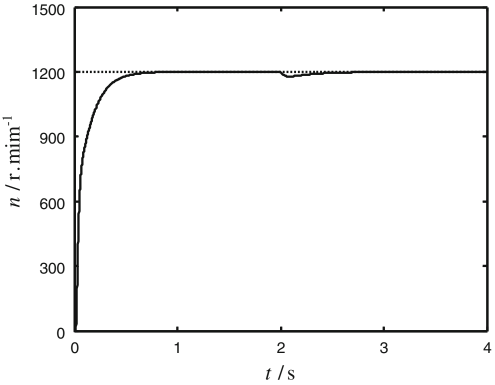


Fig. 3.6 Simulation result when load is changed

Figure. 3.5 is view of response speed when given speed shifts from 1400 r/min to 1000 r/min. The response speed can track given speed well, with high dynamic and static operation performances.

When the load of induction motor is changed from 0 to 50% rated load, the speed response to given speed of 1200 r/min is shown in Fig. 3.6, which verifies the proposed method has strong robustness to load torque disturbances.

3.5 Conclusion

In this paper, a new LSSVM inverse control method is proposed for induction motor working in vector control mode. High-performance speed control is achieved using a combined LSSVM inverse system, where the feedback of rotor flux is implemented by a function that consists of the flux-producing component of stator current and its first-order derivative. The proposed method needs only the relative degree, which is independent on model and specific parameters of induction motor. The simulation results show that the proposed method is feasible, effective and suitable for speed control of induction motor.

Acknowledgments This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 60874014 and 51077066, Science Foundation of Jiangsu Province under Grant BK2010327, and Graduate Education Innovation Project of Jiangsu Province under Grant CX09B_201Z.

References

1. Kirschen DS, Novotny DW, Lipo TA (2009) Optimal efficiency control of an induction motor drive. IEEE Trans Energy Convers EC-2(1):70–76
2. Wai RJ, Chu CC (2007) Robust petri fuzzy-neural-network control for linear induction motor drive. IEEE Trans Ind Electron 54:177–189

3. Boukas TK, Habetler TG (2004) High-Performance induction motor speed control using exact feedback linearization with state and state derivative feedback. IEEE Trans Power Electron 19:1022–1028
4. John C (1998) A new approach to dynamic feedback linearization control of an induction motor. IEEE Trans Autom Control 43:391–397
5. Huang MS, Liaw CM (2005) Speed control for field-weakened induction motor drive. IEE Proc Electr Power Appl 152:565–576
6. Zou HL, Diao XY, Zhu HQ et al (2010) Decoupling control of bearingless synchronous reluctance motor based on support vector machines inverse system. In: The 29th Chinese control conference. IEEE Computer Society, Beijing, pp 711–716
7. Marchesoni M, Segarich P, Soressi E (1997) Simple approach to flux and speed observation in induction motor drives. IEEE Trans Ind Electron 4:528–535
8. Cheng CY (2009) Sliding Mode Controller Design of Induction Motor based on Space-Vector Pulsewidth Modulation Method. Int J Innov Comput Inf Control 5:3614–3633
9. Jun H, Bin W (1998) New integration algorithms for estimating motor flux over a wide speed range. IEEE Trans Power Electron 13:969–977
10. Shang WF, Zhao SD, Shen YJ (2008) Application of LSSVM with AGA optimizing parameters to nonlinear modeling of SRM. In: The 3rd IEEE Conference on Industrial Electronics and Applications. Inst. of Elec. and Elec. Eng. Computer Society, Singapore, pp 775–780

Chapter 4
Lightweight Main Memory DB for Telecom Network Performance Management System

Lina Lan

Abstract Today telecom network is a growing complex. Although the amount of network performance data increased dramatically, telecom network operators require better performance on network performance data collection and analysis. Database is the important component in modern network management model. Since main memory database (MMDB) stores data in main physical memory and provides very high-speed access, MMDB can suffice the requirements on data intensive and real time response in network performance management system. This paper presents a novel lightweight design on MMDB for network performance data persistence. This design improves data access performance in following aspects. The data persistence mechanism employs user mode memory map provided by Unix OS. To reduce the cost of data copy and data interpretation, the data storage format is designed as consistent with binary format in application memory. The database is provided as program library and the application can access data in shared memory to avoid the cost on inter-process communication. Once data is updated in memory, query application can get updated data without disk I/O cost. The data access methods adopt multi-level RB-Tree structure. In best case, the algorithm complexity is O(N). In worst case, the algorithm complexity is O(N*lgN). In real performance data distribution scenarios, the complexity is nearly O(N).

Keywords Network management operation administration, maintenance, provisioning (OAM&P) · Performance management (PM) · Main memory database system (MMDB) · Disk-resident database system (DRDBS) · RB-Tree

L. Lan (✉)
School of Network Education, Beijing University of Posts and Telecommunications, Beijing, China
e-mail: lindalan2002@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_4,

4.1 Introduction

The database is the core component in the modern network management model. The performance of the database is important to support the large amount of data and high real time access requirement. The disk database system cannot support the performance management of network on the real time or nearly real time data access [1].

The main memory is 10^5 times more quick than disk on data access. The price of main memory is declining while the capability is increasing dramatically. Main memory database system is a good approach to support the real time data access of the application system [2].

The storage models of DRDBS (disk-resident database system) and main memory database system (MMDB) management system is much different on the data structure, algorithm, query, index etc. [3].

In this paper, based on investigation on the network performance data model, a lightweight main memory DB system is designed to provide the data access with high real time, large data set and low system consumption to support the network performance management service.

4.2 The Data Model of Network Performance Management

There are several entities in the network performance management service such as Network, LogicNE (logic network equipment), NEInstance (physics equipment), Group (Counter group) and Counter. The sample period can be assigned different in the equipments. There are many LogicNEs in a network and many NEInstances in a LogicNE. The performance data sampled from every NEInstance can be grouped into many groups. The data sampled by every period of every group is a Table. There are many Tuples in a Table. Each Tuple is mapped to a resource Instance such as CPU-1, Port-1, etc. Each column is mapped to a Counter.

According to the above analysis of the data model of performance management, the persistent class of the data is designed as the following Fig. 4.1.

In Fig. 4.1, the class persDataRoot is the root object of the physics file. Every file contains the performance management data collected from multiple networks. The network ID string is the key to index. The data from the same network can be indexed by the sample period because the sample period can be assigned as different values from different equipments. Likely, the LogicNE ID, NEInstance ID and Group ID all can be keys to index the relevant objects. The Table object is a two dimensionalities array as $R \times C$. R is the resource number of Groups, and is also the number of Tuples in the table. C is the Counter number in Groups. There are two assistant map < string,int > using the name of Resource and Counter to index the location in the Table as Table[r][c].

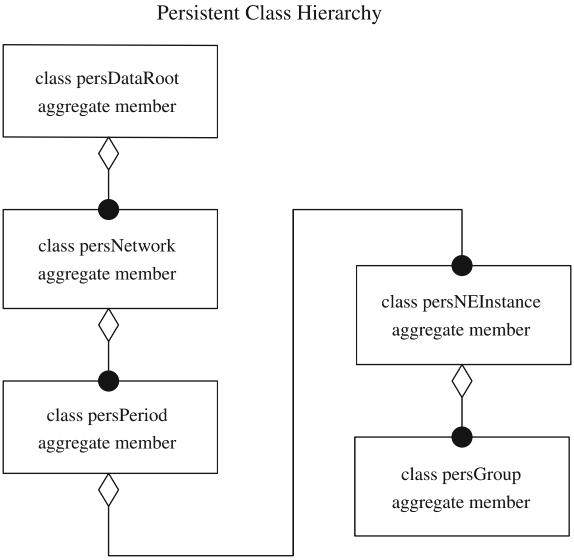


Fig. 4.1 The data model of network performance management in main memory DB

4.3 Key Design of Main Memory DB

4.3.1 Persistence Design

Main memory DB puts the "work version" of DB into main memory. This paper uses UNIX system procedure mmap() to map the DB storage file from disk to the process address space. Mmap is the wrapper of the map function provided by the file system where this file resides.

The first time we access a memory location within our segment, the page fault handling routine is called. This fault handler recognizes our segment as a mapped file and simply calls into the vnode's file system to read in a page-sized chunk from the file system. The subsequent access to memory that is now backed by physical memory simply results in a normal memory access. It is not until a page is stolen from behind the segment (the page scanner can do this) that a page fault will occur again.

Writing to a mapped file is done by updating the contents of memory within the mapped segment. The file is not updated instantly, since there is no software—or hardware—initiated event to trigger any such write. Updates occur when the file system flush daemon finds that the page of memory has been modified and then pushes the page to the file system with the file systems putpage routine [4].

Because the database which is using is in the main memory, the disk *I*/*O* operation will not influence the data access performance. Once the data is written into memory, the data is available to query process immediately. Fig. 4.2

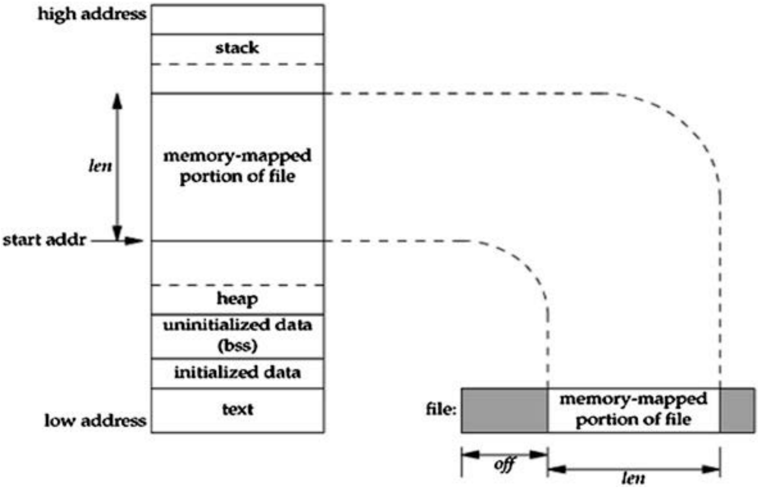


Fig. 4.2 Memory map theory

Due to the employment of memory mapping on data storage file, the data persistence work is completed by operating system.

4.3.2 Data Storage Format

The creation of intermediary objects should be avoided if possible in order to optimize the utilization of CPU and memory. To that end, the data storage format is designed as consistent with the object binary format in the virtual heap. Once the data storage file is mapped into the virtual heap, the application can use the persistent objects directly without data replication and data translation. Therefore, the code execution path of the application is shortened, and the CPU and memory cost of intermediary objects are saved.

The memory allocation of objects in MMDB differs from the memory allocation of normal objects in the application heap and stack, which employs standard memory allocators. This memory of persistent objects is allocated by Memory Map Allocator (MmapAllocator), which is specific for this MMDB design.

4.3.3 How to Create and use the Memory DB Objects

The application, which build with the MMDB shared library, can directly create and use the objects in MMDB.

The following example shows how to use the persistent objects [5].

For example, for class Counter,

1) Create the persistent object

```
Counter *pc = new Counter();
```

The above statement is to create a Counter object in heap. To create a Counter object in virtual heap must use MmapAllocator for memory allocation. In-place new is employed to construct the object at correct address in virtual heap.

For example:

```
Counter *pc = new(MmapAllocator::allocate(sizeof(Counter)))
Counter();
```

2) Create the persistent object in C ++ STL containers

```
Vector < Counter > VecCounter;
```

The above statement declares an object vector with Counters on stack. The MmapAllocator is used to replace the Allocator function in standard library to create the persistent object. For example:

```
Vector < Counter, MmapAllocator < Counter > >
```

The usage of the persistent objects created by MmapAllocator is as same as the common object. The application is easy to use the persistent object.

4.3.4 Data access method design

B-tree and B+tree are the common index technology in disk DB [6]. The *I/O* is reduced maximize, but the utilization ratio of space is only about 60% which is not unfit to main memory DB. So the index of memory DB should be designed specially. There are many index technologies such as Hash index and tree index. Hash index provides quick query and modify, but the utilization ratio of space is less than tree index. Traversing a deep tree in memory is much faster than in disk, so the tree index cannot be shallow and thick as B-Tree.

The performance data of network is appending but not modifying. Hash index has low utilization ratio of space. So tree index is much fitter to the memory DB of network performance data.

Considering the engineering realization, the RB-Tree realization of C++ standard library is high efficiency and stabilization. It can cut down the cost of development and maintenance and reduce the risk. So RB-Tree is used to be the index of memory DB in this paper.

RB-Tree is a kind of many balance trees. The basic dynamic set operation time is O(lgN). RB-Tree is a balance binary tree. The automatic compositor can achieve good results. So the standard STL map and set all use RB-Tree. The operate interface opened by map and set all have been provided by RB-Tree. The application only need call RB-Tree operation functions. Therefore, STL container (map,set) can be used for convenience to organize the store structure of DB with the existed index structure.

4.4 The Complexity Analysis of Data Access Algorithm

The data model constructed by map is a RB-Tree [7–9]. According to the data model in Fig. 4.1, DB is constructed by multiple layers of RB-Tree. The last layer node of RB-Tree is Pair < groupName,persGroup >. The key is GroupName. The value is persistent object persGroup. If there are N elements, the time complexity of search $T(N) = O(lgN)$.

4.4.1 The Time Complexity of Write in DB

The time complexity of write in DB is analyzed as following.

In case of the data is distributed equality. The size of map in each layer is equality, the size of each Table is also same. N is the Counter number of sample time duration.

$$\mathrm{N} = \mathrm{N}_1 * \mathrm{N}_2 * \mathrm{N}_3 * \ldots * \mathrm{N}_\mathrm{n} \tag{4.1}$$

In formula (4.1),

N1: Size of map in the first layer;

N2: Size of map in the second layer;

…

$\mathrm{N}_{\mathrm{n}-1}$: size of map in the n-1 layer;N_n:size of Tuple (e.g. the size of the integer array). The Counter values are stored in the array.

$$\begin{aligned}\mathrm{T(N)} &= \mathrm{N}_1 * [\lg \mathrm{N}_1 + \mathrm{N}_2 * [\lg \mathrm{N}_2 + \mathrm{N}_3 * [\lg \mathrm{N}_3 + \cdots \mathrm{N}_{\mathrm{n}-1}[\lg \mathrm{N}_{\mathrm{n}-1} + \mathrm{N}_\mathrm{n}] \cdots]]] \\ &= \mathrm{N}_1 * \lg \mathrm{N}_1 + \mathrm{N}_1 * \mathrm{N}_2 * \lg(\mathrm{N}_1 * \mathrm{N}_2) + \cdots + (\mathrm{N}_1 * \mathrm{N}_2 * \cdots * \mathrm{N}_{\mathrm{n}-1}) \\ &\quad * \lg(\mathrm{N}_1 * \mathrm{N}_2 * \cdots * \mathrm{N}_{\mathrm{n}-1}) + \mathrm{N}_1 * \mathrm{N}_2 * \cdots * \mathrm{N}_{\mathrm{n}-1} * \mathrm{N}_\mathrm{n} \\ &= \mathrm{N}_1 * \lg \mathrm{N}_1 + \mathrm{N}_1 * \mathrm{N}_2 * \lg(\mathrm{N}_1 * \mathrm{N}_2) + \cdots + [\mathrm{N}/(\mathrm{N}_{\mathrm{n}-1} * \mathrm{N}_\mathrm{n})] \\ &\quad * \lg[\mathrm{N}/(\mathrm{N}_{\mathrm{n}-1 * \mathrm{N}_\mathrm{n}})] + (\mathrm{N}/\mathrm{N}_\mathrm{n}) * \lg(\mathrm{N}/\mathrm{N}_\mathrm{n}) + \mathrm{N}\end{aligned} \tag{4.2}$$

In best case, N_n is near to N, $T(N) = O(N)$.In worst case, N_n is very small, $T(N) = O(NlgN)$.With data distribution from real network, the time complexity is between $O(N)$ and $O(NlgN)$. In practice, it is always nearly as $O(N)$.

4.4.2 The Time Complexity of Query in DB

The time complexity of query in DB is analyzed as following. The query condition includes the key in every layer of RB-Tree [10, 11]

$$T(N) = \lg N_1 + \lg N_2 + \cdots + \lg N_{n-1} + \lg N_n < O(\lg N) \qquad (4.3)$$

In worst case, T(N) = O(lgN).

In practice, the time complexity is always much better than the worst condition.

4.5 Conclusion

The lightweight main memory DB has many advantages applying in some scenes, e.g. real time network management data.

In this paper, a main memory DB is designed to support the application to manage the real time data of network performance. The novel data storage format design improves performance dramatically. The algorithm analysis shows perfect scalability on large data sets. This lightweight main memory DB can provide high performance in data intensive application with low cost under the limited computing resource.

This approach is a general advanced solution to manage the performance data from telecom management system. It has been applied in a real network management system and achieved satisfied effects.

References

1. Bohannon P, Lieuwen D, Rastogi R et al (1997) The architecture of the dali main-memory storage manager. J Multim Tools Appl 4(2):115–151
2. Yang W, Zhang J (2005) Summary of main memory DB[J]. J Xi'an Coll Posts Telecomm 10(3):96–99
3. Garcia-Molina H, Salem K (1992) Main memory database systems: an overview. IEEE Trans Knowl Data Engineer 4(6):509–516
4. McDougall R, Mauro J (2006) Solaris^{TM} Internals: Solaris 10 and open solaris kernel architecture. July 10, 2nd edition
5. Nicolai M. Josuttis (1999) C ++ Standard Library, August 06
6. Lu H,Yeung Ng Y, Tian Z (2000) T-tree or B-tree: main memory database index structure revisited, ADC 2000. In: Proceedings. 11th Australasian database conference, pp 65–73
7. Choi KR, Kim KC (1996) Real-Time computing systems and applications. In: Proceedings: third international workshop on 30 Oct-1 Nov, pp 81–88
8. Wilson PR, Johnstone MS, Neely M, Boles D (1995) Dynamic storage allocation a survey and critical review. In: international workshop on memory management, September
9. DeWitt DJ (1986) A study of index structures for main memory database management systems. In: Proceedings 12th conference on very large data bases pp 294–303
10. DeWitt DJ et al (1984) Implementation techniques for main memory database systems. In: Proc. ACM SIGMOD Conj
11. Whang KY, Krishnamurthy R (1990) Query optimization in a memory resident domain relational calculus system. ACM Trans Database Syst 15(1):67–95

Chapter 5
Multi-Stage TCAMs Architecture for IP Lookup

Weidong Wu, Wei Zhang, Liangliang Quan, Tao Yu and Tong Wu

Abstract Ternary Content Addressable Memory (TCAM)-based forwarding engines are widely used in core routers to achieve high throughput. To increase the throughput and reduce the power consumption of TCAM we propose a pipeline forwarding engine with multiple TCAMs, called Multi-stage TCAMs. Multi-stage TCAMs can perform IP lookups in parallel. A stream of IP lookup requests can be issued into multi-stage TCAMs, one every cycle, to achieve high throughput and reduce the power consumption.

Keywords Routing table · TCAM · Pipeline · IP lookup

This paper is supported by the project of Hubei provincial Department of Education No:D20101105.

W. Wu (✉) · W. Zhang · L. Quan · T. Yu · T. Wu
School of Computer, Wuhan University of Science and Technology,
Wuhan, China
e-mail: wuweidong@wust.edu.cn

W. Zhang
e-mail: zhangwei@wust.edu.cn

L. Quan
e-mail: quanliangliang@wust.edu.cn

T. Yu
e-mail: yutao@wust.edu.cn

T. Wu
e-mail: wutong@wust.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_5,

5.1 Introduction

Because of the inherent parallelism of Ternary Content Addressable Memory (TCAM), TCAM is a primary choice for many contemporary hardware architects and system designers to design high performance forwarding engines. Each cell in a TCAM can store don't-care values in addition to 0s and 1s. "Don't care" acts as wildcards. The destination IP address of an incoming packet is compared with all the prefixes in parallel. Several prefixes may match the destination IP address. Priority encoder logic then selects the longest-matching prefix. The state of the art 18 Mb TCAM can operate at a speed of up to 266 MHz and performs 133 millions lookup per second [1]. In contrast, conventional ASIC-based designs that use Trie may require multiple memory accesses for a single IP lookup. Therefore, TCAM-based solution is much faster than ASIC-based solutions for packet forwarding.

Despite these advantages, router vendors have been slow in adopting TCAM devices in packet forwarding engines; one of main reasons is the high power consumption. Current high-density TCAM devices consume as much as 12–15 Watts each when all the entries are enabled for search. Moreover, a single linecard may require multiple TCAMs to handle IP lookup on large forwarding tables. This high power consumption affects costs in two ways—first, it increases power supply and cooling costs. Second, it reduces port density since higher power consumption implies that fewer ports can be packed into the same space (e.g., router rack) due to cooling constraints. Therefore, it is important to minimize the power budget for TCAM-based forwarding engines to make them economically viable. Several strategies have been proposed to reduce TCAM power significantly by capitalizing on a feature in contemporary TCAMs that permits one to select a portion of the entire TCAM for search. There are some major concerns [2]:

(1) A partitioning method is needed here to split the entire routing table into multiple sub-tables that could be stored in separate TCAMs. It should support dynamic lookup requests distributions, efficient incremental updates, high-memory utilization and economical power dissipation.

(2) A dynamic load balancing is required for the sake of higher and robust throughput performance in parallel system because some prefixes in routing table are accessed more than other prefixes.

The subject of this paper mainly focuses on the above two issues. We will propose multi-stage TCAMs architecture for IP lookup. The main contributions of this paper are as follows: At first, we give the definition of level in routing table, and all prefixes are grouped by levels. Second, we propose a memory-balancing scheme to divide the groups into $K(= k*M)$ partitions with same size, where M is the number of TCAM chip, k is an integer. Third, we propose a load-balancing scheme to map partitions into TCAMs such that each TCAM has k partitions, and is accessed evenly. At last, we give a pipeline lookup algorithm and an efficient update algorithm. Multi-stage TCAMs is nonlinear pipeline with more entries and more exit points.

The rest of the chapter is organized as follows. In Sect. 5.2, we describe prior works. In Sect. 5.3, we propose a level-based partitioning algorithm and describe multi-stage TCAMs architecture. In Sect. 5.4, we propose the 2-step balancing scheme to evenly distribute the prefixes in routing table and lookup traffic among TCAMs. We present our performance evaluation in Sect. 5.5. Finally, we conclude our work in Sect. 5.6.

5.2 Related Works

Recently, researchers have proposed few pipeline approaches using routing table partitioning and multiple stages techniques.

Ravikumar et al. [3] proposed a two-level pipelined architecture that reduces power consumption through the prefix compaction and the partitioning technique. If there is the high-access frequency for the largest page, the power consumption increases quickly. The architecture is not suitable for the bursty access pattern. In Zheng et al. [4], exploited the parallelism among multiple TCAMs to increase the lookup throughput. The proposed scheme with four TCAMs increase the lookup throughput by a factor of 4 compared with single TCAM. In practice, all prefix are more than 8 bits, if the ID bits is more than 8 bits, we must expand the shorter prefixes so as to introduce few additional route entries, for example, the first 13 bit is selected as ID bits, this results in 25% more memory space [2]. In Akhbarizadeh et al. [5], proposed a prefix segregation scheme for a TCAM-based IP forwarding engine. In practice, the disjoint prefixes are more than 92% of the entire routing table, instead the enclosure prefixes is few. The size of TCAM1 is more than that of TCAM2.

In Lu et al. [6], developed an optimal algorithm, optSplit, for subtree splitting, and proposed an two-level TCAMs with SRAMs to achieve a significant reduction in power and TCAM size, but a IP lookup requires two TCAM searches and two SRAM accesses.

In the state-of-the-art TCAM-based pipeline forwarding engines, the lookup requests enter at the first stage, exit at last stage. In this paper, we propose a nonlinear pipeline architecture based on multi-stage TCAMs such that the lookup request can enter and exit at any stage while keeping the memory and lookup traffic balancing.

5.3 Multi-Stage TCAMs Architecture

5.3.1 Level of Prefix

Definition 1 If the prefixes in routing tables are represented as a Trie data structure, each prefix has the corresponding node in the Trie. If there are i prefixes in the path from root to the corresponding node of prefix P, we call the prefix P is in Level i.

Table 5.1 The distribution of prefixes in levels

Routing table	L0	L1	L2	L3	L4	L5	L6	L7
20060105	90813	85950	18510	3048	555	64	51	0
20070503	106602	100716	25623	3745	490	58	43	2
20080504	123500	108245	30580	4560	623	139	121	0
20090501	133284	117919	39157	8184	1300	198	0	0
20100101	144245	123934	39397	8294	1505	228	91	2

To compute the level of prefixes in routing tables, at first, we sort the prefixes in ascending order of IP address; all child prefixes follow its parent prefix. We propose the level-partitioning algorithm. Let n be the number of prefixes in a routing table, the complexity of the level-based partitioning algorithm is O($n$log$n$). For the real routing tables from routeview project [7], the level distribution is shown in Table 5.1. The maximum Level is 8. The number of prefixes in Level 1 is about ten times that in Level 2; The number of prefixes in Level 2 is about ten times that in Level 3; and so on. When we search IP address in routing table with the longest-prefix-matching algorithm, the maximum searching depth is 8, and about 98% of searching depths are no more than 2.

5.3.2 Partitioning Prefixes into Groups

We use Bit selection scheme to split the prefixes in Level 0. From the observation of routing tables [8], all prefix are more than 8 bits length. We therefore use the first 8-bits to divide the prefixes in Level 0 into 256(= 2^8) groups, that is to say, we add 256 8-bit prefixes, each prefixes in Level 0 has a parent prefix with 8 bits length. From Definition 1, each prefix in original routing table has the unique parent prefix. We can split the entire routing table (R) as follows:

$R = \mathrm{G}_1 \cup G_2 \cup \ldots \cup G_N$ such that all prefixes in G_i ($i = 1,2, \ldots,$ N) have the same parent prefix.

For $R = \mathrm{G}_1 \cup G_2 \cup \ldots \cup G_N$, there are following properties:

(1) For any prefix P_i, $P_j \in G_k$, then $P_i \cup P_j = \varnothing$.
(2) For any IP address, there is no more than one match prefix in any group G_i.
(3) For an IP address D, if $D \in P_1 \in G_i$, $D \in P_2 \in G_j$, then P_1 (P_2) is the parent of all prefixes in $G_j(G_i)$.
(4) For the longest-match prefix, IP address match with all prefixes of a group in Level 0, then in Level 1, and so on until there is no match.

The statistics of the prefixes in each group is shown in Table 5.2. The number of groups is about 10% of the total number of prefixes. In the maximum group there are about 1% of total prefixes. The average number of prefixes in groups is about 11.

Table 5.2 The statistics of groups

Routing table	No. groups	No. prefixes in groups	
		Max	Aver
20060105	16874	1944	11.8
20070503	19858	2027	11.9
20080504	23693	2018	11.3
20090501	27398	2739	11.0
20100101	29607	2679	10.7

5.3.3 Implementation of Forwarding Engine

Pipelining is an effective way to achieve high lookup rates. We introduce a pipeline forwarding engine architecture, called multi-stage TCAMs, which consists of Index Logic, prefix cache, and M TCAMs with Virtual Queue, shown in Fig. 5.1.

The function of the Index Logic is to find out the group that contains the prefix matching with the incoming IP address. Index Logic uses the first 8 bits of prefixes as the ID bits to construct a direct index table containing 256(= 2^8) entries, each of which points to a Virtual Queue. When a packet arrives, Index Logic extracts the first 8 bits of its destination IP address, computes the group ID, and delivers the IP address to the Virtual Queue with the same ID.

For each TCAM, we maintain a FIFO queue (called Virtual Queue) to store the IP address from Index Logic. Each IP address in Virtual Queue has two fields: *Next_hop*, *Group_ID*. The first field, *Next_hop*, stores the next hop by which the packet maybe forwarded. The second field, *Group_ID*, is the identification of the group that contains the match prefix.

Suppose there are M TCAMs, each TCAM is divided into blocks that can store b prefixes. Each prefix group (G_i) is stored into a portion with size(G_i)/b continuous blocks on a TCAM. An array *TCAM_entry*[i] with fields: *Next_group* and *Next_hop* is used to store the pointer from ith entry(prefix) in TCAM to the next group and the next hop of the entry(prefix).

Contemporary TCAMs have a feature that permit one to select a portion of the entire TCAM for search. In a lookup cycle, the first IP address (D) in the Virtual Queue enters TCAM and matches with all prefixes in the partition that stores the group which ID is the same as *D.Group_ID*, other portions is inactivated. Because any two prefixes in a group are disjoint, there is no more one match prefix for any IP address. Therefore, prefixes in a group can be stored into a portion of TCAM in disorder, and the priority encoder of TCAM can be removed. The lookup latency is reduced 50% [9].

IP caching is an efficient way to exploit internet traffic locality for IP lookup [10]. If an incoming IP address has cache hit, it will skip the complete lookup. Otherwise, it is sent to a Virtual Queue by Index logic. For a new flow with the same destination IP address, the first packet of the flow needs the complete lookup,

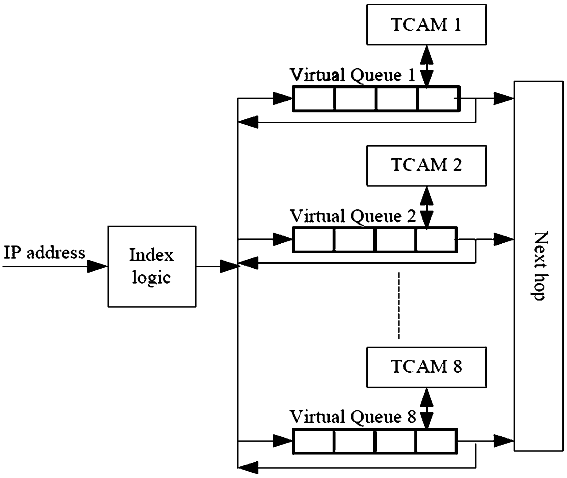


Fig. 5.1 Multi-stage TCAMs architecture

the rest of the packets of the flow are cut-through routed through cache entry lookup.

The cache update is triggered, either when there is a route update that is related to some cached entry, or after a packet that previously had a cache miss retrieves its search result from the TCAM. Any replacement algorithm can be used to update the cache. The Least Recently Used (LRU) algorithm is used as the default.

For an incoming packet, Index Logic uses the first 8 bits of its destination IP address to compute the group ID. Then the IP address is sent to the Virtual Queue with the same group ID. Each IP address is searched in the corresponding portion of TCAM. If the search result is not null, then the fields: *Next_hop* and *Group_ID* of the IP address are updated, and the IP address is sent to the corresponding Virtual Queue, otherwise, the IP address is sent out the Virtual Queue, its next hop is the search result.

For a new prefix (P_0), we find its parent prefix (P_1) and child prefix. If there is no child prefix, P_0 is inserted into the child group of the prefix P_1. If there is a child prefix, we apply a simple and widely used technique, leaf pushing [11], to expand the prefix (P_0), and the new produced prefixes are stored the child group of the prefix (P_0).

5.4 Two-Step Balancing Scheme

In multi-stage TCAMs, any group can be allocated into any TCAM. In practice, each TCAM has the same size. The prefixes in routing table should be allocated into each TCAM such that all TCAMs have the same number of prefixes. On the other hand, the distribution of packets per prefix has a heavy, Pareto-like tail [12].

Some TCAMs are accessed highly than another, it is necessary to balance the lookup traffic between TCAMs. Therefore, we propose a 2-step balancing scheme. For M TCAMs, at the first step we divide all groups into $K(= k*M)$ partitions with the same size of prefixes, where k is an integer. At the second step, we map k partitions into a TCAM based on the access frequency such that each TCAM is accessed evenly.

5.4.1 Memory Balancing

To balance memory between M TCAMs, we divide all groups into K partitions based on the size of groups. To formulate the memory-balancing problem, we use the following notations:

N:	denotes the number of prefix groups.
K:	denotes the number of partitions, $K = k * M$.
$Size(.)$:	denotes the size, i.e., the number of prefixes.
S_i:	ith partition which is the set of groups.

The memory-balancing problem can be formulated as following:

$$\min \max_{1 \le i \le K} Size(S_i)$$

Subject to

$$\bigcup_{1 \le i \le K} S_i = \{G_j | j = 1, 2, \cdots, N\}, \quad S_i \cup S_j = \varnothing \quad 1 \le i, j \le N$$

The above optimization problem is NP-complete. This can be shown by a reduction from the partition problem [13]. We use an approximation algorithm to solve it [14]. All groups $\{G_i\}$ are sorted in decreasing order of the group size. From G_1 to G_N, each group is allocated into the partition in which size is minimal. The complex of the memory-balancing algorithm is O(N*K).

5.4.2 Traffic Balancing

After the memory balancing, we map the k partitions into a TCAM based on the access frequency such that each TCAM is accessed evenly. To formulate the traffic balancing problem, we use the following notations:

M	the number of TCAM devices
$f(\cdot)$	denotes the access frequency
T_i	the set of partitions that are mapped into ith TCAM

The traffic balancing problem can be formulated as following:

$$\min \max_{1 \le i \le M} f(T_i)$$

Subject to $\bigcup_{1 \le i \le M} T_i = \{S_j | j = 1, 2, \cdots, K\}, \quad T_i \cup T_j = \varnothing \quad 1 \le i, j \le N, |T_i| = k$

The problem is the same as the memory-balancing problem. All partitions $\{S_i\}$ are sorted in decreasing order of access frequency. From S_1 to S_K, each partition is allocated into the TCAM which access frequency is minimal and which size is less than k. The complex of the traffic balancing algorithm is O(K*M).

5.5 Performance Evaluation

5.5.1 Experiments

We use the real routing tables from RouteViews Project [7] and FUNET traffic trace from the Finish University [15]. The FUNET trace could be considered as representative of Internet backbone traffic since it contains both university and student dormitory traffic.

Suppose there are eight TCAMs, and the size of cache is 1024. The prefixes in the real routing tables are partitioned into groups, shown in Table 5.2. In first step, we divide the groups into 32(= 4*8) partitions based on the size of groups. In second step, we select four partitions into a TCAM based on the access frequency. The result of 2-step balancing scheme is shown in Table 5.3. The difference of the number of prefixes between TCAMs is less than 5. The difference of traffic load between TCAMs is $< 2\%$.

5.5.2 Overall Performance

Based on the above experiments, we estimate the overall performance of Multi-stage TCAMs architecture with eight TCAMs. We add 256 8-bit prefixes to partition the prefixes in Level 0 into 256 groups. The 8-bit prefixes are used by Index logic, not stored into TCAM. Therefore, all prefixes stored in TCAMs are original prefixes of the real routing table. The total memory needed is O(m), where m is the number of prefixes in routing table.

For the longest-match prefix, it is possible for an incoming IP address to go through more than TCAMs, that is to say, an IP address needs more than one lookup. For the real routing tables in Table 5.1, the maximum number of lookups is eight, the average number of lookups per IP address is 1.6. For FUNET trace, the average number of lookups per IP address is <1.004, shown in Table 5.3. The average throughput is 8/1.004 = 7.968 packets per cycle. Because the priority

Table 5.3 The result of 2-step balancing scheme

Routing table	No. prefixes		Traffic load		Lookups per IP address		Power consumption (%)	
	Max	Min	Max (%)	Min (%)	Max	Aver	Max	Aver
20060105	13524	13521	13.8	12.1	4	1.003	8.89	0.038
20070503	16336	16334	12.8	12.3	4	1.004	6.38	0.033
20080504	18035	18032	12.9	12.0	4	1.003	3.44	0.020
20090501	20846	20844	12.7	12.2	5	1.003	2.30	0.017
20100101	21683	21681	12.9	11.8	4	1.004	2.50	0.016

encoder of TCAM can be removed; the lookup latency is reduced 50% [9]. This means the throughput of multi-stage TCAM architecture is 15.936(= 7.968*2) times that of native TCAM.

For simplification, the power consumption is defined as the total number of prefixes that match with an IP address. The ratio of the power consumption in our experiments to that of native TCAM is shown in Table 5.3. The average power consumption is <0.05% that of native TCAM. The maximum power consumption is <10% that of native TCAM.

5.5.3 Comparison with the Existing Schemes

In Lu et al. [6], proposed 2-level TCAMs architecture with Index TCAM and Data TCAM. The optSplit and PS2 algorithm is used to pack prefixes into a Data TCAM bucket, and may generate more index prefixes. Multi-stage TCAMs architecture does not generate any prefixes. In 2-level TCAMs architecture, a lookup requires two TCAM searches and two SDRAM accesses. In multi-stage TCAMs architecture, the maximum number of TCAM searches per lookup is eight, the average number is about 1.004, and multi-stage TCAMs search latency is 50% that of 2-level TCAMs because the priority encoder of TCAM is removed. The power required by 2-level TCAMs is 1/12 the native TCAM, that is the same as the maximum power of multi-stage TCAMs, and is 20 times the average power of multi-stage TCAMs architecture.

5.6 Conclusions

TCAM-based forwarding engines are widely used in core routers to achieve high throughput. To increase the throughput and reduce the power consumption of TCAM we propose a pipeline forwarding engine with multiple TCAMs. We divide the prefixes in the routing table into groups based on its parent prefix. We propose a two-stage balancing scheme to map the groups into multiple TCAMs such that memory utilization is balanced, the lookup traffic is evenly distributed between TCAMs. Multiple TCAMs can perform IP lookups in parallel. An incoming IP address enters a stage TCAM, exits another stage TCAM. A stream of IP lookup requests can be issued into multi-stage TCAMs, one every cycle, to achieve high throughput and reduce the power consumption.

References

1. CYRESS. http://www.cypress.com. Oct 2010
2. Lin D, Zhang Y, Hu CC, Liu B, Zhang X, Pao D (2007) Route table partitioning and load balancing for parallel searching with TCAMs. In: Proceedings of IEEE international parallel and distributed processing symposium, 26–30 March pp 1–10
3. Ravikumar VC, Mahapatra RN (2004) TCAM architecture for IP lookup using prefixes properties. IEEE Micro (24)2:60–69
4. Zheng K, Hu CC, Lu H, Liu B (2006) A TCAM-based distributed parallel IP lookup scheme and performance analysis. IEEE/ACM Transactions on Networking 14(4):863–875
5. Akhbarizadeh MJ, Nourani M, Cantrell CD (2005) Prefix segregation scheme for a TCAM-based IP forwarding engine. IEEE Micro 25(4):48–63
6. Lu W, Sahni S (2010) Low-power TCAMs for very large forwarding tables. IEEE/ACM Transactions on Networking 18(3):948–959
7. The University of Oregon Route Views Project (2010) ftp://routeviews.org/bgpdata
8. http:www.//bgp.potaroo.net. Nov 2010
9. Akhbarizadeh M, Nourani M (2004) Efficient prefix cache for network processors. In: Proceedings of IEEE Symposium on High Performance, Interconnects, pp 41–46
10. MacGregor MH (2003) Design algorithms for multi-zone IP address caches. In: Proceedings of high performance switching and routing, pp 281–285
11. Srinivasan V, Varghese G (1999) Faster IP lookups using controlled prefix expansion. ACM Trans Comput Syst 17(1):1–40
12. Sahni S, Kim K (2004) Efficient dynamic lookup for bursty access patterns. Int J Found Comput Sci 15(4):567–592
13. Kleinberg J, Tardos E (2005) Algorithm design. Longman Publishing, USA
14. Jiang W, Wang Q, Prasanna VK, (2008) Beyond TCAMs: An SRAM based parallel multi-pipeline architecture for terabit IP lookup. In: Proceedings of 27th conference on computer communications (INFOCOM'08), pp 1786–1794
15. http://www.csc.fi/english/funet/.Nov 2010

Chapter 6
Energy Savings in Cellular Network based on Cluster Analysis of Traffic Loads

Hongzeng He, Jingbo Sun, Yue Wang, Qi Liu and Jian Yuan

Abstract Recently, increasing attention has been paid to green communications and networks. Previous studies suggest that certain base stations could be switched off for energy saving during low traffic load periods while maintaining the service coverage. In this Chapter, we proposed an energy saving approach based on cluster analysis. The cluster analysis on original data from an operating cellular network in a big city exploits different traffic patterns of base stations. In addition, specific methods are applied to these different traffic patterns. Simulation results indicate a better energy saving performance based on cluster analysis, compared to the existing approach.

Keywords Energy savings · Cellular networks · Cluster analysis

H. He (✉) · J. Sun · Y. Wang · Q. Liu · J. Yuan
Department of Electronic Engineering, Tsinghua University, Beijing, China
e-mail: hehz09@mails.tsinghua.edu.cn

J. Sun
e-mail: sjb06@mails.tsinghua.edu.cn

Y. Wang
e-mail: wangyue@mail.tsinghua.edu.cn

Q. Liu
e-mail: liu-qi@mail.tsinghua.edu.cn

J. Yuan
e-mail: jyuan@mail.tsinghua.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_6,

6.1 Introduction

Our earth has become more warm in the last century due to the rising emission of CO_2 and other greenhouse gases. Low-carbon economy and energy saving have now become hot topics in our daily life. As the rapid development of information and communication technology (ICT) infrastructure, the ICT energy consumption cannot be neglected. According to some scientific findings recently, 3% of the wordwide energy is consumed by the ICT industry, which causes about 2% of the worldwide CO_2 emissions. In addition, power consumption of ICT is currently rising at 16–20% per year [1]. Since the ICT industry is growing more and more rapidly, energy efficiency of ICT industry should be regarded as an important issue. And recently, we are gratified to see that more and more telecommunication service providers, infrastructure manufacturers and research institutes are engaged in the development of green communications and networks (GCN) which focuses on improving the energy efficiency of ICT industry. As we know, the temporal traffic load varies violently by time during one day in each base station. But nowadays most designs of the communication networks frequently have not considered for the adaption to the variable traffic load, which means that base stations work in full power mode all day long. There will be a great waste of communication and energy resources.

In this chapter, we proposed an energy saving approach based on cluster analysis. The cluster analysis on original data from an operating cellular network in a big city exploits different traffic patterns of base stations. In addition, specific methods are applied to these different traffic patterns. Simulation results indicate a better energy saving performance based on cluster analysis, compared to the existing approach. The outline of this Chapter is as follows: Section 6.2 introduces the implement of cluster analysis and characteristics of the original data used in our work. In Sect. 6.3 we describe the details of our energy saving scheme. Section 6.4 investigates the simulation results. The benefits and energy saving effects is proposed in the present section. In Sect. 6.5, we proposed the directions of the future work and end this Chapter with some concluding discussions.

6.2 Background and Data

6.2.1 K-Means Cluster Analysis

Cluster Analysis is a conventional approach to classify a set of data into different clusters so that the data in the same cluster would show some similar characteristics in a manner. In this Chapter, we prefer the partitional clustering method which uses a high speed algorithm and is suitable for large datasets. It could determine all the clusters in less time. The k-means clustering algorithm is one of partitional methods which assigns each element to the cluster whose distance is the nearest.

6.2.2 Original Data

The original data of this Chapter are from a flourishing district of a big city in China, supported by one telecommunication operator. This region covers an area about 470.8 Km^2. It has a population of nearly 3.1 millions, which means an average density of 6,500 Km^{-2}.

The original data contains the traffic load statistics of each base station over 8 days in Erlang, including weekdays and weekends. The data is updated every hour which means that there are traffic data of more than 190 h. The traffic statistics indicate a violent fluctuation on the temporal scale, which is shown in Fig. 6.1a. We observe variations during 24 h in one day and differences of amplitude between weekday period and weekend period. For example, the traffic load of weekend period is always much lower than that of the weekday period. Figure 6.1b is a distribution map of the base station deployments based on the information provided by the original data. Each blue point represents one base station location. The total number is about several hundreds. We can see a much higher base station density in the urban space than in the rural area. As a matter of fact, there will be more redundancy in the cellular coverage that makes room for switching off some of the base stations in the low load period [2].

6.3 Energy Saving Scheme

6.3.1 Normalization

Data normalization is necessary for comparing the relative changes of the traffic loads in different base stations in a more detailed way. Our normalization steps are as follows [3]:

1. For each base station $b_{\mathrm{i}} \in \mathrm{B}$ (i = 1, 2, … N), we do averaging to the value of 8 days (192 h) that is represented by

$$M = \underset{b_i \in B}{\mathrm{mean}}\{\mathrm{erlang}(b_i, \tau)\} \quad i = 1, 2, \ldots N \tag{6.1}$$

   where b_{i} is the base station in our datasets and τ is the time period across 192 h.

2. We then calculate the variance of traffic load over time in each base station, as in

$$S = \underset{b_i \in B}{\mathrm{std}}\{\mathrm{erlang}(b_i, \tau)\} \qquad i = 1, 2, \ldots N. \tag{6.2}$$

3. We then normalize the data to the uniform margin over time of 192 h in (6.3).

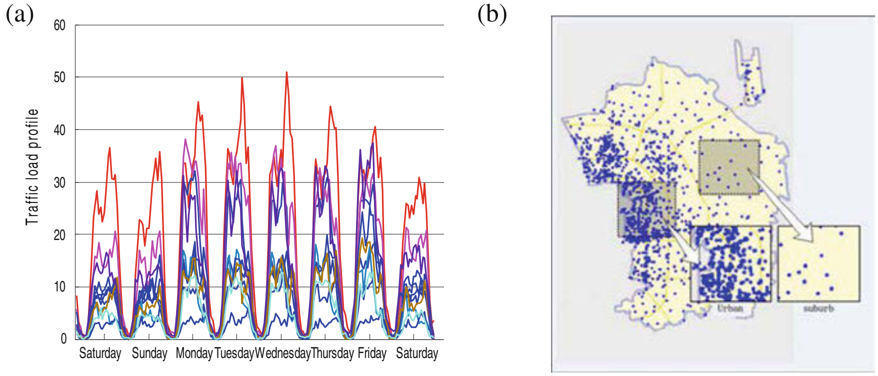


Fig. 6.1 a Temporal traffic trace of ten base stations (8 days), b Base station deployments in urban and rural space

$$\mathrm{erlang}_{\mathrm{norm\_192}} = \frac{\mathrm{erlang}(\mathrm{b_i},\tau) - \underset{\mathrm{b_i}\in\mathrm{B}}{\mathrm{mean}}\{\mathrm{erlang}(\mathrm{b_i},\tau)\}}{\underset{b_i\in B}{\mathrm{std}}\{\mathrm{erlang}(b_i,\tau)\} \quad i = 1,2,\ \ldots N.} \tag{6.3}$$

4. The differences between weekday periods and weekend periods are quite visible in Fig. 6.1. So we consider the two cases separately. One is the traffic load profile of weekday periods, and the other is of weekend periods.
5. Finally we normalize the data of the two groups to one day time separately, which can be expressed as

$$\mathrm{erlang}_{\mathrm{norm\_24\_weekend}} = \mathrm{erlang}_{\mathrm{norm\_192}}(b_i,\tau_j)/3, \quad i = 1,2,\ \ldots N \tag{6.4}$$

where $\tau_j \in [1 + 24(j-1), 24 + 24(j-1)]$ and j = 1, 2, 8, and

$$\mathrm{erlang}_{\mathrm{norm\_24\_weekday}} = \mathrm{erlang}_{\mathrm{norm\_192}}(b_i,\tau_j)/5, \quad i = 1,2,\ldots N \tag{6.5}$$

where $\tau_j \in [1 + 24(j-1), 24 + 24(j-1)]$ and j = 3, 4…7.

As a result, we get two groups of normalized data, including temporal and spatial traffic load status and base station locations' information. We observe that the traffic load during the nighttime is much lower than that in the daytime.

6.3.2 K-Means Cluster Analysis

We adopt the K-means clustering approach to the normalized data. We decide to classify all the base stations into three clusters by two K-means clustering stages.

First, we randomly choose three base stations as the original center of each cluster, designated by Z_k, k = 1, 2, 3.

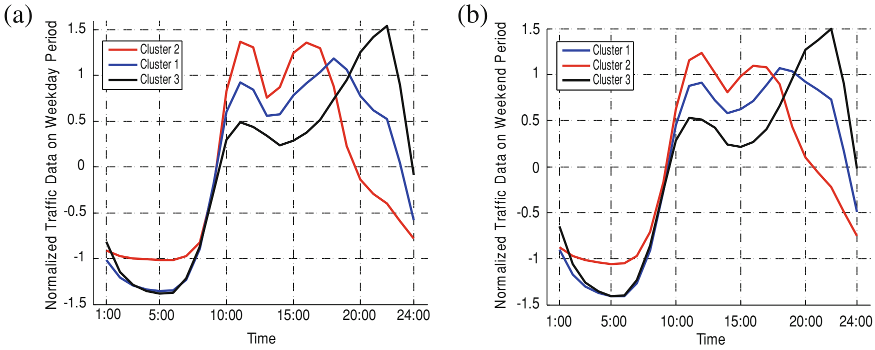


Fig. 6.2 **a** Three clusters of average traffic data on weekday period. **b** Three clusters of average traffic data on weekend period

Second, for each base station b_i, we calculate the distance to each $\mathrm{Z_k}$ and add it to the cluster X with the smallest distance by

$$x = \min \parallel \mathrm{erlang}_{\mathrm{norm\_24}}(b_i, \tau) - Z_k \parallel, \quad i = 1, 2, \ldots N \tag{6.6}$$

Third, we recalculate the center of every cluster, $\mathrm{Z'_k}$ as in

$$\mathrm{Z'}_\mathrm{k} \frac{1}{N_\mathrm{x}} \left(\sum_{b_i \in \mathrm{X}} \mathrm{erlang}_{\mathrm{norm\_24}}(b_i, \tau) \right) \tag{6.7}$$

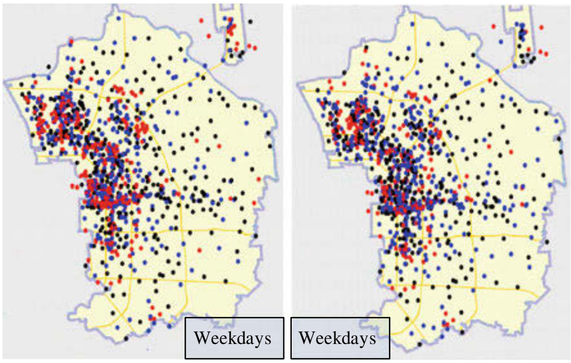
where N_X represents the number of base stations in cluster X. We then repeat stage 2 and 3 until it matches the following equation (6.8):

$$\mathrm{Z'_k} = \mathrm{Z_k}. \tag{6.8}$$

As a result of K-means cluster approach, we group all the base stations into three types. Each type has its own changing law. The average variation traces can be drawn in Fig. 6.2. Cluster 1, signified by the blue line, appears as a flatter changing pattern during all the 24 h. It implies as a matter of fact, that base stations in this cluster tackle a medium amount of traffic load with smaller variations. Cluster 2, which is painted red, exhibits higher values during the working hours from 10 pm to 17 pm except during the lunch break time. Additionally, the average load of the base stations of Cluster 3 , in black, concentrates on the after office hours from 19 pm to 23 pm. As a result, the varying patterns of the three types show significant differences in operating features and working modes of the different base stations. The traffic load information is related to the activity intensity of mobile users under coverage of the base station.

We then conclude the geographical features of base stations with the statistics we have got. Cluster 1 (blue) shows the evenness in contrast to Cluster 2 and Cluster 3, suggesting that the base stations in Cluster 2 cover areas requiring a normal

Fig. 6.3 Base stations distribution of three clusters



level communication services throughout the day, including transport hubs and other gathering places. Cluster 2 (red) which shows a high traffic level in the daytime may be connected to the working places, such as companies and banks, where is busy and bustling during business hours. Cluster 3 (black) shows a significant peak at night, suggesting that base stations in Cluster 3 cover the utility areas, for example, the residential districts, public houses and restaurants. On the weekdays, we observe higher peaks for the rapid changing of position and intensive activities on work time. On the weekends, there is a smaller slope and lower peaks by contrast. Naturally, we examine two different traffic load patterns between weekdays and weekends, and the same station appears with different patterns between weekdays and weekends. We are aware of an initial impression by projecting these base stations into a map with different colors, as in Fig. 6.3. There are some differences in the locations of each cluster. Cluster 1 (blue) shows almost a homogeneous distribution over spatial scale, consisting with its equitability on temporal scale. It is helpful to understand the relations between temporal and spatial scale. Cluster 2 (red) concentrates on the urban space of the city. In addition, there are more base stations in Cluster 2 on weekdays than that on weekends, which means a heavier load in Cluster 2 on weekdays. For the analysis of the three clusters, we have seen the individual characteristics of the three clusters that reflect the particular geographical locations.

6.3.3 Switching-Off Approach

We hope to apply the energy saving algorithm to the real networks according to the patterns we have found. In this Chapter, we use the constant energy consumption profile [4]

$$P(b_i) = P_{0c} = \text{const.} \tag{6.9}$$

The constant energy consumption profile assumes that the power consumption of each base station is a constant quantity that is independent of the variation of the traffic load. We assume that there are two power consumption states of base

Table 6.1 Switching time points

	τ_1	τ_2
Weekdays scenario		
Cluster1	1:27	8:10
Cluster2	20:45	9:00
Cluster3	00:48	8:18
Weekends scenario		
Cluster1	0:34	8:30
Cluster2	23:08	8:25
Cluster3	0:50	8:22

station: 0 and P_{0c}. An energy saving approach based on the average traffic pattern is proposed in [5] as follows:

- Let *f*(t) be the average traffic pattern in one day, which is described as a function of time *t*, with t ∈ [0, 24]. Normalize *f*(t) to the peak of 1, so that *f*(0) is the peak hour with the value *f*(0) = 1.
- Assuming that a fraction $x < 1$ of the base stations is in working mode while a fraction $1-x$ of the base stations is switched off during a period when the traffic load is below certain threshold.
- Let τ_1 and τ_2 be two time points of the switching-off period, with $f(\tau_1) = f(\tau_2)$. The optimization is given by (6.10)

$$\begin{array}{ll} \min & C(\tau_1, \tau_2) = P_{0c}[T - (\tau_2 - \tau_1) + f(t_1)(\tau_2 - \tau_1)], \\ \text{s.t.} & f(\tau_1) = f(\tau_2), \end{array} \tag{6.10}$$

where C(τ1, τ2) represents the average energy consumed per base station in a day under this approach and T reflects the 24 h in one day .

We apply this approach to the different traffic patterns of the three clusters to control the on–off states of the base stations separately. It provides us with more precise calculations on the energy savings. We determine the best τ_1 and τ_2 with the biggest energy saving ratio using the optimal method, as is in Table 6.1.

Then we calculate the total energy savings of the improved approach by (6.11)

$$\text{Net}_{\text{saving}} = 1 - \frac{C(\tau_1, \tau_2)}{P_{0c}T}. \tag{6.11}$$

We have introduced some simplifying assumption about the network environment. For example, the covering radius of active base stations could be extended to tackle the entire service requirement. However, some problems may emerge due to the extension of base station coverage, such as same frequency interferences. As a result, certain technologies should be employed to protect against these problems, for example, smart antenna is a new technology that can reduce the interferences of cellular networks by adjusting the antenna configurations. As a result, the coverage areas of the base stations are reorganized to avoid interferences in the overlapping coverage areas.

Table 6.2 Average energy saving ratios

	Energy saving ratio (%)	Proportion in number (%)
Weekdays scenario		
Cluster1	25.24	41.12
Cluster2	38.37	19.73
Cluster3	25.08	39.15
Total	27.77	100
Weekends scenario		
Cluster1	23.68	38.95
Cluster2	31.13	15.85
Cluster3	24.52	45.20
Total	25.24	100

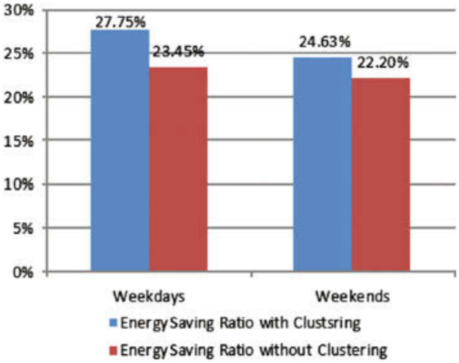


Fig. 6.4 Comparison of performance of two approach in energy saving

6.4 Simulation Results

We applied the energy saving method to the scene of real network and accessed the energy saving ratios of our method on different conditions. Analysis results are summarized in Table 6.2. We observe different energy saving levels of each cluster. We then inspect the benefits of the approach based on cluster analysis compared to those without clustering, which are shown in Fig. 6.4. Here we can see that an additional part of about 3–4% is saved if we introduce the clustering analysis to energy saving algorithm.

6.5 Conclusions

We proposed an energy saving approach based on cluster analysis. The cluster analysis on original data from an operating cellular network in a big city exploits different traffic patterns of base stations. In addition, specific methods are applied to these different traffic patterns. Simulation results indicate a better energy saving performance based on cluster analysis, compared to the existing approach.

Acknowledgments This work is support by the National Basic Research Program of China (973 Program) under grants 2007CB307100 and 2007CB307105.

References

1. Hérault L, Strinati Calvanese E (2010) Holistic approach for future energy efficient cellular networks. Elektrotechnik and Informationstechnik 127:314–320
2. Oh E, Krishnamachari B, Liu X, Niu Z (2010) Towards dynamic energy-efficient operation of cellular network infrastructure. IEEE Commun Mag
3. Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: explorations in urban data collection. Pervasive 6(3):30–38 Sept
4. Dufková K, Bjelica M, Moon B, Kencl L, Le Boudec, J-Y (2010) Energy savings for cellular network with evaluation of impact on data traffic performance. Wireless Conference (EW), European, pp 916–923
5. Marsan MA, Chiaraviglio L, Ciullo D, Meo M (2009) Optimal energy savings in cellular access networks. In: Proceedings of GreenComm

Chapter 7
A New Intelligent Model for Short Time Traffic Flow Prediction via EMD and PSO–SVM

Gu Yue-sheng, Wei Ding and Zhao Ming-fu

Abstract Accurate and reliable short time traffic flow forecasting is one of the most important issues in the traffic information management. Due to the nonlinear and stochastic of the data, it is often difficult to predict the traffic flow precisely. Hence, a new hybrid intelligent forecasting approach based on the integration of empirical mode decomposition (EMD), particle swarm optimization (PSO) and support vector machine (SVM) is proposed for the short time traffic flow prediction in this paper. The advantages of the proposed method are that the combination of EMD and PSO–SVM can deal with the nonlinear and stochastic characteristics of the original data well. The forecasting rate may be enhanced using this new technique. Seven-hundred and twenty samples of the practical traffic flow data were applied for the validation of the proposed prediction model. The analysis results show that the proposed method can extract the underlying rules of the testing data and improve the prediction accuracy by 10% or better when compared to SVM approach. Thus, the new EMD–PSO–SVM traffic flow forecasting model provides practical application.

Keywords Traffic flow · Short time prediction · EMD · PSO · SVM

7.1 Introduction

Traffic flow prediction has become a popular research topic in the field of intelligent transport system. Accurate and real-time traffic flow prediction is the key fact for traffic control and traffic flow guidance. Short-term traffic flow prediction can

G. Yue-sheng (✉) · Z. Ming-fu
Henan Institute of Science and Technology, Xinxiang 453003, China
e-mail: hz34567@126.com

W. Ding
NanYang institute of technology, Nanyang 470004, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_7,

forecast about the traffic flow state of the next several minutes to provide real-time effective information for travelers, realize the dynamic route guidance, save travel time, relieve the traffic congestion, reduce pollution, save energy and other purposes [1]. Therefore, it is imperative to implement short time traffic flow prediction.

Some mature prediction models have been applied to short time traffic flow prediction since the 1960s. In general, they can be divided into four kinds of prediction methods, i.e., traditional statistical theory based approach, neural network-based method, nonlinear theory-based and emerging technologies-based strategies. Due to the fact that the traffic system is a complex nonlinear system with time-varying and high uncertainties, the prediction accuracy of existing models can not be satisfactory. The forecasting accuracy needs to be improved for real practice applications [1–7]. Since the integration of different analysis techniques can provide better performance than independent use, a new hybrid approach to short-term traffic flow prediction based on empirical mode decomposition (EMD) and artificial intelligence is proposed in this work. This method has been marked with the advantages of the good nonlinear signal process ability of the EMD and the powerful learning ability of the support vector machine (SVM). Meanwhile, to achieve the structural parameter optimization of the SVM, the particle swarm optimization (PSO) is employed to obtain good generalization ability of the prediction model. By using the practical dataset for experimental analysis, the results show that the new method can predict short-term traffic flow effectively and the prediction rate is higher than the independent use of the SVM.

This paper is organized as follows: in Sect. 7.2, the proposed method for short time traffic flow forecasting based on the combination of EMD, PSO and SVM is described. The application of the proposed method is presented for short time traffic flow forecasting in Sect. 7.3. The performance of nonlinear signal process using EMD is described. The effectiveness of the proposed method is valued by analyzing the real traffic data. Conclusions are drawn in Sect. 7.4.

7.2 Hybrid Intelligent Model

Due to the interference of internal and external excitations, the short-term traffic flow is a kind of typical non-stationary signal. The different signal components of short-term traffic flow exhibit various characteristics, and produce different effects under the influence of change trend of traffic flow. The general trend of traffic flow is determined by deterministic signal, and the uncertain interference signals make the actual traffic flow present fluctuations near the general trend. EMD is a new approach to deal with nonstationary signal. According to different scales of fluctuation, non-stationary signal is decomposed step by step into some (Intrinsic mode functions) IMFs. Each IMF includes signals of different bands from high to low, and has its unequal features. Moreover, the adaptive decomposition is presented by EMD based on the inherent characteristics of the signal [8–11]. Therefore, these different modes can reflect the essence and potential rule of the traffic flow more clearly.

In the analysis of short-term traffic flow, the EMD is first used to decompose the actual traffic flow to remove the disturbance signals. Then, the predicted mode for each IMF is established using SVM, and the PSO is applied for the model optimization. Finally, the traffic flow is obtained by adding up the predictive values of each SVM models.

7.2.1 Empirical Mode Decomposition

EMD [10] is a useful advanced signal processing technique for the analysis of nonlinear signals. EMD has the ability to decompose a signal into a number of monocomponent signals, named as IMFs [10]. IMFs represent simple oscillatory modes embedded in the signal [11]. An IMF is a function that satisfies the following definitions [11]:

(1) In the whole analysis dataset, the number of extrema and the number of zero-crossings must either equal or differ at most by one.
(2) At any point, the mean value of the envelope defined by local maxima and the envelope defined by the local minima is zero.

To extract IMFs from a signal x, all the local extrema are first identified. Then, a cubic spline line connects all the local maxima as upper envelope and all the minima as lower envelope. The mean of upper and lower envelope is subtracted from x to obtain h_1. Checking h_1 for the IMF conditions, if it satisfies the conditions it is an IMF, otherwise upper and lower envelopes are found for the h_1 and the process is repeated till the first IMF c_1 is obtained. Subtracting c_1 from x and the result obtained is now treated as new original signal and the above process is repeated to get the second IMF. The process is continued till no more IMF can be extracted. Thus, at the end of the EMD decomposition we obtain

$$x = \sum_{i=1}^{N} c_i \, + \, r_N, \tag{7.1}$$

where, r_N is the final residue and c_i (i = 1, 2, …, N) is the ith IMF.

7.2.2 PSO–SVM

Since there may be a certain correlation between the current state of the traffic data and the future state, which may be difficult to describe using analytical methods, the SVM [12] is applied to learn about this relationship. The SVM, which has the ability to find the decision function from low training set sizes, has been widely used as a learning algorithm in a wide variety of applications. The concept of the kernel trick allows SVM to perform regression and prediction even for nonlinear cases. In this paper, the SVM with RBF kernel is used for the traffic condition forecasting. Moreover, to improve the prediction model robust, the PSO [13] algorithm is adopted to optimize the SVM boundary parameter C and kernel parameter. The proposed forecasting processes are given as follows:

Step 1: pretreat traffic data to standardize the data format.

Step 2: extract nonlinear features from input data in the form of IMF by EMD.

Step 3: train SVM by each IMF, optimize SVM kernel parameters and boundary parameter by PSO, and sum each SVM model output to obtain the prediction result.

Step 4: test the performance of the SVM prediction model, and provide the test result as the basis for a valid traffic management decision. The following indices are selected for the evaluation performance of the prediction:

(1) Mean absolute error (MAE)

$$\mathrm{MAE} = \frac{1}{N}\sum_{t=1}^{N}\left|y_t - \widehat{y}_t\right|, \tag{7.2}$$

(2) Mean square error (MSE)

$$\mathrm{MSE} = \frac{1}{N}\sqrt{\sum_{t=1}^{N}\left|y_t - \widehat{y}_t\right|^2}, \tag{7.3}$$

(3) Mean absolute percent error (MAPE)

$$\mathrm{MAPE} = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \widehat{y}_t}{y_t}\right|, \tag{7.4}$$

(4) Mean square percent error (MSPE)

$$\mathrm{MSPE} = \frac{1}{N}\sqrt{\sum_{t=1}^{N}\left|\frac{y_t - \widehat{y}_t}{y_t}\right|^2}, \tag{7.5}$$

A flow chart of the proposed prediction method for short-term traffic flow is illustrated in Fig. 7.1.

7.3 Experimental Analysis

In order to validate the performance of the proposed algorithm, the traffic information is recorded for 5 days in real practice application in this paper. 720 data sets are prepared for the traffic forecasting procedure. 576 sample data of the first

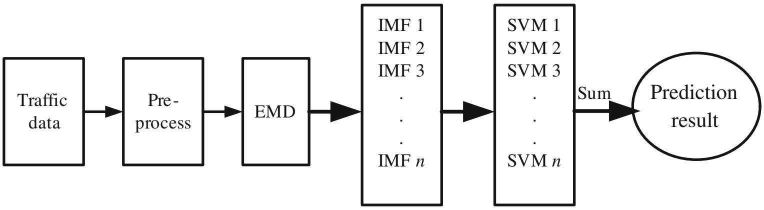


Fig. 7.1 The short-term traffic flow system based on EMD–PSO–SVM

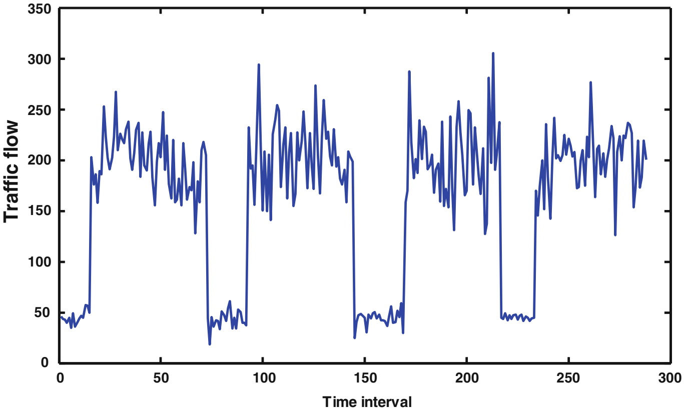


Fig. 7.2 The original traffic flow time series

4 days are used to train the prediction model, and the rest 144 samples are for test. A portion of the traffic flow time series is shown in Fig. 7.2.

As mentioned above, the EMD–PSO–SVM is proposed to forecast the traffic flow. The original data is first decomposed into 6 IMFs by EMD. The time spectra of the IMFs are shown in Fig. 7.3. Then, the PSO–SVM is used to get the prediction component of each IMF, and thus, their sum indicates the final short time traffic flow prediction value. Figure 7.4 gives the performance of the proposed method for traffic flow forecasting. We have compared the performance of the proposed method with the independent use of SVM. One can note that the proposed method has increased 10% of prediction accuracy compared with the SVM.

The prediction performances of the EMD–PSO–SVM and the PSO–SVM models are compared in Table 7.1. The comparison results show that the proposed method for short time traffic flow prediction is more effective than the PSO–SVM. By the EMD processing, the nonlinear elements are depressed and thus the forecasting error is decreased by 0.71% or better. One can note that the EMD plays an effective role in the improvement of short time traffic flow prediction.

The prediction performances of the EMD–PSO–SVM and the EMD–SVM models are compared in Table 7.2. It can be seen from Table 7.2 that the PSO

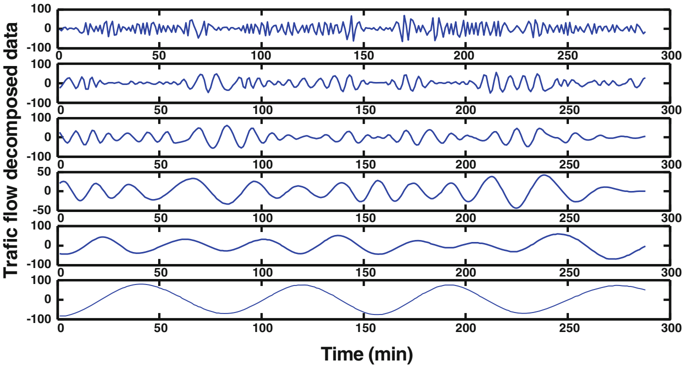


Fig. 7.3 The six IMFs of the original traffic data

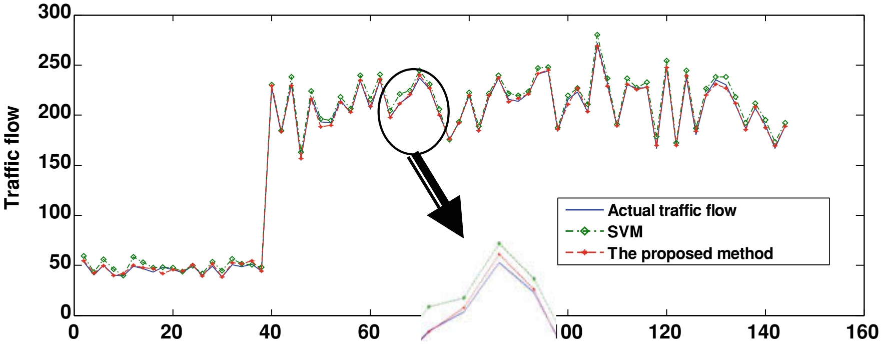


Fig. 7.4 The performance of the proposed method for traffic flow forecasting

Table 7.1 The traffic flow prediction results of EMD–PSO–SVM and PSO–SVM

EMD–PSO–SVM model				PSO–SVM model			
MAE	MSE	MAPE	MSPE	MAE	MSE	MAPE	MSPE
1.31%	2.02%	1.57%	1.66%	2.31%	2.73%	2.85%	2.61%

Table 7.2 The traffic flow prediction results of EMD–PSO–SVM and EMD–SVM

EMD–PSO–SVM model				EMD–SVM model			
MAE	MSE	MAPE	MSPE	MAE	MSE	MAPE	MSPE
1.31%	2.02%	1.57%	1.66%	2.57%	3.11%	3.27%	3.09%

optimization plays an important role in the improvement of the short time traffic flow prediction. With the PSO processing, the forecasting error is decreased by 1.09% or better.

7.4 Conclusions

Intelligent Transportation management relies on precise traffic flow forecasting. It is necessary to employ advanced data mining approaches to excavate the hidden knowledge of the traffic data. This paper presents a new hybrid intelligent model for the short time traffic flow forecasting. This new method combines the advantages of the nonlinear analysis of EMD and supervised learning of SVM to mine distinct and potential patterns of the traffic data. Moreover, the PSO algorithm is applied to optimize the SVM parameters. The experimental test results have proven that the presented prediction approach is feasible and efficient for short time traffic flow forecasting. The prediction rate of the proposed EMD–PSO–SVM is much better than the model with no EMD processing. Thus, the proposed method has application importance.

References

1. Zahra Z, Mahmoud P, Hossein SM (2010) Application of data mining in traffic management: case of city of Isfahan. In: Proceedings of the 2010 2nd international conference on electronic computer technology, vol 2010, pp 102–106
2. Nejad S, Seifi F, Ahmadi H, Seifi N (2009) Applying data mining in prediction and classification of urban traffic. In: Proceedings of the 2009 WRI world congress on computer science and information engineering, vol 3, pp 674–678
3. Zhao X, Jing R, Gu M (2008) Adaptive intrusion detection algorithm based on rough sets. J T Singhua Univ (Sci & Tech) 48:1165–1168
4. Li Z, Yan X, Yuan C, Zhao J, Peng Z (2011) Fault detection and diagnosis of the gearbox in marine propulsion system based on bispectrum analysis and artificial neural networks. J Marine Sci and Appl 10:17–24
5. Wen Y, Lee T (2005) Fuzzy data mining and grey recurrent neural network forecasting for traffic information systems. In: Proceedings of the 2005 IEEE international conference on information reuse and integration, vol 2005, pp 356–361
6. Hauser T, Scherer W (2001) Data mining tools for real time traffic signal decision support and maintenance. In: Proceedings of The IEEE international conference on systems, man, and cybernetics, Vol 3, pp 1471–1477
7. Park B, Lee D, Yun H (2003) Enhancement of time of day based traffic signal control. Proc IEEE Int Conf Syst Man Cybern 4:3619–3624
8. Luo X, Niu G, Pan R (2010) Short-term traffic flow prediction method based on EMD and artificial neural network. Comput Eng Appl 46:212–214
9. Huang NE, Wu ML, Qu WL et al (2003) Applications of Hilbert–Huang transform to non-stationary financial time series analysis. Appl Stoch Models Bus Ind 19:246–268

10. Huang NE, Shen Z, Long SR et al (1998) The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc R Soc Lond A 454:903–905
11. Parey A, Tandon N (2007) Impact velocity modelling and signal processing of spur gear vibration for the estimation of defect size. Mech Syst Signal Process 21(1):234–243
12. Vapnik V (1995) The nature of statistical learning theory, 1st edn. Springer, Berlin
13. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of IEEE international conference on neural networks, Piscataway, vol 1, pp 1942–1948

Chapter 8
Evolutionary Algorithm Based Power Distribution Optimization for Visible Light Communication

Zhitong Huang, Jupeng Ding and Yuefeng Ji

Abstract Visible Light Communications (VLC), which is expected to provide wireless communications and illumination facilities together, is earning increasing attention. However, it is hard to obtain satisfyingly uniform signal quality at the receiving terminal even for locations within the same room by Komine et al. in Asia-Pacific conference on communication (APCC 05), IEEE Press, Perth, pp 294–298 (2005). In this letter, an evolutionary algorithm-based optimization scheme is proposed as a candidate approach for VLC to reduce the variability of the received power. The presented results based around the use of a commercially available detector with a FOV $= 45^{\circ}$. show that the dynamic range of received power can be reduced to 34% against the peak optical power from 52% while the impact on illuminance function is negligible.

Keywords Visible light communication · Evolutionary algorithm · Illumination · SNR optimization

Z. Huang (✉) · J. Ding · Y. Ji
Key Laboratory of Information Photonics and Optical Communications (BUPT),
Ministry of Education, Beijing University of Posts and Telecommunications,
Beijing 100876, People's Republic of China
e-mail: hzt@bupt.edu.cn

J. Ding
e-mail: jupeng7778@163.com

Y. Ji
e-mail: jyf@bupt.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_8,

8.1 Introduction

Visible light communications (VLC) using illumination fixtures is earning increasing attention and research which is inspired by multiple inducements [1]. Solid-state lighting (SSL) has made dramatical advances in heightening brightness, improving electrical–optical power conversion efficiency etc. On the other side, highly congested radio frequency spectrum and high energy consumption make it almost impossible for the traditional radio frequency (RF) and 60 GHz communication systems to independently satisfy the constantly increasing demand for wireless data transmission, such as Mobile TV and CMMB.

Although multiple contributions to VLC have been proposed, overcoming the limitations imposed by multi-path transmission channel is still one of the most challenging aspects in the design of a VLC system. The channel characteristics are decided by the room size, the material properties of each surface the radiation is incident upon and the indoor objects, such that a single VLC mobile terminal may have obvious performance variation when implemented in different locations.

In our previous research [2], the quantitative analysis of dissatisfactory optical power has been provided. As an extension of previous work, a tailored evolutionary algorithm-based optimization scheme is proposed which controls the relative optical power of each LED to obtain even power coverage at all locations toward uniform system performance characteristics.

8.2 Indoor VLC System Model

8.2.1 *Diffuse Conventional Model*

In order to calculate the total power incident at the detector including the reflection portion, all the surfaces of the indoor environment are divided into a number of small reflecting surface elements, which are considered to have a pure Lambertian reflection characteristic. Up to third order reflections are taken into account and higher order reflections are ignored due to their small contribution. The system environment deployed in this work consist of an empty room with dimensions $x = 4$ m, $y = 8$ m, $z = 3$ m. (Fig. 8.1), the walls, ceiling and floor of the room have a reflectivity of $\rho_{\text{walls}} = 0.8$, $\rho_{\text{ceiling}} = 0.8$ and $\rho_{\text{floor}} = 0.3$.

8.2.2 *Received Signal Power*

In the adopted illumination configuration, $I = 21$ lighting lamps are evenly distributed over the ceiling. Each of the lamp is filled with 49 (7 × 7) LEDs. The space between LEDs is 4 cm. In this paper, it is assumed that an LED has a

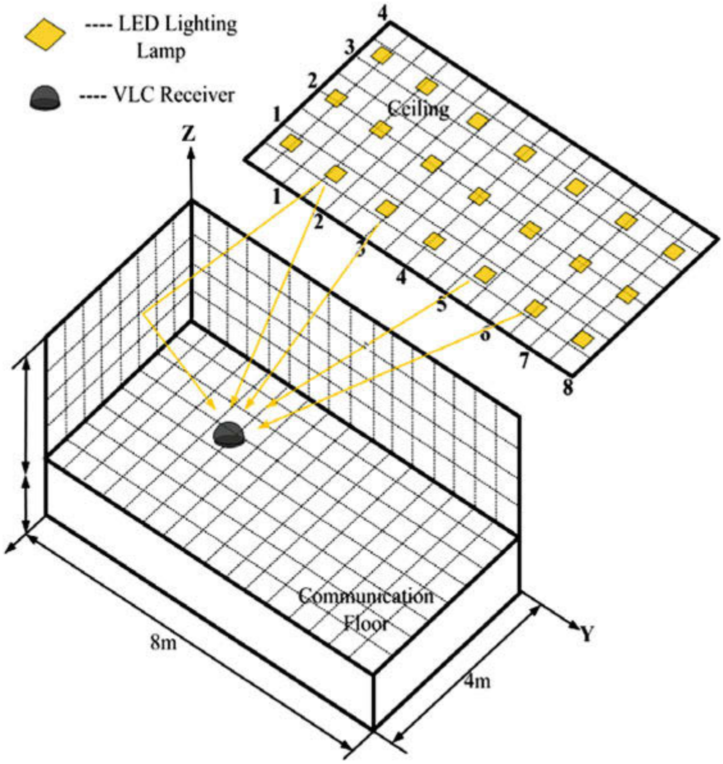


Fig. 8.1 Indoor VLC application scenario with 21 LED lighting lamps fixed on the ceiling

Lambertian radiation pattern [3], where the radiant intensity depends on the angle of irradiance. The half-power angle (HPA) and the center luminous intensity of an LED is $\Phi_{1/2} = 54\,°.$ and $I\ (0) = 32.69$ cd, respectively, measured in [3]. The analysis of received signal power is carried out with $J = 1{,}568$ receivers uniformly placed on the communication floor (CF), 1 m above ground. Each receiver orients vertically upwards, has a field of view (FOV) 45° and is with an active detection area $A = 1\ \text{cm}^2$.

Furthermore, the received power is given by the channel direct current (DC) gain on directed path$H_d(0)$, reflected path $H_{\text{ref}}(0)$ and transmitted optical power$P_t =\ 174$ mW.

$$P_r(R_j) = \sum_{i=1}^{I} \left\{ P_t H_d(0; S_i, R_j) + \int_{A_{\text{sur}}} P_t H_{\text{ref}}(0; S_i, R_j) \right\}. \tag{8.1}$$

where S_i is the ith LED lamp, R_j is the jth receiver and A_{sur} is the area sum of reflection surfaces. Respectively, the channel DC gain on directed path between S_i and R_j can be obtained from

$$H_d(0;S_i,R_j)=\begin{cases}\frac{(m+1)A}{2\pi D_d^2}\cos^m(\phi)T_s(\psi)g(\psi)\cos(\psi), & 0\le\psi\le\Psi_c\\ 0, & \psi>\Psi_c\end{cases}. \tag{8.2}$$

Where ϕ is the angle of irradiance, ψ is the angle of incidence, m is the Lambertian index ofS_i, A is the physical area of a detector in a photo diode, D_d is the distance betweenS_i and R_j, $T_s(\psi)$, is the gain of an optical filter and Ψ_c represents the field of view (FOV) for each receiver. And$g(\psi)$ is the gain of optical concentrator, where$g(\psi)=n^2/\sin^2(\Psi_c)$ for $0\le\psi\le\Psi_c$ and $g(\psi)=0$ otherwise.

In order to calculate the DC gain on reflected path, all the walls of the test room are divided into a number of differential elements with area dA and reflection coefficient ρ. The differential elements are viewed as generalized Lambertian sources, which emit diffusely into different directions from their center with a Lambertian pattern. In this paper, up to three order reflections are taken into account, so the DC gain on reflected path can be given by:

$$H_{\text{ref}}(0;S_i,R_j)=\sum_{k=1}^{3}\left(\sum_{l=1}^{N}H^{(k-1)}(0;S_i,\delta_l)H^{(0)}(0;\delta_l,R_j)\right). \tag{8.3}$$

where N is the amount of differential elements and k is the reflection order. The specific expression $H^{(0)}(0;\delta_l,R_j)$can be obtained according to Eq. 8.2 while the counterpart of $H^{(k-1)}(0;S_i,\delta_l)$can be given iteratively from

$$H^{(k-1)}(0;S_i,\delta_l)=\sum_{m=1}^{N}H^{(k-2)}(0;S_i,\delta_l)H^{(0)}(0;\delta_l,\delta_m). \tag{8.4}$$

where δ_l and δ_m stands for two independent differential elements.

8.2.3 Illuminance Characteristics

Another measure of VLC system quality is the horizontal brightness on the working surface. When no reflection is added, the horizontal illuminance is derived as follows: $E_h=I(0)\cos^m(\phi)\cos(\psi)/R^2$ where φ is the angle of irradiance, R is the angle of incidence and R is the distance between the source and the receiver surface [3]. The m is the Lambertian index, which is decided by the HPA of each LED as m = minus; ln2/ln (cos$\Phi_{1/2}$).

For estimating the effects of reflection on the total brightness distribution, the luminous flux of each reflective element has to be identified as a secondary source. The total luminous illuminance flux of each element can be calculated by: $F=\rho_e E_{\text{he}}A_e$ where $\rho_e$is the reflective index of the differential element, E_{he} is the surface illuminance of the differential element and A_e is the area of the reflective differential element such that, the overall horizontal illuminance of the CF can be given as:

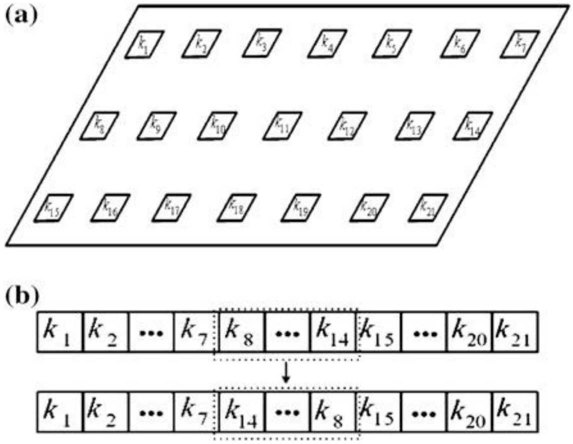

(a)
k_1 k_2 k_3 k_4 k_5 k_6 k_7
k_8 k_9 k_{10} k_{11} k_{12} k_{13} k_{14}
k_{15} k_{16} k_{17} k_{18} k_{19} k_{20} k_{21}
(b)
k_1 k_2 ••• k_7 k_8 ••• k_{14} k_{15} ••• k_{20} k_{21}
k_1 k_2 ••• k_7 k_{14} ••• k_8 k_{15} ••• k_{20} k_{21}

Fig. 8.2 **a** Non-optimized power distribution: Min.184 μW, Max.386 μW, Ave. 314 μW. **b** Optimized power distribution: Min.104 μW, Max.158 μW, Ave. 133 μW

$$E = E_h + \sum_{i=1}^{N} I_i \cos(\phi_i)\cos(\psi_i)/r_i^2. \tag{8.5}$$

where ϕ_i is the irradiance angle of the ith reflective differential element, ψ_i is the incidence angle from the ith reflective differential element, r_i is the distance from the ith reflective differential element to the illuminated surface and $I_i = (m_e + 1)F/2\pi$ is the maximal luminous intensity of the reflective differential element. In this paper, the overall reflective differential elements are viewed as pure Lambertian diffusers with $m_e = 1$.

8.3 Proposed Evolutionary Algorithm

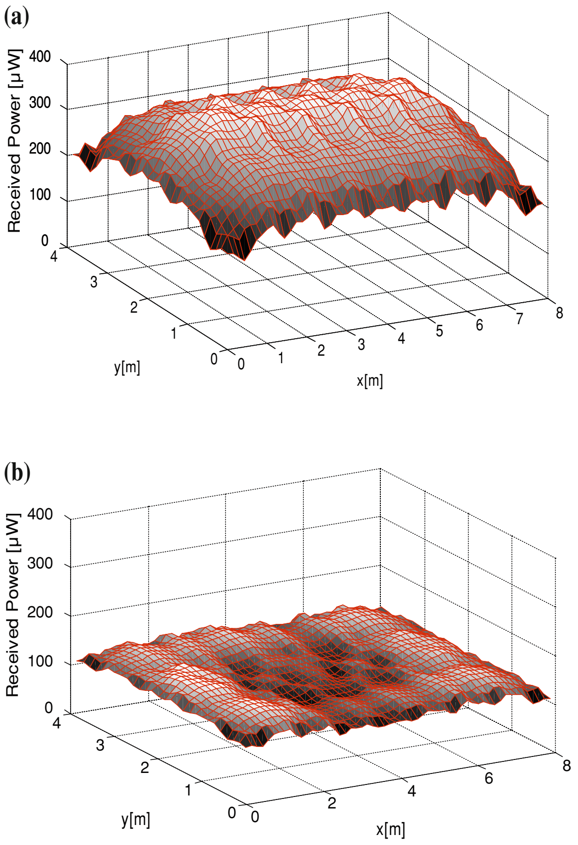
8.3.1 Optimization Factors

Due to the limitation imposed by the complex VLC channel characteristics, it is quite different for the performance of different receiver located on the CF, as shown in Fig. 8.2. On the other side, the use of intelligent technique has proven to be beneficial in mitigating some of these constraints. Assuming that all LEDs lamps are individually scaled by a factor $0 < k_i \leq 1$(called optimization factor), the instantaneous signal optical power got at a given receiver can be rewritten as:

$$P_r(R_j) = \sum_{i=1}^{I} \left\{ k_i P_t H_d(0; S_i, R_j) + \int\limits_{A_{\text{sur}}} k_i P_t H_{\text{ref}}(0; S_i, R_j) \right\}. \tag{8.6}$$

Such that it is possible to find a set of factors $k_1 k_2 \cdots k_I$, which will allow the J receivers to obtain the same or very close optical power as: $P_r(R_1) \approx P_r(R_2) \approx P_r(R_3) \approx \ldots \approx P_r(R_J)$. The respective relationship between the optimization

Fig. 8.3 **a** A set of optimization factors applied to the LED lamps. **b** Mapping relationship between genotype structure and optimization factors



factors and the LED lamps is shown in Fig. 8.3a. And it can be seen that optimization factors k_7 and k_{14}, k_8 and k_{15} are physically close to each other in application, but far apart in the chromosome as in the top of Fig. 8.3b. Alternatively, a modified structure, as in the bottom of Fig. 8.3b, is applied to alleviate this issue. Specific modification is that optimization factors $k_8, k_9, \ldots, k_{14}$ are placed in a descending order to construct the genotype structure.

8.3.2 Objective Function

While the chromosome structure sets the relationship between the optimization factors and the genes, the objective function, or fitness function, is used to connect the optimization aim and the phenotypic appearance. The objective function of the proposed evolutionary algorithm is given by:

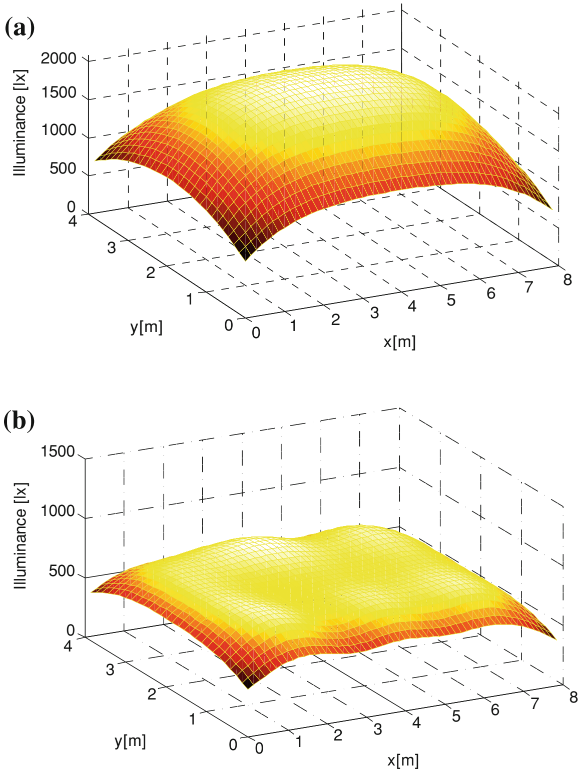


Fig. 8.4 **a** Non-optimized illuminance distribution: Min. 713 l×. Max. 2087 l× Ave. 1592 l×. **b** Optimized illuminance distribution. Min. 365 l×. Max. 780 l× Ave. 664 l×

$$O(\alpha_n) = \left(100 - 100\left(\frac{\max P_r(R_j) - \min P_r(R_j)}{\max P_r(R_j)}\right)\right)\%. \tag{8.7}$$

where $\max P_r(R_j)$ and $\min P_r(R_j)$ are the maximum and minimum of the received optical signal power on the CF after application of an individual, such that the individuals of the population that provided the most uniform power coverage are given the highest chance of selection for reproduction into the next generation.

8.3.3 Optimization Operators and Termination

For reducing the complexity, we adopt the mature operators, including the roulette selection operators, the double point crossover operator and the mutation operator to implement the algorithm [4, 5]. And to each gene, we maintain the crossover rate $\rho_c = 0.7$ and the mutation rate $\rho_m = 0.05$ of a population of 200. Based on the fast convergence characteristic of evolutionary algorithm, the optimization procession is allowed to run for 5,000 generations before the latest fittest individual is decided.

8.4 Performance Evaluation

Using our established environment and VLC system design, the performance of the proposed algorithm is evaluated by simulation. Figure 8.2a shows the received signal power at each receiver location, varying between 184 and 386 nW, respectively, equating to 202 nW, or 52% power deviation from the peak. The horizontal illuminance varies between 713 and 2087 1×, as shown in Fig. 8.4a. Upon the application of the tailored evolutionary algorithm, the power now ranges from 104 to 158 nW equating to a 44 nW, or 34% power deviation, as seen in Fig. 8.2b. Under the influence of the optimization the horizontal illuminance now varies between 365 and 780 1×, as shown in Fig. 8.4b. From this figure, sufficient illuminance, 300−1500 1× by ISO, is still obtained at all the places of the CF.

8.5 Conclusion

In this work, a tailored evolutionary algorithm is proposed to control the transmitters (i.e. LED lighting lamps) to optimize the received optical signal power distribution. The simulation results presented show that this algorithm is capable of reducing the dynamic range of the power distribution with the affect to the illumination function of VLC system being negligible.

Acknowledgments This research was supported by NSFC (No. 60932004), the National 973 Program (No. 2011CB302702), National 863 Program (No. 2009AA01A345) and Fundamental Research Funds for the Central Universities (BUPT2009RC0126), P. R. China.

References

1. Delgado F, Quintana I, Rufo J, Rabadan JA, Quintana C, Jimenez RP (2010) Design and implementation of an Ethernet-VLC interface for broadcasting transmission. IEEE Commun Lett, 14(12):1089–1091
2. Ding J, Huang Z, Ji Y (2010) Independent reflecting element interaction characterization for indoor visible light communication based on new generation lighting. Chinese Opt Lett 8(11):1182–1186
3. Komine T, Lee JH, Haruyama S, Nakagawa M (2005) Adaptive equalization for indoor visible-light wireless communication systems. In: Asia-Pacific conference on communication (APCC 05), IEEE Press, Perth, pp 294–298
4. Back T, Hammel U, Schwefel HP (1997) Evolutionary computation: comments on the history and current state. IEEE Trans Evol Comput 1(1):3–17
5. Vucic J, Kottke C, Nerreter S, Langer KD, Walewski JW (2010) 513 Mbit/s Visible light communications link based on DMT-Modulation of a white LED. IEEE J Lightwave Technol 28(24):3512–3518

Chapter 9
XMPP-Based Solution for Accessing Server with Dynamic IP

Jianyi Wang, Zhiqiang Ma, Bo Li, Yiran Guan, Xianyi Liu and Guifu Yang

Abstract This Chapter presents an XMPP-based solution for accessing a server of which the IP is dynamic or private. Two XMPP clients, one that stays with the server together inside the Intranet and another one is on Internet with the client that is required to access the server, retrieve and provide the IP address of the server, so that the client side can initiate a TCP communication channel using the server IP as target address. This solution also supports the server has not public IP, or creating VPN-like channel.

Keywords XMPP protocol · Dynamic IP · Server

Supported by the Research Foundation of Science and Technology Department of Jilin Province, China (No. 20100172).

J. Wang (✉) · Z. Ma · Y. Guan · X. Liu · G. Yang
School of Computer Science and Information Technology, Northeast Normal University,
Changchun 130117, People's Republic of China
e-mail: mygaara@gmail.com

Z. Ma
e-mail: mazq@nenu.edu.cn

Y. Guan
e-mail: guanyr@gmail.com

X. Liu
e-mail: liuxy506@gmail.com

G. Yang
e-mail: guifu.yang@gmail.com

B. Li
College of Electronic Science and Engineering, Jilin University,
Changchun 130012, People's Republic of China
e-mail: libo086@gmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_9,

9.1 Introduction

Along with the exhaustion of IPv4 address, Dynamic Host Configuration Protocol (DHCP) and Network Address Translation (NAT) technique are using widely. There methods relieve the pressure of lacking IP address by using IP address rotating, but in the same time, they prevent computers from being servers natively [1–3].

As a DHCP [4] client, the computer can retrieve a dynamic IP for every time it asks when it needs to connect the Internet. Meanwhile, the address IP cannot be kept always. After an expired duration or the computer is off-line, it should release the IP. Another time when the computer asks an IP address, it can retrieve one, but it can not be sure the same address. Without a static IP address, the server can not be reached by any computer around the Internet to initialize a connection that needs the server's IP address.

Even worse, the computer after a NAT server has not a public IP. The service in such computer maps to a public TCP/UDP port in the NAT server. When a guest computer in Internet accesses the public port, the requirement is delivered to a port on the inner IP of that computer inside a private network behind the NAT server. If the NAT server itself has not a static IP, the situation turns to be more complex. Such difficulty happens when a NAT is served for a small business in China Communication network.

DDNS (Dynamic DNS) [5] helps to resolve this problem in wireless and other scenarios [6, 7]. This solution needs the NAT server or the computers behind it pay for registration on the supplier. As soon as the DDNS client connects to the Internet and get an IP address, it ask the application program installed on it to notify the supplier's registration server its IP address currently. Then the registration server modifies the records on DNS server. Thus, the new lookups requirements is mapping to current IP address. Once the IP changing, the DDNS client sends another message to modify the DNS records. This solution needs pay and it depends on installing a proprietary application program.

In this paper, we present a free charge solution without installing proprietary to permit a computer without static IP acting as a server. This solution is based on XMPP protocol, a free Internet instant chatting protocol, so we introduce XMPP briefly as a background before the method details is presented.

XMPP is a standard RFC3920 from IETF, as one of the most important four instant message protocols [8]. It is free open public protocol, and there are many clients running on the Internet. No center server is needed, so that the servers up to the standard of XMPP can connected together with the XMPP protocol. Therefore our solution has choices widely for XMPP clients.

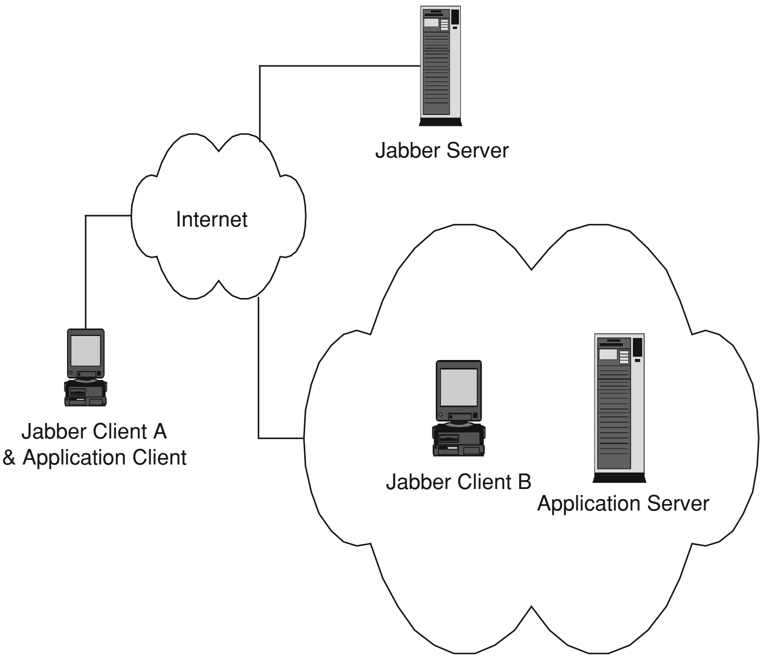


Fig. 9.1 Structure of network

9.2 Method

We present the solution using an example network with an application server and a XMPP client needed. The IP address of the application server can be exposed with the XMPP client, therefore the application server can be accessed this IP address.

Structure of Network. Structure of Network is shown in Fig 9.1. A guest computer (singed as XMPP Client A & Application Client) required to access the application server is on the other side of Internet. A XMPP client is installed on it. There are no more requirements about the XMPP client. It can be any one downloadable on Internet, or it can be built-in the application discussed that needs the application server IP address to initialize a connection.

A XMPP server is running on the Internet, not controlled by our solution. This server provides normal XMPP service, such as registration, id verification, etc.

A computer or program acts as a XMPP client, signed as XMPP Client B. It is in the same intranet cloud with the application server, in which way it is implied that XMPP Client B can obtain the IP of Application Server. They can run in same hardware machine, or XMPP Client B can consult an IP table that storing Application Server has stored in. The ways are many which are not discussed in details in this paper.

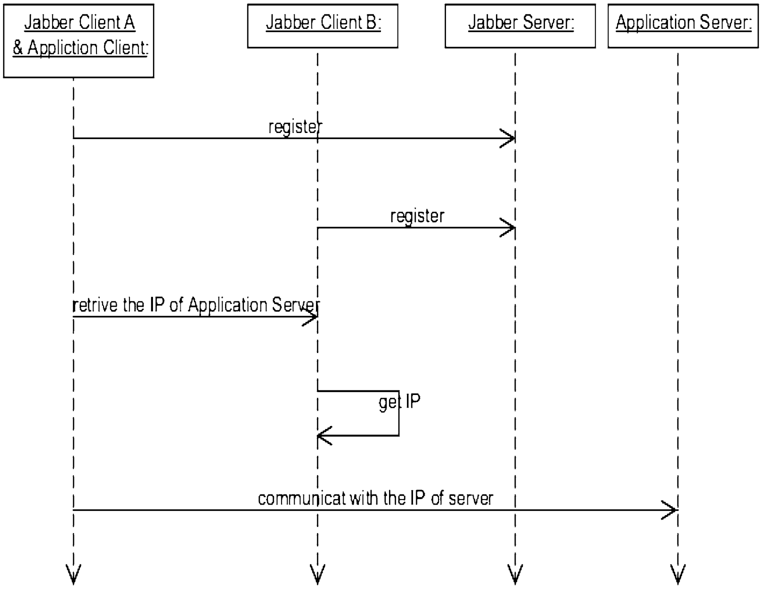


Fig. 9.2 The sequence of communication

The Application Server is asked to be a server, but it has not a static IP address. A dynamic IP has been arranged to the application server, and the IP may be rearranged at any time. For discussing easily, we assume that the IP address is a public one. Later, we will introduce the way to resolve the problem having not a public IP.

Process of Communication. Before all of the processes are required, the Application Server has obtained an IP address. Then the sequence to create the connection of communication is shown in Fig 9.2.

Step 1, XMPP Client A, which is Application Client as well, registers itself to XMPP Server with a certain ID. Any computer on Internet has the ability to access a XMPP server even without a public IP.

Step 2, XMPP Client B, which is in the intranet, registers itself to XMPP Server with another ID. The XMPP server that serves XMPP Client A and The one that servers XMPP Client B are not necessary to be the same. The solution can work so long as they are connected with XMPP protocol. It is neither necessary that XMPP Client B has a public IP.

Step 3, XMPP Client A sends a message to XMPP Client B. In the message, a protocol between XMPP Client A and XMPP Client B is defined to specify some commands from XMPP Client A asking the opposite side to execute something. In this situation, XMPP Client A asks XMPP Client B to retrieve the IP address of Application Server and return it to XMPP Client A. The command can be binary if it is encode with base64. In this context, it is “/IP” in plain text with a slash prefix.

Step 4, based on some kind of mechanism, XMPP Client B knows the IP of Application Server, so that it can get the IP with a private function without communicating with Application Server. As an easy way, XMPP Client B can run in the same machine with Application Server, therefore they have a same IP address. But it is not necessary, and any method to retrieve the IP is acceptable for this solution.

Step 5, with the IP retrieved, Application Client, which is XMPP Client A as well, can create the initial connection to Application Server.

After the process based on XMPP protocol, the program Application Client knows the IP of Application Sever, therefore it can initiate the normal communication with Application Server.

9.3 Applying Cooperated with Other Technique

Additionally, some more complex scenario can be supported by this solution.

Application Server is behind a firewall. If Application Server is behind a firewall, the private IP of Application Server can be mapped to a NAT server's public port. Once the public port and the public IP of NAT server for Application Server is sent to XMPP Client A, which is Application Client, the communication channel can be built by the knowledge of two sides of public ports and two side of public IP addresses. What is need further more is a new command to retrieve the public port. This command is no difference with retrieving public IP address in the answering mechanism.

XMPP Clients act as a reverse proxy. XMPP Client B can cat as a reverse proxy in the same machine with Application Server. In this way, any computer outside in the Internet can access XMPP Client B instead of accessing Application Server directly. XMPP Client A acts a proxy for Application Client, so that the communication process can be create and kept automatically.

If the XMPP clients are wrapped with a LAN adapter driver program, the users on Application Server and Application Client consider the communication channel as VPN channel, just like no XMPP protocol exist, and it is not necessary of setting XMPP ID.

9.4 Conclusions

Two XMPP clients help retrieving the IP of an application server to create a communication channel between any computer on Internet and a server without static IP.

This solution works when the server behind a NAT server has no public IP. It also supports to create a VPN-like channel invisible to the final user.

Acknowledgments The authors are most grateful to State Key Laboratory of Electroanalytical Chemistry, Changchun Institute of Applied Chemistry, Chinese Academy of Sciences for the cooperation with the experiments. The authors also wish to express thanks to Dr. Li Niu for the valuable discussion and recommendation.

References

1. Comer D (2006) Internetworking with TCP/IP, 5th edn. Pearson Prentice Hall, Upper Saddle River
2. Tanenbaum S (2003) Computer Networks, 4th edn. Prentice Hall PTR, Upper Saddle River
3. Comer D, Stevens DL (2001) Internetworking with TCP/IP. vol III Client-server programming and applications, Linux/POSIX sockets ed. Prentice Hall, Upper Saddle River
4. Cepeda and International Business Machines Corporation, International Technical Support Organization (2000) Beyond DHCP: work your TCP/IP internetwork with dynamic IP, 2nd edn. IBM, Austin
5. Wikipedia, Dynamic DNS, http://en.wikipedia.org/wiki/Dynamic_DNS. Accessed 14 Sept 2010
6. Hershkovitz S et al (2010) System and a method for remote monitoring customer security systems, ed: Google Patents
7. Weng C (2010) Design and implementation of a self-adaptable networked DVR
8. Saint-Andre EP (2004) Extensible messaging and presence protocol. Jabber Software Foundation 10

Chapter 10
Development of New Test Bench for Vacuum Ejector and Vacuum Regulator Based on ISO Standard

Chong Liu, Weiqing Xu and Maolin Cai

Abstract In this paper, we developed two new test benches for vacuum ejector and vacuum regulator, which based on ISO standards, are suited for the current status of the domestic production of vacuum components. These benches have a simple structure, can be used easily and are designed with a more reasonable circuit. The high-response flow transducer used in the circuit was designed on pneumatic power to make the result more accurate. The bench, not only provides a new platform for characteristic testing of vacuum components, but also greatly facilitates the standardization of the production of vacuum components.

Keywords Vacuum ejector · Vacuum regulator · Test circuit · Lab windows/CVI

10.1 Introduction

With the development of modern industry, pneumatic products have developed rapidly with their unique advantages. Today, in industrially technologically developed countries in Europe and America, the ratio of using hydraulic and pneumatic products has reached 6:4, while in Japan the ratio it reached 5:5 more.

C. Liu (✉) · W. Xu · M. Cai
School of Automation Science and Electrical Engineering,
Beihang University, Beijing 100191, China
e-mail: lcyj120206@163.com

W. Xu
e-mail: xwqyyl@163.com

M. Cai
e-mail: caimaolin@buaa.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_10,

Meanwhile, SMC, FESTO, NORGREN and other large pneumatic product manufacturers, and more than ten million kinds of R & D products [1] have emerged.

With the development and progress of technology in our country, a number of inventions such as STNC, JELPC, ABC and other famous manufacturers have also emerged. However, China's air companies are currently stuck at the stage of imitation of foreign products, and they have weak research and development, lower innovation ability, and limited means of processing which resulted in poor product quality, low prices and uncompetitiveness. Meanwhile the lack of appropriate testing methods and test equipment is a critical factor that limits the quality of products to improve to the higher level. For this, Beihang University joined with the State Bureau of Quality and Technical Supervision to develop a number of new test benches for pneumatic components.

The test benches for vacuum ejector and vacuum regulator are some of the successful standard benches. The standard test platform that combined with the new ISO 6358 standards revisions, are based on the basic characteristics, using the most advanced sensor—the high-response flow transducer.

10.2 Test Objects and Their Properties

10.2.1 Vacuum Ejector and its Characteristics

The device used to complete the characteristic test of the vacuum ejector include the exhaust characteristics and vacuum degree-inhalation flow characteristics.

The exhaust characteristic of the vacuum ejector, represents the relationship among the maximum degree of vacuum, air consumption, maximum inspiration flow and the supply pressure. Maximum vacuum degree is the vacuum degree when the vacuum port is fully closed. Air consumption is the flow through the supply of irrigation. (ANR). Maximum inspiration flow is the flow from the vacuum suction port(ANR) when the vacuum port is open to the atmosphere. Vacuum degree-inhalation flow characteristic is the relationship between the vacuum degree and the inhalation flow, when the supply pressure is 0.45 MPa and the vacuum port is under the status of changing and not closed [2].

10.2.2 Vacuum Regulator and its Characteristics

The device used to complete the characteristic test of the vacuum ejector include the flow characteristic, pressure characteristic and vacuum pressure-input signal characteristics.

The principles of flow characteristics are: the cut-off valve set on the upstream of the mouth sets the pressure of the regulator, adjusts the pressure of the regulator to an initial setting pressure when the cut-off valve is closed; and flow characteristic is the relationship between the inhalation flow and the setting vacuum degree.

The principle of pressure characteristic: from a set point, the output pressure of electrical-pressure proportional valve of the vacuum side fluctuates within a certain range, the variation of setting side pressure caused by the pressure change of vacuum side is the pressure characteristic of the regulator.

The principle of vacuum pressure—input signal characteristic: close the setting side cut-off valve, control the vacuum regulator to open gradually from zero to maximum, then close it in the same way; the relationship of the signal between the vacuum pressure and the vacuum regulator's input signal is the vacuum pressure-input signal characteristic.

10.3 Structure of the Test System

10.3.1 Mechanical Structure

The performance test platform consists of two sections, stainless steel, sheet metal, experimental table which is made of common structures. Electromagnetic valve, electrical—flow proportional valve, electrical-pressure proportional valve, high-response flow transducer developed according to pneumatic power and other components form a standard test circuit connected by the hose.

In order to prevent noise, we designed the special back-cover, meanwhile designed some dedicated quickly plug port, which make the connection among the components conveniently, the test circuit picture same to the fact connection support the test operator more convenient. The hardware of the test bench is IPC supported by ADVANTECH. Together with the 22-inch display makes the test curve clearly easy to read. The band of the test procedure's introduction which is below the LED display reminds the operator to test according to the steps.

These two test benches construct compactly, operate easily, with a cleaner appearance, modern and strong, as shown in Fig. 10.1.

10.3.2 Test Circuit and Test Procedures

Test circuit and test procedures of Vacuum ejector [3]

The test circuit of vacuum ejector is as shown in Fig.10.2.

The test steps of the vacuum ejector's exhaust characteristic are:

Control Solenoid valve (E) through the circuit, regulates the opening size of the electrical-flow proportional valve (M), making the flow reach the settings.

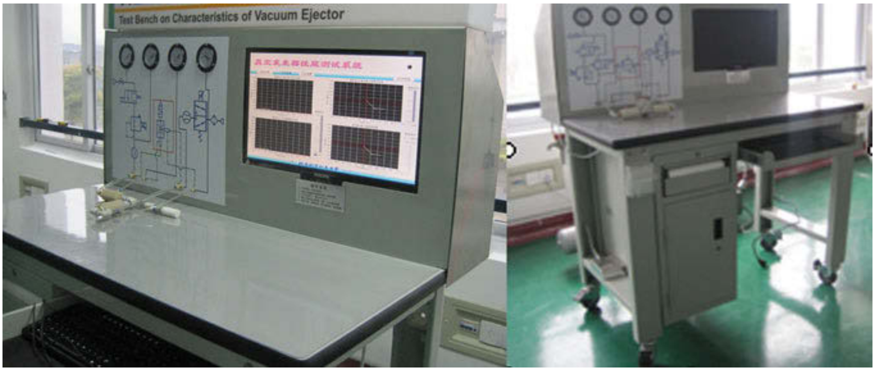


Fig. 10.1 Test benches on characteristics of vacuum ejector and vacuum regulator

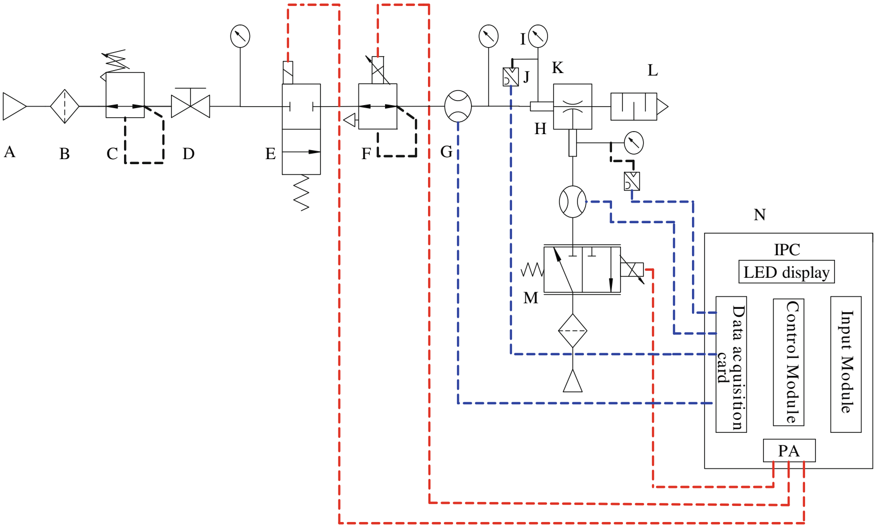


Fig. 10.2 Test circuit of vacuum ejector. *A*-gas source; *B*-filter; *C*-regulator; *D*-cut-off valve; *E*-solenoid valve; *F*-electrical-pressure valve; *G*- high-response flow transducer; *H*-pressure measuring tube; *I*-pressure gauge; *J*-pressure sensor; *K*-vacuum ejector; *L*-silencer; *M*-electrical-flow proportional valve; *N*-control center

Set the output air pressure range of the electrical-pressure proportional valve (F). Start to test, IPC (N) control the change of the electrical-pressure proportional valve's (F) pressure automatically, meanwhile, collects dates, displays, saves and generates the reports automatically after the test.

The test steps of vacuum degree-inhalation flow characteristic are:

Control Solenoid valve (E) through the circuit, regulates the electrical-flow proportional valve (F), setting the output pressure on 0.45 MPa.

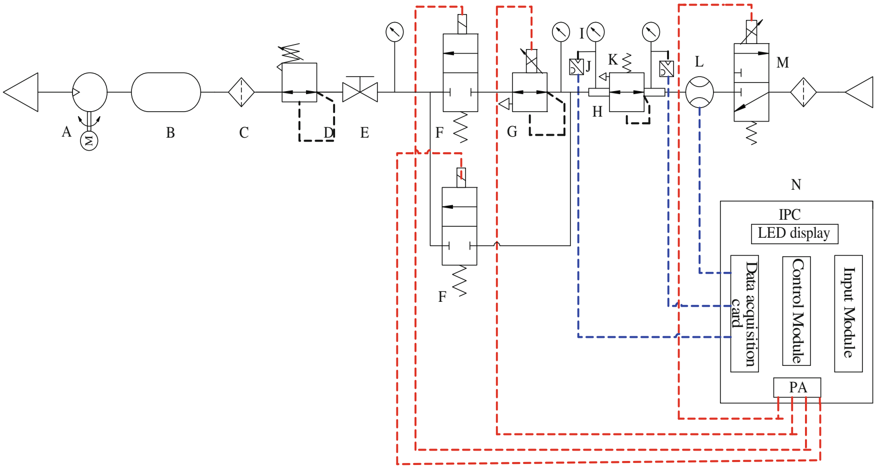


Fig. 10.3 Test circuit of vacuum regulator. *A*-vacuum pump; *B*-cylinder; *C*-filter; *D*-regulator; *E*-switch; *F*- solenoid valve; *F*-electrical-pressure valve; *G*- high-response flow transducer; *H*-pressure measuring tube; *I*-pressure gauge; *J*-pressure sensor; *K*-vacuum regulator; *L*- high-response flow transducer; *M*-electrical-flow proportional valve; *N*-control center

Set the output flow range of the electrical-flow proportional valve (M).
Start to test, IPC (N) controls the change of the electrical-flow proportional valve's (M) flow automatically, meanwhile, collects dates, displays, saves and generates the reports automatically after the test.

10.3.3 Test Circuit and Test Procedures of Vacuum Regulator

The test circuit of vacuum regulator is as shown in Fig. 10.3 [4].

The test steps of vacuum ejector's exhaust characteristic are:

Control Solenoid valve (F) through the circuit, closes the electrical-flow proportional valve (M), setting the output pressure by regulating the electrical-pressure proportional valve (G).
Set the vacuum pressure of the vacuum regulator (K) and the opening size range of the electrical-flow proportional valve (M).
Start to test, IPC (N) controls the change of the electrical-flow proportional valve's (M) flow automatically, meanwhile, collects dates, displays, saves and generates the reports automatically after the test.

The test steps of vacuum regulator's pressure characteristic are:

Control Solenoid valve (F) through the circuit, setting the output pressure by regulating the electrical-pressure proportional valve (G).

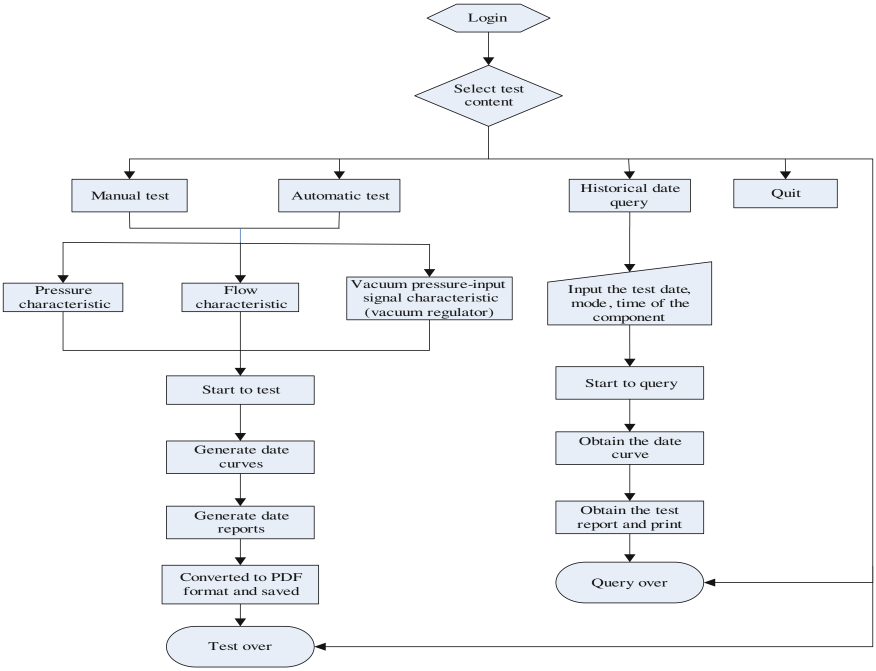


Fig. 10.4 Realization of the function of the test platform

Regulate the electrical-flow proportional valve (M) and measured vacuum regulator (K) to set the output of the vacuum pressure, setting the change, range of the electrical-pressure's(G) output pressure divided into three sections.
Start to test, IPC (N) control the change of the electrical-pressure proportional valve's (G) output pressure automatically, meanwhile, collects dates, displays, saves and generates the reports automatically after the test.

The test steps of vacuum regulator's vacuum pressure—input signal characteristic are:

Control Solenoid valve (F) through the circuit, close the electrical-flow proportional valve (M), setting the output pressure through regulating the electrical-pressure proportional valve (G).
Set the opening size change range of the electrical-pressure proportional valve (G).
Start to test, IPC (N) controls the change of the electrical-pressure proportional valve's (G) output pressure automatically, meanwhile, collects dates, displays, saves and generates the reports automatically after the test.

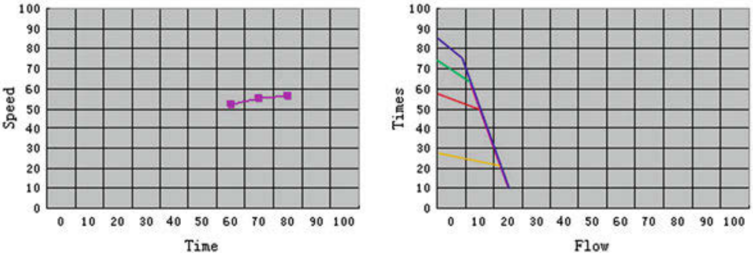


Fig. 10.5 Basic working function interface

10.3.4 Signal Acquisition and function realization

Realization of the function of the test platform as shown in the process in Fig. 10.4.

The test benches develop a series of standard modules based on the flexible software-Lab Windows/CVI. The working platform and its operation is convenient, humanistic and reliable, and it is easy to modify the program [5]. Software designed can achieve the following requirements:

The test parameters set and interface controlled;

The data acquisition module parameters' collected and filtered;

Calculate the performance parameters according to the collected data;

Monitor the test system parameter, judge the working conditions, abnormal security alarm.

Record storage and real-time display;

The historical data query and search;

The test report's generation and printing.

Achieve the above functions,which are the basic working interface as shown in Fig.10.5.

10.4 Summary

This series of test platforms developed along the needs of the enterprise, according to the latest ISO standards, and designed as the opening structure and test platform in Lab Windows/CVI, can test the characteristics of the vacuum ejector and vacuum regulator. The test platform supports a new convenient and effective way of testing, which must promote the improvement of pneumatic components and make the pneumatic industry develop better and better.

Acknowledgment The test benches' research is supported by special funding of the State Bureau of Quality and Technical Supervision.

References

1. Liu C, Shi Y, Xu W, Cai M (2010) Development of general characteristic testing platform for pneumatic components [J]. Mach Tool Hydraul 38(13):91–93
2. SMC (China) Co., ltd. Modern practical pneumatic technology [m]. China Machine Press, Beijing
3. Liu C, Cai M (2010) Proposal of power determination and measurement of compressible air. Appl mech mater 34–35:551–556. Online available since 2010/Oct/25 at www.scientific.net
4. ISO 6358-1989 (1989) Pneumatic fluid power-components using compressible fluids-determination of flow-rate characteristics [S]
5. Zhao D, Tao G (2010) Development of characteristic testing platform for pneumatic components based on modular [J]. Chin Hydraul Pneum 10:78–81

Chapter 11
Research and Implementation of Sensor Network Gateway in Green Internet of Things

Yongtao Meng, Yongjun Zhang, Kai Zhang and Wanyi Gu

Abstract With the development of Internet of Things, the Sensor Network becomes more intelligent and practical for efficient and low-carbon green communications, such as environment monitoring, intelligent building, etc. While the problem that it is difficult to connect mutually between sensor network and existing transmission network has affected its development. In this paper, we propose the Sensor Network Gateway (SNGW) with multi-interface access to adapt the various sensor networks. In order to explain the principle better, we take Zigbee access as the example and analyze the structure according to the position of the SNGW in the protocol architecture. And at the end of this paper, we perform the implementation of SNGW with ARM9-based Embedded System.

Keywords IOT · Sensor network · Gateway · Zigbee

11.1 Introduction

With the proposal and application of the Internet of Things, the Government pays more attention to promote the development of Internet of Things actively. A new round of information technology will drive our society into an era of omniscient IOT communication. With the development of the Internet of Things, we will

Y. Meng (✉) · Y. Zhang · K. Zhang · W. Gu
Key Laboratory of Information Photonics and Optical Communications,
Ministry of Education, Beijing University of Posts and Telecommunications,
Beijing 100876, People's Republic of China
e-mail: myt_163@163.com

Y. Zhang
e-mail: yjzhang@bupt.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_11,

manage production and living in a more meticulous and dynamical way, in order to achieve the state of intelligence and improve the utilization of resource and quality of life and the relationship between human and nature. It is consistent with our development strategy of information technology to research and apply IOT technology at this stage.

The Internet of Things is a novel paradigm that is rapidly gaining ground in the scenario of modern wireless telecommunications. The basic idea of this concept is the pervasive presence around us of a variety of things or object, such as radio frequency identification tags, sensors, actuators, mobile phones, etc. [1]. However, with the rapid development of the IOT, the number of the sensors increases geometrically and the types and interface of the sensors will be more complicated, and the sensors form a large and widely distributed networks which is one part of the IOT. And the transmission pressure of the integration information from the sensors increase gradually. As a result, a bottleneck appears between millions of the information from the sensors and the transmission network. So, in order to resolve the problem, we have to use Sensor Network Gateway (SNGW) which is able to solve this problem of the information transmission among different networks.

As we all know, the traditional definition of the Internet Gateway (IGW) is to selectively information relay from one sub-network to another and to perform protocol conversion where necessary. The IGW performs the network interconnection at the Transport Layer and is the most complex internet equipment, just applying between two different higher layer protocols. The structure of the IGW is similar to the router and can be both used for WAN and LAN. In the use of different communication protocol, data format or language, or even completely different architecture between two systems, the IGW seems like a translator. The IGW re-packages the information after receiving to suit the terminal system. Meanwhile, the IGW can also provide filtering and security features. Most IGW run on the Application Layer which is on the top of OSI architecture.

11.2 Sensor Network Gateway

The Internet of Things is not entirely new network architecture. It should be a new generation of integrated network based on the existing sensor and transmission network and application of industry. At the same time it should appear as the extended and supplementary of the nerve endings of the ubiquitous network. The basic network architecture is composed of Application Layer, Transmission Layer and Perception Layer [2]. If we introduce the deployment of Aggregation Layer, the typical IOT application architecture can be divided into four layers as shown in Fig. 11.1 and they are Application Layer, Transmission Layer, Aggregation Layer and Perception Layer.

Application layer: according to the command of application, it computes and analyses the millions of the information and then provides various services. That

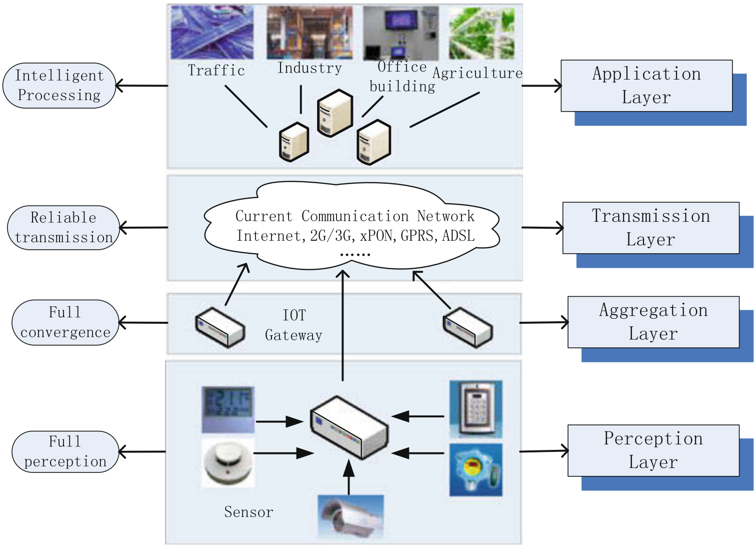


Fig.11.1 Typical IOT application architecture

is, Data processing and services providing are two major purposes of the application layer.

Transmission layer: the transmission channel is set up by the current communication network, such as Internet, 2G/3G, etc.The aim of this layer is to transfer data between aggregation layer and application layer in a large area or long distance through the transmission network. Long-range wired and wireless communication technologies, network techniques are necessary in this layer.

Aggregation layer: in this layer, the information collected from the perception layer will be pretreated. Before transferring the information to the upper layer, we have to unite the transmission format because the transmission layer can only identify the data with the standard format. In this paper, the SNGW device is in this layer, and this part is most significant than others [3].

Perception layer: low-cost, reliability and easy to deploy are three important factors of the perception layer. In this layer, it collects the data through data acquisition device at first, and then the data is transferred to the next layer by wire or wireless ways according to different devices. The intelligence of this layer depend on the capability of computing and analysis of aggregation layer.

As a bridge to connect sensor networks with traditional communication networks, SNGW can provide the functionalities of protocol conversion and device management to achieve the integration between various perceived access and core network. Specifically, SNGW has the following characteristics:

Receiving various types of sensor data from front terminal and performing unified management. As the perceptive module of front terminal, the Gateway allows to input various types of data and communicate with different kinds of interface. Not only the digital input, but also the collection of the analog can be permitted. The Gateway is able to identify and specify the information at the same time. There are several kinds of interface which accesses the Gateway, such as serial port, Ethernet and USB, etc.

Intelligent processing and feedback control. The Gateway achieves storage, management and control of various kinds of sensor information and provides basic information management of database, fault, security, voice and video for the background platform. And another aim is to judge and treat the feedback control signal based on the intelligent analysis of the receiving information. Then according to the command of the application, it sends control signals to the front terminal. Finally, the mechanism is formed automatically and interactively.

Above all, the Gateway is the carrier of the aggregation layer and also the key device with generality and standardization.

11.3 Design of Sensor Network Gateway Module

11.3.1 Module of the Gateway Requirements

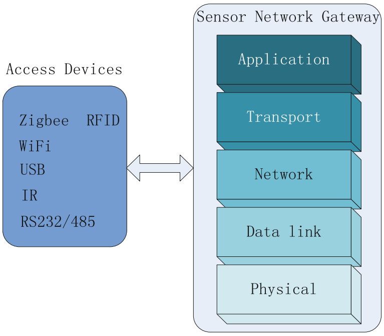
According to the position of the gateway in the IOT application architecture, it should support internal data collection and aggregation from the sensors and transfer the data information to the transmission network, such as Internet, 2G/3G networks, xDSL networks and other network interfaces and the requirements are shown as follows:

Data collection: the first function of the SNGW system is to receive data from sensor terminals or commands from transmission network. And then transfer data to the other networks transparently and correctly.

Protocol conversion: RFID technology and Zigbee network communication protocols are used in wireless sensor network, while the internet network is based on TCP/IP protocol. As a result, the SNGW has to acquire the message packet from the sensor nodes by short-distance wireless communication or by twisted pair or coaxial cable directly. And then it uses the xDSL, 3G and other network interfaces to send packets to telecommunication or Internet. Therefore, SNGW should analyze and re-package the sensor data after receiving it, and then capsulate and send the re-packaged data based on telecommunication protocols [4].

Management and control: as an independent system, the SNGW should also support the capability of management and control. On one hand, the Gateway has to manage the information from the sensor network, and store or forward the information to the other sensor network. On the other hand, the gateway receives the commands from the remote server; it should translate the commands and then send them to the sensor nodes so that the devices in the sensor network can execute the command correctly.

Fig. 11.2 The SNGW layers



11.3.2 Design of Sensor Network Gateway Module

The SNGW should contain all layers of the networks involved and converts application information between appropriate formats. And the SNGW is implemented according to the proper architectures to perform the interconnection transition. These are several reasons why the gateway has to contain all layers of the network. First the gateway converts frames with consistent addressing schemes at the data-link layer and the router deals with packets at the network layer. While sub-networks differ in their higher layer protocols, especially in the application layer, or the communication functions of the bottom three layers are not sufficient for coupling. This is the obvious difference between SNGW and usual getaway.

We give the module and analyze the protocol conversion of the SNGW in this part. We divide the Gateway into five layers shown in Fig. 11.2 which are composed of Physical Layer, Data Link Layer, Network Layer, Transport Layer and Application Layer.

In perception layer, there are several kinds of way of communications, such as Zigbee, RFID, USB, Wi-Fi, RJ-45, RS232/485, etc. because of the difference of the communication between different devices, the gateway has to provide the various access interfaces, and that is, the gateway can identify and analyze the data from different interfaces of the sensors. In Fig. 11.2, the sensing devices are classified according to their internal protocol structure.

In order to gain insight into the architecture of the Sensor Network Gateway, we take Zigbee for example. In December 2000 IEEE IEEE802.15.4 set up a working group to define a low complexity, low-cost and low-power consumption low-rate wireless connectivity that would be used in suitable fixed, portable or mobile devices. The wireless connectivity technology is ZigBee. Therefore, ZigBee technology's physical layer and link layer protocol used mainly IEEE802.15.4 standard [5]. IEEE 820.15.4/ZigBee protocol stack architecture is shown in Fig. 11.3.

A SNGW is intended to provide an interface between ZigBee and IP devices through an abstracted interface on the IP side. The IP device is isolated from the ZigBee protocol by that interface, see Fig. 11.4. The Gateway translates both

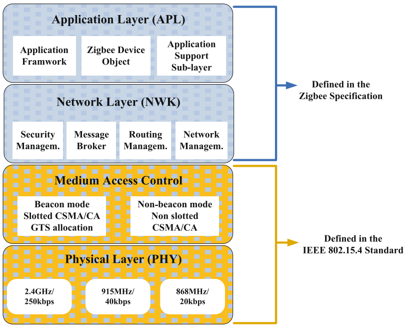


Fig. 11.3 IEEE 820.15.4/ ZigBee protocol stack architecture

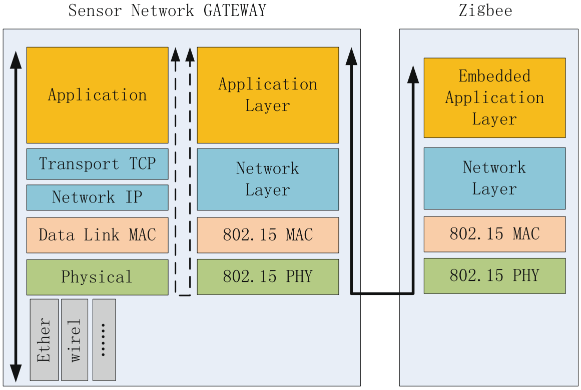


Fig. 11.4 Sensor Network Gateway architecture

addresses and commands between ZigBee and IP. The IP stack is terminated at the Gateway as is the ZigBee Stack; the Gateway provides translation between the respective stacks [6].

Simply, whether Zigbee, RFID or USB communication, we can divide the application into several layers and the Gateway can analyze the data in every layer by packaging and re-packaging the frames. And finally the SNGW can package the information which can transfer on the transmission layer.

11.4 Implementations

This section presents some applications about the SNGW system.

11.4.1 Sensor Nodes

The hardware architecture of the sensor nodes are shown in Fig. 11.5. We take Zigbee device, serial device and USB device as the terminal sensors. The Zigbee

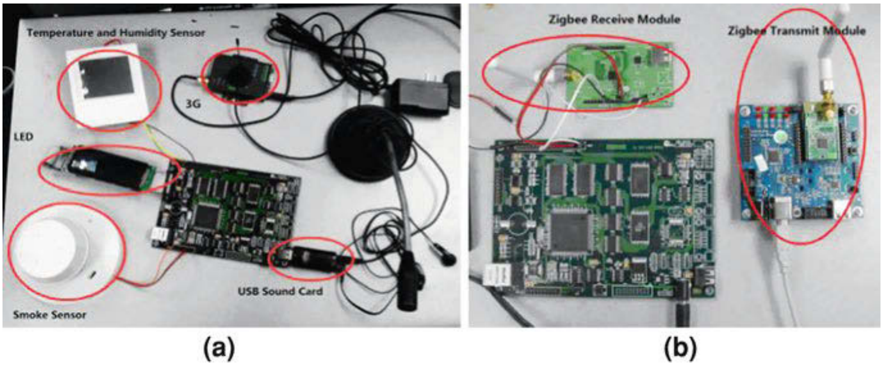


Fig. 11.5 The hardware architecture of the sensor nodes. **a** Serial devices and USB device; **b** Zigbee device

node uses CC2430 as the wireless communication module, integrated with temperature data collection module.

We use IAR system Embedded Workbench to develop the CC2430. The sensor nodes collect the data of temperature, and then the data can transfer after packaging according to Zigbee protocol by built-in wireless channel. The receiver receives the wireless data, and re-package until the data can transfer in physical layer through serial port. So the data is reported to the gateway by the serial port finally. Besides, the sensor with serial device can transfer the data by the serial port directly. And the USB device uses its own protocol to exchange the information with SNGW as the Zigbee device. In this application, we use USB Sound Card as the Bidirectional voice channel.

11.4.2 Sensor Network Gateway

The hardware structure of SNGW is shown in Fig. 11.6. The Gateway uses AT91RM9200 platform with 180 MHz CPU, 32 M DSRAM, 4 M NOR flash and 64 M NAND flash as the processor. The storage and memory are respectively and we use ARM-Linux and GCC as the operating systems and programming environment.

(1) CPU; (2) SDRAM; (3) NAND Flash; (4) NOR Flash;

The main function of the gateway is to read data from serial port, write data to the serial port and forward sensed data. We communicate with SNGW system by Hyper Terminal serial port. The data from the serial port and A/D conversion can been seen in the screen shown in Fig. 11.7. As shown in Fig. 11.7a, the data collected by the sensor devices is sent to gateway through serial port. In Fig. 11.7a, 1 the audio value; 2 control air-condition ON/OFF according the temperature; 3, 4 and 5 The voltage value from the AD chip; 6 alarm signal; 7 the data of

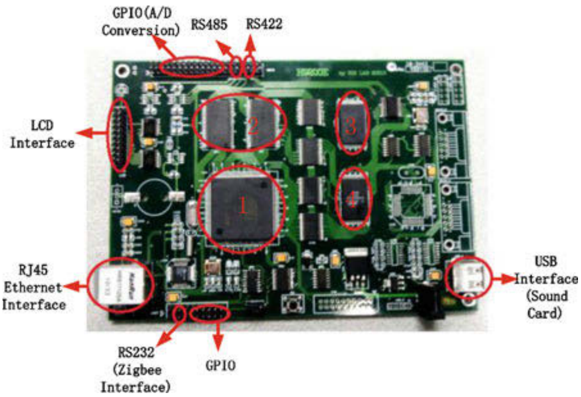
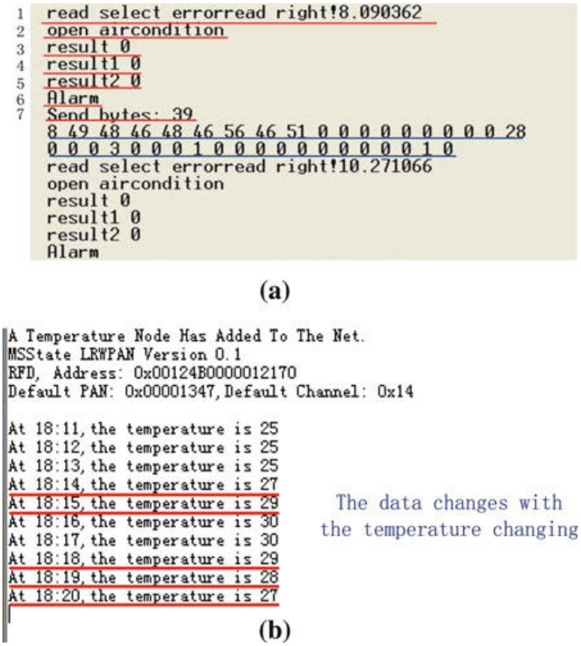


Fig. 11.6 The hardware structure of Sensor Network Gateway



```
read select errorread right!8.090362
open aircondition
result 0
result1 0
result2 0
Alarm
Send bytes: 39
8 49 48 46 48 46 56 46 51 0 0 0 0 0 0 0 0 0 28
0 0 0 3 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0
read select errorread right!10.271066
open aircondition
result 0
result1 0
result2 0
Alarm
```

(a)

```
A Temperature Node Has Added To The Net.
MSState LRWPAN Version 0.1
RFD, Address: 0x00124B0000012170
Default PAN: 0x00001347,Default Channel: 0x14

At 18:11,the temperature is 25
At 18:12,the temperature is 25
At 18:13,the temperature is 25
At 18:14,the temperature is 27
At 18:15,the temperature is 29
At 18:16,the temperature is 30
At 18:17,the temperature is 30
At 18:18,the temperature is 29
At 18:19,the temperature is 28
At 18:20,the temperature is 27
```

The data changes with the temperature changing

(b)

Fig. 11.7 The data from the SNGW platform with linux operating system. **a** data from sensors; **b** data from Zigbee

temperature, AD value and alarm. In Fig. 11.7b, the data collected by the Zigbee node is sent to the gateway through the wireless channel and the receiving chip connect to the Gateway by serial port. After turn on the Zigbee module, it shows “A temperature node has added to the net” and assigns the address and the channel. We can obviously see the change of the temperature if the environment is changing. After transplanting Linux operating system into the embedded ARM9 system, we need to implement the data transmission, protocol conversion and command agent functions to meet the requirements mentioned above. After the

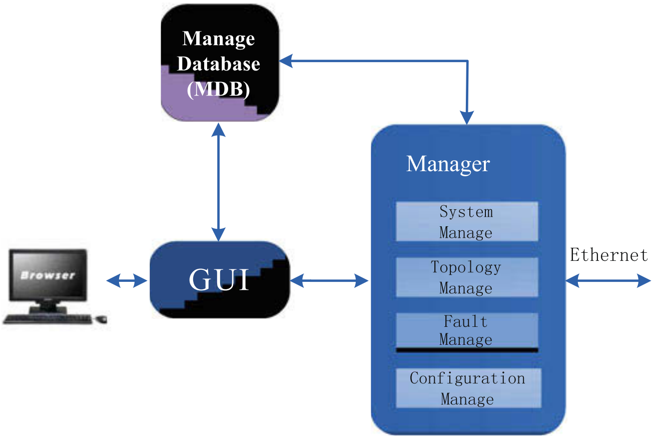


Fig. 11.8 The structure of the application platform

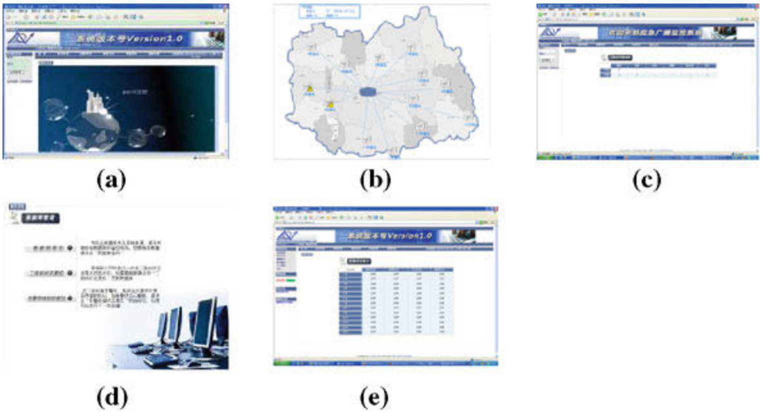


Fig. 11.9 Application functions. **a** system management; **b** topology management; **c** data collection; **d** warning management; **e** database management

system is turned on, the Linux operating system boots first, and then the main program will initialize the applications and establish the socket connection by setting IP address and MAC address. In this design we provide Ethernet with the remote server.

11.4.3 Application Service

In this application, we take JAVA as the main technical, and the other auxiliary programming technical contain JavaBean, JavaApplet, JavaScript, Ajax and Socket, etc. The main modules deployed in the application server shown in Fig. 11.8.

In TCP Server, we implement data receiving and processing functions which supports Ethernet communication channels. The information is uploaded to the server by Socket communication technology. The manager processes and stores the information, and then transfers to the GUI. Finally, the information is displayed on the browser.

The application platform provides the functions including system management, topology management, data collection, fault monitoring, database management, etc. as shown in Fig. 11.9.

11.5 Summary

The SNGW becomes more and more important to adapt to a various access methods. Sensor networks/IOT will be the expansion of present network, eventually becoming part of the next generation Internet to achieve global interoperability. In this paper, we present the SNGW with multi-access, and analyze in detail with Zigbee access system. Besides, in the implication we use several kinds of sensor devices in order to perform the multi-access. According to different requirements, the SNGW can widely use in various areas, such as smart home, industrial control, environment monitoring, etc. While there is still much work waiting for research for the more comfortable life in future.

Acknowledgments This work was supported by the Key Project of the Chinese National Programs for Fundamental Research and Development (2010CB328204).

References

1. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. Comput Netw 54(15): 2787–2805
2. Fangchun Y (2009) Internet of things should be more concerned the development of industrial applications. Things networking technology and industrial development Forum
3. Han D, Zhang J, Zhang Y, Gu W (2010) Convergence of sensor networks/internet of things and power grid information network at aggregation layer. 2010 International conference on power system technology: technological innovations making power grid smarter, Hangzhou, China, pp 1–6
4. Zhu Q, Wang R, Qi C, Liu Y, Qin W (2010) IOT gateway: bridgingwirelss sensor networks into internet of things. Embedded and Ubiquitous Computing (EUC), IEEE/IFIP 8th International Conference, Dec 2010, pp 347–352
5. Clerk Maxwell J (1892) A treatise on electricity and magnetism, vol 2, 3rd edn. Clarendon, Oxford, pp 68–73, IEEE 802.15.4 Standards-Part 15.4. 2003.10
6. Sveda M, Trchalik R (2007) ZigBee-to-internet interconnection architectures. System, 2007. ICONS'07, Second International Conference, p 30, April 2007

Chapter 12
LDPC Coding Proposal for Pulsed-OFDM Modulation for WPAN Systems Using UWB Communication in Indoor Propagation Channels

C. T. Manimegalai, R. Kumar and S. B. Sumith Babu

Abstract In this chapter, we describe a combined approach where low density parity check (LDPC) codes are used to reduce the complexity and power consumption of pulsed orthogonal frequency-division multiplexing (pulsed-OFDM) ultra-wideband (UWB) systems. The proposed system will use LDPC codes to achieve higher code rates without using convolution encoding and puncturing thereby reducing the complexity and power consumption of pulsed-OFDM system. The LDPC-pulsed-OFDM system will achieve channel capacity with different code rates and will have good performance in different channel fading scenarios. The proposals from pulsed-OFDM system is used where pulsed signals could spread the frequency spectrum of the OFDM signal. The performance of LDPC-pulsed-OFDM system for wireless personal area networks (WPAN) is analyzed for different UWB indoor propagation channels (CM3 and CM4) provided by the IEEE 802.15.3a Standard activity committee. To establish this, we present a design of LDPC-pulsed-OFDM system using the digital video broadcasting-satellite-second generation (DVB-S2) standard and provide the simulation results for the different code rates supported by LDPC codes.

C. T. Manimegalai (✉)
Assistant Professor, Department of Electronics and Communication Engineering,
SRM University, Chennai, India
e-mail: ctm@ktr.srmuniv.ac.in

R. Kumar
Professor, Department of Electronics and Communication Engineering,
SRM University, Chennai, India
e-mail: rkumar@ktr.srmuniv.ac.in

S. B. S. Babu
M.Tech, Department of Electronics and Communication Engineering,
SRM University, Chennai, India
e-mail: sumithbabusb@yahoo.co.in

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_12,

Keywords Low density parity check (LDPC) · Orthogonal frequency-division multiplexing (OFDM) · Pulsed orthogonal frequency-division multiplexing (pulsed-OFDM) · Ultrawideband (UWB) · Digital video broadcasting-satellite-second generation (DVB-S2) · Wireless personal area networks (WPAN)

12.1 Introduction

The upsurge of wireless communication devices in our lives shows no sign of languor. The growing demand for high quality media and high-speed content delivery drives the pursuit for higher data rates in communication networks. Wireless personal area networks (WPANs) are used to convey information over relatively short distances of about 10 m among a relatively few participants. Unlike WLANs, WPANs connections involve little infrastructure. This allows small, power efficient, inexpensive solutions to be implemented for a wide range of devices. LDPC codes have the advantage of achieving near-channel capacity for different code rates. LDPC codes are the codes that offer error detection and correction capabilities close to theoretical limit [2]. Higher code rates can be achieved easily and hence reduce the complexity, power consumption and cost of the system implemented using LDPC codes. Also, UWB technology is used for short- and medium-range wireless communication networks with various throughputs including very high data rate applications. UWB communication systems use signals with a bandwidth that is larger than 25% of the center frequency or more than 500 MHz. The main issue of spectrum scarcity is overwhelmed by ultra-wideband technology. UWB communication systems have advantages, including robustness to multipath interference and inherent support for location-aware networking and multiuser access [3, 4]. UWB communications transmit in a way that does not interfere largely with other more traditional narrowband and continuous carrier wave which uses the same frequency band.

OFDM technique and its variations are widely used in several narrow-band systems. Pulsed-OFDM is a major UWB system that uses OFDM modulation in the UWB spectrum. The pulsation of the OFDM signal spreads its spectrum and provides a processing gain that is equal to the inverse of the duty cycle (less than one) of the pulsed subcarriers [1]. A pulsed-OFDM signal can easily be generated by up-sampling the output of an inverse fast Fourier transform (IFFT) module in a normal OFDM system. Also, a low-complexity receiver is achieved for the pulsed-OFDM system that exploits the spreading gain provided by the pulsation to enhance the performance of the system in multipath fading channels [1]. In this paper, we propose an enhancement to the pulsed-OFDM system where the complexity of achieving higher data rates using convolution encoding and puncturing technique is replaced by an LDPC encoder. The new approach is an combined form of the benefits of LDPC codes, UWB and pulsed-OFDM technology.

Chapter 2 discusses about the pulsed-OFDM signal generation and its key concepts that is used in the proposed system. Chapter 3 describes the proposed system model and how it reduces the complexity and power consumption when compared to Pulsed-OFDM system. Chapter 4 presents the simulation results of the proposed system for channels CM3 and CM4 with different code rates.

12.2 Pulsed-OFDM-System

In the pulsed-OFDM scheme [1], the pulsed-OFDM signal can be generated by up-sampling the digital baseband OFDM modulated signal before sending it to a conventional DAC. The up-sampling is done by inserting $K-1$ zeroes between samples of the signal. The resulting pulsed-OFDM signal is then a pulse train with a duty cycle of $1/K$. The main difference between pulsed-OFDM and normal OFDM is the up-sampling operation after the IFFT. The up-sampling factor K needs to be smaller than or equal to an upper limit $K_{\max}$ given by (12.1), where ω is the subband bandwidth, B_{c} $_{\mathrm{i}}$s the coherence bandwidth of the channel and Tspread is its maximum delay spread. Here, $\lfloor x \rfloor$ denotes the largest integer that is smaller than x. For a given channel, the optimum K is in the range K = 1,…, $K_{\max}$.

$$K_{\max} = \left\lfloor \frac{\omega}{B_c} \right\rfloor = \left\lfloor \omega T_{\text{spread}} \right\rfloor. \tag{12.1}$$

In the multiband-OFDM (MB-OFDM) approach in [5], the available UWB spectrum is divided into several subbands of smaller bandwidth. An OFDM symbol is transmitted in each subband, and then, the system switches to another sub-band. Quadrature phase-shift keying (QPSK) modulation is used for OFDM. The transmitted signal in this scheme is given by

$$x(t) = \sum_{r} \sum_{k=0}^{M-1} b_k^r \, e^{j2\pi k f_0 t} \, p\left(t - rT_p\right) \, e^{-j\frac{2\pi c(r)t}{T_s}}. \tag{12.2}$$

where M is the number of subcarriers in each OFDM symbol, and p(t) is a low-pass pulse with duration Tp. The QPSK symbol that is transmitted in the rth time slot and over the kth subcarrier is denoted by b_k^r. The subcarrier spacing is denoted by f0 and is equal to 1/Tp. Sequence c(r) controls frequency hopping between subbands. The MB-pulsed-OFDM signal can be presented with a similar formula as the MB-OFDM signal in (12.2). Here, p(t) is a train of pulses with duty cycles less than one [1], i.e.,

$$p(t) = \sum_{n=0}^{N-1} s(t - nT). \tag{12.3}$$

where $s(t)$ is a monopulse with duration T_s, and T is the pulse separation time, which is larger than T_s. The number of monopulses is denoted by N and is the same as the number of subcarriers for the OFDM modulation. This number will be chosen, such that the total bandwidth of the pulsed-OFDM signal becomes equal to that of the non-pulsed-OFDM signal [1]. The selection of up-sampling factor K in a given scenario can be done once a suitable design criterion is chosen. This is explained in [7], using the concept of outage capacity [8, 9]of the pulsed-OFDM system in fading channels. The advantage of this approach is that it leads to results that can be applied, regardless of the choice of coding, interleaving and modulation schemes [1].

12.3 LDPC-Pulsed-OFDM

LDPC codes are a class of linear block codes developed by Robert G. Gallager in 1963. LDPC codes have easily parallelizable encoding and decoding algorithms. The parallelizability is ‘adjustable’ providing the user an option to choose between throughput and complexity. The function of the encoder is to add extra redundant data for given uncoded data. This extra redundant data, called as parity data is useful in detecting the errors that are introduced during the data transmission through a channel. LDPC encoder along with BCH encoder block is used for generating the parity data in DVB-S2 systems. In this approach, the parity-check matrix of the LDPC code with code rate R is obtained from the DVB-S.2 standard.

12.3.1 System Parameters

To transmit information, the Pulsed-OFDM system uses convolutional coding and puncturing to achieve a rate of 2/3, followed by OFDM modulation with $M = 32$ subcarriers. In the LDPC-Pulsed-OFDM system we use LDPC codes to achieve a specific code rate, followed by OFDM modulation. Figure 12.1 shows the new system transmitter and receiver. The input signal is assumed to be scrambled and is fed to the LDPC encoder. The encoder uses 100 iterations to encode the scrambled input signal. QPSK mapping sets the constellation points for the encoded symbols. This helps in error detection and correction. The signal is then passed through a serial-to-parallel converter to separate the diversity branches. Each branch is separately demodulated using FFT algorithm. A 32-point IFFT is used at the transmitter followed by up-sampling with a processing gain of $K = 5$. Similar to other OFDM systems, a cyclic prefix (CP) added after the IFFT at the transmitter and discarded from the received signals before the FFT in each branch eliminates inter-symbol interference and inter-channel interference in all branches. At the

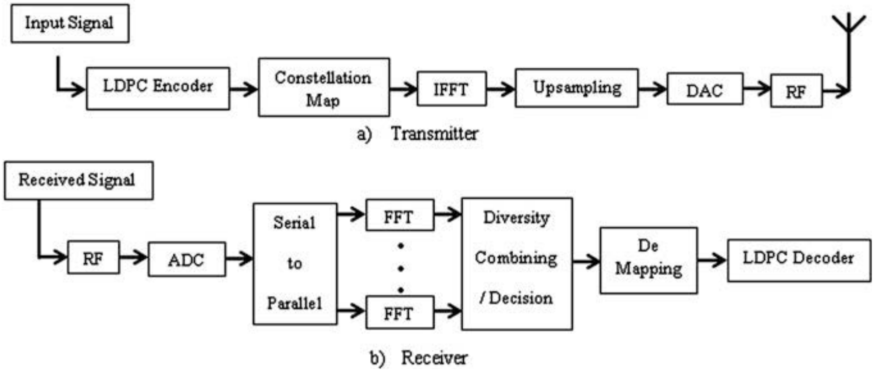


Fig.12.1 System Model for the proposed LDPC-Pulsed-OFDM System. **a** Transmitter. **b** Receiver

receiver, the diversity branches are combined using equal gain combining followed by constellation de-mapping and LDPC decoding.

12.3.2 System Performance

To compare the performance of the LDPC-Pulsed-OFDM and Pulsed-OFDM systems, a complete simulation of the system over the channel models described in the IEEE 802.15.3a UWB channel modeling report [10]. Two channel models (named CM3 and CM4) are presented to model the channels at 10 m. Here, the simulation results of both channels at extreme fading conditions are presented. Figure 12.2a–g shows the results over the CM3 and CM4 channel under log normal fading conditions. In this figure, the bit error rate is plotted versus the signal-to-noise ratio for both systems. The simulation results show that the LDPC-Pulsed-OFDM system performance is stable for different code rates and achieves a BER nearly 10-5 for SNR up to 6 dB using QPSK. The performance of LDPC-Pulsed-OFDM is better in additive white Gaussian noise (AWGN) channel and would achieve BER nearly 10-7 for SNR up to 16 dB using amplitude-phase-shift keying (APSK).

12.3.3 Power Consumption

The power consumption of a very-large-scale-integration (VLSI) chip is determined by its clock rate and the supply voltage and capacitance of the circuit. As the number of components is reduced, the power consumption of the VLSI chips will be less. This system is implemented with lower complexity and power consumption compared to the existing baseline system.

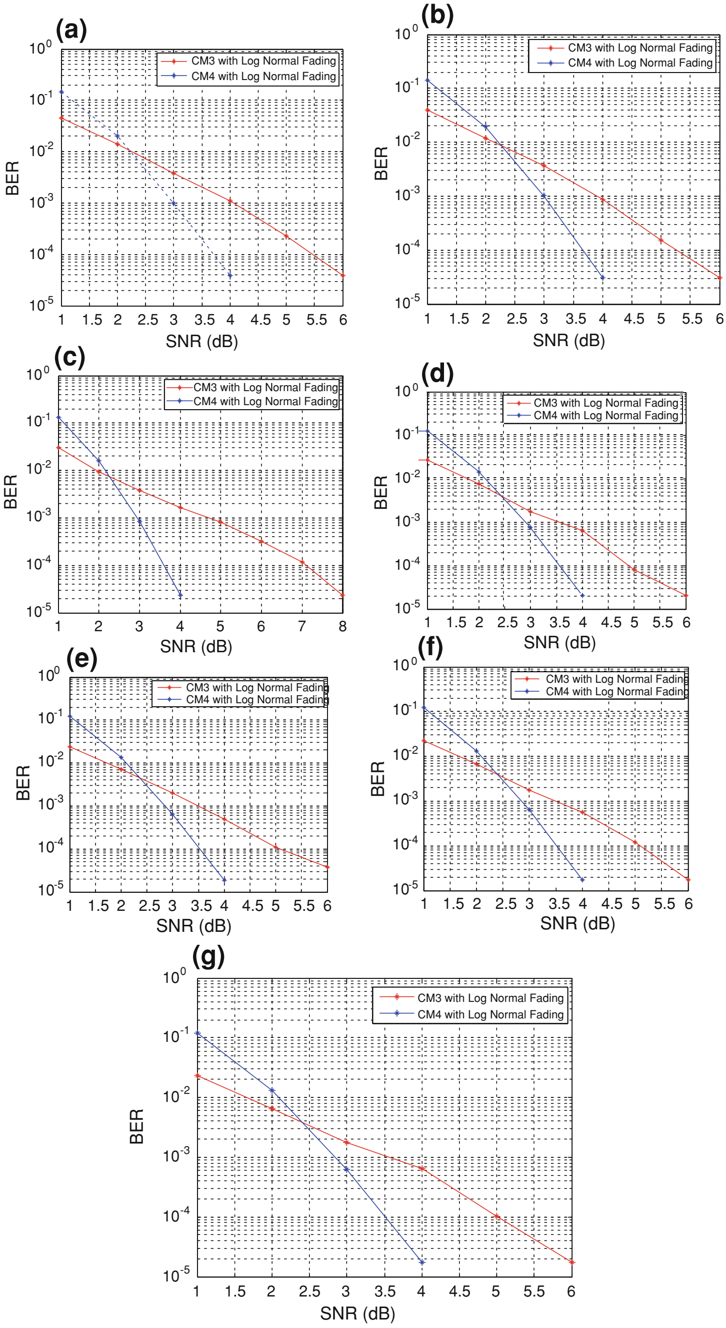


Fig. 12.2 BER versus SNR in dB for the LDPC-Pulsed-OFDM system in CM3 and CM4 channels **a** CM3 and CM4 for code rate 2/5 **b** CM3 and CM4 for code rate ½ **c** CM3 and CM4 for code rate 2/3 **d** CM3 and CM4 for code rate 3/4 **e** CM3 and CM4 for code rate 5/6 **f** CM3 and CM4 for code rate 8/9 **g** CM3 and CM4 for code rate 9/10

Table 12.1 IEEE 802.15.3a UWB channel parameters

Model parameters	CM3	CM4
Λ [1/ns] (cluster arrival rate)	0.0667	0.0667
λ [1/ns] (ray arrival rate)	2.1	2.1
Γ (cluster decay factor)	14.00	24.00
γ (ray decay factor)	7.9	12
$\sigma 1$ [dB] (stand. dev. of cluster lognormal fading term in dB)	3.5	3.5
$\sigma 2$ [dB] (stand. dev. of ray lognormal fading term in dB)	3.4	3.4

12.3.4 Channel Parameters

The IEEE 802.15.3a UWB channel parameters that is used for the simulation is given below in Table 12.1.

12.4 Simulation Results

The performance of LDPC codes is measured in terms of bit error probability versus signal-to-noise ratio. The simulation results of LDPC-Pulsed-OFDM for the different code rates supported by LDPC DVB-S.2 standard are presented. The proposed system is analyzed for UWB indoor propagation channels CM3 and CM4 under log normal fading with the code rates 2/5, 1/2, 2/3, 3/4, 5/6, 8/9 and 9/10. The log normal fading characteristics incorporate the worst channel conditions possible. The simulation results in Fig. 12.2a–g show that the LDPC-Pulsed-OFDM system achieves a bit error rate nearly 10-5 for the above indoor channels. The frame size of 300 and 256 bits per block is used for all of the above code rates under extreme line of sight channel conditions. SNR of 6-8 dB is achieved for CM3 and SNR of 4 dB for CM4 using QPSK. Higher values of SNR can be achieved by using different modulation schemes.

12.5 Conclusion

LDPC-Pulsed-OFDM is a combination of the benefits of LDPC codes and pulsed-OFDM, utilizing the ultra-wideband spectrum to efficiently achieve a comparable performance under different code rates and achieves a bit error rate nearly 10-5. The system provides frequency spreading and diversity in multipath fading channels. By replacing the convolution encoder and puncture in the Pulsed-OFDM system and using LDPC encoder, we designed a system for the WPAN utilizing the UWB channel conditions with reduced complexity, reduced power consumption. Since, QPSK is used; the maximum achievable SNR is 6–8 dB. To enhance this system, amplitude-phase-shift keying (APSK) could be used, to achieve SNR up to 16 dB for different code rates. Also, data rates of more than 1Gbps could be achieved using MIMO.

References

1. Saberinia E, Tang J, Tewfik AH, Parhi KK (2009) Pulsed-OFDM Modulation for Ultrawideband Communications. IEEE Trans on veh Tech 58(2):720–726
2. Gallager, Robert G (1963) Low-density parity-check codes. MIT Press, Cambridge
3. Win MZ, Scholtz RA (2000) Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications. IEEE Trans Commun 48(4):679–689
4. Win MZ, Scholtz RA (1998) On the robustness of ultra-wide bandwidth signals in dense multipath environments. IEEE Commun Lett 2(2):51–53
5. Batra et al. (2003) Multi-band OFDM: Merged proposal #1, Merged Proposal for the IEEE 802.15.3a Standard, San Francisco, CA. IEEE 802.15 work group official web site [Online]. Available: http://grouper.ieee.org/groups/802/15/pub/2003/Jul03/
6. Balakrishnan A, Batra D, Dabak A (2003) A multi-band OFDM system for UWB communication. In: Proc IEEE Conf Ultra Wideband Syst Technol pp 354–358
7. Saberinia E, Tewfik AH (2004) "Outage capacity of pulsed-OFDM ultra wideband communications. In: Proceedings Joint IEEE Conf UWBST/IWUWBS, Tokyo, Japan 323–327
8. Bieglieri E, Proakis J, Shamai S (1998) Fading channels: Information theoretic and communications aspects. IEEE Trans Inf Theory 44(6):2619–2692
9. Ozarow LH, Shamai S, wayner AD (1994) Information theoretic considerations for cellular mobile radio. IEEE Trans Veh Technol 43(2):359–378
10. Foerster J et al. Channel Modeling Sub-Committee Report Final. 802.15 work group official web site [Online]. http://grouper.ieee.org/groups/802/15/pub/2003/May03/

Part II
Data Management and Database System

Chapter 13
Study of Voice Conversion Information Systems

Xiaoning Li, Yingjuan Sun and Zhuo Zhang

Abstract Voice conversion is based on the combination of proposed independent target speaker pitch between the sources of the transfer information systems. Because, although the conversion is in the field by modifying the residual signal oise time interval to achieve, it is not on cloud/unvoiced decisions need and make the same silent treatment, which, in the expression part of the, IoSE equivalent pitch information, and in the silent period.

Keywords Voice conversion · Information systems · Prosody transformation

13.1 Introduction

Advances in low-power components and system design have brought general-purpose computation into watches, wireless telephones, PDAs and tablet computers. Parameters in the speaker, in addition to personality characteristics reflect channel spectral envelope information parameters [1], the rhythm of the speech signal functions pack speaker of the status and wealth. Prosodic parameters

Supported by Education Department of Jilin Province, the "Eleventh Five-Year" key scientific and technological research Head ([2010] No. 383).

X. Li (✉) · Y. Sun
Collegye of Computer Science and Technology, Changchun Normal University,
Changchun, People's Republic of China

Z. Zhang
Changchun City Experimental High School, Changchun, People's Republic of China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_13,

include pitch contour, phoneme duration, energy and other parameters. Power management of these systems has traditionally focused on sleep modes and device power management. In a system we normally operate the CPU/SDRAM at 266/133 MHz above 1.65 V and at 66/33 MHz above 0.9 V, typically providing a 13:1 SOC core power range over the 4:1 performance range [2].

In this study, consider the main speaker and target speaker pitch between the source of the transfer because, although the conversion is in the field by modifying the residual signal IoSE time interval to achieve, it is not on cloud/unvoiced decisions need and make the same silent treatment, which, in the expression part of the, IoSE equivalent pitch information, and in the silent period, IoSE reflected signal suddenly development time. In the past, voice conversion algorithm, frequent change of pace only part of the pitch period of voiced and voiceless, including neglected prosodic information. Prosodic features as a parameter IoSE, more fully into account useful information on the speech characteristics [3] of speech signal conversion rhythm.

13.2 Information Systems

The CPU clock is generated by a clock divider outside of the PLL feedback path, and the circuit design allows to divide ratios to be changed without glitches on the output clock. This is because based on gaussian mixture model (GMM) [4]. The spectrum for voice conversion algorithm is used to match the weighted average of function, likely to cause conversion of the speech spectrum too smooth, formant peaks weakened, and the bandwidth expansion.

Viterbi algorithm-based voice conversion for each frame spectral characteristics of the GMM parameters to find the best components for conversion, a single component of the transfer function, avoiding the transfer function of the weighted treatment. In addition, the spectral distortion measure is a common way to measure the spectrum distortion degree method. This paper Itakura spectral distance measure is a statement in the four cases, after the traditional GMM-based Viterbi search algorithm and the algorithm based on the Itakura spectral distance converted.

The ratio of the control chart is found, it can be seen from the chart, in either case. The base in the Viterbi search algorithm is better than conversion based on GMM algorithm. But can also be seen that the opposite sex has a superior voice conversion performance conversion between the voice in the performance of the same sex. This is special because of the opposite sex sign from the larger parameter space, the relative ratio of energy conversion in the past will be larger, and the effect will be more evident in some ways. Compared with the Fourier transform, wavelet transform is a space (time) and frequency of the local transformation, which can effectively extract information from the signal. By dilation and translation functions such as operations on the function or signal can be multiscale refinement analysis; Fourier transform solves many difficult problems that

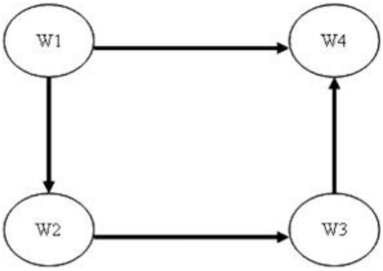


Fig. 13.1 Ways conversion system

can not be solved. Contact the application of wavelet mathematics, physics, computer science, signal and information processing, image processing, seismic exploration, and other disciplines. Mathematicians believe that wavelet analysis is a new branch of mathematics, it is a functional analysis, Fourier, like coherence analysis, numerical analysis of the perfect crystal; signal and information processing experts believe that wavelet analysis is the time-scale analysis and multi-resolution analysis is a new technique, which signal analysis, speech synthesis, image recognition, computer vision, data compression, seismic exploration, atmospheric and ocean wave analysis and other aspects of studies have meaning and application of scientific results.

Figure in the case of a voice (voice script "Variable capacitor") of the spectral envelope of the comparison chart. It can be seen from the chart, GMM algorithm based on the conversion results are poor, badly flat speech spectrum slip, peak decreased, while the Viterbi algorithm based on the performance of a good conversion. Performance and voice format have been strengthened.

Experiment was divided into four sub-tasks, namely the transformation from male to male (W1−W4), male to female conversion (W1−W2), female to female's turn for (W2−W3), female to male conversion (W3−W4). Figure 13.1 is W1−W4 case, source statement "socialism" before the conversion and the target language after the spectrum in terms of voice control chart.

As can be seen from the graph out, based on the GMM algorithm and Viterbi search algorithm based on voice conversion has obvious effects, converted spectrogram are more recent the target voice spectrogram. But Viterbi search algorithm based on the language spectrum trend on the chart in GMM-based algorithm is better than spectrogram that is lower the inter-frame GMM algorithm dynamic distortion of speech spectrum, and achieved good results.

Y. Meyer constructed a real chance of wavelets, and established cooperation with S. Mallet unified method of wavelet construction—multi-scale analysis, wavelet analysis began to develop, including Belgium, written by female mathematicians I. Dubieties, *"Ten Lectures on Wavelets"* the popularity of wavelet played an important role in promoting. Fourier transform, the window fourier transform (Gabor transform), compared with a good time–frequency localization properties, which can effectively extract information from the signal through dilation and translation functions such as computing functions or signals on the detailed analysis of multi-scale (multistage analysis), to solve the Fourier transform cannot solve the many difficult problems, which wavelet transform known as the “mathematical microscope”, which is a milestone in the history of harmonic analysis of the development progress.

Six high-tech electronics and information technology is an important area in which important aspects of image and signal processing. Today, signal processing has been the work of contemporary science and technology, an important part of the purpose of signal processing is: accurate analysis, diagnosis, coding and quantization, fast delivery or storage and precision to reconstruct (or recover). From the mathematical point of view to signal and image processing can be unified as a signal processing (image can be seen as two-dimensional signal), the wavelet analysis to many of the many applications, can be attributed to signal processing problems. Now, for the nature of change over time, the signal is stable (stationary random process), is still the ideal tool for dealing with Fourier analysis. But in practice the vast majority of the signal non-stable (non-stationary random process), and especially for non-stable signal wavelet analysis tools can be applied.

Proposed a voice conversion based on Viterbi search operator France, through the establishment of the target speech frames to describe the voice transition probability matrix signal frame timing information, then use Viterbi search algorithm to find the characteristic parameters of speech signal frames the best GMM components to complete the spectral characteristics of speech signal parameters of the conversion, in the rhythm conversion is too strong on the residual signal in the statistical analysis of the excitation pulse sequence match so as to achieve a change of pitch information. Experimental results show the proposed.

This allows CPU frequency scaling over a wide range with low latencies. B. System Clocking The 405LP [5] CPU clock is further divided down to generate clocks for the internal and external busses: the high-speed processor local bus (PLB) in a CPU-intensive application like media decoding we find that lowering bus frequencies from their maximum performance points can result in system-wide energy savings without impacting real-time performance. However, power management policies are typically defined to only scale the CPU and PLB (memory) frequencies, leaving the lower speed busses at fixed frequencies. This avoids the need to reprogram bus controller parameters and IP cores that have some sensitivity to peripheral bus frequencies. In the 405LP design changing the CPU and bus clocking scheme requires updating anywhere from one to three control registers and may also require reprograming the SDRAM controller and other bus controllers.

13.3 Voice Conversions

The voice conversion algorithm, to overcome the traditional GMM based voice conversion operator law of the spectrum caused by the dynamic distortion of the shortcomings of frame, but also improved the weighted average of the speech spectrum caused by the problem of too smooth, so that the converted voice formants have been strengthened.

But code word vector quantization (VQ) codebook based on the form of matching limited, limited to set parameters, resulting in discontinuous parameters, greatly reducing the voice quality. Stygian, who proposed based on GMM of the matching function of the continuous, the minimum mean square error criterion to estimate the parameters of the matching parameters to maintain the continuity of space and improve the quality of synthetic speech, and demonstrates how the matching function VQ is based on a special form. The discrete form of matching function has been through a continuous process [6], the matching function is based on early VQ's code of this, is a discrete form. It will be a source and target feature space quantization of the feature space, respectively, from the histogram, each code word between the two codes match probability matrix formation, thus completing the source to the target feature space mapping feature space.

$$T = \sum_{i=1}^{\mathrm{n}} wi \times qi. \tag{13.1}$$

Formula 13.1 is a source of voice conversion before and after the target voice pitch frequency trajectory comparison chart, the converted language sound track in the pitch frequency of apparently close to the target speech. Sensory testing is a subjective listening test of a speech signal. In the voice conversion system performance tests, ABX test method is a commonly used test method, which is used to distinguish between different words A and B, respectively, the source and target speaker voice, X said converted voice. In the experimental test, subjects were asked to judge if X is closer to A or B. In this 'closer to' experiment, where 10 subjects of the converted voice were asked to do ABX test. Results are shown in Table 13.1. From the test results, based on voice conversion algorithm Viterbi search GMM is superior to the voice conversion algorithm based on the same time, and objective evaluation consistent results is that the transition between the opposite sex is better than the same between the conversion [7].

In recent years, with the development of electronic technology, smart mobile devices in the actual raw living is more widely used in practical applications, the urgent need for faster McNair, convenient and small man–machine interface, and the traditional keypad or touch panel equipment in this regard does not give satisfactory answers. At the same time, automatic speech understanding Do technologies; Automatic Speech Recognition (ASR) technology has been rapidly the development of some simple speech recognition system can be applied to the embedded level has been Stage, such as telephone voice dialing, smart toys and robot control. Because of the language Itself is the most commonly used form of communication humans, therefore, embedded speech recognition technology, Will become the future for smart mobile devices, human–computer interaction is an important choice. In order to address the high recognition rate, complex models with limited hardware resources, the conflict between shields, we use a secondary search algorithm. Using this method, because the model in a phase of low

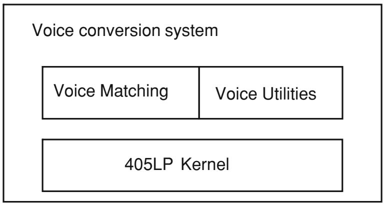


Fig. 13.2 Voice conversion system

complexity, saving hardware resources, improves the recognition speeds. In the second stage, be recognized very few number of entries, which can be highly accurate model of knowledge triphone does not perform. English pronunciation in the process, due to changes in tone between the situation of phonemes are more a result of the use of triphone model to more accurately reflect the synergy between the english phonemes pronunciation, better recognition effect. However, due to the large number of triphone models of in the n phonemes, need a triphone model to describe such a complex model in practice, cannot afford, so methods using a certain state of poly class is required. In this article, use the decision tree (DT) clustering and data-driven (DD) state the method of combining clustering, design triphone model structure for the 3-state output probability density function for the 8 mixture of the GMM. Clustering obtained using decision tree clustering 1635 model and then use data-driven method for the 674 together to form the second-order segment recognition model. In the data-driven clustering, we define two triphone models and divergence from the model are as follows

DPM policies are data structures registered with and interpreted by the DPM implementation in the kernel. Policy activation is controlled by an application-specific, executable policy manager, also provided by the system designer. The policy manager is optional, as some systems may be effectively power managed by a single policy installed at system initialization.

Figure 13.2 shows the different components and their interactions in a DPM-enabled system. Several researchers have arrived at the conclusion that an optimal power management policy will require the active participation of power-aware tasks in an operating system capable of task-/event-specific power and performance management [8]. However, we believe that in many systems only a small number of application programs will ever be modified to be power-aware. Therefore, the system must be effectively able to handle mixed workloads of power-aware and conventional tasks. We also do not believe that a single task running on a general purpose system should be able to unilaterally set the system operating point, unless that task also has the authority to act as a complete power-policy manager for the system.

The DPM task state mechanism for task specific power management was arrived at as a compromise between competing concerns for simplicity, flexibility, performance and optimum power efficiency. Under DPM, the various task states are recognized as separate operating states of the system. The task state index of each task is stored in the task structure, and is not interpreted by the core OS.

Instead, DPM policies map operating points to the various task states, and whenever a task is scheduled the context switch invokes the DPM layer to actuate the operating point associated with the task's DPM task state. The task state mechanism allows privileged, power-aware tasks to indirectly set their own operating point or the operating points of other tasks by changing task state assignments [9].

Only the policy manager determines the task state to operating point assignment, however, by choosing to activate an appropriate DPM policy. IV. DPM STRATEGIES we argue that DPM is a useful abstraction because of its ability to easily implement a wide variety of effective power management strategies. In the following we describe several strategies that we have implemented and tested on experimental platforms based on the 405LP. Single-policy strategies the simplest DPM strategies require only a single policy and no run time policy manager. An example is the idle scaling (IS) strategy. We are particularly interested in strategies that combine load scaling, which has proved to be "good enough" for general-purpose applications without real-time constraints, and task specific requirements. We refer to these hybrid systems as application scaling (AS) strategies. Video decoding is commonly used as an example of an important workload that is difficult to power manage without application participation, so we developed an AS strategy for an experimental [10], multi-threaded MPEG4 video/audio decoder for Linux.

13.4 Conclusions

This power-aware video decoding approach [11] does not require extensive analysis and prediction of processing requirements or the implementation of new scheduling mechanisms, but still achieves good results. We present this AS example to illustrate an interesting use of DPM's policy mechanisms without which the application would need significant additional complexity to make it power-aware. We made two simple changes to the video decoding threads to make it easier to power manage this application with our LS policy manager. We modified the video decoder to begin processing the next frame immediately upon completion of the current frame, rather than idling.

This paper presents a fixed-point DSP-based voice commands embedded knowledge of english other system that uses two-stage CDHMM model identification, the first phase of a 20 state triphone model, quickly identify the second phase of the state with 674 tri accurately identify the phone model. Share and model by model number of the state the author of innovation: And other methods improve the recognition rate.

Spectrum of normal human speech signal is composed by a number of frequency groups, each spectrum signals of spectrum groups showed convex envelope shape, the amplitude at its center frequency at the largest of its neighboring frequency signal amplitude is gradually reduced.

The envelope signal by the network processing, while maintaining its central location at the same frequency spectrum, can reduce the packet the width of the network, thus increasing the distance between groups of different frequencies to help in speech recognition to extract feature signals. The following speech to a real spectrum as the net input, display the network. Network voices translate to the perception of the model function.

In the white noise environment, the Chinese “North” spectrum. The signal input to the network, the network processing, and the output signal. The results show that the network model in addition to filtering noise, but also identify the main frequency section of the signal, making the spectrum of the signal more clearly.

References

1. Rao KS, Yegnanarayana B (2006) Prosody modification using in-stants of significant excitation [J] . IEEE Trans Audio Speech Lang 14(3):972–980
2. Wu CH, Chsia C, Liu TH, Wang JF (2006) Voice conversion usingduration-embedded bi-HMMs for expressive speech synthesis [J]. IEEE Trans Audio Speech Language Process 14(4):1109–1116
3. Wang G, Liang, Weiqian, Liu J et al (2005) In: Moderate vocabulary English speech recognition system embedded on a chip [J].Qinghua Daxue Xuebao/Journal Tsinghua University 45(10):1393–1396
4. Elovitz HS, Johnson R, McHugh A et al (1976) Letter-to-soundrules for automatic translation of English text to phonetics [J]. IEEE Trans Acoustics Speech Signal Process 24:446–459
5. Morrison D, Rui-L Wang, I De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres [J]. Speech Commun 49(2):98–112
6. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods [J]. Speech Commun 48(9):1162–1181
7. Hong-Xing Li, Ling-Xia Li, Jia-Yin Wang et al (2004) Fuzzy decision Ma-King based on variable weights [J]. Math Comput Modell 163–179
8. Ramamohan S, Dandapat S (2006) Sinusoidal model-based analysis and classification of stressed speech [J]. IEEE Trans Audio Speech Lang Process 14(3):737–746
9. Marchand Y, Damper RI (2000) A multi-strategy approach to improving pronunciation by analogy [J]. Comput Linguist 26(2):195–219
10. Hasan MM, Nasr AM, Sultana S (2005) An approach to voice conversion using feature statistical mapping [J]. Appl Acoust 66(5):513–532
11. Nwe TL, Foo SW, De Silva LC (2003) Speech emotion recognition using hidden Markov Models [J]. Speech Commun 41(4):603–623

Chapter 14
Necessity and Programs in Student Information Management System

Ning Li, Quanrui Wang, Guohong Gao and Hui Ye

Abstract With the further expansion of college enrollment, and the growing size of schools, the management of students becomes more complex and onerous. School development based on the student information management is becoming higher and higher, to facilitate the school management to ensure that student information is secure, accurate, and should rely on advanced information technology to solve various problems encountered in the management proposed for centralized management of university student information management system design ideas.

Keywords Student information management system · Necessity · System design goals · Needs analysis · System development strategy

14.1 Introduction

With the further expansion of college enrollment, more and more students have access to institutions of higher learning with the growing size of the school, making the student management more complex and onerous. With an enormous amount of information, we need to have student information management system to enhance the efficiency of student management. High-efficiency and accurate management system for students' information is a necessity for the improvement of the quality of teaching methods in high schools with the enlargement of

N. Li (✉) · Q. Wang · G. Gao · H. Ye
Institute of Information Engineering Henan Institute of Science
and Technology, Xinxiang 453003, China
e-mail: 30240547@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_14,

information and the aggravation of assignments on students' management [1]. In view of such problems as few business and dispersed management existed in students' management in high schools, management system for high school students' information should be established for the improvement of students' management and information's security as well as the lessening of duplicate task. Therefore, the systematic management of student information, standardization, and automation is an inevitable requirement.

14.2 The Necessity of System Development

14.2.1 From the Student Perspective

Student information is on behalf of the students' identity such as students' basic information, academic achievement, enrollment changes, glory, rewards and punishments, and student fees and other informations [2]. The data reflect a situation of the students' performance at school, so that a student should have only relevant information to their counterparts, whose accuracy, uniqueness and safety are essential. This requires that the management of information must be centralized, rigorous and timely. However, the traditional and man-made managing ways have been used for long time, which lies in many defects with low efficiency and poor confidentiality. Moreover, plentiful files and data may arise with time lengthening out. As a result, this brings about much difficulty in searching, updating and maintenance. Coping students' information with computer shows an incomparable advantage, compared with artificial management, such as swift index, convenient search, high credibility, great storage volume, excellent confidentiality, long validity, low cost and so forth. Such advantages can greatly improve the managing efficiency and they are as well the main ways of the realization of scientization and normalization in information management.

14.2.2 From the Perspective of Student Information Management

Student information management includes student record management, integrity management, performance management, rewards and punishments management, funding management, student fees case management, employment management and so on. This information is primarily employed by the Admissions Office, Office of Academic Affairs, Student Affairs Office, the Treasury, the Communist Youth League and other departments, but Counselor of the Department is mainly responsible for the information analysis and consolidation [3]. If we use a manual way to collate above data every time,which will lead to waste of human and material resources and is not conducive to the late management. So, for convenient, fast and accurate grasp of the situation of students, it is very much necessary to develop a systematic and integrative student management. This system is primarily based on computer as a tool to disengage administers from the tedious computing by gathering, confirmating, processing, operating, analysing, managing and maintaining the information needed for students management, ensuring it more energy to engage in researching implication on students managing policy, in researching improvement on managing method and in supervision checking on the efficiency of management, and then achieving a desired effect overall [4]. In the current information age, each school needs a practical management system to standardize the management rules about the school, students, data statistics and analysis, which will greatly boost the managing level, optimize the resources, reduce the cost by a wide margin and make efficiencies as many as possible.

14.3 The Development of System Design

14.3.1 System Design Goals

System development is to achieve the overall objective of the systematic management of student information, standardization and automation [5]

1. Easy to set professional
2. Easy to class management
3. Easy to curriculum
4. Easy to complete new enrollment information, registration, older students, enrollment information query, report printing
5. When students need to query results, they have an easy acess to print information required from the database

14.3.2 Needs Analysis

(1) The system function

Student performance management system is mainly to facilitate the efficient management and online information access platform. Students can access information through the system, and administrators can manage all information. The main functions of this system are the following:

1. Student management: In order to facilitate student to add, delete, modify and query information
2. Course Management: Administrators can modify curriculum and other relevant information by filling out the form

3. Performance management: Administrator can add and modify student performance information in the database.
4. Class management: Administrators can use this function to add, delete, modify and query the class information.
5. User management: Users can add, delete, edit and view the landing of the program user, and the super administrator can set user permissions.
6. Information Sharing Function: make individual information for each student shared to other department where necessary.

(2) The database needs analysis

In order to achieve the objectives and functional requirements of student management information system, as shown below, we designed data items and data structures:

1. The student basic information: student number, name, sex, date of birth, the class, department, date of admission, telephone number, home address, etc.;
2. Class of information: class number, department, head teacher, classrooms, etc.;
3. Department of the Ministry of Information: Department name, phone, basic information, course number, course name, course type, course description, etc.
4. The student achievement information: student number, name, course of study, results and so on.

14.3.3 Choice of Development Tools

(1) Microsoft Visual Studio 2008

Visual Studio is a complete set of development tools, used to build ASP.NET Web applications, XML Web Services, desktop applications and mobile applications. Visual Basic, Visual C + + , Visual C # and Visual J # all use the same integrated development environment (IDE). With this IDE we can share tools and help to create mixed-language solutions. In addition, these languages use. NET Framework features. With this framework, we can use the simplified version of ASP Web applications and XML Web Services development of key technologies.

(2) Microsoft SQL Server 2005

SQL Server 2005 is a comprehensive database platform, with integrated business intelligence (BI) tools, providing enterprise-class data management. SQL Server 2005 database engine provides a more secure and reliable storage for relational data and structured data, so you can build and manage high availability and high performance for business data applications. Overview SQL Server 2005 features.

SQL Server 2005 data engine is the core of the enterprise data management solutions. Furthermore, SQL Server 2005 combines the analysis, reporting, integration, and notifications. This allows your business to build and deploy cost-effective BI solutions and helps your team through the scorecard, Dashboard, Web services and mobile devices push data applications to all areas of business. The close integration with Microsoft Visual Studio, Microsoft Office System and the new development kit (including the Business Intelligence Development Studio) makes SQL Server 2005 different. Whether you are a developer, database administrator, information worker or decision maker, SQL Server 2005 can provide you with innovative solutions to help you benefit more from the data.

(3) Database objects

Table is not the only object of SQL Server 2000 database. Specific stored datas or entities operated can be referred to as objects. For example, the key, constraint, index and so on.

The main ones are the following:

1. Table: the most common and most frequently used for organizing and storing data objects, like the ranks of arrangement.
2. The virtual table defined by the View and logic: exported by one or a few basic tables.
3. Stored procedure Trans-act-SQL code packages: executed in SQL Server 2000 S-side, callable and reusable.
4. Triggers: a special stored procedure, associate forms, the implementation of data integrity.

14.3.4 System Development Strategy

(1) The use of advanced B/S structure, (Browser/Server) B/S structure, i.e, the browser and the server architecture, it is with the rise of Internet technology a structure for the change or improvement of the C/S structure. In this structure, the user work interface is achieved through the WWW browser, very small part of business logic is achieved in the front (Browser), but the main business logic is implemented on the server side, thus formating the so-called three-tier 3-tier structure. This greatly simplifies the client computer load, reducing the cost and effort to maintain and upgrade system, reducing the overall cost of the user (TCO). Look at the current technology, local area network set up network application of the B/S structure; and through Internet/Intranet mode, database applications, it is relatively easy to grasp and the cost is lower. It is the one place in the development and can enable different people, from different locations, in different access methods (such as LAN, WAN, Internet/Intranet, etc.) to access and manipulate a common database; it can effectively protect and manage access to data platform, the server database is also very safe. B/S architecture management software is convenient, fast and efficient.

(2) Net advantages and the breakthrough improvements in the. Net environment are the following: it uses a unified Internet standards (e.g XML) to a different system and docking; This is the Internet's first large-scale highly distributed

applications framework; using a management process called "coalition", which can manage service programs running on the platform comprehensively and provide them with strong security background;

Major breakthrough for ASP.NET to Asp:

(1) Because of different operating mechanism, asp is an interpreted programming framework, whose core is vbs and js,whose restrictions on the two scripting languages decide deficiencies of the asp and which cannot do the underlying operating like a traditional programming language. So if you need some operations, such as socket, file, etc., and have to resort to using other traditional programming languages, such as C++, VB, JAVA and other components of the preparation, and because interpreted, so it is greatly reduced in the operating efficiency as for The ASP.NET, it is a compiled programming framework, whose core is NGWS runtime, in addition to that it can be used as a programming language like asp vbs and js, you can also use VB and C # to write, which determines its powerful features, that is, low-level operations can be carried out without the help of many other programming languages.

(2) Efficient data processing, ASP.NET not only brings about ADO.NET, but also the SQL Managed Provider in ASP.NET; we have two ways to connect database: ADO.NET Managed Provider and the SQL Managed Provider, among them, the first way can connect any ODBC or OLEDB data center, while the second way can connect MS SQL Server.

14.3.5 Security Policy

Security of the system is the issue designers must consider, so they should make full use of the operating system and database system security in order to combine it with the security of the applications system, while also taking some special measures to improve system security.

1. Login authentication. In addition to ensure the user's normal access to the login page of the system, but also it can prevent users from trying to bypass the login page to access the system of non-normal.
2. Access control. In addition to that the system users must set the password, it has special requirements for the permissions assigned. System uses the functional module, roles and the user three-tier distribution rights.

   Firstly set information of modules permission, which define what module functions can be provided by the user, and define permissions code. Secondly, according to business needs, set different roles, each role is relatively independent on sub-modules permissions. Finally, set the role the user belongs to, that is, provide the user the ultimate authority. This setting can improve the flexibility of distribution of competences to modify the permissions of a class, users only need to modify permissions to their respective roles, and it can ensure that different users have different functions.

3. Log management. The system provides complete logging operation, used for fault diagnosis, finding the problem, operating log to record the user's major operating commands, the operator, IP address, operating hours and so on.

14.3.6 System Test

Analysis of test results:

(1) The different roles landing test

Login process the user's login name, password and verify the role of the normal landing.

Passing user login information normal landing, access control is normal.

(2) The system management module test

Users can add the information of role and the user name normally, information inputted into validation module also works.

(3) Profession, class, students, curriculum and module test results

Information can be properly entered and displayed in the browser window, the basic filter can achieve the desired effect, delete and modify functions normally, but not perfectly, the basic functions can be properly realized.

14.4 Conclusions

This Web service-based student information management system has the following characteristics: It uses Browser/Server three-tier architecture, so that the system has good maintainability and reusability. In the process of developing this system, ASP + IIS + SQL Server mode is used, this mode will clear the display and the logic of separation, making the code easier to manage, suitable for large-scale development projects. The middle layer uses a database connection pool technology to speed up data processing speed with database server, but also to speed up the response speed of the client. Background database use the SQL, which is more powerful feature, in addition to that it can handle a variety of platforms included in the database management system kernel, but also including data replication, database system management, Internet gateway support, online analytical processing, multimedia support and a variety of parallel processing capabilities.

References

1. Lin C (2005) Visual Basic 6.0. People Post Press, UK
2. Gong P (2005) Oriented Programming Series. Higher Education Press, Beijing
3. Zhang X, Xu M (2009) C # based and case development explain. Tsinghua University Press, Beijing, pp 344–403
4. Hao A, Xu Y, Kang H, Hong W (2005) SQL Server 2005 Essentials and experimental guidance. Tsinghua University Press, Beijing
5. Duan X, Zhou S (2008) College students need for integrated information management analysis. Shanxi Coll Soc Sci 3:120–121

Chapter 15
The Establishment of Student Management Application Platform with ASP Technology

Wenlong Wan, Peixi Deng, Wenxian Xiao and Yulan Li

Abstract The original students' management system generates many problems such as low efficiency, poor security and produces a large number of redundant files and data which are difficult to find update and maintain. The article aims to design and develop college students management application system, which is based on Web with ActiveX Server Pages (ASP) program to explore the approaches to constructing the student management system in the college campus network, to provide a theoretical basis for the realization of network management of student information, and eventually realize efficient, remote and interactive management for college students through the campus network.

Keywords ASP · Student management · Networking · College

Student Management is an important part of college management. Management of students has been using the manual method for a long time, which leads to many shortcomings such as low efficiency, poor security, and producing a large number of redundant files and data which are difficult to find, update and maintain. As the process of social information continues to advance, computer networks increasingly penetrate deeply into the group of college students, so the college student management faces new challenges and opportunities, and the traditional management methods cannot meet the new situation. How to apply information technology into the management of university students is an issue needed to be solved quickly. Using ASP technology combined with VbScript, JavaScript scripting language, HTML hypertext language and ActiveX components for to create dynamic, interactive and efficient Web server application programs for

W. Wan (✉) · P. Deng · W. Xiao · Y. Li
Henan Institute of Science and Technology
Henan Xinxiang 453003, China
e-mail: wanwenlong@hist.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_15,

college students management information, that is to say, ASP student management system, will enable students management information resources to be fully shared, improve efficiency, achieve information interaction and remote management.

15.1 About ASP

ActiveX Server Pages (ASP) is the server-side scripting environment developed by Microsoft, which can be combined with HTML pages, ASP commands and ActiveX components to create dynamic, interactive and efficient WEB server application programs [1]. The establishment of multi-user mode enables ASP system to achieve the sub-level management, and to solve database structure and the scientific distribution of permissions is the key to establish a multi-user mode [2]. The commonly used multi-user mode at present is:

Mode 1: the users are divided into several fixed levels; the permissions distribution of the same level is consistent.
Mode 2: all users are managed regardless of level; the permissions of all users can be different.

Using the ActiveX Data Objects (ADO) component in the database access components, ASP can be connected to the database and Web pages to achieve data transmission. ADO component can access relation-based, text-based hierarchy-based or other types of databases. The operation of the database can be achieved by means of calling ADO object in ASP. For example, the establishment of Conn.asp files to complete the connection with databases:

```
SET Conn=Server.CreateObject (``ADODB.Connection'')
Connstr=``Provider=Microsoft.Jet.oledb.4.0; data source=''
& Server.MapPath
Conn.Open Connstr
```

Database is composed of a number of tables, whose creation must be set based on the model of multi-user. The table established for the mode I only needs to simply set three fields such as the user name, password, and permission, in which the data type of the user name and password field is set as String sub-type, while the data type of permissions field is set as the Integer subtypes.

The setting of users and the assignment of user permissions are operated by the administrators who own user management permissions through the Web page [3]. For mode I, it can be completed with the following source:

```
    <!--#include file=``conn.asp''-->
    <% if request.form.count <>0 then
    Sql=``Select * From Managing Users ''
    Set Rs=Server.Createobject(``Adodb.Recordset'')
    Rs.Open Sql, Conn,3,2
```

```
Rs.Addnew
Rs (``User Name'')=request.form (``user'')
Rs (``password'')=request.form (``pwd'')
Rs (``permission'')=request. form (``Purview'')
Rs.Update
Rs.Close
End if
%>
```

For mode II, source code (assuming permission entries is 10) is available

```
For i=1 to 10
Rs (``permissions''&i) = request. form (``purview''&i) = 1
Next
```

In place of `Rs (``permission'') = request. form (``Purview'')` statement in the source mode I to complete.

ASP achieved the identification of the user by means of WEB pages, for example: the Login. asp can be regarded as the login to screen user identification, and a text box can be used to transmit the data on the user name and password to the current page in the way of the Post method to carry out the user authentication[4???]. the source code is as follows:

```
    <!--#include file=``Conn.asp''-->
    <%
    If request.form.count < >0 then
    Sql=``Select * From administrative user where username = '''&
Trim (Request.Form (``User'')) &`` ' ''
    Set Rs = Server.Createobject(``Adodb.Recordset'')
    RS.Open Sql,Conn,1,1
    Here test whether it is the empty set is saved
      If Rs (``password '') = Request. Form (``Pwd'') then
    Session (``User'') = Request. Form (``User'')
    Session (``Purview'') =Rs (``permissions '')
    Response. Redirect ``index.asp''
    else
    response. redirect ``Error.asp''
      end if
    end if
    %>
```

This is the source code of user identification corresponding to the model I. “Permission” fields value recorded by a user is assigned to Session object. WEB pages determine the user’s level by the identification of Session (``purview''); for model II, the following code is needed to replace the Session (``purview'') = rs (``permissions'') statement

```
for i = 1 to n
session (``purview'' & i) = rs (``permissions'' & i)
next
```

where the value of n is the number of all permissions items, this source code assigns the value of each permissions item of some user to Session object, and then WEB pages determine whether the user has an administrative rights by means of the recognition of Session (``purview?'')

The user accesses the corresponding operation page after the screen of the identification, that is to say, users with different permissions will enter a different user interface; it is also possible to set a unified interface, in the operation of each task, to first identify the user rights, and only users with the permission can operate accordingly. Either way, the source code which is used for user identification is necessary to add in the front of user operation pages [5].

For mode I the following source code can be used: (assuming the authority code to operate the current page is ``1'')

```
    <% if Session (``User'') =`` '' then Response. Redirect
``Login. asp''
    If Session (``purview'') < >``1'' then Response.Redi-
recr ``Error.asp'' %>
```

For mode II the following source code can be used: (assuming the permission item of current page operation is ``permission 1'')

```
    <% if Session (``User``) =`` `` then Response.Redirect
``Login.asp``
    If Session (``purview1``) =false then Response. Redi-
rect ``Error. asp``%>
```

Here is to determine whether the user has this authority by means of identification of the object of the session, and redirect the users who do not have the authority to operate the current page with the Response object's Redirect method. If the authentication process does not validate user permissions, the data record opening the current user in the process of user identification is needed to identify in WEB pages.

```
    < % if Session (``User'') = ``'' then Response. Redirect
``Login. asp''

    If Session (``purview'') < >``1'' then Response.Redirecr
``Error. asp'' %>
```

15.2 A Student Information Database

The establishment of student management system needs a strong database support. ASP using ADO components (ActiveX Data Objects) enables network system developers to easily connect databases with the Web page and to achieve data transmission between Web pages and databases. The common database is generally Microsoft Access or SQL Server.

15.2.1 Create Database (Take Microsoft Access for Example)

First, establish the database structure. Subtotal student information in the form of a data table, and build a separate tabulation for information in common, such as name, gender, class and ethnicity. Tables should be built, respectively, for non-common information, in which table records must be distinguished by a unique index, such as ID, Student ID and so on [6].

Second, define the data format. According to the data requirements in the data table define data formats for each field, such as "Name" is defined as "text" format, "Student Number" is defined as "digital" format and "birth date" is defined as "Date/Time" format.

Third, establish the linking between tables. Table needing to establish a connection must have one or more fields in the table corresponding with the connection object field or multiple fields.

15.2.2 Database Management and Maintenance

Take regular backups of the database data; download the database from the server to the local computer for restoring the system when it is necessary; important data can be stored in permanent storage medium.

Database maintenance can be completed by one or more users with different administrative privileges, the data entry, modification and deletion can be remotely operated, also in the local computer remote operation consistent with the characteristics of ASP can be used, which can ensure the integrity of data. But if scripting vulnerability exists, garbage data will be caused easily, and the local operation will easily disrupt the connections between data tables.

15.2.3 Data Security Measures

(1) Use the implicit script, so that the browser cannot display its operation on the database.

(2) Hide the database address, conduct encryption on the convert format of the database, so that visitors cannot find or properly download and open the database.
(3) Verify and resolve the problem of scripting vulnerability.
(4) Add secure authentication script for the management page using the built-in ASP Objects "Session", forcing the operation of the data to pass secure authentication.

15.3 The Database Processing of Web Page

The management of remote database needs to realize in the form of WEB pages, using ASP host script VbScirpt with ASP objects, methods, processes, etc., the database can be created, added, modified, deleted, searched and other operations [7].

15.3.1 Database Link

ASP can use ActiveX Data Objects (ADO) components for the preparation of a compact simple script to link to databases Open Database Connectivity (ODBC)-compliant and data sources Object Linking Embedding (OLE) DB-compatible.

```
Create a linked object: SET Conn = Server.CreateObject
(``ADODB.Connection'')
Create a data engine: Connstr = ``Provider = Microsoft.
Jet.oledb.4.0; data source = '' & Server.MapPath (data
source)
Open Data Link: Conn.Open Connstr
```

15.3.2 Data Retrieval, Adding, Modification and Deletion

As to data retrieval, you can use the ASP Connection object's Execute method to obtain records set, as follows: `SQL = ``Select [*] From [sheetname]''`

Set RS = Conn.Execute (SQL)

Or by creating a RecordSet object to get recorded set, the object RecordSet of ActiveX Data Objects (ADO) components is the data collection to save the database search results [8], the code is as follows:

```
SQL = ``Select [*] From [sheetname] [where condition]''
SET RS = Server.Createobject (``ADODB.Recordset'')
RS.Open SQL, Conn, [a], [b]
```

Standard SQL statements determine the operation content of the data, so the designer may set the appropriate SQL according to the need, but SQL statements must be written before the data retrieval for reference while retrieving data records. The above SQL statement [*] indicates the section name to open, [sheet name] as the data table will be retrieved, [where condition] as the conditions of retrieval, which is optional. All records will be stored into the RS set if there is no retrieval condition, such as:

SQL =“Select name, student number, gender, from student information where gender = ‘M’ ”

When data records need to add, modify and delete, the change of SQL statements is only necessary. If to retrieve the data records by the method of creating the RecordSet object, you can set the parameters [a] and [b] to define the reading ways and permissions to open the data table, to use RS.addnew, RS.update, etc. in the data processing to realize the adding, modification and other operations of data record.

15.4 The Mode of Classification Management

ASP student management system can be divided into three levels of management, namely: super administrator, general manager and student users.

1. Super Admin administrative privileges: set general manager, assigning administrative rights; set management mode, management structure; review the information the general manager add or modify; check management reports submitted by the general manager and so on.
2. General Manager administrative privileges: add, delete, modify, grade, class and student information; more access conditions, summary, statistics, retrieval of student information; output student information statements; browse, review and respond to feedback submitted by students; submit management reports to the Super Management Members.
3. The student user’s browsing permissions: check their own basic information on grades, scholarships, incentive records; export all the information report; information transmission between student users; feedback information to the administrator.

15.5 The Interaction of Data and Network Interaction

Students management systems built using ASP can easily achieve the interaction of data and network interaction, the share of student management data not only provides the convenience of managers, but also creates a good platform for information exchange among students, and between students and administrators.

The system can set a few of general managers, each responsible for updating the data and management within their own permissions, which reduces the work pressure of student management staff, and its operating records can be recorded into the database by ASP and submitted to the super administrator, as the assessment basis for super administrator to assess the general administrator; general manager will allocate the student management resources to student users, so that students can easily visit the related information, general manager can collect information on student user feedback and make appropriate responses or feedback to the super administrator to take full advantage of position of the information network; student user can use this system to view the system resources general manager allocated to their own, promoting the smooth flow of student management information, student users can achieve horizontal communication among students and transfer of information via the ASP page embedded within by means of message boards, BBS and other student management system. It can also be feedback problems in the management of students and situations to the administrator through this system, so as to make ASP student management system become the contact and bridge between students and administrators.

References

1. Wang W (2000) HTML and XML markup language and comparative analysis[J]. Mod Lib Inform Technol (5):23–24
2. Yang J, Ma Y (2010) BBS ASP-based address book system design and implementation[J]. Comput Learn (04):19–21
3. Lin X (2010) Based on B/S structure NCRE online registration system design and implementation. Inform Comput (Theory) (04):82–83
4. Hui L (2010) ASP + Access network based on examination management system. J Comput Program Skills Mainten 12:50–51
5. Zhiguo S (2005) ASP.NET Application Tutorial[M]. Tsinghua University Press, Beijing, pp 20–180
6. Zhan W (2005) Long sual basic database development classic case analysis[M]. Tsinghua University Press, pp 246–253
7. Liping L, Ge MW, Yong L (2002) XML-depth Analysis[J]. J Liaon Techn Univer 21(2):207–208
8. Wu H, Chen X (2010) Mode of three-layer structure based file management system design and implementation[J]. Comp Appl 8

Chapter 16
Development and Implementation of Enterprise Management System Based on J2EE

Jia Xiaoyun and Zhao Xiao

Abstract Based on the practical design of enterprise management system, this paper illustrates the design idea of enterprise management system, such as design procedure, logical division of layer and function, system efficiency enhancement, and code quality improvement and puts forward how to establish a high-efficient, safe and scalable enterprise management system.

Keywords J2EE (Java 2.0 Enterprise Edit) · JDBC (Java Database Connectivity) · Application server

16.1 Introduction

With rapid development of e-commerce, each branch of the commercial society needs to apply enterprise application program to jointly accomplish its work, for example, enterprises conduct online transaction, enterprise e-commerce web site makes online payment and settlement of accounts used in banking system, enterprise conducts allocation and purchase online, government makes bidding and tendering of urban planning on the web site [1]. In the near future, enterprise application program will have a strong influence on the development of social economy.

J. Xiaoyun (✉) · Z. Xiao
College of Electrical and Information Engineering,
Shaanxi University of Science and Technology,
710021 Xi'an, China
e-mail: sunnyjingyi@126.com

Z. Xiao
e-mail: 39458893@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_16,

The key to develop enterprise application program is not only timeliness but also convenient deployment, flexible migration, easy upgrading and renewal of procedure, etc. Therefore, application developers face two basic requirements: rapidity and high efficiency. What is the key to rapidly develop and deploy a high-quality application program? It is "Architecture" or programming model. In the computer field, "Model"and "Architecture"occupy important positions [2]. A good model can always get twice the result with half the effort, and a good architecture can not only improve development efficiency through logical layer division but also supply convenience for upgrading and migration. Enterprise Management System based on J2EE can solve the problem very well.

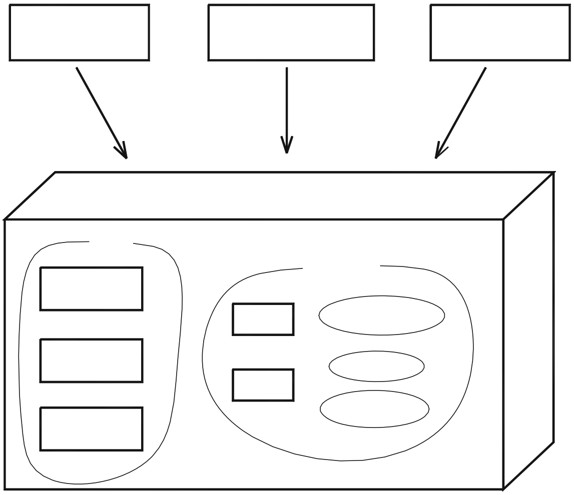
16.2 Implementation Scheme Based on J2EE

16.2.1 Architecture and Hierarchical Structure

No more talking about the disadvantage of traditional C/S Architecture, let us discuss the well-known application program of three layer or multilayer. The three-layer application program had a more early origin than the Web, but Web greatly promoted the development of three-layer application program. In the beginning, people used ASP, JSP, PHP script programs and to implement the interaction between browser and server. The structure encapsulates script program of simple transaction logic, so that the program may run in the Web server and accomplish the tasks, such as database access, security certificate, numerical analysis and so on and then return the result in HTML format to the browser [2]. The Web server acting as the middle layer accomplishes business logic analysis. But the three-layer application program has the following disadvantages: First, the script program containing lots of business logic analysis has low execution efficiency. Second, script code is embedded to HTML label, although many new technologies supply simplified programming, the development debugging efficiency is still not acceptable. Third, the code could not be reused or migrated. Fourth, the program in the form of script code supplied to users is against copyright protection. In addition, with rapid development of e-commerce, people have high requirements for application program, such as RPC; load balancing, concentrated and high-efficient affairs and safety disposal, existing system of integrated enterprise, the most important software reuse and distributed computation, all the new requirements need a more complex architecture, a more subdividable layer and service middle layer with stronger function to support [5]. Server component architecture based on J2EE is the best option for enterprise application.

Both three-layer and multi-layer architecture need a strong interface, where server components run and the Middleware accomplish most of enterprise calculating work. Application server is the running environment for server component. Application server can make much group work, such as transaction processing, safety, thread scheduling, database connection pooling, communication among components etc. Architecture of server component is as seen in Fig. 16.1).

Fig. 16.1 Architecture of server component



Because the application server accomplishes much group work, enterprise application developers may concentrate on design and application of business logic, nothing to worry about "pipe work" required by running critical commercial application. This is a guarantee to be rapid and efficient.

Hence, the architecture adopts multilayer of J2EE composed by client layer, WEB layer, EJB layer and data layer [3]. It has many advantages, such as definite roles and responsibilities, convenient operational maintenance, strong independence, high safety, easy programming, cross-platform etc.

First layer: Client layer

This is a browser layer in charge of the interaction between system and client, for example, to show query result and collect the information input by client in HTML language.

Second layer: WEB layer

This is composed of WEB components like JSP, SERVIET, JavaBean and run by WEB container. It mainly takes charge of invocation of EJB layer and simple logic of some other client. It should be taken into consideration that WEB application layer should only comprise simple client logic, like effective simple judgment input,and not application logic [2].

Third layer: EJB enterprise component layer

Enterprise component layer is run by EJB container to support EJB, JMS, JTA services and technologies. It is the core layer of the whole system, where the enterprise application logic is implemented. On one hand, it transforms object for database record to analyze and design the data by object-oriented method, but on the other hand, it supplies invocation interface of application logic to WEB layer. Fortunately, we do not need many low-level programmers owing to EJB container [4]. In order to ensure the high-efficiency and independence of the system, the layer actually is divided into several sublayers, and then we will discuss it more in the subsequent chapters. We choose Web Logic for the application server of this layer and then we shall describe Web Logic of EBA in detail.

Fourth layer: data layer

It stores physical data, and provides relation data model to EJB layer or WEB layer. This layer is the lowest layer and realized by mature relational database system. Oracle 8i is used in practice [5].

16.2.2 Function Module

This system is developed based on daily management of one Power Supply Bureau Design Division. We understand that the main function of the system is composed by management of blueprint, receiving and dispatch of files and information through the investigation in our own company. Because system client is WEB browser, it is necessary to have a consideration for safety. Except the three management functions, additional function of account management is required. Each management function is divided into log-in, query, modification etc.

16.3 Related J2EE Technology

J2EE comprises of various technologies which supplement each other and implement different functions. Here we introduce several technologies used in management system design.

16.3.1 JavaServlets Technology

It is doable to compile HTTP Servlet in use of JavaServlets which extends server function and responds to user request, and it is similar to traditional EGI.

16.3.2 JavaServerPages Technology

JSP is similar to an ASP page. User can embed short Java code into HTML label. Application program dynamically generates HTML, WML or XML by these codes. JSP also supplies user-defined label library which can separate Java code from TIML label very well for debugging , division and cooperation [6]. At last JSP is transformed to Servlet for implementation and the purpose of JSP is to make script code of server in a more visual way.

16.3.3 EnterpriseJavaBean Technology

Enterprise Bean comprises of the methods of implemented business logic which run in J2EE server and the client accomplishes enterprise analysis by these methods. EJB 2.0 normatively defines three Enterprise Bean:Session Bean, Entity Bean and Message-driven Bean,these three EnterpriseBean execute their functions respectively, for example, EntityBean can directly access database without compiling any SQL language, which will be introduced in the back. It is notable that EnterpriseJavaBean and JavaBean are completely two different concepts, JavaBean is one class, but a EJB maybe comprising of several interfaces, class files and some files describing component properties [7].

16.3.4 JDBC API

JDBC API is a technology of database access of Java, by which user can make database operations in EJB or JSP/Servlets, and he can access those databases connected with ODBC in use of JDBC-ODBC also.

16.3.5 Java Naming and Directory Interface

Any Java application program can use Java Naming and Directory Interface (JNDI) to access user information (user status, telephone, email etc.), machine information (network address, machine setup etc.) and various services. JNDI names those objects such as Web component, EJB component, database resource, files, system, machine and so on. These names are bound to a complete object through name and directory service supplied by J2EE server, and then application program can position these objects by those names.

16.4 Implementation Example

Take direct database access of WEB layer for example:

In the design management system of one Power Supply Bureau Design Division, when a user looks through drawing sheets or material contents, whether screen matches database or not that is not of vital importance, on the contrary, quick display and re-getting information is very important. In order to query lots of data items from database, we do not use EJB but JDBC directly, thus eliminate current payments of EJB (such as JavaRMI, affairs management, data serialization and so on). If we update data, we should use EJB, because many people might update the same record. If we store data by JDBC directly, it is not ensured of integrality, consistency and reproducibility of data[6].

We make a utility class mydb2 for JDBC database access, and the program is as follows:

Utility class mydb2.java

```
package mybean;
import java.sql.*;
public class mydb2 {
  String sDBDriver = “sun.jdbc.odbc.JdbcOdbcDriver”;
  String sConnStr = “jdbc:odbc:xzy”;
  private Connection conn = null;
  private Statement stmt = null;
  ResultSet rs = null;
public mydb2() {
    try {          Class.forName(sDBDriver);       }
    catch(java.lang.ClassNotFoundException e) {
      System.err.println(“mydb(): “ + e.getMessage());        }    }
public ResultSet executeQuery(String sql) throws SQLException{
  rs = null;
conn = DriverManager.getConnection(sConnStr“sa”,“xzy”);
  stmt = conn.createStatement();
  rs = stmt.executeQuery(sql);
  return rs;    }
public void executeUpdate(String sql) {
  stmt = null;
  rs = null;
  try {conn = DriverManager.getConnection(sConnStr);
  stmt = conn.createStatement();
  stmt.executeQuery(sql);
  stmt.close();
  conn.close();      }
  catch(SQLException ex)
  {   System.err.println(“aq.executeQuery: “ + ex.getMessage());
  }   }
public void closeStmt(){
  try{     stmt.close();       }
  catch(SQLException e){         e.printStackTrace();           }    }
public void closeConn(){
  try{        conn.close();      }
  catch(SQLException e){    e.printStackTrace();    }
}  }
```

BookData.jsp calls JDBC method of mydb2.java to accomplish the query function of technical information [7]. No more discussion about the details here.

The above example explains how to adopt JDBC to directly read database on the condition that high-efficient access data is more important than obtaining the up-to-date data, consequently, eliminating the current payments of EJB and meet speed requirement.

16.5 Conclusion

In the society with rapid development of e-commerce and information technology, application developer must cost less money and resource to develop enterprise application program faster, based on J2EE. With the use of its good design, systemic control database, application server and operation system, it is convenient to migrate to other database, and achieve the lowest cost and highest efficiency of the whole system to meet the requirements of all enterprises.

Acknowledgment This work was supported by the special research projects of Education Department of Shaanxi Province (No. 2010JK432), China.

References

1. Zhao K, Huang Z (2001) Java2 Class Base. Engineering Industry Press, Beijing
2. Sestoft Peter (2007) Java Precisely London. The MIT Press, Massachusetts
3. ED Roman (2010) Mastering EJB, vol 2. Wiley Longmen, New York
4. Sun Microsystems Inc.,(2009) Core Servlets and JSP. Scotts Valley, California
5. BEA Systems Inc (2010) Programming WebLogic Enterprise JavaBeans. BEA Systems Inc, London
6. BEA Systems Inc (2008) Programming WebLogic JSP. BEA Systems Inc, London
7. Bales D (2008) Java Programming with Oracle JDBC, Sebastopol. O'Reilly Publishing, New York

Chapter 17
Dilution of Position Calculation for MS Location Accuracy Improvement

Szu-Lin Su, Yi-Wen Su, Chien-Sheng Chen and Chyan-Tay Hwang

Abstract Geometric dilution of precision (GDOP) represents the geometric contribution of observation errors to the positioning accuracy. GDOP is defined under the assumption of equal measurement variances. GDOP was originally used as a criterion for selecting the optimal geometric configuration of satellites in global positioning systems, which presents that the smaller the value the more precise the location. In this paper, we apply GDOP concepts to select appropriate base stations (BSs) in cellular communication systems. The proposed BS selection criterion performs better than the random subsets of four BSs chosen from all seven BSs. After BS selection, the proposed geometrical methods provide high accuracy of mobile station (MS) location estimation for time of difference arrival schemes. The results show that the poor geometry problem can be eliminated and the location accuracy can be significantly improved. From simulation results, the performances of MS location strongly depend on the relative position of the MS and BSs. Therefore, it is very important to select a subset with the most appropriate BSs rapidly and reasonably before positioning.

Keywords Geometric dilution of precision (GDOP) · Time of difference arrival (TDOA) · Angle of arrival (AOA) · Non-line-of-sight (NLOS)

S.-L. Su · Y.-W. Su
Department of Electronic Engineering, National Cheng Kung University,
Tainan, Taiwan

C.-S. Chen (✉)
Department of Information Management, Tainan University of Technology,
Tainan, Taiwan
e-mail: t00243@mail.tut.edu.tw

C.-T. Hwang
Tainan University of Technology, Tainan, Taiwan

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_17,

17.1 Introduction

Geometric dilution of precision (GDOP) is widely employed as the criteria to select the right geometric configuration of the measurement units. The smaller GDOP value means the better geometric configuration which brings the more accurate location estimation. If more measurements are available, the optimal measurements selected with the minimum GDOP can prevent the poor geometry effects and have the potential of obtaining greater. GDOP computation assumes that the pseudo-range errors are mutually independent and identically distributed [1]. GDOP can be approximately inversely proportional to the volume of the tetrahedron formed by the ends of unit user-to-satellite vectors [2, 3]. The maximum volume method only requires low computing time which selects a subset with maximum volume of tetrahedron. However, it is not universally acceptable to use this method because there is no guarantee that the optimal subset of the four visible satellites will be found. The conventional matrix inversion method can guarantee the optimal subset but presents a computational burden.

The various schemes to determine the mobile station (MS) location in wireless communication systems include angle of arrival (AOA) [4], time of arrival (TOA) [5] and time difference of arrival (TDOA) [6] techniques. The accuracy of MS location can be strongly affected by the relative geometric configuration of base stations (BSs). To determine the optimal set of BSs, which can yield superior MS location estimation accuracy, GDOP effect must be taken into account. If the geometric relationship of the BSs relative to the MS is poor, the location estimation of MS performs much worse.

For TDOA schemes, we use the subset with the minimum GDOP to estimate the MS location in cellular communications system. The most commonly used approach for calculating GDOP value is to use matrix inversion and only a subset with minimum GDOP is chosen for positioning. By using the BS selection criterion, the results imply that an improvement in MS location accuracy is very obvious. Simulation results show that the proposed BS selection criterion always produces more accurate location estimates than the random subsets of four BSs. It is enough for selecting four BSs for the compromise between completeness of data and simplification of computation.

17.2 Proposed Geometrical Methods for TDOA Schemes

We have proposed the geometrical positioning methods utilizing the intersection of TOA circles and AOA lines to estimate MS location [7, 8]. In this paper, we also expanded the proposed methods to locate MS for TDOA schemes. After using the proposed BS selection criterion, MS location can be estimated by these geometrical methods.

The TDOA technique is based on the measurement of the time difference of arrival of the travel signal sent by the MS and received by the multiple BSs.

From the viewpoint of geometric approach, TDOA value can be used to form a hyperbola. The constant time difference between two BSs defines a hyperbola, with foci at the BSs, on which the MS must lie. At the intersections of these hyperbolas can be given the position of the MS. The range difference between the *i*th and the serving BS (BS1) can be expressed as

$$r_{i1} = r_i - r_1 = \sqrt{(x - X_i)^2 + (y - Y_i)^2} - \sqrt{(x - X_1)^2 + (y - Y_1)^2} \tag{17.1}$$

where r_i is the distances between BSi and the MS, (x, y) is the MS location and (X_i, Y_i) is the *i*th BS location.

In order to achieve high accuracy with less effort, the proposed positioning methods in [7, 8] can be applied to determine MS.

17.2.1 Distance-Weighted Method

Step 1. Find all the intersections of these hyperbolas.
Step 2. The MS location $(\bar{x}_N, \bar{y}_N)$ is estimated by averaging these intersections, where

$$\bar{x}_N = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and } \bar{y}_N = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{17.2}$$

Step 3. Calculate the distance d_i between each intersection (x_i, y_i) and the average location $(\bar{x}_N, \bar{y}_N)$.

$$d_i = \sqrt{(x_i - \bar{x}_N)^2 + (y_i - \bar{y}_N)^2}, \quad 1 \le i \le N \tag{17.3}$$

Step 4. Set the weight for the ith intersection to $(d_i^2)^{-1}$. Then the MS location (x_d, y_d) is determined by

$$x_d = \frac{\sum_{i=1}^{N} (d_i^2)^{-1} x_i}{\sum_{i=1}^{N} (d_i^2)^{-1}} \text{ and } y_d = \frac{\sum_{i=1}^{N} (d_i^2)^{-1} y_i}{\sum_{i=1}^{N} (d_i^2)^{-1}} \tag{17.4}$$

17.2.2 Threshold Method

Step 1. Find all the intersections of these hyperbolas.
Step 2. Calculate the distance d_{mn}, $1 \le m$, $n \le N$, between any pair of intersections.
Step 3. Select a threshold value D_{thr} as the average of all the distances d_{mn}.
Step 4. Set the initial weight, I_k, $1 \le k \le N$, to be zero for all intersections.

If $d_{mn} \le D_{thr}$, then $I_m = I_m + 1$ and $I_n = I_n + 1$ for $1 \le m, n \le N$.

Step 5. The MS location (x_t, y_t) is estimated by

$$x_t = \frac{\sum_{i=1}^{N} I_i \cdot x_i}{\sum_{i=1}^{N} I_i} \quad \text{and} \quad y_t = \frac{\sum_{i=1}^{N} I_i \cdot y_i}{\sum_{i=1}^{N} I_i} \tag{17.5}$$

17.2.3 Sort Averaging Method

Steps 1–3 are the same as those of the distance-weighted method.
Step 4. Rank the distances d_i in increasing order and re-label the intersections in this order.
Step 5. The MS location $(\bar{x}_M, \bar{y}_M)$ is estimated by the mean of the first M intersections.

$$\bar{x}_M = \frac{1}{M}\sum_{i=1}^{M} x_i,\ \bar{y}_M = \frac{1}{M}\sum_{i=1}^{M} y_i\ (M = 0.8 \times N) \tag{17.6}$$

17.2.4 Sort-Weighted Method

Steps 1–4 are the same as those of the sort averaging method.
Step 5. The MS location is estimated by a weighted average of the first M intersections with weight = $\left(d_i^2\right)^{-1}$.

$$x = \frac{\sum_{i=1}^{M} \left(d_i^2\right)^{-1} \cdot x_i}{\sum_{i=1}^{M} \left(d_i^2\right)^{-1}} \quad y = \frac{\sum_{i=1}^{M} \left(d_i^2\right)^{-1} \cdot y_i}{\sum_{i=1}^{M} \left(d_i^2\right)^{-1}}\ (M = 0.8 * N) \tag{17.7}$$

17.3 Calculation of GDOP for TDOA Schemes

Originally, GDOP concept has been widely used to indicate the geometric effect of GPS satellite configurations. GDOP is addressed as a quality measure in satellite positioning. The purpose of the BS selection algorithm is to minimize the GDOP to improve the MS position accuracy. In the TDOA measurements, the accuracy varies with the error, as well as the relative positions of the MS and BSs. If the measurement errors are uncorrelated and have equal variances, GDOP for TDOA schemes can be defined as [9]

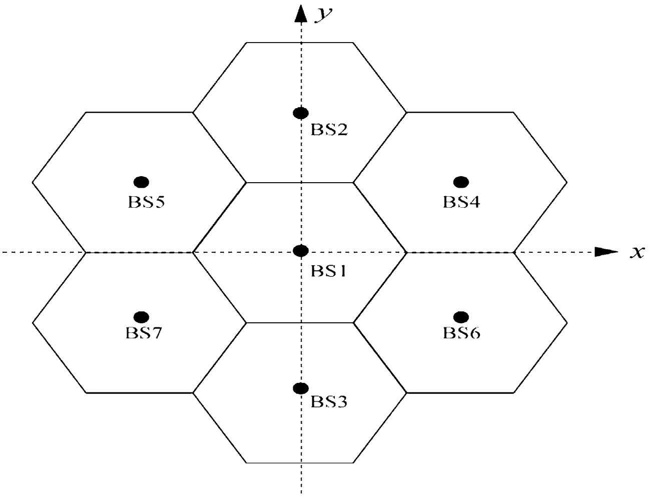


Fig. 17.1 Seven-cell system layout

$$\text{GDOP} = \sqrt{\operatorname{trace}(H^T H)^{-1}}. \tag{17.8}$$

The geometry matrix H for TDOA schemes is $H = \begin{bmatrix} \partial r_{i1}/\partial x & \partial r_{i1}/\partial y \\ \partial r_{i2}/\partial x & \partial r_{i2}/\partial y \\ \vdots & \vdots \\ \partial r_{ij}/\partial x & \partial r_{ij}/\partial y \end{bmatrix}$.

17.4 Proposed BS Selection Criterion

With the increasing in the number of BSs, the computation complexity will increase. The redundant measurements increase the computational overhead and are not able to improve the location accuracy. To further reduce the computational complexity and enhance the performance of location estimation, GDOP concept is a good idea to reduce the number of measurements. The measurements are divided into several subsets and location solution of the minimum GDOP subset can be found. Only a subset of the four measurements selected from among seven for location process in cellular communication systems, as shown in Fig. 17.1. Those BSs are the ones with the minimum GDOP.

The detail of the proposed BS selection criterion is as follows: choose n measurements taken from seven BSs to generate different subset in cellular communication systems, which are divided into $C(7, n)$ possible subsets. GDOP is computed for all subsets of n measurement units and the BSs belonging to the subset with minimum GDOP are the selected BSs at every instant. According to minimum GDOP, n measurement units of this subset are used to find out the MS location solution.

17.5 Simulation Results

Computer simulations were performed to investigate the accuracy improvement through selections of the four BSs. We attempt to improve the performance of the MS location estimate in cellular communication systems. We consider a center hexagonal cell (where the serving BS resides) with six adjacent hexagonal cells of the same size, as shown in Fig. 17.1. Each cell has a radius of 1 km and the MS location is uniformly distributed in the center cell [10]. The serving BS, that is, BS1, is located at (0, 0). The dominant error for wireless location systems is usually due to the NLOS propagation effect.

The NLOS propagation model is based on the uniformly distributed noise model [11], in which the TOA measurement error is assumed to be uniformly distributed over $(0, U_i)$, where U_i is the upper bound. Based on the above BS selection criterion, the most straightforward location method employs the BSs with minimum GDOP to estimate the MS location. Different methods based on GDOP to select the best subset of four BSs to estimate the MS location. The most commonly used measure of the positioning accuracy is the root mean square (RMS) error. Figure 17.2 was performed to examine how the proposed BS selection criterion compares with the subset selecting four BSs randomly when the upper bounds are varied. Four randomly selected BSs with poor geometry perform extremely worse location estimation and the accuracy of MS location can be strongly affected by the relative geometry between BSs and MS. Four randomly selected BSs with bad geometry yield poor location estimation and the proposed BSs criterion provides precise MS location estimation even in severe NLOS conditions.

The second NLOS propagation model is based on a biased uniform random variable [12], in which the measured error of TOA between the MS and BSi is assumed to be $\eta_i = p_i + u_i \cdot q_i$, where p_i and q_i are constants and u_i is a uniform random variable over [0, 1]. The error variables for all BSs are chosen as follows: $p_{\mathrm{i}} = 150$ m, $q_i = 200$ m. Figure 17.3 shows cumulative distribution functions (CDFs) of average location error for different subsets. The subset with minimum GDOP always provides much better location estimation than the other subsets with four BSs taken from seven BSs randomly regardless of the different methods. In order to improve the positioning accuracy, the BS selection with minimum GDOP criterion can be used and optimal geometric configuration is obtained.

The final NLOS propagation model is based on the circular disk of scatterers model (CDSM) [11]. Figure 17.4 shows the CDF of the average location error of the minimum GDOP subset and using all seven BSs method. The radius of the scatters of CDSM is assumed to be 100 m. The positioning precision of using all seven BSs slightly overmatched that of the minimum GDOP subset with four BSs.

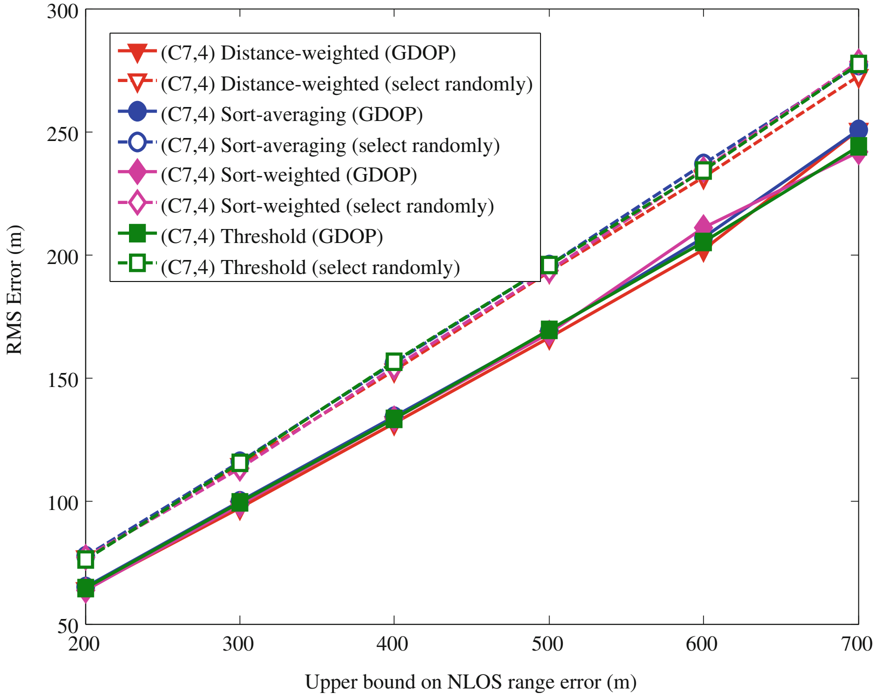


Fig. 17.2 Comparison of RMS errors when NLOS errors are modeled as uniformly distributed

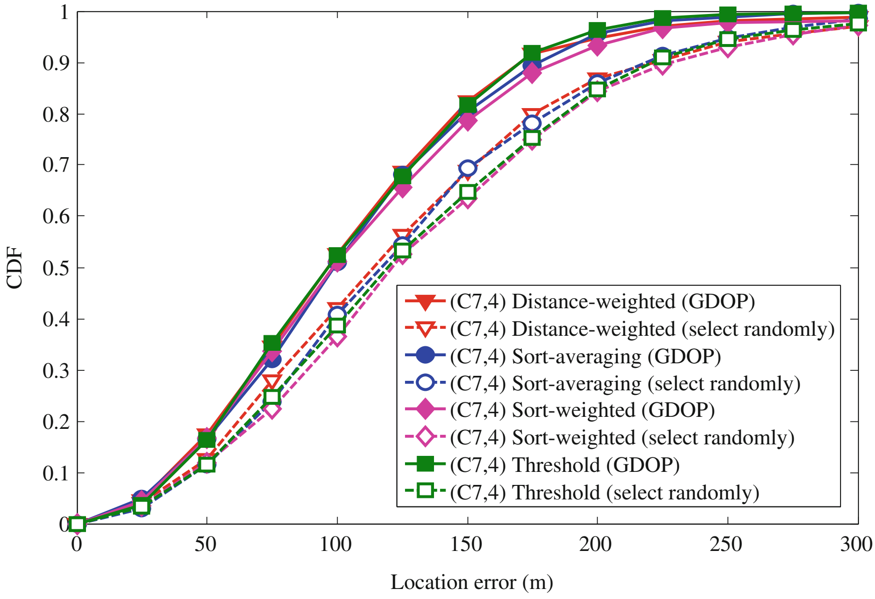


Fig. 17.3 Comparison of error CDFs when NLOS errors are modeled as biased uniform random variables

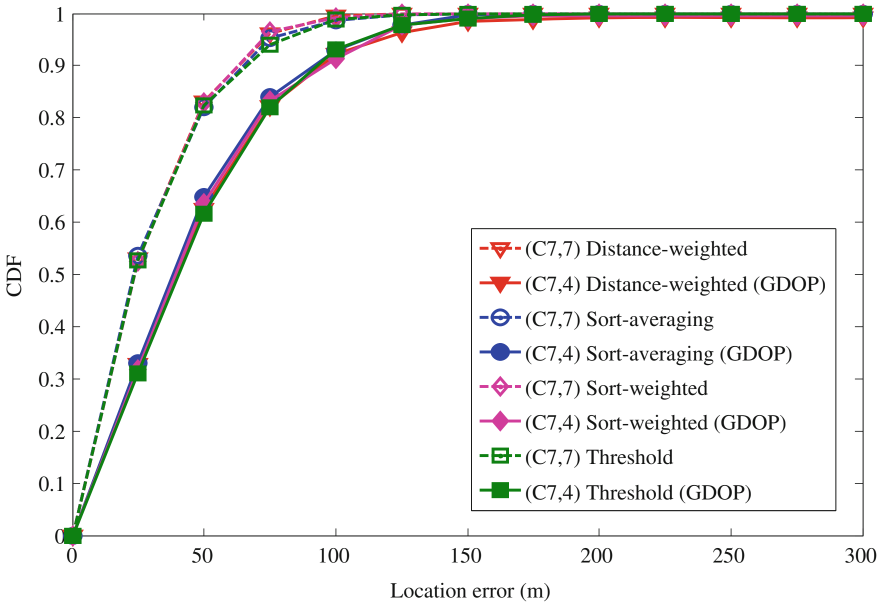


Fig. 17.4 Comparison of average location error using all seven BSs and the subset with minimum GDOP

17.6 Summary

GDOP is a scalar, dimensionless expression and can be relatively simple under the assumption of equal measurement error variances. In order to enhance the performance of location estimation, the selection of BSs with minimum GDOP criterion can be employed to determine the MS location in cellular communication networks. In our simulations, only four BSs with best geometry among seven BSs are chosen to determine the MS location. By selecting the minimum GDOP subset of four BSs, the accuracy of MS location estimation can be improved. It can also be seen that the subset with minimum GDOP can obtain the huge decrease of the positioning errors.

References

1. Kaplan ED, Hegarty CJ (2006) Understanding GPS: principles and applications, 2nd edn. Artech House Press, Boston
2. Hsu Y (1994) Relations between dilutions of performance and volume of the tetrahedron formed by four satellites. In: Proceedings of IEEE position location navigation symposium, pp 669–676
3. Parkinson W, Spilker U (1996) Global positioning system: theory and applications. AIAA Press, New York

4. Chen CS, Su SL, Huang YF (2009) Hybrid TOA/AOA geometrical positioning schemes for mobile location. IEICE Trans Commun E92-B(2):396–402
5. Yang Y, Miao L (2004) GDOP results in all-in-view positioning and in four optimum satellites positioning with GPS PRN codes ranging. In: Proceedings of IEEE position location navigation symposium, pp 723–727
6. Cong L, Zhuang W (2005) Nonline-of-sight error mitigation in mobile location. IEEE Trans Wireless Commun 4(2):560–573
7. Krizman KJ, Biedka TE, Rappaport TS (1997) Wireless position location: fundamentals, implementation strategies, and sources of error. In: Proceedings of IEEE vehicular technology conference, vol 2, pp 919–923
8. Chan YT, Ho KC (1994) A simple and efficient estimator for hyperbolic location. IEEE Trans Signal Process 42(8):1905–1915
9. Venkatraman S, Caffery J, You HR (2004) A novel TOA location algorithm using LOS range estimation for NLOS environments. IEEE Trans Veh Technol 53:1515–1524
10. Chen CS, Su SL, Huang YF (2011) Mobile location estimation in wireless communication systems. IEICE Trans Commun E94-B(3):690–693
11. Al-Jazzar S, Caffery J, You HR (2002) A scattering model based approach to NLOS mitigation in TOA location systems. In: Proceedings of IEEE vehicular technology conference, vol 2, pp 861–865
12. Venkatraman S, Caffery J (2004) Hybrid TOA/AOA techniques for mobile location in non-line-of-sight environments. In: Proceedings of IEEE wireless communications and networking conference, vol 1, pp 274–278

Chapter 18
Research on Application of Data Warehouse to Port Cross-Border Transportation

Xu Qi and Jin Zhihong

Abstract As the gateway for a nation to perform international business, the good development of the cross-border transportation has a great impact on the sustainable economic improvement of a port city and its hinterlands. In that process, the informationization plays a significant role. This paper chose the 20 national ports in the Yunnan Province in China and its adjacent ports in the GMS (Greater Mekong Subregion) as the investigation object, and made an analysis of multi-dimensional data requirement according to the related management departments. Based on that, the paper designed a systematic structure of decision support system for port cross-border transportation under multimodal circumstances using the data warehouse technology. Then, the author realized that the OLAP and multi-dimensional analysis report formed by establishing the DW made decisions on location for international logistics hub by using the model libraries, and supplied the related departments with decision support by the data mining technology. The system revealed the feasibility and effectiveness of the application of the technology in practice.

Keywords Port cross-border transportation · Decision support · Data warehouse (DW) · On-line analytical processing (OLAP) · Data mining (DM)

18.1 Introduction

As the gateway for a nation to perform international business and hub for international transportation, port is an important international logistics node. The positive development of port cross-border transportation (named PCBT below) is

X. Qi (✉) · J. Zhihong
College of Transportation Management, Dalian Maritime University,
Dalian 116026, China
e-mail: cogi@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_18,

an engine to promote the foreign trade of a port city and its economic hinterland. As China's 'big province of ports', Yunnan Province has greatly improved its port infrastructure conditions. The port informationization construction has also revealed rapid development. However, the information that the existing systems supply for PCBT managers is only for data query and statistics, but powerless to perform a deeper level of service such as DM and intelligent-aided decision. The related departments of PCBT need a new-typed data processing system that can support data analysis and decision. This system can at least perform functions such as: (1) integrating massive port logistics data that are separated in different departments; (2) providing decision support of business intelligence for related management departments. All that mentioned above are the advantages of DW and DM.

Lei [1] applied the DW management system to the actual operational management of great highway transportation enterprises; Liu and Dongyuon [2] proposed a space–time model which is suitable for analyzing data of highway management; Wang et al. [3] constructed a logistics decision support system based on DW; Liu [4] made analysis of requirements for the construction of DW and proposed the corresponding logic model by the example of rail sales system; Meng et al. [5] constructed a multi-dimensional analysis model for a certain port, and discussed the role, constructing mode and application prospect of the DW technology in port logistics analysis. Hugh et al. [6] constructed a DW system based on the company of BCBSNC in North Carolina, whose structure and processing procedure provides the best practice example for other companies to develop DW.

As we can see from the research fruits mentioned above, there have been some successful applications of DW in transportation. However, most of these are based on one company or port, under one mode of transportation. The application of DW to PCBT which combines the specialty of cross-border transportation, comprehensiveness of multimodal and advantages of DW is still vacant. Thus this paper chose the 20 national ports in the Yunnan Province and the corresponding ports in the neighborhood as the investigation object analyzed the characteristics of cross-border transportation, and made the DW requirements analysis according to different management departments. Based on that, a PCBT decision support system structure which is suitable for all those related management departments under many transportation modes was then designed. At last, the function of OLAP and multi-dimensional reports was realized by DW, the decision of logistics node location was made by model library, and the aided decision support for the related departments was reached by DM technology.

18.2 Analysis of Cross-Border Data Requirements

Long transportation distance and duration are the characteristics of cross-border transportation. During the process, it is inevitable to experience port inspection, quarantine, customs, border identity identification and other links. This to some degree increased the transportation costs [7]. The contents of data that the related departments focus are different. Thus, it is necessary to consider their data requirements, respectively.

18.2.1 Transportation Management Department

As to the transportation management department which is represented by the Yunnan Communications Department, what they need include:

- Understand the actual situation of difference between the actual service ability and demand of port infrastructure; identify what are the bottlenecks among the infrastructure channels that restrict the cross-border transportation development.
- As to a certain transportation channel, know the proportion of cross-border transportation amount that it entails, and when there is traffic congestion there.
- By forecasting the cross-border transportation amount, judge the traffic development situation between Yunnan and its adjacent countries.

18.2.2 Foreign Trade Supervision Department

As to the foreign trade supervision department which is represented by the Department of Commerce of Yunnan Province, what they need include:

- By forecasting the cross-border transportation amount, judge the traffic development situation between Yunnan and its adjacent countries.
- Understand the general situation of cross-border trade in Yunnan, so as to provide decision support for foreign regulators when judging cross-border trade development trend of Yunnan.
- Through the advantage of DM in the cross-border transportation DW, accurately predict the prospective new economic growth point.

18.2.3 Customs and Inspection Departments

As to these departments, how to implement more effective supervision on staff, goods and vehicles that are involved in the foreign trade will be a main topic for them. Thus what they need include:

- Implement data tracking and monitoring on the moving, releasing and other disposal about goods, personnel and vehicles.
- Based on the analysis of the data in the DW, consider how one can increase the convenience of customs clearance, and reduce the proportion that the clearance process time takes up in the whole cross-border transportation.

- Through the DW, analyze the influence of existing tax mechanism on foreign trade, so that customs can adjust accordingly to promote foreign trade business.

18.2.4 Port

As to a port organization, how to improve its efficiency is the main thesis. Thus, what they need include:

- Analyze the average customs clearance duration in a specific time period (e.g., a quarter, one day), duration, efficiency of each operation in all operation points.
- By the historical data in the DW analyze the operational process in each operation points; optimize them using the function of decision support that the DW has.

18.3 The Systematic Structure of Decision Support System for Port Cross-Border Transportation

According to the requirements that were mentioned in the previous section, the structure of port cross-border decision support system for Yunnan was designed as shown in Fig. 18.1. The structure includes three levels: source data level, data processing level and application level.

18.3.1 Source Data

In the source data level, integrate all the data that are scattered in different departments, including customs clearance data, trade data from department of commerce, transportation data from the communications department, all departments of inspection and quarantine of specific business data and other relevant data. By the data collection platform, these data are interrelated, integrated and stored in the database system. These data are used as the data source for the warehouse extraction.

18.3.2 DW System

The data collected from the data source level will experience the process of ‘extraction-transformation–loading’ (ETL). Meta-data are simultaneously generated and stored into the DW system. The DW system is the core of decision

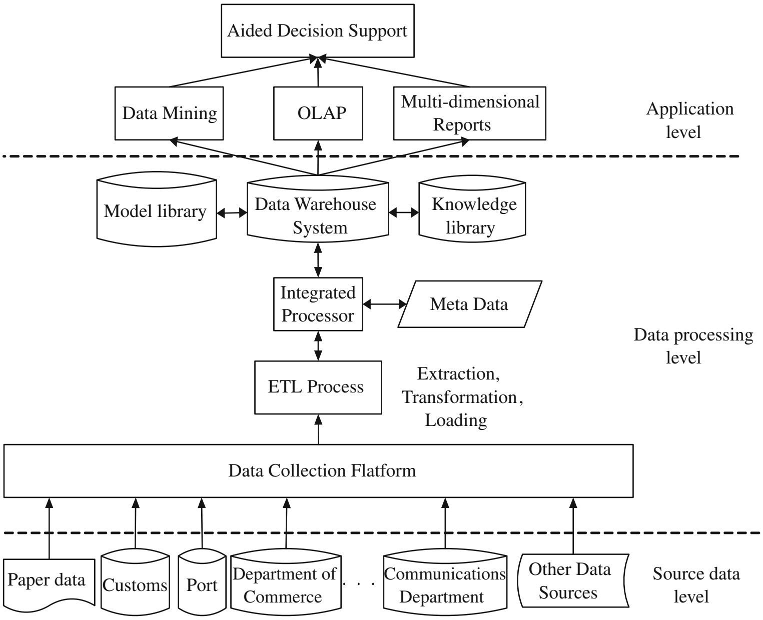


Fig. 18.1 The architecture of cross-border transportation DW decision support system

support system. With the OLAP analysis and DM tools, the role of 'decision support' can be easily realized.

As PCBT involves two kinds of data, which are 'passenger transportation' and 'goods carriage', we adopt the structure of the constellation as the multi-dimensional data organization means for Yunnan port cross-border DW. The structure is shown in Fig. 18.2. There are two fact sheets: freight fact sheets and passenger fact sheets. Dimension sheets of port, time, transportation mode, vehicle, freight information and customs broker information are connected to these two fact sheets by foreign keys.

18.3.3 Model Library and Knowledge Library

The model library in the PCBT decision support system has three basic functional models, which are statistical model, predictive model and simulation model. Among them, the main function of statistical model is finishing statistical historical passenger traffic volume and OD flow, customs clearance data. The main function of the prediction model is forecasting the future development trend of cross-border transportation business based on historical data in the statistical module. The simulation model is the simulation to the actual situation of ports, such as the visualization of the comprehensive transportation network and

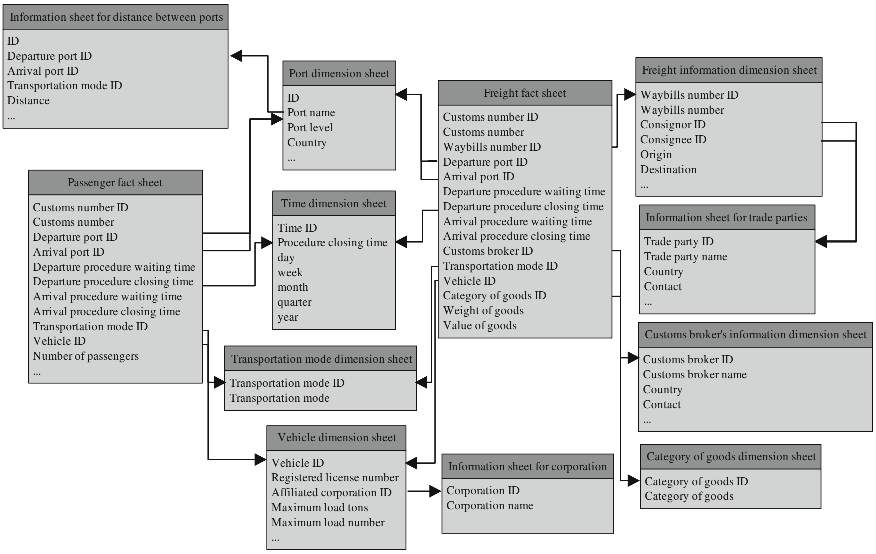


Fig. 18.2 Illustration of different perspectives from related departments of cross-border transportation system

historical OD flow. From the perspective of application, model libraries mainly include network model, transport model, location model and inventory model.

Knowledge library is an intelligent tool to offer solutions for dynamic and complex cross-border transportation problems for policy makers. During the decision-making process, knowledge library interacts with the model in model library and data in the DW, so as to realize the combination of qualitative and quantitative analysis, and help policy makers to clarify decision objects, establish and modify decision model. This knowledge and rules provided in the system mainly include empirical knowledge rules, such as customs clearance knowledge, goods knowledge, port knowledge and some rules in the process of model building and choosing. Besides, they also include knowledge and rules obtained in the process of OLAP.

18.3.4 OLAP Multi-Dimensional Analysis and DM

According to specific cross-border transportation problem, OLAP multi-dimensional analysis can analyze data sets after cutting from different angles. Its basic action includes slice, dice, roll up and drill down, pivot, etc. Through the combination with the model and means in model library, the OLAP analytical capability can be greatly increased.

The emphasis of OLAP is to provide multi-dimensional views of relevant data, so as to analyze historical data. DM is inclined to automatically search for mode and useful information hidden in historical data, to mine knowledge from the DW and put them into knowledge library. The results of OLAP multi-dimensional analysis can be used as the basis of DM, and the DM is a deeper level of knowledge discovery based on the multi-dimensional analysis [8].

18.4 The Realization of Decision Support System for Port Cross–Cross Transportation

This paper will establish multi-dimensional set using MS SQL Server 2005, construct the corresponding data source for the multi-dimensional set using ODBC meta-data administrator, conduct OLAP and DM on the interface of Visual Studio 2005. Based on that, the decision support system for PCBT can be realized.

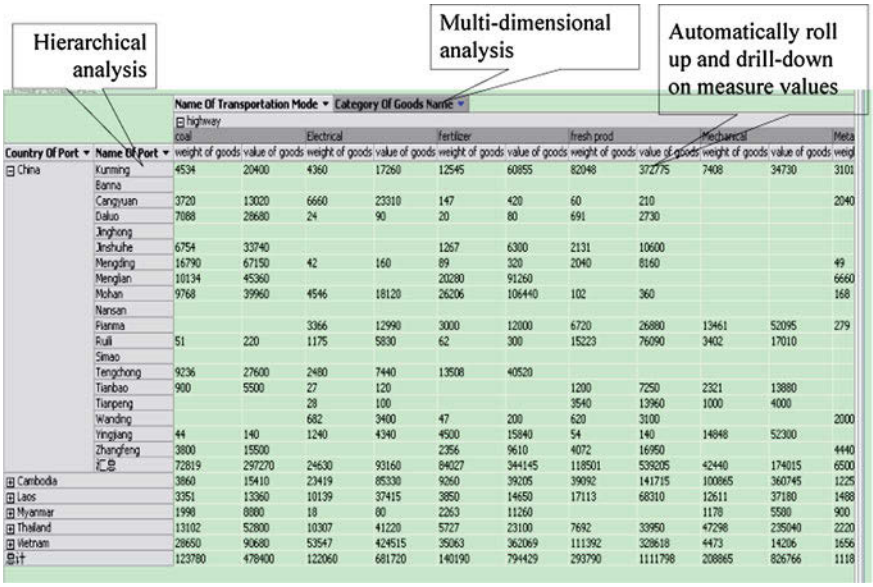
18.4.1 Application Show of OLAP and Multi-Dimensional Analysis Reports

For example, in order to inquire the transportation data of ports in GMS from the perspective of category and transportation mode, one just need to edit the fields required, and then the results will show up in the form of Fig. 18.3. As shown in Fig. 18.3, once ‘country of port’ is unfolded, you can see that the weight and value of goods of Yunnan’s 20 ports are displayed. The effect that Fig. 18.3 shows must have the technical support of a DW system, as traditional database cannot reach the purpose of multi-dimensional analysis well. By folding and unfolding the field selected, one can easily realize the operation of slice, dice, roll up and drill down, pivot, so that the target of specific data query is achieved.

18.4.2 Application Show of System Model

For example, to choose a port in Yunnan to establish international logistics hub. We use the rate-of-flow method location model in the model library. The basic process is as follows: (1) establish transport flow network and collect transport amount data; (2) establish the impedance function; (3) construct the traffic volume distribution model; (4) compute the node transfer amount; (5) construct the location model; (6) compute the location result.

During the process, the system is connected to the MapInfo software, and then the cross-border transportation network model of ports in Yunnan is constructed as shown in Fig. 18.4. Data including transportation speed, cost, distance and capacity between each port of different transportation mode are included in this model.


Hierarchical analysis
Multi-dimensional analysis
Automatically roll up and drill-down on measure values

		Name Of Transportation Mode ▾	Category Of Goods Name ▾									
		⊟ highway										
		coal		Electrical		fertilizer		fresh prod		Mechanical		Meta
Country Of Port ▾	Name Of Port ▾	weight of goods	value of goods	weight of goods	value of goods	weight of goods	value of goods	weight of goods	value of goods	weight of goods	value of goods	weig
⊟ China	Kunming	4534	20400	4360	17260	12545	60855	82048	372775	7408	34730	3101
	Banna											
	Cangyuan	3720	13020	6660	23310	147	420	60	210			2040
	Daluo	7088	28680	24	90	20	80	691	2730			
	Jinghong											
	Jinshuhe	6754	33740			1267	6300	2131	10600			
	Mengding	16790	67150	42	160	89	320	2040	8160			49
	Menglian	10134	45360			20280	91260					6660
	Mohan	9768	39960	4546	18120	26206	106440	102	360			168
	Nansan											
	Pianma			3366	12990	3000	12000	6720	26880	13461	52095	279
	Ruili	51	220	1175	5830	62	300	15223	76090	3402	17010	
	Simao											
	Tengchong	9236	27600	2480	7440	13508	40520					
	Tianbao	900	5500	27	120			1200	7250	2321	13880	
	Tianpeng			28	100			3540	13960	1000	4000	
	Wanding			682	3400	47	200	620	3100			2000
	Yingjiang	44	140	1240	4340	4500	15840	54	140	14848	52300	
	Zhangfeng	3800	15500			2356	9610	4072	16950			4440
	汇总	72819	297270	24630	93160	84027	344145	118501	539205	42440	174015	6500
⊞ Cambodia		3860	15410	23419	85330	9260	39205	39092	141715	100865	360745	1225
⊞ Laos		3351	13360	10139	37415	3850	14650	17113	68310	12611	37180	1488
⊞ Myanmar		1998	8880	18	80	2263	11260			1178	5580	900
⊞ Thailand		13102	52800	10307	41220	5727	23100	7692	33950	47298	235040	2220
⊞ Vietnam		28650	90680	53547	424515	35063	362069	111392	328618	4473	14206	1656
总计		123780	478400	122060	681720	140190	794429	293790	1111798	208865	826766	1118

Fig. 18.3 Illustration of the OLAP multi-dimensional analysis

The impedance function is as Eq. (18.1) shows.

$$c_i(x) = \gamma_d(k_i + \mathrm{op} \cdot \mathrm{L_i}) + \gamma_t t_i \left[1 + \alpha_i \left(\frac{x_i}{C_i} \right)^{\beta_i} \right] \mathrm{vot} \tag{18.1}$$

In the equation, $c_i(x)$ is the generalized cost for section i of a channel; γ_d is the weight of distance cost; γ_t is the weight of time cost; k_i means the fixed cost for section i; op stands for the operational cost of a unit length of channel; L_i is the distance of section i; t_i means the free transportation time on section i; C_i is the capacity of section i; vot is the constant for time value; α_i, β_i are the coefficients of the impedance function.

The traffic volume distribution model is shown in (18.2)–(18.5).

$$\min \sum \int_0^{x_i} c_i(x)dx \tag{18.2}$$

$$\text{s.t.} \sum_k f_k^{rs} = q_{rs} \quad \forall r, s \tag{18.3}$$

$$x_i = \sum_r \sum_s \sum_k f_k^{rs} \delta_{i,k}^{rs} \quad \forall i \tag{18.4}$$

$$f_k^{rs} \geq 0 \quad \forall r, s \quad \forall k \tag{18.5}$$

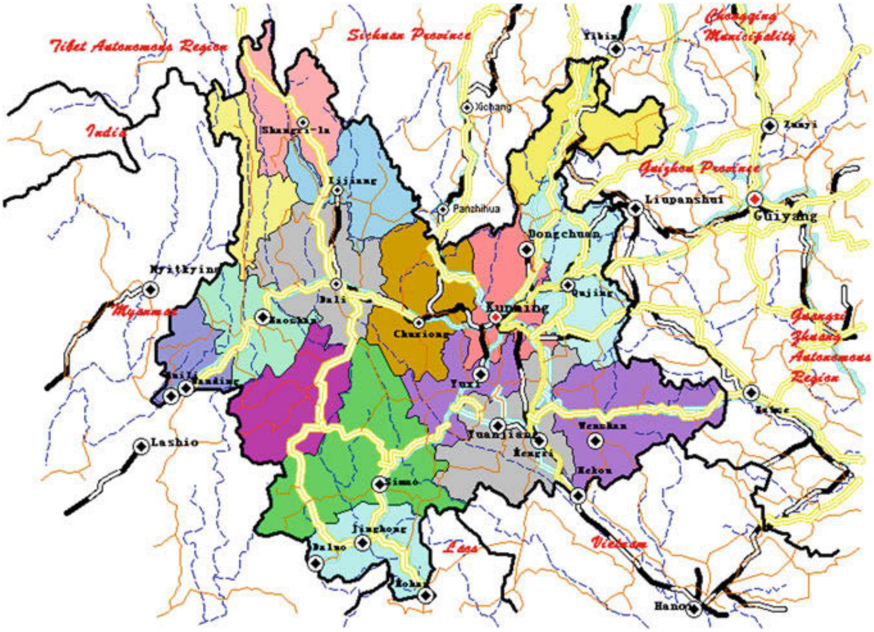


Fig. 18.4 The network model of the cross-port transportation for Yunnan

In the traffic volume distribution model, x_i means the traffic volume on section i; $c_i(x)$ is the generalized cost for section i; f_k^{rs} means the traffic volume on the kth route where the origin is r and destination is s; C_k^{rs} is the generalized cost on the kth route where the origin is r and destination is s; $\delta_{i,k}^{rs}$ is a binary variable, when section i is on the kth route where the origin is r and destination is s, $\delta_{i,k}^{rs}$ equals 1, otherwise $\delta_{i,k}^{rs}$ equals 0.

Last, we use the rate-of-flow method location model, as in (18.6)–(18.8).

$$\max \sum_i \sum_j Q_{ij} x_i \tag{18.6}$$

$$\text{s.t.} \sum_i x_i = 1 \tag{18.7}$$

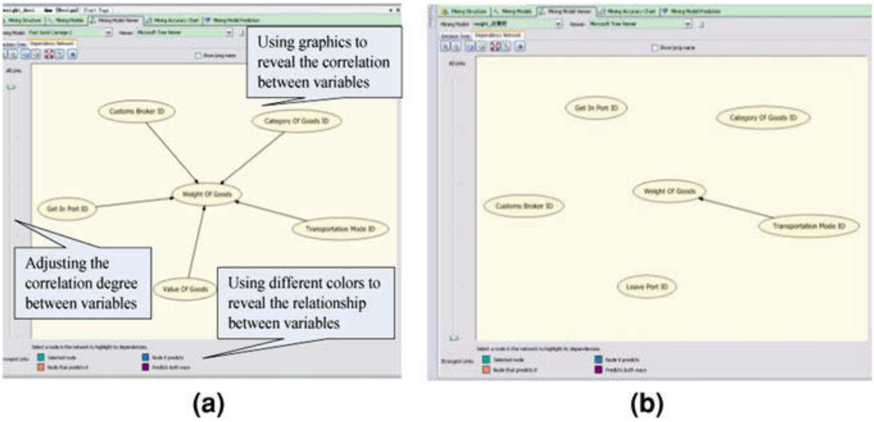
$$Q_{ij} = q_{ijo} \tag{18.8}$$

Therein, Q_{ij} stands for the traffic flow of the ith candidate node on the jth channel; x_i is a binary variable, if the ith node is chosen as the hub, x_i equals 1, otherwise, it equals 0; q_{ijo} is the quantity of shipments that is assigned to the ith node on the jth channel.

We choose Kunming, Ruili, Hekou and Mohan as the candidate port for the international logistics hub. Using the model we proposed before, the annual

Table 18.1 The annual amount of transfer for each candidate nodes and the location results

Candidate node	Kunming	Ruili	Hekou	Mohan
Transfer amount (10,000 tons)	16,278	11,352	13,682	9,204
x_i	1	0	0	0


Using graphics to reveal the correlation between variables

Adjusting the correlation degree between variables

Using different colors to reveal the relationship between variables

Fig. 18.5 Illustration of DM using the decision-tree algorithm

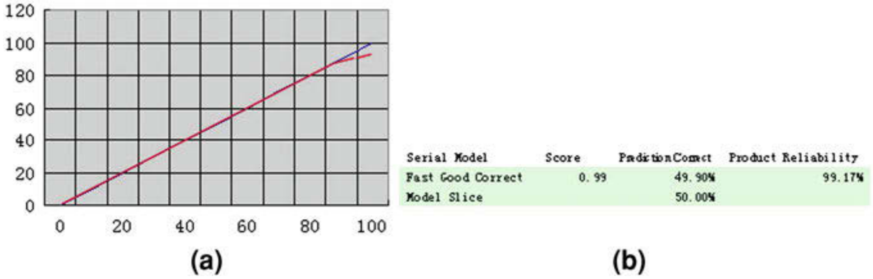
transfer amount and the location result are illustrated in Table 18.1. Therein, 1 means chosen for hub, 0 means not.

By using the rate-of-flow method location model, we propose choosing Kunming as the international logistics hub.

18.4.3 Application Show of DM

For example, we use the decision-tree algorithm to do the DM on the historical freight data in the GMS, so as to see whether there is any undiscovered information. We choose ‘weight of goods’ as the analysis object, and include transportation mode, category of goods etc. as possibly related columns. After processing, the dependency network is shown in Fig. 18.5.

In Fig. 18.5a shows, all of the relevant columns point to ‘weight of goods’. That is to say, they will affect the ‘weight of goods’ to a certain extent. Then we drag the slider to the bottom, namely the strongest link, we get Fig. 18.5b. At this time, only the ‘transportation mode ID’ points to ‘weight of goods’. This means that transportation mode has the greatest influence on traffic volume. This is quite consistent with our inspiration from the fact that the highway traffic volume takes up a great proportion of the total volume.



Serial Model	Score	PredictionCorrect	Product Reliability
Fast Good Correct	0. 99	49. 90%	99. 17%
Model Slice		50. 00%	

Fig. 18.6 The mining accuracy chart for testing the data mining results-lift chart

On the other hand, we also need to analyze the accuracy of DM by electronic interface. Thus, we need to use DM accuracy charts. There are two kinds of accuracy charts, one is lift chart, and the other one is classification matrix. Figure 18.6 is the display of lift chart for the DM accuracy analysis.

As shown in Fig. 18.6a, the diagonal line represents results that an ideal model can be produced, with exactly perfect predictions; the curve line is the result for the DM. The closer the two lines are, the closer the effect of DM to the ideal model. Figure 18.6b shows the relationship of the two lines, and provides the degree of satisfactory for the mining results. In the illustration, the mining results obtain 99 points, which is a very high score.

18.5 Conclusion

As an emerging data processing technology, DW has its special advantages. Its OLAP and the DM tool can greatly raise the ability for processing and application of data information. Based on Yunnan's participation in the GMS, this paper made a data requirement analysis on cross-border transportation and designed a structure for cross-border transportation decision support system based on DW. The structure integrates technology such as DW system, OLAP and DM, so as to strengthen the system intelligence. Through the application of the DW system, more reliable and comprehensive decision support information can be provided. The application show proves the feasibility and effectiveness of the technology in the practice. However, the DW technology in China is still in the stage of research and preliminary application. Building a more perfect PCBT decision support system needs more theoretical support and practical experience.

Acknowledgments This paper was partly supported by the National Nature Science Foundation of China (70572104); the Doctoral Foundation of Ministry of Education of China (20070151002); Transportation construction projects in Western China (2009318791007).

References

1. Jian L (2007) Research on application of DW system to the road passenger transportation. Technoecon Manag Res 2:54–55
2. Xingjing L, Yang D (2003) Multi-dimension analysis model and its achieving method on spatial DW of highway management. Chin Civ Eng J 36:99–104
3. Wang Y, Wang Z , Lu Y, Qian X (2005) Logistics decision support system for multimodal transportation based on DW. J Jilin Univ Technol (Nat Sci Ed) 35:641–645
4. Jingrong L (2002) Application of DW technology in Chinese railway transportation. Railw Transp Econ 24:24–25
5. Meng Y, Wang J, Huang Y, Yang B (2008) Applying data warehousing technology to port logistics analysis. J Shanghai Mar Univ 29:64–69
6. Hugh JW, Celia F, Ariyachandra T (2008) DW governance:best practices at Blue Cross and Blue Shield of North Carolina. Decis Support Sys 38:435–450
7. Li L (2009) Analysis of demands on transit transport. Logist Eng Manag 31:86–87
8. Tie Z, Chen Q, Ruizhao Y (1999) An efficient parallel algorithm for mining association rules. J Comput Res Dev 36:948–953

Chapter 19
Database of City Sound Spaces and its Intelligent Applications

Ji Qing

Abstract Limitation of traditional computer tools about city sound-planning is obvious. Some of them depend on a geographic information system to represent sound environment. Some of them have been designed to evaluate the propagation of traffic noise. Both cannot provide practical sound-planning references for the real projects. For improvement, a new database seems necessary first. Benefitting from the researches about how neighboring sound spaces interact on each other, we designed two kinds of data forms to integrate all useful data into this database. Sound event form is considered as the basic unit to compose a sound space. Sound space form is used to represent each sound space. With the help of this database, complicated sound spaces can be represented effectively. Furthermore, many intelligent applications about city sound-planning can be realized. For example, users can diagnose the sound-planning of their own project by making inquiry in this database.

Keywords Database · Sound space · Intelligent application

19.1 Data Resources

In the field of city sound environment, by the knowledge gained from the previous research projects, we can find out emerging conditions and internal social dynamics of most of the sound phenomena that happen in a single sound space

J. Qing (✉)
State Key Laboratory of Subtropical Architecture Science,
South China University of Technology, Guangzhou 510640,
Guangdong, China
e-mail: sboxi@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_19,

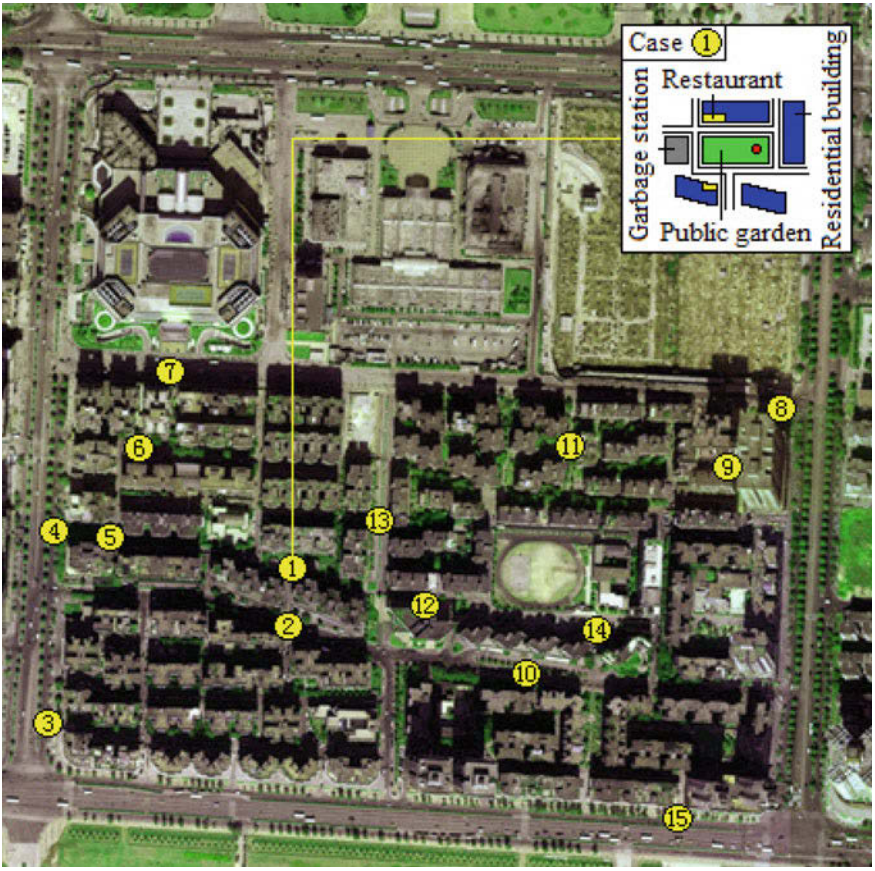


Fig. 19.1 Aerial view of the Tian-He-Nan residential area and the organization way of sound spaces (take case 1 as example)

easily, but researchers still know little about the mechanism of interaction between two neighboring sound spaces. This shortcoming limits our capacity to solve those complicated sound-planning problems [1]. Before developing all kinds of intelligent applications of city sound-planning, to create a new database, which can simulate the organization ways of our city sound spaces, seems inevitable [2].

In order to reflect our city sound reality, all the data needed by this database has been acquired by some investigations and some physical measurements in site. Specifically, in Canton Tian-He-Nan residential area, 15 cases have been selected as research objects (Fig. 19.1). This residential area was developed from 1980s to 1990s [3]. All social living patterns in this area have been stable. These cases reflect typical organizations of sound spaces nowadays in Canton city. Working on site, four kinds of data had been acquired, including investigation of local residents, acoustic measurement, Observation of sound behaviors and two-minutes' sound recording. These data has been collected every hour in a normal day.

So each case will have 24 suites of data to represent the evolution of its whole sound environment in a day with high quality.

Remark 1 For example, case 1 concerns the organization of four kinds of sound spaces, public garden, residential building, restaurant and garbage disposal.

By analyzing all raw data indoors, all the useful data concerning the interaction between neighboring sound spaces can be integrated into the database by two kinds of data forms, sound event form and sound space form.

19.2 Sound Event Form

In this new database, all notable sound events have been considered as the basic units to construct a sound space. Currently, there are 33 types of notable sound events in the database, including nine sound requirement/contribution events, 19 sound contribution events and five sound requirement events.

Besides text details, six values belong to three interaction channels from sound requirement and sound contribution, and have been recorded in the database to describe these notable sound events. These three interaction channels of neighboring sound spaces are degree of permeation, degree of public and degree of nature. All the values for degree of permeation are presented by dB(A) [4]. Among them, the values that belong to sound contribution are obtained by acoustic measurement (Leq) in site. The values that belong to sound requirement are determined by specialists on considering the requirement of the national norms. All the values for the degree of public and the degree of nature are obtained by the sound-reaction test technique, since these two interaction channels can only be represented by qualitative data. After listening to the sound record of a sound event, the testers must judge to which level the degree of public and the degree of nature are belonging to (Five reference values –2, –1, 0, 1, 2 had been set at first to represent the level from low to high). The judgment from the people who carry out this sound events will used as the reference value of sound requirement. The judgment from the people who experience this sound events frequently will be used as the reference value of sound contribution. To those sound events in which human behavior is not included, only the reference value of sound contribution can be obtained. Taking the sound event "pedestrian" as example, six reference values of this sound event are shown in Table 19.1. In database, the abbreviations (I_R, P_R, N_R, I_C, P_C, N_C) have been used to symbolize these reference values.

19.3 Sound Space Form

In this database, each sound space will has its own data form. In this form, the sound event dominating sound requirement and the sound event dominating sound contribution in every hour of a day will be recorded. These sound events will determine

Table 19.1 Sound event form (take “pedestrian” as example)

Sound requirement	Degree of permeation	I_R	Leq = 55 dB(A)
	Degree of public	P_R	0.53
	Degree of nature	N_R	−0.70
Sound Contribution	Degree of permeation	I_C	Leq = 51 dB(A)
	Degree of public	P_C	1.07
	Degree of nature	N_C	0.60

Table 19.2 Sound space form (take “garden with pavilion” as example)

Hour	Sound event	Sound requirement			Sound contribution		
		I_R	P_R	N_R	I_C	P_C	N_C
0–7	Insect voice				43	–2	2
7–8	Insect voice/Morning exercise	55	0.47	1.53	43	–2	2
8–9	Play mah-jong/Pedestrian	55	0.53	0.73	54	−1.33	–1.1
9–10	Play mah-jong/Pedestrian	55	0.53	0.73	54	1.33	–1.1
10–11	Play mah-jong	55	0.53	0.73	54	1.33	–1.1
11–12	Pedestrian	55	0.53	-0.7	51	1.07	0.60
12–13	Pedestrian	55	0.53	-0.7	51	1.07	0.60
13–14	Play mah-jong	55	0.53	0.73	54	1.33	–1.1
14–15	Play mah-jong	55	0.53	0.73	54	1.33	–1.1
15–16	Play mah-jong	55	0.53	0.73	54	1.33	–1.1
16–17	Pedestrian	55	0.53	–0.7	51	1.07	0.60
17–18	Pedestrian	55	0.53	–0.7	51	1.07	0.60
18–19	Rest/Insect voice	40	–0.6	1.47	43	–2	2
19–20	Rest/Insect voice	40	–0.6	1.47	43	–2	2
20–21	Rest/Insect voice	40	–0.6	1.47	43	–2	2
21–22	Rest/Insect voice	40	–0.6	1.47	43	–2	2
22–23	Insect voice				43	–2	2
23–24	Insect voice				43	–2	2

Note: “–2” to “2” means the level from the lowest to the highest

the values of three interaction channels from sound requirement and from sound contribution (all the values can be transferred from the sound event forms). In order to improve the applicable ability, these sound space forms are adjustable. Users can regulate the appearance time and appearance type of the sound events following the reality of their own projects. Table 19.2 is an example of sound space form.

19.4 Diagnose the Sound-planning of a Project

To diagnose the sound-planning of a project by the help of this database is a kind of intelligent application. It can be realized in the database directly.

An example can be used to explain this application of diagnosis. This example is based on a supposed project of urban sound-planning. Now, we are going to deal

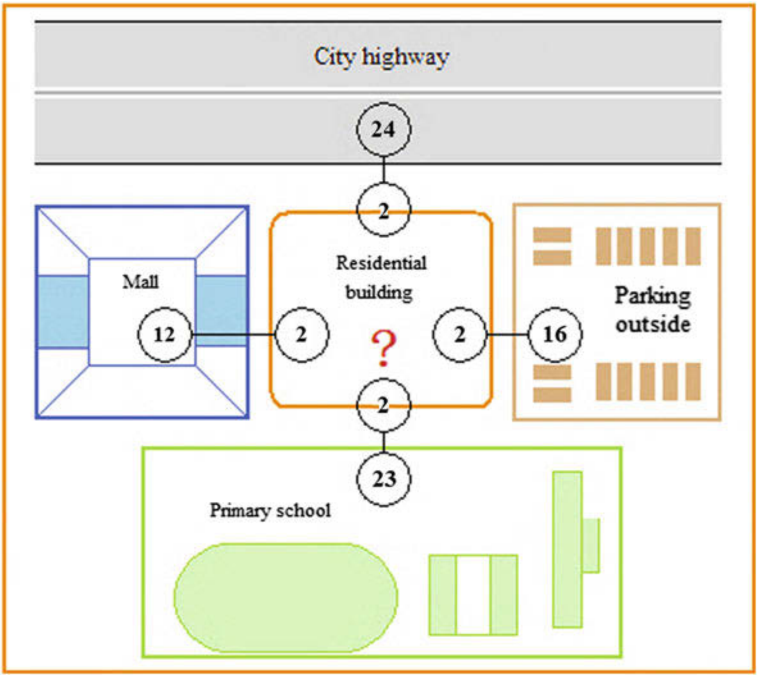


Fig. 19.2 A supposed project of sound-planning

with a vacancy, which is circled by four different sound spaces like in Fig. 19.2, a traffic road at the north, a parking outside at the east, a primary school at the south and a mall at the west. The proprietor of this vacancy wants to build some residential buildings here, but the designer finds out that the sound situation here is not optimistic. So they want some advice from the specialist of sound environment. They want to make sure whether it is reasonable to build the residential building in this vacancy? If they insist to do so, how can they prevent all the potential sound conflicts?

In order to realize this purpose of diagnosis, we can use the sound space forms that this project concerns in the database. By matching these forms following the neighboring relations existing, we can get the diagnostic result. In detail:

First, we consider all the sound space forms to which this project is concerned in the database. In this example, five sound spaces forms ought to be cited: residential building (2), mall entrance (12), parking outside (16), primary school (23) and city highway (24). The numbers following are codes of these sound space in the database. In the real project, before using them, we ought to regulate the appearance time and appearance type of the sound event following the real situation in these forms. Since this example is supposed, so they need not be regulated.

Second, we demand the computer to match these forms following the existing neighboring situation. For each neighboring situation, there exists two times'

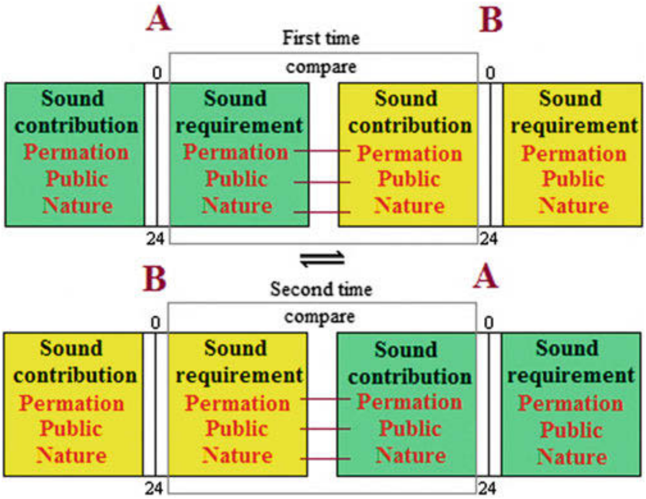


Fig. 19.3 To evaluate the matching level of two neighboring sound spaces by comparing their sound space forms

matching. For example, a neighboring situation concerns the sound space "*A*" and the sound space "B" (Fig. 19.3). For the first time, we match the sound requirement of the sound space "*A*" and the sound contribution of the sound space "*A*", and obtain three values of difference from three interaction channels. For the second time, we match the sound contribution of the sound space "*A*" and the sound requirement of the sound space "*A*", and obtain the other three values of difference from three interaction channels also. After two times' matching, we will have six values of difference for each hour. The smaller the value of difference means the higher the matching level. On contrary, the bigger the value of difference means the lower the matching level.

This process seems less complicated, but the computer can finish it easily. In detail, since the database we used is created by Microsoft Access, we can use the calculation function (consulting form) provided in Access to realize it. This example concerns four neighboring relations: residential building/mall entrance, residential building/parking outside, residential building/primary school and residential building/city highway.

Finally, a diagnosis report can be issued. The substance of this report is to list all the potential sound conflicts happened in each pair of neighboring sound spaces, which can be found easily on considering their values of difference. The bigger the value of difference is, more serious the sound conflicts may exist. From these big values of difference, we can easily know what time a potential sound conflict appears and which interaction channel it comes from. The user can determine the sifting standard. Generally speaking, for degree of public and degree of nature, the conflict situation that has a value of difference bigger than 2, ought to be pointed out. For degree of permeation, the conflict situation that has a value of difference over 15dBA

Table 19.3 A part of the diagnosis report (shows the matching result between one pair of sound spaces, residential building and mall entrance)

Potential sound conflicts between residential building (2) and mall entrance (12)

Hour	Sound events concerned Requirement/ Contribution (Belong to which sound space)	Interaction channel concerned	Value of difference
7–8	Eat a meal (2)/Calm (12)	Degree of permeation	15 dBA
11–12	Pedestrian chatting/calm (2)	Degree of public	3.27
13–14	Sleep (2)/crowd chatting (12)	Degree of public	3.54
13–14	Crowd chatting (12)/calm (2)	Degree of public	3.80
13–14	Crowd chatting(12)/calm (2)	Degree of nature	2.63
14–15	Sleep (2)/single car parking (12)	Degree of nature	2.86
14–15	Sleep (2)/single car parking (12)	Degree of public	2.74
19–21	Crowd chatting (12)/calm (2)	Degree of public	3.80
20–21	Rest (2)/crowd chatting (12)	Degree of permeation	19 dBA
21–22	Rest (2)/single car parking (12)	Degree of permeation	18 dBA
21–23	Pedestrian chatting/calm (2)	Degree of public	3.27
23–24	Sleep (2)/discharge goods (12)	Degree of nature	3.13
23–24	Sleep (2)/discharge goods (12)	Degree of permeation	16 dBA

Note: when “residential building” and “mall entrance” are in neighborhood, there exist two kinds of potential sound conflicts mainly: 1, sleep and rest of the residents are disturbed easily by sound events of mall at noon and in the evening. 2, Since the residential building is calm in some hours, some sound events of the clients may be restrained

ought to be pointed out. Table 19.3 is a part of the diagnosis report of this example, it shows the matching result between “residential building/mall entrance”. By this diagnosis report, specialist of sound environment can provide a very precise sound-planning reference to the designer of the project. From this report, the designer can know which neighboring relations are optimistic and which neighboring relations will cause big sound problems. Further more, the designer can locate all the potential sound conflicts easily and accurately, since the report can point out at what time and which interaction channel they are concerned. Following these indications, the designer can eliminate or weaken these potential sound conflicts by using the traditional methods, such as adding some acoustic construction, installing a new sound space as transition, applying some sound effects, planting some vegetables or having the aid of topography. After all, this diagnosis report can make the process of sound-planning more reasonable and more regular, so that a specialist of sound environment can join an urban planning project more effectively.

19.5 More Intelligent Applications in Near Future

To optimize the organization ways of sound spaces is another intelligent application in sound-planning, which can be realized with the help of this database. In near future, some best organization ways of sound spaces about user’s own urban-planning project can be provided by the computer automatically. Frankly speaking, this intelligent application is an extension of the diagnosis function that we mentioned above. But it cannot be realized in database directly, we must use a series of calculations to finish the optimization steps in a computer program. In this program, a data channel must be established to connect the database, because the process of optimization must use sound space forms also. Basing on some systems' engineering knowledge such as matrix of relation and linear optimization equation, the program can calculate all possible organizations' way of the sound spaces concerned in user's project and find out the best ones.

More and more intelligent applications will be developed to improve our ability to deal with the complicated sound problems in our city. Most of them cannot leave the effective data support, that is the value of this database.

References

1. Augoyard J-F (1999) Du Bruit à L'environnement Sonore Urbain. J. Données urbaines n°3. France: Paris, Ed Anthropos
2. Thibaud J-P (2002) Comment Observer Une Ambiance? J. France: AMBIANCES ARCHITECTURALES ET URBAINES, pp 77–81
3. Balaÿ O (1999) Les Indicateurs de L'indentité Sonore d'un Quartier. CRESSON, France
4. Balaÿ O, Arlaud B (1999) La Representation de L'environnement Sonore Urbain à L'aide d'un Système d'Information Géographique. CRESSON and LISI, France

Chapter 20
Fast Processing in Support Vector Machine for Large-Scaled Data Set

Xilong Qu, Linfeng Bai and Hong You

Abstract Support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The original SVM algorithm was invented by Vladimir Vapnik and the current standard incarnation (soft margin) was proposed by Corinna Cortes and Vladimir Vapnik. A new reduction strategy is proposed for training support vector machines with large-scale data set based on the analysis of the nature and difficulties in training SVM.

Keywords Support vector machines · Fast processing · Large-scale data

20.1 Introduction

Vapnik and others invented support vector machines (SVM) [1] which is a new pattern classification method. SVM is based on the principle of structure risk minimization (SRM) and can effectively avoid the curse of dimensionality and local minimum problems in classical learning methods. The parameters of the

X. Qu (✉)
School of Computer and Communication, Hunan Institute of Engineering,
Xiangtan 411101, China
e-mail: quxilong@126.com

L. Bai
School of Information Engineering, Henan Institute of Science and Technology,
Xinxiang 453003, China

H. You
Training Center of Engineering, Hunan Institute of Engineering,
Xiangtan 411101, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_20,

maximum-margin hyperplane are derived by solving the optimization. There exist several specialized algorithms for quickly solving the QP problem that arises from SVM, mostly reliant on heuristics for breaking the problem down into smaller, more manageable chunks. A common method for solving the QP problem is Platt's Sequential Minimal Optimization (SMO) algorithm, which breaks the problem into two-dimensional subproblems that may be solved analytically, eliminating the need for a numerical optimization algorithm.

It still has good generalization ability under small sample conditions. SVM has got very good results in the area of solving the classification, regression and density estimation problems in machine learning, and has been successfully applied to practical problems of text recognition and speech classification [2], but the training time being very long is a big drawback [3].

20.2 Data Processing Strategy of Support Vector Machine

Since the 1960s, scholars began to study the problem of machine learning based on the data. During the 1990s, Vapnik and his colleagues created a statistical learning theory (SLT), making data-based machine learning to become a more complete theory, and on this basis, finally created a class of effective generic machine learning algorithms—Support vector machines [1]. The application of kernel as a universal technology, can extended to other learning systems. Currently, the main kernel machine algorithm including support vector machine algorithm [2], kernel Fisher classifier [4], kernel principal component analysis [5], kernel independent component analysis [6] tec. Among them, typical algorithm is SVM.

When using SVM to solve practical problems, selecting the appropriate kernel is a key factor [3]. Currently there are a number of ways for the construction of kernel functions. Ling Zhang [7] proved that for a given sample, kernel function must exists. There are still no uniform guidelines yet on the kernel function and parameter selection [8]. Though selected still by experiment, cross-validation method [8] and leave-one method [9] in a limited data set for parameter optimization are commonly used. Literature [10] made a useful discussion on selecting parameters of kernel function, but principally used are the cross-validation method [11] and leave-one method [12] in a limited data set for parameter optimization. In this chapter, using the principle of structural risk minimization [2] and projection analysis to make a quantitative analysis on the kernel function to decision function's influence and to guide the kernel's option, has some innovative. This paper focuses on Vapnik proposed standard support vector machine, but after a few changes also applicable to various support vector machines.

The introduction of the concept of rigid body in mechanics, with the radius of positive and negative samples in N-dimensional space respectively were ζ^{+} and ζ^{-} within the super balls, and the distribution was independent and identical, then the set of positive and negative samples were manifested as two rigid bodies, the type of centroids, respectively were M_{+} and M_{-}. According to statistical theory, when

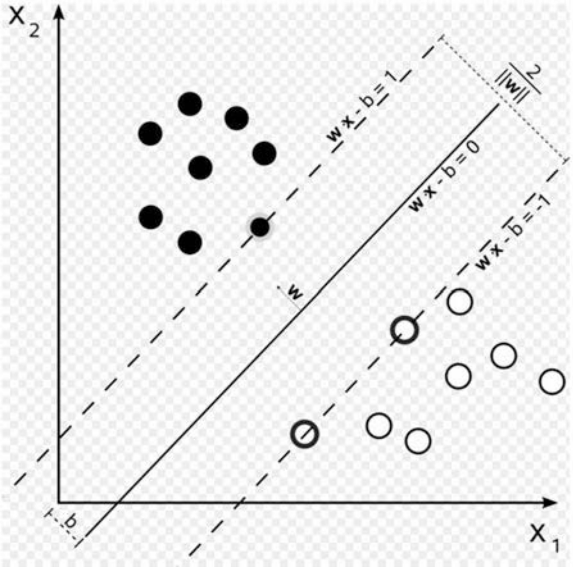


Fig. 20.1 Sample class centroid

the sample set is larger, can use sample set centroid to approximate rigid body centroid. Positive and negative sample set centroid M_+ and M_- are calculated by:

$$M_+ = \sum_{i \in S^+} \Phi(x_i)/l_+, \ M_- = \sum_{i \in S^-} \Phi(x_i)/l_- \tag{20.1}$$

where l_+ and l_- and are positive and negative samples, S^+ and S^- were positive and negative sample sets.

If two sample sets after mapping to the Hilbert space can be separated, then the optimal separating hyperplane is located between two centroids (Fig. 20.1), namely, the two centroids lying on both sides of the optimal hyperplane, and various types of support vector distribute between the class centroid and the optimal hyperplane.

From the view of clustering point, types of rigid body centroid can be regarded as types of cluster center. Suppose two centroid were on the side of optimal separating hyperplane, from the sample distribution, which will occur only when overlapping regions of positive and negative set samples are more, leading to two sample sets cannot be separated. Similarly, according to samples distribution also, one can obtain that the support vector is between the optimal hyperplane and the centroid.

20.3 Processing Algorithms

Call the matrix composed of $K_{ij}(i,j = 1, 2, \ldots, l)$ as kernel matrix K, directly through formula to calculate cos θ_i needed to traverse all K_{ij}. The dimension of the nuclear matrix is l^2, and its computation time complexity is $O(l^2)$. If the training algorithm uses the classical convex quadratic programming algorithm (such as Newton method, quasi-Newton method, etc.), its computational complexity is $O(l^3)$, and the optimization process also needs to traverse the nuclear matrix, the time on support vector preselected will not exceed the training time. However, with the current SVM fast training algorithm (such as SMO), because such an optimization process of algorithm does not necessarily traverse the nuclear matrix, in the case the training set is large, it even appears that the time-consuming status of preselected support vector is longer than time-consuming of training.

$$\cos\theta_i = \begin{cases} p(e - \frac{1}{l_1}a_{1i} + \frac{1}{l_3}b_{1i} - f)/(h_i - \frac{2}{l_1}a_{1i} + e)^{1/2}, & i \in S_1 \\ p(e - \frac{1}{l_1}a_{2i} + \frac{1}{l_3}b_{2i} - f)/(h_i - \frac{2}{l_1}a_{2i} + e)^{1/2}, & i \in S_2 \\ p(g - \frac{1}{l_3}c_{1i} + \frac{1}{l_1}d_{1i} - f)/(h_i - \frac{2}{l_3}c_{1i} + g)^{1/2}, & i \in S_3 \\ p(g - \frac{1}{l_3}c_{2i} + \frac{1}{l_1}d_{2i} - f)/(h_i - \frac{2}{l_3}c_{2i} + g)^{1/2}, & i \in S_4 \end{cases} \tag{20.2}$$

Among

$$e = \sum_{j \in S_1} a_{1j}/l_1^2, \quad f = \sum_{j \in S_1} b_{1j}/l_1 l_3, \quad g = \sum_{j \in S_3} c_{1j}/l_3^2, \quad p = 1/(g - 2f + e)^{1/2} \tag{20.3}$$

use formula 20.2 and 20.3 to calculate the cosine of the samples, not only avoid time cost in repeat counting Kij, but also avoid storage in the nuclear matrix. Set the time to calculate a nuclear function as a unit, then fast algorithm time complexity is $O((l_1 + l_3)l)$, it ensures the feasibility of preselection strategy in large-scale data set.

20.4 Experiment Simulations

The simulation of the UCI data proves the effectiveness of this method. Since classical SVM training speed is slow, experimental Libsvm based SMO algorithm is used with hardware environment CPU-AMD2600, memory-512 M, programming environment using MS VC ++6. The following results with the average to do five times, the time units as milliseconds are obtained.

Adult data set classification experiment, to verify the effectiveness of presentation set of sample centroid.

For sample set, select Adult data set of UCI [7], Adult data set has been widely used to test the classification algorithm. Adult data sets totally divided into nine groups, we use the a5a, a6a data sets to do the experiment, the sample is 123-dimensional vector. Kernel function select RBF kernel, $K(x_i, x_j) = \exp$

$(-|x_i - x_j|^2/(2\sigma^2))$, the parameters $\sigma = 0.05$, $C = 1$, $\varepsilon = -0.1$. Experimental setting $r_+ = r_- = r$.

The training set sizes of 5a and a6a respectively were 6414 and 11220 even in the case $r = 0.05$, the results of the training also includes most of the support vector, the classification accuracy rate of test set also has no change, while time-consuming of sample pre-selection is only 1/20 of when $r = 1$. This fully shows use part of sample to approximate calculate the centroid of sample can essentially shorten the time of sample pre-selection in the basic premise of not lose generalization ability of SVM. From the experimental data when $r = 1$, also shows that if do not reduce the number of elements which need to calculate in the nuclear matrix, the sample pre-selection method will be not worth the candle. When the order of magnitude of data sets between 103–105, we generally take $r = 0.1$ for the experience value, for the data set which the order of magnitude greater than 105, can reduce the values of r.

20.5 Conclusion

Seen from these two experiments, set the empirical value $r = 0.1$and $\varepsilon = -0.1$, the scale of sample set after preselected is about 60% of the original scale. Total training time (including preselection time-consuming) is <70% of direct training time-consuming, its ability to promote basically unchanged, shows the method of preselection support vector in this paper is effective.

Acknowledgments Authors gratefully acknowledge the Projects Supported by Scientific Research Fund of Hunan Provincial Education Department (09C271, 08A009 and 10C0569) for supporting this research. Project supported by Provincial Natural Science Foundation of Hunan (10JJ6099) supports the research. Project supported by Provincial Science & Technology plan project of Hunan (2010GK3048) supports the research. This research is supported by the construct program of the key discipline in Hunan province.

References

1. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn, vol 20. http://www.springerlink.com/content/k238jx04hm87j80g/
2. Joachims T, (1999) Transductive inference for text classification using support vector machines. In: Proceedings of the 1999 international conference on machine learning (ICML 1999), pp 200–209
3. Meyer D, Leisch F, Hornik K 2003 The support vector machine under test. Neurocomputing 55(1–2):169–186. http://dx.doi.org/10.1016/S0925-2312(03)00431-4
4. Zheng Z, Yang J (2006) Support vector machines learn new methods of large-scale training data. J Comput Eng Des (13):2425–2431
5. Aizerman EB, Rozonoer L (1964) Theoretical foundations of the potential function method in pattern recognition learning. Autom Remote Control 25:821–837
6. Boser BE, Guyon IM, Vapnik VN 1992 A training algorithm for optimal margin classifiers. In: Haussler D (ed) 5th Annual ACM workshop on COLT, ACM Press, Pittsburgh, pp 144–152

7. C-W Hsu, C-C Chang, C-J Lin (2003) A practical guide to support vector classification. Technical Report Department of Computer Science and Information Engineering, National Taiwan University
8. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. Advances in Neural Information Processing Systems 9, NIPS 1996, 155–161, MIT Press
9. Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
10. Ferris MC, Munson TS (2002) Interior-point methods for massive support vector machines. SIAM J Optim 13(3):783–804
11. Luo Yu, Wende Yi, Danchen Wang (2007) Sample reduction strategy for support vector machines with large-scale data set. Comput Sci 10(34):260–265
12. Wikipedia Support vector machine. http://en.wikipedia.org/wiki/Support_vector_machine

Chapter 21
Research and Design of Web Management System for Data Manipulation Base on the Internet of Things

Qiyao Xiang, Yongjun Zhang, Wanyi Gu and Jiarui Kou

Abstract Internet of Things (IOT) has been called as the third wave of the world's information industry following the computer and the Internet. Its essence is interconnecting and sharing information,i.e.,the automatic identification of things through the Internet with RF automatic identification technology. Although without human intervention, it still needs to provide a management platform to realize the artificial management, check and fault handling, etc. This paper mainly puts forward a model of web-based management system with the technology of IOT, in order to improve the operation, management and maintenance ability, and ensure the normal operation of the monitoring, and improve environmental safety and reliability.

Keywords IOT · Management system · Web-based · UDP · Data process

21.1 Introduction

Internet of things (IOT) is also known as sensor networks, the concept of which was made in 1999. The definition of IOT is, in accordance with the agreed protocol, connecting anything with Internet through RF automatic identification

Q. Xiang (✉) · Y. Zhang · W. Gu · J. Kou
Key Laboratory of Information Photonics and Optical Communications,
Ministry of Education, Beijing University of Posts
and Telecommunications, Beijing 100876, People's Republic of China
e-mail: qiyaox@bupt.edu.cn

Y. Zhang
e-mail: yjzhang@bupt.edu.cn

W. Gu
e-mail: wyg@bupt.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_21,

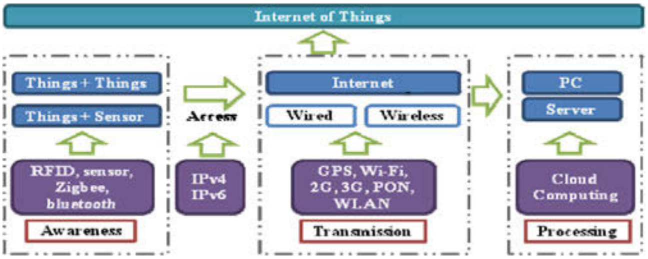


Fig. 21.1 Three platforms of IOT

technology (RFID), infrared sensors, global positioning system, laser scanner and other information sensing device, to exchange and share the information, in order to achieve a network with the function of intelligent identification and management. In this network, the goods can communication with each other without people. The essence is to realize the automatic identification of items and information interconnection and sharing through RFID and computer network [1].

In the exchange of things and things without human intervention, there must be an interface that can be provided for human to manage and check the objects. On the other hand, the automation management needs a strong management system to handle. If there is a abnormal situation happening, it can also send anomaly information automatically to solve the malfunction in time.

IOT should have three support platforms, namely the perception platform, the transmission platform and the processing platform. RFID and sensor networks solve the problem of the identification of object and information perception. More problems such as address assignment, information representation, transmission, storage and processing, security and value-added services require more technology to support [2]. Now it has been put forward to the cloud computing to solve these problems. But cloud computing still remains mature and standardized, so nowadays we can use web-based management system to provide the processing platform of IOT. The main technologies of each platform are showed in Fig. 21.1.

21.2 Structure of Web-Based Management System

As the scale and complexity of the network is increasing, a simple unified structure is needed for network management, so web-based management system appeared. Web-based management system is a kind of information management system with Browser/Server (B/S) structure. It just needs a web browser to be installed in the client. It makes the administrator to achieve the management function through any browser which connects to the Internet with its strong distribution [3]. In the web, we can query, update or handle the data and problem of the monitored devices. It is not required for the client to update the software. Simply by increasing the web

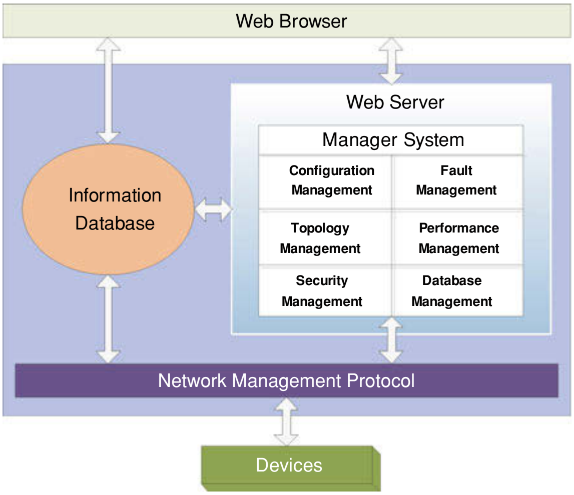


Fig. 21.2 Web-based management system

pages to extend network management function, the cost of maintenance and management can be reduced tremendously. In summary, it has advantages as follows [4]:

The remote control can be conveniently used.
It is easy to learn and use with the consistent management interface.
The independent platform.
It has quick access to the network equipment system through the web when there is an emergency situation.

The traditional web-based management system mainly includes the following several modules: configuration management, fault management, topology management, performance management, security management and database management, which is shown in Fig. 21.2 [4].

As shown in Fig. 21.2, the network management protocol is the base of the management system. Protocol is the rules for anyone to communicate with others. In the Internet, in order to assure the communication between different devices, they must comply with different agreement that can accomplish many tasks. Protocol is defined as followed [5]:

Message format, e.g., how much data can be held in each message.
The method to the destination's path information sharing by intermediate devices.
The method to process information for update between intermediate devices.
The process to initiate and terminate communication between hosts.

Nowadays, Transmission Control Protocol/Internet Protocol (TCP/IP) is the most fundamental Internet protocol. TCP/IP defines electronic devices' standards about how it connects with the Internet and how the data transmit between them. TCP/IP is a four-layer system structure. The top layer is for transmission control protocol, it is responsible for gathered information or put the documents into smaller bag. The lower layer is for Internet protocol, it handles each packet's address to send these packages to the right destination.

In TCP/IP, TCP and User Datagram Protocol (UDP) are the two most widely used transmission protocols. UDP is a simple and connectionless protocol, which has the advantage of providing low cost data transmission; because it transmits the datagram just try its best. TCP is a connection-oriented protocol, which generates additional costs to achieve additional functions, such as original sequence processing, reliable transmission and flow control. Each TCP segment packaged in the header of application layer data has a cost of 20 bytes, while each UDP segment has a cost of 8 bytes [5].

21.3 Requirement of IOT Management System Module

In Internet of things, the monitored device is a gateway, which collects different information such as temperature, humidity, smoke, voltage and so on, through many kinds of sensor devices, and then sends the information to the server to manager through the network transmission equipments and lines. When the server receives the information, data are processed as needed. In summary, there are three important layers from bottom to top: transfer protocol, data processing and GUI showing.

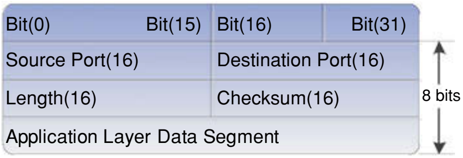
Transfer protocol. In the Internet of things, the underlying device needs to send information to the upper in real-time, so the traffic in the network is too much. In order to reduce the congestion caused by the cost of frequent communication, we can choose this low cost data transfer protocol, UDP, as the underlying communication protocol of the system. Figure 21.3 is an example of UDP segment [5].

As showen in the Fig. 21.3, the application layer data segment is used to encapsulate the information collected by the gateway or the orders send to it. We need to develop a unified format for up and down link protocol, need to consult a server-port and client-port to receive and send data. Then data's capsulation and decapsulation will be completed automatically.

Data processing. When we receive data, a series of data processing need to be done: data collecting, data analysis, history management, trend analysis and so on. In consideration of the management server of the traditional web-based management system's basic modules and the specific needs of this paper, the management system in Internet of things can be divided into the following modules: user management, data management, fault management, topology management and emergency management.

User management. This module's main function is the user login and security-related identification, including the sign-in system verification, exiting the system to sign-off, password changing, the user information query, the user group information query, add or remove users and groups and so on. It can restrict different users to access and manipulate the system through different levels by setting administrative privileges.

Fig. 21.3 UDP segment and header fields



Bit(0)	Bit(15)	Bit(16)	Bit(31)
Source Port(16)		Destination Port(16)	
Length(16)		Checksum(16)	
Application Layer Data Segment			

Data management. This module is the most basic in the manage server because all the functions for management and processing are based on real-time, accuracy and completeness of data. It mainly includes data collecting and saving, history management, data analysis, database backup and query. The system collects data through the socket and then classifies and stores them into different fields of database, creates historical records, analyzes and processes data by the established mathematical model, takes back up of the database automatically and periodically to prevent data loss caused by system crashes.

Fault management. To reflect the status of each managed device, when the system initializes, the module adds the fault information of each device from database to each management object. When the system is running, real-time alerts is also notified to the administrator by this module.

Topology management. This module is responsible for direct management of network resources and timely performance about the operational status of the managed device. It collects the topology information of the network, and changes the physical objects into the graph nodes to provide an easy form for administrator.

Emergency management. This module is a post-processing of data management and fault management, mainly used to classify the fault information to make different emergency alert measures according to the different alert levels. The alarm levels can be divided into three or more, which is dependent on the practical requirement.

GUI showing. GUI, short for Graphical User Interface, is a user interface for computer operations displayed by graphical way. It closely connects administrator and user with the system, when accessed through the web browser. Administrator can easily view and manage the operation within the system through GUI, and real-time information, alarm status and processing status of the managed divice, can also maintain and backup the database.

21.4 Design for IOT Management System

Based on the analysis for the system requirements above, we can propose a system model structure for this research topic as shown in Fig. 21.4.

Process design. Next, it is mainly to explain the detailed design for the system in the way of data processing. It is divided into the following steps.

When the server program is running, the socket port for the program of accepting data opens, and begins to listen whether there is data coming to the port.

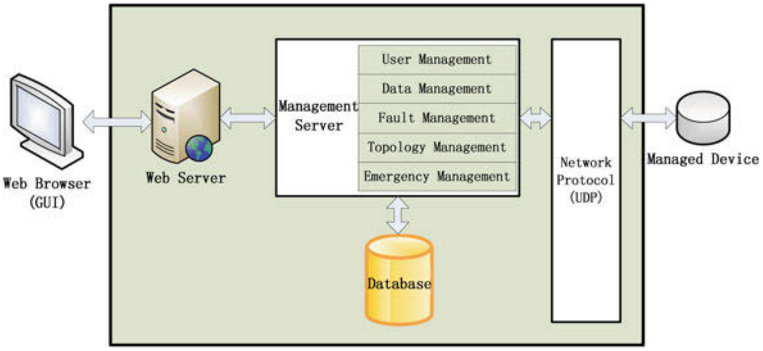


Fig. 21.4 Structure of web-based management system for IOT

When data arrives, the receiver program parses the packet in accordance with prior proposed protocol format. The packets are broken down into individual data blocks, including device identification, IP address and environmental information. Then the data are saved in the table of database.

Data are read from database to judge whether there is any alarm. If there is no alarm, analysis of the working conditions, and the environmental trends are made to make early warning. If there is alarm, determination about alarm level, and the severity of the alarm is analyzed to make the appropriate emergency treatment. Then, the treatment above is saved into the database as historical records.

According to the results of data processing the appropriate warning or alarming is made, to warn early or issue a directive immediately. If alarm level is divided into three, the ways to deal with the alarm are suggested as follows: simple treatment with no report, alarm through the warning box or email or send message immediately to the relevant administrator. The program flow charts are shown in Fig. 21.5.

Implementation technology. The realization of web-based management system includes two parts: GUI and background business logic. The whole development is based on Java, which can be divided into several parts of technology, such as Applet, Servlet, JavaBean, JDBC and so on [6]. The main technologies of system are listed in Table 21.1.

21.5 Conclusion

With the development of Internet of things, it will increasingly drive intelligent businesses, so within the next few decades, the idea of earth wisdom will be realized. Then, the trillions of objects on the earth will be joined into the Internet, and the applications in Internet will be greatly enriched; all kinds of things that exist can be checked in the Internet and communicated by all kinds of applications. So with the flourish development of IOT recently, we more and more need such a

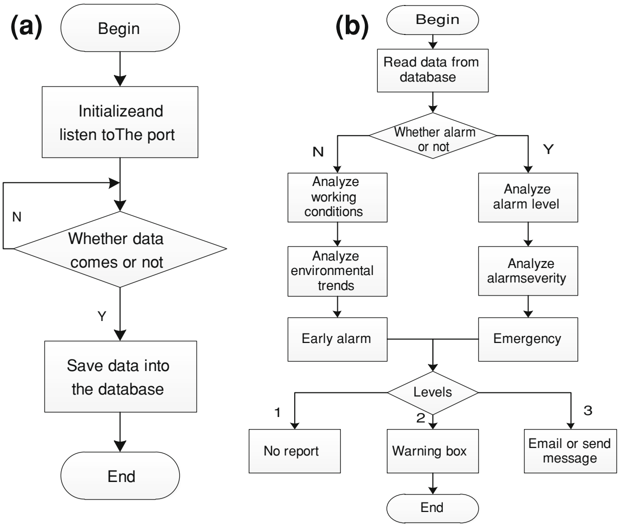


Fig. 21.5 **a** Data collecting flow chart. **b** Data handling flow chart

Table 21.1 Main technologies

Technology	Main function	Used module
JSP	Design foreground display	GUI
Socket	Communication	Data collecting
JDBC	Connect to database	Data reading and saving
JavaBean	Encapsulate	Data transmission
Applet	Dynamic diagram	Topology management
Servlet	Main business logic	Forwarding

standard protocol-based unified system to manage the complex network equipment. It can improve the ability for operation, management and maintenance to ensure the safety and reliability of network and devices. Hence, it can speed up the coming era of Internet of things.

Acknowledgments This work was supported in part by the Key Project of the Chinese National Programs for Fundamental Research and Development (2010CB328204) of advanced optical network lab in Beijing University of Post and Telecommunication.

References

1. Xing L (2009) Internet of things: history has given us new opportunities in post-crisis era [J/OL]. Inform Construction (10)
2. Shi J (2009) Perceive China promotes accelerated development of Internet of things of China [J]. Commun Manage Technol 10(5):1–3

3. Guo J (2006) Network management [M]. 2nd edn. Beijing University of Posts and Telecommunications Publishing House, Beijing, pp 1–20
4. Wang G (2005) Design and implementation the Web and SNMP based network management system [D]. Northeast University, Shenyang, pp 24–30
5. Graziani R, Johnson A (2009) CCNA exploration companion guide: network fundamentals [M]. Cisco Press, America, pp 29–30
6. Horstmann CS, Cornell G (2008) Core Java Volume I: fundamentals, 8th edn [M]. Prentice Hall PTR, Englewood Cliffs, pp 376–437

Chapter 22
Game Analysis of Multi-Strategy Between Government and Suppliers in Green Supply Chain

Changfei Jin and Lijun Mei

Abstract In order to study the roles that the government and the suppliers played in the green supply chain, this paper constructed the game model of multi-strategy between government and suppliers. Game theory was then used to seek the optimal mixed strategies of the government and suppliers under equilibrium conditions. It is shown that government can adjust the economic policy according to the cost of government supervision, the penalties on non-green products, the environmental damage costs of non-green products, and so on. On this basis, this paper provided theoretical ideas for the government to improve the efficiency of the green supply chain.

Keywords Green supply chain · Game theory · Mixed strategy

22.1 Introduction

With the depletion of natural resources and worsening environmental pollution, environmental issues have attracted more and more attention from many regions and countries around the world. Circular economy mode of green supply chain management has become an inevitable trend. Therefore, the theory and practice of green supply chain management has paid government agencies, enterprises,

C. Jin (✉) · L. Mei
School of Economy and Trade, Hunan University, Changsha 410079,
Hunan Province, People's Republic of China
e-mail: changfeijin@gmail.com

L. Mei
e-mail: bmljunlove@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_22,

Table 22.1 The multi-strategy set of the government

	Strategies		
Government	Non-supervision (N)	Supervision	
		Payment mode (P)	Withdrawal mode (W)

and scholars attention [1–3]. van Hoek [4] considered that we need to examine our "ecological footprint" to lower the environmental damage of business . Sheu et al. [5] presented an optimization-based model to deal with integrated logistics operational problems of green supply chain management. Zhu et al. [6] examined the links between green supply chain management initiatives and performance outcomes, and offered the suggestions for further research on the implementation of green supply chain management. Nwe et al. [7] presented an approach integrating LCA indicators and dynamic simulation for green supply chain design and operation. Although we already know that the government and suppliers play an important role in the green supply chain management systems, it is rare that the articles discuss the decision problems in the quantitative way between government and suppliers in the green supply chain based on game theory. This paper applied the analysis of game theory in green supply chain. According to the establishment the multi-strategy model and the analysis of results of the game equilibrium on both sides, it can provide the basis for government regulatory decisions.

22.2 Model and Assumptions

22.2.1 Construction of the Model

The players of the game model in the green supply chain are the "Government" and "Suppliers". The Government means the environmental protection agencies which supervise and manage suppliers on behalf of the public interest. There are some strategies of government in green supply chain game model, such as, supervision and non-supervision ("non-supervision" is written "N"). The supervision method has two kinds, "Payment mode" (written "P") and "withdrawal mode" (written "W"). The Payment mode is the traditional government management, that is to say, when the government supervises suppliers, if the suppliers do not take environmental protection measures and produce non-green products, the suppliers will pay for the environmental damage caused by non-green products. "Withdrawal mode" is the latest management. The government cooperates with retailers in the green supply chain. Government certifies suppliers and publishes substandard suppliers, and the retailers do not purchase the products which are produced by the substandard suppliers. Therefore, in this case, if the suppliers do not take environmental protection measures, they will not profit. The multi-strategy set of the government in green supply chain is shown in Table 22.1.

Table 22.2 The multi-strategy set of the suppliers

Suppliers	Strategies		
The first stage	Green (G–G)	Non-green	
The second stage		Green (N–G)	Non-green (N–N)

The suppliers may take environmental protection measures and produce green products in the green supply chain. Also they may not take environmental protection measures and produce non-green products. We assume that the process of taking environmental protection measures consists of two stages. Therefore, suppliers strategies are divided into three kinds. One kind of strategy is producing green products throughout both stages (written "G–G"); another kind is that the producing green products during the latter stage and producing non-green products during the first stage (written "N–G"); the other is producing non-green products throughout both stages (written "N–N"). The multi-strategy set of the suppliers in green supply chain is shown in Table 22.2.

22.2.2 Assumptions of the Model

We simplify some complex conditions without changing the essence of conditions and make the following assumptions for this model.

It is assumed that the costs of the government supervision are c. If the government supervises the suppliers by P strategy, when suppliers produce non-green products during the first stage, the government imposes fees as p on suppliers; when in the second stage without implementing environmental protection measures, suppliers not only receive penalty fees as p, but also pay additional penalties cost as kp, k is the penalty coefficient $(k > 0)$.

When suppliers adopt green environmental protection measures, the suppliers will be exempted from payment of fees and penalties. If the government supervises the suppliers by ψ strategy,when the suppliers produce non-green products, the retailers cannot purchase products from the suppliers, and the profits of the suppliers are 0 in that stage.

We ignore the suppliers fixed costs. If the government does not supervise the suppliers, when suppliers take the green environmental protection measures, the profits of the suppliers are u, and there are positive external effects v for the government; and when suppliers do not take the green environmental protection measures, the profits of the suppliers are $(u + t)$, then there are negative external effects (–t) for the government $\left(p < t < \frac{2p+kp}{2}\right)$.

It is assumed that players of the game are risk neutral, and the parameters $u, v, t, p > c > 0$.

The matrix of strategy and payoff for government and suppliers is shown in Table 22.3.

Table 22.3 The matrix of strategy and payoff for government and suppliers

Suppliers	Government		
	N (x)	P (y)	W (1–x–y)
G–G (i)	$2u$, $2v$	$2u$, $2v$–c	$2u$, $2v$–c
N–G (j)	$2u + t$, v–t	$2u + t$–p, v–$t + p$–c	u, v–c
N–N (1–i–j)	$2u + 2t$, –$2t$	$2u + 2t$–$2p$–kp, –$2t + 2p + kp$–c	0, –c

22.3 Equilibrium Analysis

It is assumed that each player knows the other's strategy set and the payoffs. Therefore in the short-term equilibrium, this game is the complete information static game. We find out that government and suppliers do not have pure-strategy Nash equilibrium, and only have mixed-strategy Nash equilibrium. The government's mixed-strategy is expressed $S_g = \{x, y, 1 - x - y\}$. This means that the probability of government to non-supervise the suppliers is x, the probability of government using the Payment Mode of supervision methods is y, and the probability of withdrawals mode is $(1 - x - y)$. The suppliers mixed-strategy is expressed $S_s = \{i, j, 1 - i - j\}$. In the same way, this means that the probability of suppliers taking the G–G pure-strategy is i, taking the N–G pure-strategy is j, and taking the N–N pure-strategy is $(1 - i - j)$.When suppliers take the mixed-strategy $S_s = \{i, j, 1 - i - j\}$, the government's expected profits function of each pure-strategy N, P, and ψ is expressed as E_g^N, E_g^P and E_g^W respectively, so we obtain the expected profit functions as follow:

$$E_g^N = i(2v) + j(v - t) + (1 - i - j)(-2t). \tag{22.1}$$

$$E_g^P = i(2v - c) + j(v - t + p - c) + (1 - i - j)(-2t + 2p + kp - c). \tag{22.2}$$

$$E_g^W = i(2v - c) + j(v - c) + (1 - i - j)(-c). \tag{22.3}$$

When the expected profits of the pure-strategy N, P, and ψ are no different for government, the suppliers' optimal mixed-strategy is worked out. That is to say, when $E_g^N = E_g^P = E_g^W$, the mixed-strategy $S_s^* = \{i^*, j^*, (1 - i - j)^*\}$ of suppliers is the optimal. Therefore, from the Eqs. 22.1–22.3, we obtain that:

$$i^* = 1 - \frac{c(pk + p - t)}{kpt}. \tag{22.4}$$

$$j^* = \frac{c(pk + 2p - 2t)}{kpt}. \tag{22.5}$$

$$(1 - i - j)^* = \frac{(t - p)c}{kpt}. \tag{22.6}$$

In the same way, when government takes mixed-strategy $S_g = \{x, y, 1 - x - y\}$, the suppliers' expected profits of each pure-strategy G–G, N–G, and N–N is expressed as E_s^{GG}, E_s^{NG} and E_s^{NN} respectively, so we obtain the expected profit functions as follows:

$$E_s^{NN} = 2u. \tag{22.7}$$

$$E_s^{NG} = x(2u + t) + y(2u + t - p) + (1 - x - y)u. \tag{22.8}$$

$$E_s^{GG} = x(2u + 2t) + y(2u + 2t - 2p - kp) + (1 - x - y) \times 0. \tag{22.9}$$

When expected profits of the pure-strategy G–G, N–G ,and N–N are no different for suppliers, the optimal mixed-strategy of government is worked out. That is to say, when $E_s^{GG} = E_s^{NG} = E_s^{NN}$, the mixed-strategy $S_g^* = \{x^*, y^*, (1 - x - y)^*\}$of government is the optimal. Therefore, from Eqs. 22.7–22.9, we obtain that:

$$x^* = \frac{u}{u + t}. \tag{22.10}$$

$$y^* = 0. \tag{22.11}$$

$$(1 - x - y)^* = \frac{t}{u + t}. \tag{22.12}$$

Therefore, the optimal mixed-strategies of suppliers and the government are expressed S^*. In such a case, the maximum expected profits of suppliers $\left(E_s^*\right)$ and the government $\left(E_g^*\right)$ are, respectively, shown as follows:

$$\begin{aligned} S^* &= (S_s^*, S_g^*) \\ &= \left(\left\{1 - \frac{c(pk + p - t)}{kpt}, \frac{c(pk + 2p - 2t)}{kpt}, \frac{(t - p)c}{kpt}\right\}, \left\{\frac{u}{u + t}, 0, \frac{t}{u + t}\right\}\right). \end{aligned} \tag{22.13}$$

$$E_s^* = 2u. \tag{22.14}$$

$$E_g^* = \frac{2vt - tc - vc}{t}. \tag{22.15}$$

According to the game analysis above, we can obtain some significant conclusions as follows.

From Eqs. 22.4–22.6, we find that when other parameters are constant, if the values of p and k increase, that is to say, if the government imposes more fees on suppliers that produce the non-products, then the probabilities of both G–G and N–N strategy of suppliers will be smaller, and the probability of N–G strategy of suppliers will be larger.

The increase of government regulation costs (c) can cause the probability of the G–G strategy of suppliers to reduce.

From Eqs. 22.10–22.12 to, we can obtain that the probability for government regulation is inversely proportional to t, namely when suppliers adopt the non-green measures, the negative external effects for the government will increase, then the government will strengthen the supervision of suppliers and urge suppliers to adopt green environmental protection measures.

From Eq. 22.11 we can find that when the government takes the supervision strategy, the best way of supervision methods is using withdrawals mode.

From Eqs. 22.14 and 22.15, when both players have reached mixed-strategy Nash equilibrium, the suppliers' optimal expected profits are only related to the benefits of selling the green products.

The government's optimal expected profits are proportional to v and inversely proportional to t and c, namely the greater the positive external effects for the government of green products and the lower the costs of government supervision, the more the profits of the government.

In summary, the government needs to reduce its supervision costs to make more benefits. We suggest the suppliers increase the benefits of selling the green products by proper method, and we suggest the government use the withdrawals mode of supervision methods to urge suppliers to take green environmental protection measures.

22.4 Conclusions

This paper constructed the milt-strategy game model for government and suppliers in green supply chain. We obtained the equilibrium mixed-strategy and optimum solution of the game for both players. From the equilibrium mixed-strategy of the game, we found out that those factors, including the cost of government supervision, the external social benefits of green products, payment costs, penalties, and so on, have an impact on strategies and benefits of the government and suppliers. According to the analysis results, it is shown that the government plays a very important role in the green supply chain management, and if the government wants to improve the efficiency of the green supply chain and improve the proportion of suppliers taking green environmental protection measures, it should reduce the supervision costs appropriately and use the withdrawals mode of supervision methods.

Acknowledgments This study was supported by National Outstanding Youth Science Foundation of China (70925006) and the Key Scientific and Technological Project of Hunan Province (2009FJ1003, 2008FJ1006).

References

1. Gibbs D, Deutz P (2005) Implementing industrial ecology? Planning for eco-industrial parks in the USA. Geoforum 36(4):452–464

2. Tsai WT, Chou YH (2004) Government policies for encouraging industrial waste reuse and pollution prevention in Taiwan. J Clean Prod 12(7):725
3. de Brito MP, Carbone V, Blanquart CM (2008) Integrating the Kano model into a robust design approach to enhance customer satisfaction with product design. Int J Prod Econ 114(2):534–553
4. van Hoek RI (1999) Power, value and supply chain management. Supply Chain Manage Int J 4(3):129–135
5. Sheu JB, Chou YH, Hu CC (2005) Supply chain management and sustainability: procrastinating integration in mainstream research. Transport Res Part E Logist Transp Rev 41(4): 287–313
6. Zhu QH, Sarkis J, Lai KH (2007) Initiatives and outcomes of green supply chain management implementation by Chinese manufacturers. J Environ Manag 85(1):179–189
7. Nwe ES, Arief A, Halim I, Srinivasan R (2010) Green supply chain design and operation by integrating LCA and dynamic simulation. Comput Aided Chem Eng 28:109–114

Chapter 23
Study of Data Integration Model on Securities

Guorong Xiao

Abstract The existence of heterogeneous data sources brings great inconvenience to realize the exchange visits to data in securities. Therefore, it becomes a meaningful research topic to solve the problem of realizing convenient and flexible exchange visits. This paper combines the data representation format of xml generally used in current securities data, and constructs a xml data model, which can implement structured data of relational type as well as describe unstructured data and self-describing semi-structured data.

Keywords Securities · Data integration · Model · Xml

23.1 Introduction

With the rapid development of securities, competition in the securities industry has become increasingly fierce. This requires a lot of data, and needs to discover its operating rules and future trends. Securities trading system has accumulated vast amounts of data including internal data, such as financial status, financial status of customers trading status, transaction status and position status, and corporate external data, such as the number of customers, customer preferences, stock market information and more. The data on securities firms and exchanges has a very high value. The challenge faced by enterprise application program is how to collect and use these data sources to dig and analysis, and develop a long-term

G. Xiao (✉)
Department of Computer Science and Technology,
GuangDong University of Finance, Guangzhou, China
e-mail: newducky@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_23,

planning by the analysis of results, how to achieve correct data through different information platforms, how to integrate a large number of available data and transform them into information assets as soon as possible. These information assets allow enterprises to make more pointed responses to the market and customer needs, improve the company's management and competitive advantage and keep a leading market competitiveness and constantly open to new commercial opportunities.

Currently, the securities industry's massive data consists of three parts:

Business data, business data is mainly in the securities transaction process. Among them, the trading system data is the most important. It consists of transactions in real-time trading system, it is the main basis for data mining, customer analysis and CRM system.

Market data, market data is provided by the Shenzhen and Shanghai Stock Exchange, the securities are issued in the opening period of the transaction data, it is the key data for data analysis.

Related network information data, it is mainly related to the network information data published by various media relating to securities, including satellite, television, radio, Internet, books and magazines and other mass media. Internet is the media that covers the most amount of information, the information mainly in text form,such as user discussion and thoughts on the stock, the information often exists in the customer feedback e-mail or forum.

Key business data and market data released by the stock exchanges, satellite systems through the network or sent to the securities companies, these data have established a database format, and collected more easily.

Therefore, we focus on the text collected from the network. These data are divided into two categories: one is the structured data, such as numbers and symbols; while the other numbers or information cannot be unified structure that we call unstructured data. Now, with the network technology, especially internet and intranet technology's fast development, the number of unstructured data increases. According to statistics, the current amount of data is in unstructured form, such as web pages, technical documentation and e-mail. Therefore, in the securities-related information and data collection, in addition to face structured data, the current disorder more often needs to deal with unstructured data. The text data source is varied, so the model in the design of data collection, the need to maximize the integration of various information sources, to improve the recall of information, other information collected in the repetitive, non-related aspects should be dealt with, otherwise there will be duplicate search results, non-related, or even false information.

From the perspective of database research, web site information can be viewed as a larger, more complex database. Each site on the network is a data source, each data source is heterogeneous, and thus each site information is different, which constitutes a large heterogeneous database environment [1, 2]. If you want to use these data for data mining, first of all, we must study the integration of heterogeneous data between sites problem, only the data of these sites are integrated together to provide the user with a unified view of data, and we should obtain the required things.

Network-oriented data integration has become a difficult problem to solve. The emergence of XML and web data mining could solve the problem and opportunities. XML allows different sources of structured data to combine easily, thus making the search of a variety of incompatible database to be possible, so as to solve the data integration work. Also, because XML data is based on self-description, the data can be described without an internal exchange and processing [3].

23.2 Solution Based on XML Technology

XML is a common language specification established by W3C organization on February 1998. It is a simplified subset of SGML, especially designed for web application program. To make it possible to exchange data based on XML [4], we must realize the XML data's access in database, integrate XML data with application program and then make it combine with existing business rules.

The following illustrates how to use XML model to represent an object. For the stock data, we could create financial data files for a enterprise stock, including the information of "exchange", "name", "symbol", "price", etc. First, we need to create the stock object, which represented a certain enterprise financial report. A complete presentation of stock object is showed below:

```
<Stock >
<Exchange>Shanghai</Exchange>
<Symbol> 600050</Symbol>
<Name> China Unicom </Name>
<Date>2011-03-09</Date>
<Change>0.02</Change>
<Open>6.05</Open>
<High>6.09</High>
<Low>6.00</Low>
<Price>6.02</Price>
<PreviousClose>6.04</PreviousClose>
<PE>98.5</PE>
<Rate> 0.35%</Rate>
</Stock>
```

The above example is simple, but it is enough to illustrate that using XML data model can completely represent the company stock data in traditional relational database. The data model has a nature of self-describing with a self-defining label, so it is particularly suitable to describe those data objects without display mode or with unpredictable mode. It is a common data model of heterogeneous data integration system, which logically and uniformly represents heterogeneous data from various data sources.

In fact, XML data model can not only represent these structured data, as shown in the following example but it can also be used to represent semi-structured data of XML which is increasingly and extensively applied in the network and some other unstructured data. While building model with traditional relational database for the latter two situations, it will face great trouble.

23.3 Using Xml Data Model to Represent Data

The following illustrates how to use XML data model to represent the semi-structured data [5]. Considering the situation in securities with many trading accounts, we made some simplification to the problem under the premise without affecting the description. Assuming each transaction user in this securities can own several stocks. Each user can use a transaction account to buy stocks, all funds were divided into two parts, one is stock position, another is cash. Each of the stock position will indicate the stock name, the current price, holding number, stock value and profit and loss situation.

Using xml data model can easily represent the above information. Firstly, create a main label of “account” to represent the trading account, including three sublabels of “customer”, “positiondata” and “funds”. The sublabel of “customer” represents the transaction user’s basic information. The sub-label of “positiondata” represents the group of stocks holding, which contains sublabels of “item” representing each stock. The sublabel “funds” contains transaction user’s funds information.

The complete XML presentation of the above example is shown below:

```
<? Xml version=”1.0” encoding=”GB2312”?>
<Account>
<Customer>
<Name>Bill Buckram</Name>
<Cardnum>033953533136</Cardnum>
<City>Guangzhou</City>
</Customer>
<Positiondata>
<Item>
<Stcode>600030</Stcode>
<Name>CITIC Securities</Name>
<Quantity>3500</Quantity>
<UnitPrice>15.3</UnitPrice>
<Value>53550</Value>
<Cost>30000</Cost>
<Profit>23550</Profit>
</Item>
<Item>
<Stcode>600036</Stcode>
```

```
<Name>China Merchants Bank</Name>
<Quantity>5000</Quantity>
<UnitPrice>15</UnitPrice>
<Value>75000</Value>
<Cost>30000</Cost>
<Profit>45000</Profit>
</Item>
<Item>
<Stcode>600519</Stcode>
<Name>Kweichow Moutai</Name>
<Quantity>1000</Quantity>
<UnitPrice>185</UnitPrice>
<Value>185000</Value>
<Cost>150000</Cost>
<Profit>35000</Profit>
</Item>
</Positiondata>
<Funds>
<Subtotal>313550</Subtotal>
<cash>538950</Tax>
<Total>852500</Total>
</Funds>
</Account>
```

It can be found from the above examples that using XML data model represents structured data in relational type as well as describes unstructured data and self-describing semi-structured data. Thus, XML data model can be considered as a common data model integrated by heterogeneous data to integrate these heterogeneous data.

XML is the core of many emerging technologies today, for example, grid computing and autonomic computing. XML in the application of these technologies is significant and database vendors continue to explore more efficient XML.

23.4 Conclusion

This paper introduced the current situation and historical progress of data integration of securities. The constantly emerging XML brings us some innovative and effective solutions for dealing with the problem of heterogeneous data integration. It provides us a powerful and stable platform for processing problems of information representation and information access from heterogeneous data integration system. For a heterogeneous data integration system, it needs to face a variety of data sources, each of which has its own characteristics. Except for difference in data model, some of the data sources do not have a fixed model but easily vary in structure. In addition, some of them even contain a number of unstructured data, such as voice and graphic. The data model is a very flexible model, and its basis XML is being widely used in a variety of data environment. Being regarded as a semi-structured data model, it can conveniently describe the data from variety of data sources, particularly the self-describing data, which is incomparable by other data models.

References

1. Lei Q (2002) Research of heterogeneous database integration system and the prototype of its supporting tools. Master's theses of maritime affairs, University of Dalian, 3
2. Liu J (2003) Prototype implementation of distributed intelligent heterogeneous data integration support system. Master's theses of maritime affairs, University of Dalian, 3
3. Li J, Zhang J, Zhou M, Geng G (2002) Research on method of XML-based technology of enterprise disparate data integration. Comput Eng 28(9):63–74
4. Bright MW et al (1992) A taxonomy and current issues in multi database systems. IEEE Comput 25(3):50–59
5. Brandin C (2003) XML data management, information modeling with XML, May 27

Part III
Digital Image Processing

Chapter 24
Trademark Image Retrieval Algorithm Based on SIFT Feature

Shijie Jia, Nan Xiao and Zeng Jie

Abstract Image matching is an important way to implement trademark image retrieval. In this paper, we proposed an algorithm based on Scale Invariant Feature Transform (SIFT) for automatic trademark image search. Firstly, feature points were detected and described through calculating the gradient histogram of nearby region. Secondly, Euclidean distances were calculated to obtain the similarity between the SIFT features extracted from the two corresponding images. Experiment illustrated that the proposed method was robust to scale, viewpoint, occlusion and noise interference.

Keywords Trademark retrieval · Image matching · SIFT · Feature detection

24.1 Introduction

Trademark image retrieval is one of the important ways to check the repeatability of trademarks, which aims at examining the repeatability and similarity among trademark images [1]. "Some opinions on questions regarding trademark administration [2]" issued by trademark administration indicates that estimation of repeatability or similarity among trademarks is under the conditions below: (1) According to the registered trademarks rather than the ones practically used.

S. Jia (✉) · N. Xiao · Z. Jie
College of Electrical and Information, Dalian Jiaotong University,
Dalian 116028, China
e-mail: jsj@djtu.edu.cn

S. Jia
Faculty of Electronic Information and Electrical Engineering,
Dalian University of Technology, Dalian 116023, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_24,

(2) The subjective criterion is judged by consumers' normal attention, and the synthesized estimation is adopted by combining whole comparing with salient parts comparing. As it is complicated to realize trademark image retrieval by hand, we proposed image matching algorithm for automatic trademark image retrieval [3].

The task of image matching is to find the correspondences between two images with the same scene at different viewpoints. It is a basic research field in computer vision and also the fundamental research on some computer vision applications, such as depth recovery, camera calibration, motion analysis and three-dimensional reconstruction [4]. Along with the development of technology, image matching has become a very important method in image processing. There are many image matching algorithms among which a common one is directly matching on gray images. These kinds of methods are simple but easily affected by the environments and with more calculation cost. In order to overcome these limitations of the above method, some shape-based matching methods are proposed. (e.g. shape context). These kinds of methods extract features from the candidate images, and adopt the geometric transforms according to similarities and some restriction conditions on these candidate images [5]. Although the related matching method attains a high precision, it costs much more computation and hard to meet real-time requirement. On the other hand, feature-based image matching technology combines and transforms the features to form the easily matching and stable feature vectors. Consequently, the image matching is regarded as a feature matching one, and then the feature matching is regarded as feature vector clustering in feature space. As an outstanding representation of feature-based matching, Scale Invariant Feature Transform (SIFT) proposed by David G [6]. Lowe obtains widespread application. In brief, SIFT extracts local features, which finds extreme points in scale-space, and these feature points should be invariant to position, scale, rotation, zoom and affine transform.

This paper employed SIFT feature to implement trademark image retrieval. The second section analyzed SIFT feature extraction method. Section 24.3 showed and discussed the image matching results using SIFT method. The last part concluded the paper and gives an outlook in the future.

24.2 SIFT Algorithm

SIFT algorithm is based on image feature scale selection. It builds multi-scale space, and detects the same feature points in different scales. Then locate the feature point and choose their scales to reduce zooming effects. Meanwhile, it obliterates the low contrast points and some edge response points and extracts feature descriptor with rotation invariant to realize the affine transformation. This algorithm contains four steps: (1) building scale space and finding the candidate points; (2) locating the key points and eliminating the instable points; (3) finding the orientations of the key points; and (4) extracting feature descriptors.

Gaussian function $G(x, y, \sigma)$ is the only possible scale-space kernel to realize scale transform. Therefore, the scale space of an input image, $I(x, y)$ is defined as a function, $L(x, y, \sigma)$ that is produced from the convolution of a variable-scale Gaussian, with the input image:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{24.1}$$

where * is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2 + y^2)/2\sigma^2} \tag{24.2}$$

The difference-of-Gaussian function, $D(x, y, \sigma)$ which can be computed from the difference of two nearby scales convolved with the input image, then repeat the above process through down-sampling by a factor of two until the size of the processed image decreases to one certain threshold (e.g. 32 × 32).

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \tag{24.3}$$

Each sample point is compared to its neighbors in the current image and other neighbors in the adjacent scales. The position and scale of the local extrema (key points) are obtained. Least square approximation is applied by using Taylor expansion (up to the quadratic terms) of the scale-space function (24.4), $D(x, y, \sigma)$ and computes the extrema of the multi-quadric to find the exact location and scale of the key point. The final position and scale of the key point can be to an accuracy of subpixel level.

$$D(x) = D + \frac{\partial D^T}{\partial X} X + \frac{1}{2} X^T \frac{\partial^2 D}{\partial X^2} X \tag{24.4}$$

SIFT algorithm appoints dominant directions through gradient orientation distribution of the points nearby the key points. Peaks in the orientation histogram correspond to dominant directions of local gradients. The follow-up descriptors are constructed according to the dominant directions. The formulas of gradient absolute value and orientation are as below,

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{24.5}$$

$$\theta(x, y) = \tan^{-1} \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \tag{24.6}$$

In order to achieve the rotation invariant, the coordinates of the descriptor and the gradient orientations are rotated relative to the key point orientation when constructing the descriptor. A key point descriptor is created by first computing the gradient magnitude and orientation at each image sample point in a region around the key point location. A 4 × 4 array of histograms with eight orientation bins forms a 128-dimension vector which is defined as a descriptor of one feature point.

The feature vector extracted is invariant to scale and rotation. Moreover, it can reduce the illumination effects if it is normalized.

24.3 Applications of SIFT in Trademark Image Retrieval

Trademark image retrieval aims at making sure whether there exists confusion between a given trademark and anyone in the trademark data set. The potential similar features in the data set could be exactly extracted with SIFT. The considerable questions should be advanced in the process of trademark image retrieval based on SIFT:

(1) In order to avoid extracting insignificant features and improving retrieval speed and retrieval precision, noise reduction should be adopted.

(2) The number of feature points significantly effects retrieval result as well, more feature points decrease the retrieval speed, while less feature points easily cause missing detection. Consequently, normalization should be applied to the number of feature points when constructing the trademark data set. There are about 100 stable feature points in a normal 150×150 pixels trademark image. Different scale and type images produce different number of feature points. For images with the same content, the larger scale one produces more feature points. For images with the same size, the one that has more texture produces more feature points; however, images with more graphics only produce feature points in the edges and corners. So during the process of building feature data set, the image size should be normalized and then classify the images according to the complexities. In order to ensure that all the images in the data set produce stable and approximately same number of feature points, different SIFT parameters should be applied to different categories.

24.4 Experiments

24.4.1 Feature Matching Under Different Thresholds

Figure 24.1a shows the test images, and Fig. 24.1b is the feature extracting image, where the ellipses in the images are the SIFT feature matching regions, and the centre of the ellipses are the two-dimension coordinate position. The long axes of the ellipses represent the scale of the keypoint, and the ellipses orientation is the orientation of keypoint. Figure 24.1c is the matching image under the threshold $d = 1.2$. (Limited by the space of the paper, images under other thresholds are not listed). Different numbers of features and matching time are shown in Table 24.1, where n is the number of the features and$t$ is the matching time. The experiment

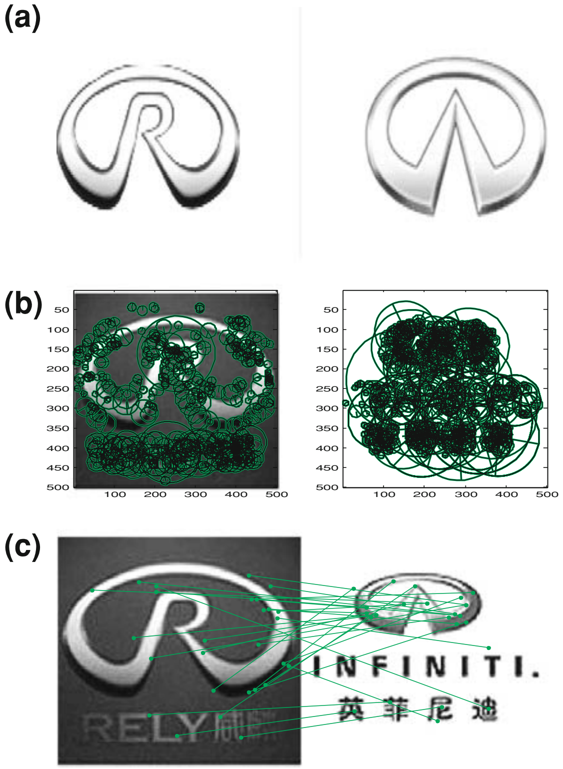


Fig. 24.1 **a** Original images; **b** SIFT feature extracting; **c** SIFT matching (threshold = 1.2)

result indicated the higher the threshold is, the less number of matching points are. But they are more stable. The matching time is less than 1.5 s which met the real-time requirement.

24.4.2 Trademark Retrieval Experiment Results

The experiment chose the trademark images as the retrieval images, and designed 8–12 basic images for each image. The trademark data set contained 2,200 images which include 200 normal images and 2,000 unrelated ones. The confidence level was set as 0.01, which meant the number of the most similar images was 22, and they were listed according to the similarities between images. The most similar image was used to analyze and sampled with SIFT descriptors.

The experiment results were illustrated in Fig. 24.2. All the right results are returned by SIFT descriptors, which showed good performances of SIFT descriptor in the trademark image retrieval.

Table 24.1 Number of features and matching time under different thresholds (*D*)

D	1.2	1.4	1.5	1.8	2.0	2.2	2.5	2.8	3.0	3.2
N	287	198	156	129	116	89	71	58	54	48
T(s)	1.125	1.016	0.543	1.141	0.69	1.418	1.281	1.344	1.515	1.375

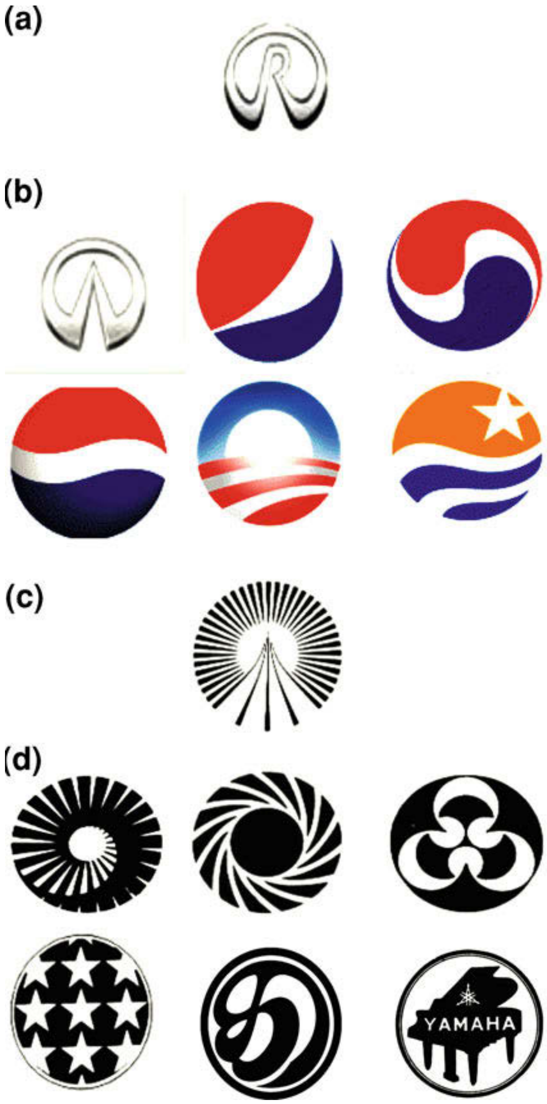


Fig. 24.2 **a** Input image; **b** trademark image retrieval results; **c** input image; **d** trademark image retrieval results

24.5 Conclusion

This paper studied the SIFT image matching algorithm. SIFT feature is a local feature which possesses rich information and it is well used in fast and precisely matching of magnanimous features. Stable feature points are also found through appropriate threshold selection. The disadvantage of SIFT descriptor is its weakness to describe the contour of images, which is easy to cause missing detection and decreases the recall ratio. For this problem, it could be better to combine SIFT with other features, such as colour, shape and texture.

Acknowledgments This work was sponsored by the Project of National innovation fund for technology based firms (No. 09c26222123243).

References

1. The Trademark Law of the People's Republic of China http://sbj.saic.gov.cn (in Chinese)
2. Kim YS, Kim WY (1998) Content-based trademark retrieval system using visually salient feature. Image Vis Comput 16(12):931–939
3. Guo L, Huang Y-Y, Yang J-Y (2005) Using shape and spatial information in trademark retrieval. Comput Softw Appl 22(1):93–95 (in Chinese)
4. Kong X-D, Qu L, Gui G-F, Liang D (2004) Dense stereo matching based on epipolar constraint and edge points detecting. Comput Eng 2004(30):40–41, 179 (in Chinese)
5. Zhao H SIFT feature matching technology. http://www.qiji.cn/epint/abs/3059.html. (in chinese)
6. David GL (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2:91–110

Chapter 25
Automatic Product Images Classification With Two Layers of SVM Classifier

Shi-jie Jia, Xiang-wei Kong and Man Hong

Abstract To achieve visual-based automatic product image classification is a great need of e-commerce development. In this chapter, we presented a two layers of SVM classifier to combine each spatial level of pyramid histogram of words (PHOW) and pyramid histogram of orientated gradients (PHOG) descriptors. In the first layer, each of six chi-square kernel SVM classifiers is trained with a spatial level of PHOW and PHOG, then the probability outputs of these SVM classifiers are concatenated into feature vectors for training another SVM classifier with a Gaussian RBF kernel. Experiments on the product image dataset (PI 100) demonstrated the effectiveness of our scheme for the tasks of product classification.

Keywords Product image classification · SVM · PHOW · PHOG

S. Jia (✉) · X. Kong · M. Hong
Faculty of Electronic Information and Electrical Engineering,
Dalian University of Technology, Dalian 116023, China
e-mail: jsj@djtu.edu.cn

S. Jia
College of Electrical and Information,
Dalian Jiaotong University, Dalian 116028, China

M. Hong
Information Engineering College,
Dalian Jiaotong University, Dalian 116052, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_25,

25.1 Introduction

The world has stepped into a new economy era of e-commerce. Shopping online is becoming really easy and fast and has brought huge growth in shopping on the internet in the last decades. However, “One picture is worth a thousand words”, using human employees to classify product images is labor-intensive and hard to be accurate and complete. There is a great need to find the way to implement the automatic classification of online product based on images features. In this chapter, we focus on product category, a particular type of image classification, which categorizes images by the product types, such as piano or guitar. It is one specific application of content-based image classification, most methods of which mainly use supervised learning feature-based modeling with intermediate semantic analysis to achieve classification results. Although the product images appear “tamer” than usual natural images (where the object dominates the image and there is little underlying background clutter), our task is still a challenging problem for the large number of categories and intra-class variations. Reference [1] explored the feasibility of tagging products through supervised image classification and achieved accuracies between 66 and 98% on 2- and 3-classes classification tasks. Reference [2] proposed a fast supervised image classifier which is based on class-specific descriptor and Image-to-Class distance and achieved 84% for 30 product classes. However, such a small number they tested was far from real application. In fact, for a large number of classes, single descriptor cannot be optimal in all situations to alleviate the effect of intra-class variations.

As each classifier uses different features and levels, the errors with each classifier should be somewhat uncorrelated. Consequently, combining the results of the classifiers should produce an improved classifier.

The rest of our chapter is organized as follows. Section 2 describes our approach, including the image representation, kernel-based SVM Classifier and the details of combination strategies. Experiment setup and typical results are described in Sect. 3. The final part concludes with suggestions for future research.

25.2 Approach

25.2.1 Architecture

The architecture of our approach is illustrated in Fig. 25.1.

In the first layer, each of six chi-square kernel SVM classifiers is trained with a spatial level of pyramid histogram of words (PHOW) and pyramid histogram of oriented gradients (PHOG) descriptors, then the probability outputs of these SVM classifiers are concatenated into feature vectors for training another SVM classifier with a Gaussian RBF kernel.

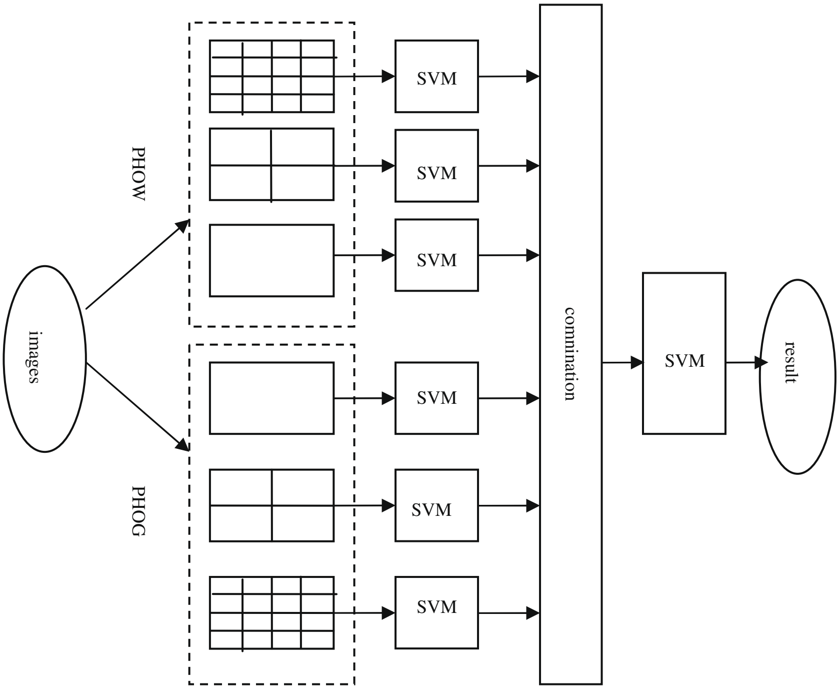


Fig. 25.1 The architecture of our approach

25.2.2 Image Representation

25.2.2.1 PHOW

The bag-of-words (BOW) model [3] treats an image as a collection of orderless descriptors extracted from local patches, quantifies them into discrete "visual words" and then computes a compact histogram representation for semantic image classification. Lazebnik et al. [4] proposed an extended BOW model, which works by partitioning the images into increasingly fine subregions and computing histograms of local features found inside each subregion. The local features are extracted with dense sampling and represented with SIFT descriptor [5]. Therefore, an image is represented as a PHOW descriptor. Figure 25.2 shows the representation of bag-of-word.Toy example of constructing a three-level pyramids is illustrated in Fig. 25.3 [6].

25.2.2.2 PHOG

Histogram of orientated gradients (HOG) [3] is a useful shape descriptor which is based on evaluating well-normalized local histograms of image gradient

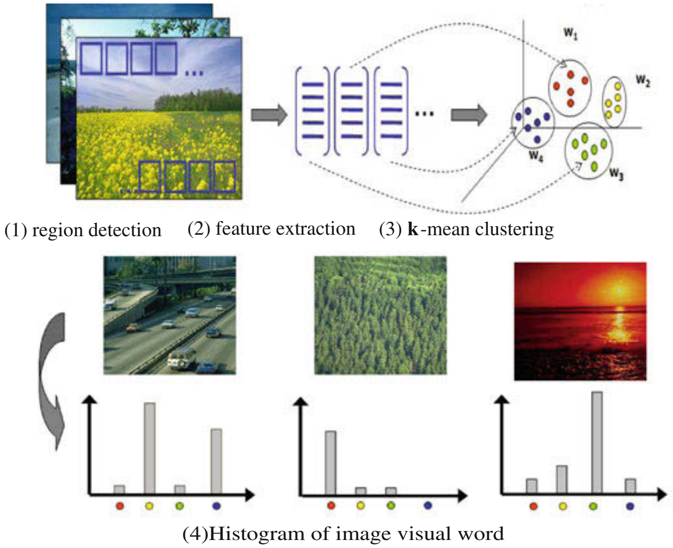


Fig. 25.2 The representation of bag-of-word (1) region detection (2) feature extraction (3) k-mean clustering (4) Histogram of image visual word

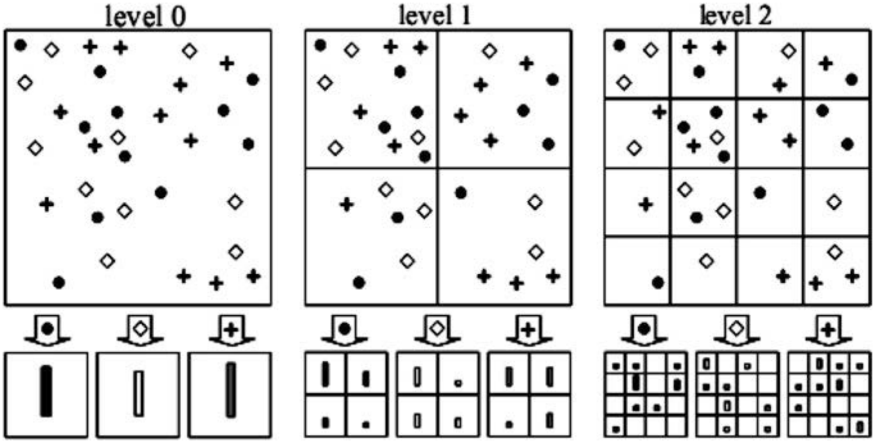


Fig. 25.3 Toy example of constructing a three-level pyramids

orientations in a dense grid. HOG has taken the spatial distribution of the image into account. However, it is ignored that the combination at different spatial scales has an space effect on the performance of retrieval and classification. In view of this, Bosch et al. [4] proposed the PHOG descriptor which is captured by titling the image into regions at multiple resolutions and consists of a histogram of orientation gradients over each image subregion at each resolution level. PHOG feature

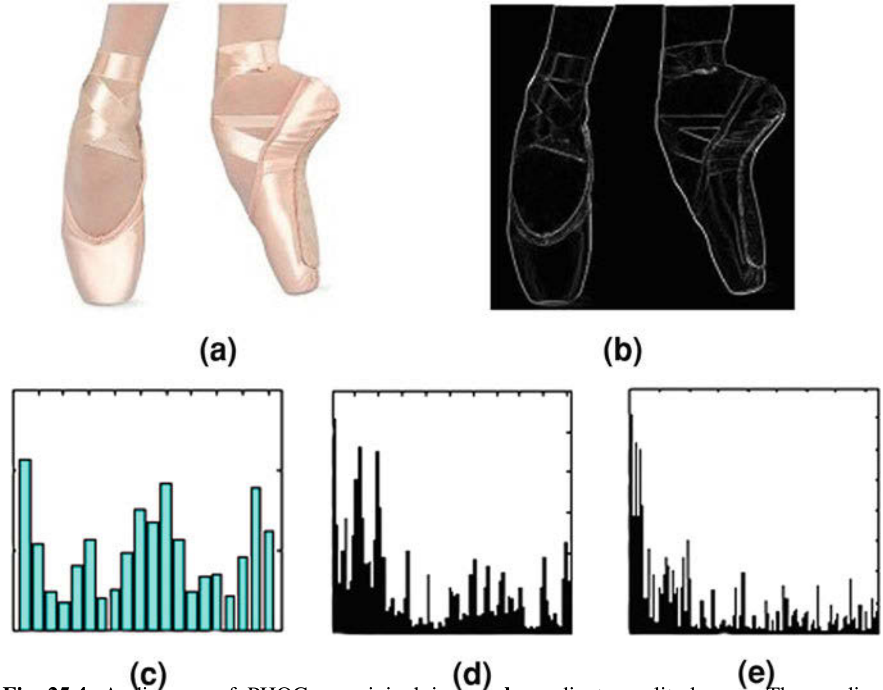


Fig. 25.4 A diagram of PHOG. **a** original image. **b** gradient amplitude. **c**–**e** The gradient histograms for $l = 0, 1, 2$, respectively)

extraction is as follows: (1) Convert the RGB image into a gray image and calculate the horizontal and vertical gradients with the gradient operator. (2) Quantize the gradient direction into K bins. Extract the edge of the gray image and take these pixels gradient magnitudes as the weight to calculate the histograms of K bins. A diagram of PHOG is illustrated in Fig. 25.4.

25.3 Support Vector Machines

In support vector machines, the data representation is implicitly chosen through the so-called kernel $K(x_i, x_j)$, which implicitly maps examples x to a feature space given by a feature map $\varphi(x)$ via $K(x_i, x_j) = \varphi(x_i) * \varphi(x_j)$. This kernel defines the similarity between two examples x_i, x_j. Through the 'kernel trick', classifiers can be learnt and applied without explicitly computing $\varphi(x)$. In the training phase, SVM solves the quadratic programing problem for (25.1) to get the optimum classification superhypeplane.

$$W(\partial) = -\sum_{i=1}^{l} \alpha_i + \tfrac{1}{2} \sum_{i,j=1}^{l} \alpha_i \, \alpha_j \, y_i y_j K(x_i, x_j)$$

$$\sum_{i=1}^{l} y_i \partial_i = 0 \tag{25.1}$$

where x_i, x_j indicates the training examples and y_i, y_j are their classification labels. The final decision classification of hyperplane is :

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \right) \tag{25.2}$$

The choice of the kernel largely influences the performance of the algorithm. For an ideal kernel, examples in the same class should have high kernel value while examples in different classes should have low kernel value. The choice of kernel type are closely related to the types of input patterns. In our scheme, chi-square kernel was employed for the base SVM classifiers and RBF kernel for the high level SVM classifier. The two kernel types are as follows:

(1) RBF kernel

$$k(x, y) = \exp(-|x - y|^2 / 2\delta^2) \tag{25.3}$$

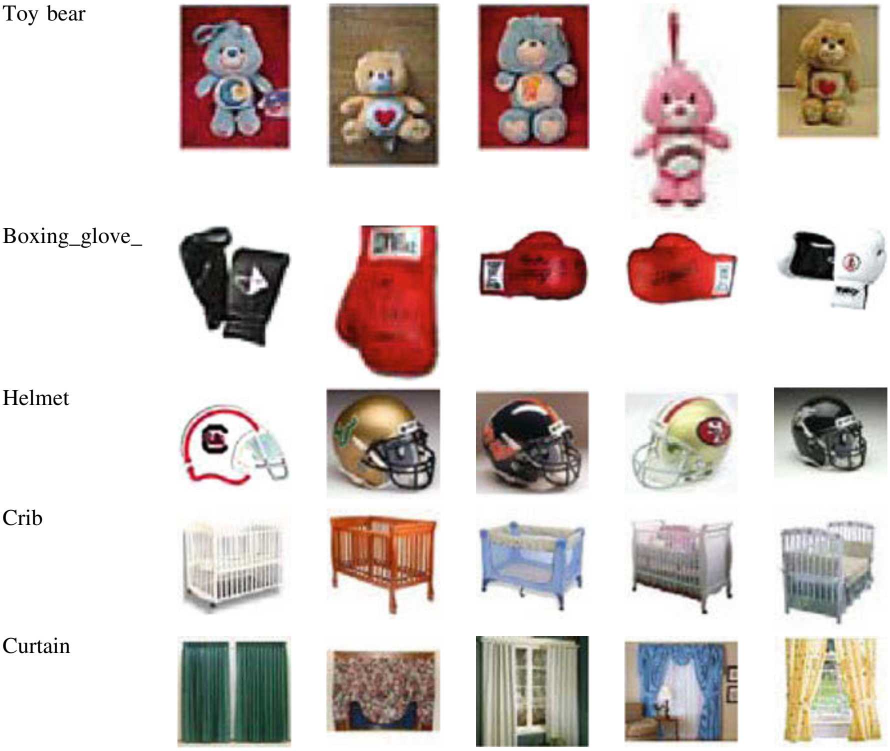
(2) Chi-square kernel

$$k(x, y) = \exp(-\gamma \chi^2 (x, y)^2 / d)$$

$$\chi^2(x, y) = \sum_k - \frac{(x_k - y_k)^2}{x_k + y_k^2} \tag{25.4}$$

where indicate the examples and x_k, y_k are their discrete distribution, respectively. δ and d are the so-called band-width parameters of kernel functions and should be selected via cross-validation. (d is usually set to the mean chi-square distance between all pairs of training samples).

Table 25.1 Some product images of Microsoft image set (PI 100)



25.4 Experiment

25.4.1 Experiment set

We tested our algorithms on Microsoft research's product image categorization data set (PI 100) [7], which was collected from the MSN shopping web site (http://www.shopping.msn.com/). PI 100 contains ten thousands low resolution (100 × 100) images in 100 categories. Each image contains a single object or one dominant object in relatively stable forms, just as most product images appear on the Web. Table 25.1 shows some sample images from PI 100.

The experiments were performed on an Intel Pentium CPU 2.66 GHz computer running Windows XP and MATLAB 2010 with 2 GB RAM. The LibSVM [8] implementation Toolbox with one-against-one strategy is used to train the multi-class classifier. Libsvm provides a multi-class probability estimate by combining all pairwise comparisons, in which the parameter ‘b’ is set to 1 [9]. We randomly select 30 training images per class and test on the remaining images, reporting the average accuracy for all the classes.

Table 25.2 The average accuracies of classification on PI 100

Features	Level	Average accuracy (%)
PHOW	1	45.8
	2	72.4
	3	78.1
PHOG	1	41.8
	2	66.7
	3	73.7
Combination with two layer SVM classifier		83.4

For PHOW descriptor, the sampling interval is set to 8 pixels, each 16 × 16 pixels block formed a 128-dimensional SIFT feature vector [10]. The optimal setting of pyramid level L is 3. The PHOW is normalized to sum to unity taking into account all the pyramid levels.

For PHOG descriptor, Performance was found not to be very sensitive to the number of bins K, but $K = 20$ orientations bins for Shape 180, and $K = 40$ for Shape 360 were found to be optimal [6]. The pyramid levels were set to 3. The PHOG is normalized to sum to unity taking into account all the pyramid levels, which ensures that the images with more edges are not weighted more strongly than others [6].

25.4.2 Results and Discussion

Table 25.2 shows the average classification accuracies of the base SVM classifier training with the different levels of PHOW and PHOG.

The results indicate that the classification performances of PHOW were slightly better than PHOG in general, and all descriptors perform better as the spatial level increased, the variation of accuracy from level 1 to 2 is much more than that from level 2 to 3. Obviously, the average accuracies of combining with two layer SVM classifier were improved considerably (by 5.3–9.7% points) than even the most discriminative individual channel.

25.5 Conclusion

In this chapter, we proposed an effective two layer SVM classification scheme for product image classification, combining pyramid spatial levels of two complimentary descriptors PHOW and PHOG. Experiments result indicated that it is necessary to use a combination of several base classifiers to make effective use of the complementary fearures, which provides a way forward for the improvement of product classification. As no single feature is sufficient for handling diverse intravariation among broad categories, future research should focus on designing more discriminative robust visual features combining with multiple spatial levels as well as combining strategies of different classifiers.

Acknowledgments This work was sponsored by National Science Foundation of Funded Project (No. 70890083) and the Project of National innovation fund for technology based firms (No. 09c26222123243).

References

1. Tomasik B, Thiha P, Turnbull D (2009) Tagging products using image classification. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, Boston, pp 792–793
2. Jia S, Kong X, Fu H, Jin G (2010) Auto classification of product images based on complementary features and class descriptor. J Electron Inform (China) 10:2294–2300
3. Sivic J, Zisserman A (2003) Video Google: a text retrieval approach to object matching in videos. In: ICCV '03: Proceedings of the nineth IEEE international conference on computer vision, p 1470
4. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the IEEE computer society conference of computer vision and pattern recognition (CVPR'06), vol 2, New York, pp 2169–2178
5. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Comput Vision 60:91–110
6. Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In: CIVR '07: Proceedings of the 6th ACM international conference on image and video retrieval. ACM press
7. Microsoft research: product image categorization data set (PI 100). http://research.microsoft.com/en-us/people/xingx/pi100.aspx, 12, 2009
8. Chang C, Lin C (2001) LIBSVM: a library for support vector machines
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection [C]. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1, San Diego, pp 886–893
10. Wu TF, Lin CJ, Weng RC (2004) Probability estimates for multi-class classification by pairwise coupling. J Mach Learn Res 5:975–1005

Chapter 26
Genetic FCMS Clustering Algorithm for Image Segmentation

Chunyu Zhang, Pengfei Wang and Cuiyin Liu

Abstract FCM is populated in image segmentation for its simplicity and easily realization. The classic FCM segmentation used only the gray value for segmentation, and is liable to stuck at local values, and the result is relied on cluster center of initial selection. In this paper, we present a Genetic fuzzy c-means (GFCMS) algorithm that incorporates spatial information for segmentation. The first improvement is to use the spatial information of pixel in FCM algorithm. The second improvement is to use the genetic algorithm for searching the global optimum. The results of the experiment validates that the algorithm has better adaptability and gets the correct global optimum.

Keywords Fuzzy k-means · Genetic algorithm · Spatial information · Clustering

26.1 Introduction

Image segmentation plays an important role in a variety of applications in high level image processing. Segmentation of structures from images is an important step for image analysis that can help in visualization, automatic feature detection,

C. Zhang (✉) · C. Liu
College of Computer Science, Sichuan Panzhihua University,
Panzhihua 610007, Sichuan, China
e-mail: chunyu.z@126.com

C. Liu
e-mail: liucuiyin@163.com

P. Wang
International Baccalaureate Precollege Program 2009,
The High School Affiliated to Nanjing Normal University, Nanjing 210003, China
e-mail: chnxzwb@139.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_26,

image-guided surgery, and also for registration of different images. There are various kinds of methods for image segmentation. Fuzzy c-means (FCM) clustering [1, 2] is an unsupervised technique that has been successfully applied to image segmentation. The FCM has been reported by Dunn in [3, 4], and proposed by Bezdek in [5].

FCM algorithm clusters the image by minimizing the inner class variance in the feature space. The feature usually used is the gray value of pixel. The conventional FCM segmentation algorithm utilizes only the gray value of image and does not consider the spatial information, so that the segmentation adaptive capacity is limited. Keh-Shih Chuang in [6, 7] modified the method with altering the membership weighting of each cluster based on spatial information. This scheme greatly reduces the effect of noise and biased the algorithm toward homogeneous clustering. However, the Clustering algorithm is used to trap at local extreme in the process of optimizing the clustering criterion [4, 8].

The aim of study is to propose a new segmentation method for FCM segmentation. Our new scheme spatial information and the genetic algorithm (GA) approach to clustering with fuzzy c-means. Not only does the genetic algorithmic framework prove to be effective in coming out of local optima, but it also brings considerable flexibility into the segmentation procedure [5, 6, 9, 10]. In order to avoid local minima of the FCM function, we hybrid the genetic algorithm with the FCM in our approach. Furthermore, we adopt the scheme in [1, 11]. The advantages of the new methods are as follows: firstly, the adaptive capability is enlarged; secondly, the global optimum is achieved.

The paper is organized as follows:

In Sect. 26.2, an overview of the Fuzzy c-means (FCM) algorithm, and introduce the modified FCM with spatial information.
In Sect. 26.3, the genetic algorithm, which is parallel optimum, and FCMS is combined, and the detailed realization of the hybrid algorithm is given.
In Sect. 26.4, we apply the hybrid algorithm to several test images and draw comparisons between the standard FCM algorithm and GFCMS. Finally, Sect. 6 gives the conclusions.

26.2 Segmentation with FCM Algorithm

26.2.1 The Conventional FCM for Segmentation

FCM is a unsupervised clustering algorithm that has been successfully applied to image segmentation. A image can be represented in various character space, and the FCM can assign patterns into each category by using fuzzy memberships.

Let $X = \{x_1, x_2, \ldots x_\mathrm{N}\}$ be a finite Set and let $2 \leq \mathrm{C} < \mathrm{N}$ be an integer, and the X can be clustered into c prototype. Each $x_i \in R^n$ is a feature vector consisting of feature values of the pattern. The features could be gray value, local character, and spatial information of image. The X will be clustered into group of like object, and the image will be segmented. The algorithms iteratively optimize the value with the minimization of an objective function defined as equation:

$$J(X; U, V) = \sum_{i=1}^{c}\sum_{k=1}^{N} (\mu_{ik})^m \, ||x_k - v_i||^2 \tag{26.1}$$

where,

$$V = [v_1, v_2, \ldots, vc_c], \quad v_i \in R^n \tag{26.2}$$

$$\mathrm{dist}_{(x_k, v_i)} = ||x_k - v_i||^2 \tag{26.3}$$

$V = \{v_1, v_2, \ldots, v_n\}$ represents the unknown prototype which is known as the cluster center, and m is a weighting exponent for controlling the result of clustering, and $m = 2$ is often used in application. μ_{ik} represents the fuzzy membership of pixel x_k in the group of vi. ||·|| is a norm metric for the distance of two feature vectors. $\mathrm{dist}_{(x_k, v_i)}$ is a measure of the distance from x_k to v_i which is specified by Euclidean distance or mahalanobis distance. In this paper, the Euclidean distance metric is used in all work reported here. U is fuzzy matrix and with the constraint is

$$\sum_{k=1}^{c} \mu_{ik} = 1, \quad 1 \le k \le c \tag{26.4}$$

The problem can be seen as a nonlinear optimization problem. The stationary points of the objective function (26.1) can be found by adjoining the constraint (26.4) to J by means of Lagrange multipliers:

$$\bar{J}(X : U, V, \lambda) = \sum_{i=1}^{c}\sum_{k=1}^{N} (\mu_{ik})^m ||x_k - v_i||^2 + \sum_{k=1}^{N} \lambda_k \left(\sum_{i=1}^{c} \mu_{ik} - 1 \right) \tag{26.5}$$

Getting the partial derivative of $\mu_i(x_k)$ and V to zero, the membership functions and clusters are updated by the following:

$$\mu_i(k_k) = \frac{(1/||x_k - v_i||)^{1/(m-1)}}{\sum_{j=1}^{c} (1/||x_k - v_j||)^{1/(m-1)}} \quad (k = 1, 2, \ldots, n, j = 1, 2, \ldots, C, \text{ for all } k,j) \tag{26.6}$$

$$v_j = \frac{\sum_{k=1}^{n} [\mu_i(x_k)]^m x_k}{\sum_{k=1}^{n} [\mu_i(x_k)]^m} \quad (\text{for all } j) \tag{26.7}$$

By iteratively updating the fuzzy membership with (26.6) and the centers with (26.7), the algorithm converges to a local minimum object function of J.

26.2.2 The improved FCM algorithm

The conventional FCM segmentation algorithm does not consider the spatial information of pixel, and this method is sensitive to noise. In [3, 6, 7], an improved FCM is present to exploit the spatial information for correct clustering. A spatial function is introduced and considered in the FCM. The new method FCMS is constructed.

Spatial information is defined as: $h_{ij} = \sum_{k \in NB(x_j)} u_{ik}$, where NB($x_j$) is the 5×5 neighborhood of this pixel. The fuzzy membership matrix is modified with spatial information as

$$u'_{ij} = \frac{u_{ij}^{p}\, h_{ij}^{q}}{\sum_{k=1}^{c} u_{ij}^{p}\, h_{ij}^{q}} \tag{26.8}$$

So the algorithm is divided into two steps. The first step is the same as the conventional FCM [12–14]. The fuzzy membership matrix and cluster prototype are firstly calculated, and then incorporated with the spatial information for achieving the new fuzzy membership value. The modification in FCM will be testified insensitive to the noise in application example.

26.3 Genetic FCMS Algorithm for Segmentation

The FCM algorithm is liable to the local optima. In order to avoid this deficiency, the paper genetic algorithm (GA) method is adopted. The genetic algorithm randomly searches and optimize the true value by the principles of evolution and natural principle. In GA, the parameters of the search space are encoded in the form of strings (called chromosomes). At first a population is randomly created, which represents different point in the search space. A fitness function is defined for representing the degree of goodness of each of the chromosome. Based the natural principle of the survival of the fittest, a few of the string are selected to go into the next generation. Reproduction (selection, crossover, mutation) operators are applied on these strings and yield a new population. The iteration continues for a fixed number of generations or value of improvement between two consecutive iterations is less than the minimum amount of improvement specified.

The genetic fuzzy C-means with spatial information algorithm used for image segmentation is descried in detail as follows:

(1) Cluster number

In order to minimize operator interaction and improve speed of parallel search in the optimum process, in the first step we adopt an automated method for determining the initial values of the centers proposed in [5, 9, 10]. The cluster number is estimated by the kernel method.

(2) Fitness function

The fitness of a chromosome indicates the degree of goodness of the solution it represents. The fuzzy clustering metric iteratively optimizes the value with the minimization of an objective function defined as the Eq. (26.1). In this paper, J is selected as the fitness function. The objective is therefore to minimize the J for achieving proper clustering. The value of fitness function of genetic string is greater, and the genetic string is more adapt to survival.

(3) Selection operator

The MR image segmentation is our object. Low contrast infrared image is more difficult for segmentation. In order to avoid the less variance of fitness function, and approximate probability of selection, a linear scaling is applied to the raw reciprocal J values.

$$J_m = f - f_{\min}^k + \xi^k \tag{26.9}$$

ξ^k is selection pressure regulation value, increased along the reduced number of iterations. Best chromosome is chosen to the next generation, it is to help to speed the convergence of GA [15]. In this paper, an "elite" algorithm is in selection process for selecting the best two members (lowest J) from the population at generation k and put into the $k+1$ generation. The roulette wheel selection process is used to select the remaining (p – 2) members of the new generation based on the fitness value.

(4) Crossover operator

The creation of new genomes from existing ones during reproduction is the process of crossover. A single crossover during reproduction has been tried with less success and is not reported here due to space limitations. P_c of cross probability variable from the iteration based the fitness.

$$P_c = \begin{cases} P_{c1} - \dfrac{(P_{c1} - P_{c2})(f_{\max} - f)}{f_{\text{avg}} - f}, & f \geq f_{\text{avg}} \\ P_{c1,} f < f_{\text{avg}} \end{cases} \tag{26.10}$$

where $f_{\text{avg}}$is the average fitness of each generation in population, $f_{\max}$ represents the max fitness of each generation in population, f is fitness of the cross individual, in general P_{m1} = 0.1, P_{m2} = 0.001.

(5) Mutation operator

The role of the mutation operator is to introduce new genetic material to the gene pool, thus preventing the inadvertent loss of useful genetic material in earlier phases of evolution. The mutation operator in GFCMS flips each bit of the bit string with a small probability P_m ($P_m = 0.03$). The probability of mutation is created randomly. if the probability of mutation is less than the P_{m}, then the mutation is implemented in a bit selected randomly.

Taking the partial derivative of Eq. (26.5) with respect to $\mu_i(x_k)$and v to zero results in three necessary conditions for J to be at a minimum. Using these conditions, the steps of our GFCMS algorithm can then be described as follows:

Step 1:
Provide initial clustering number by the kernel-function method, the initial population is created based on the clustering number. A genetic string represents combination of clustering prototype. For example, string (23, 58, 90, 135) represents four clustering number and each allele represents the gray threshold for segmentation.
Step 2:
For each chromosome, the fuzzy membership matrix, new center should be calculated. Among them, the fuzzy membership matrix was attained by Eq. (26.6), and the new cluster center was attained by Eq. (26.7).
Step 3:
New fuzzy membership matrix with spatial information is calculated by Eq. (26.8).
Step 4:
With the new fuzzy membership and cluster center, the fitness value is calculated by Eq. (26.1).
Step 5:
Genetic operators (Selection, crossover and mutation) are applied on the whole population, new generation population is produced for next iteration.
Step 6:
If the fitness value difference between two iteration is less than a given minimum number, or the iteration number is more than a given number, the algorithm stops. The least fitness value of gene string is selected as the result.

26.4 Experiment and Conclusion

Please punctuate a displayed equation in the same way as ordinary text but with a small space before the end punctuation (Fig. 26.1).

26.5 Summary

In this paper, we proposed a novel segmentation method which hybrids the modified FCM (FCMS that incorporates the spatial information into the membership function) with genetic algorithm to segment the image by the minimum object function. We adopt the kernel method to estimate the cluster number without prior conditions. The GFCMS is applied on the MR image, the result

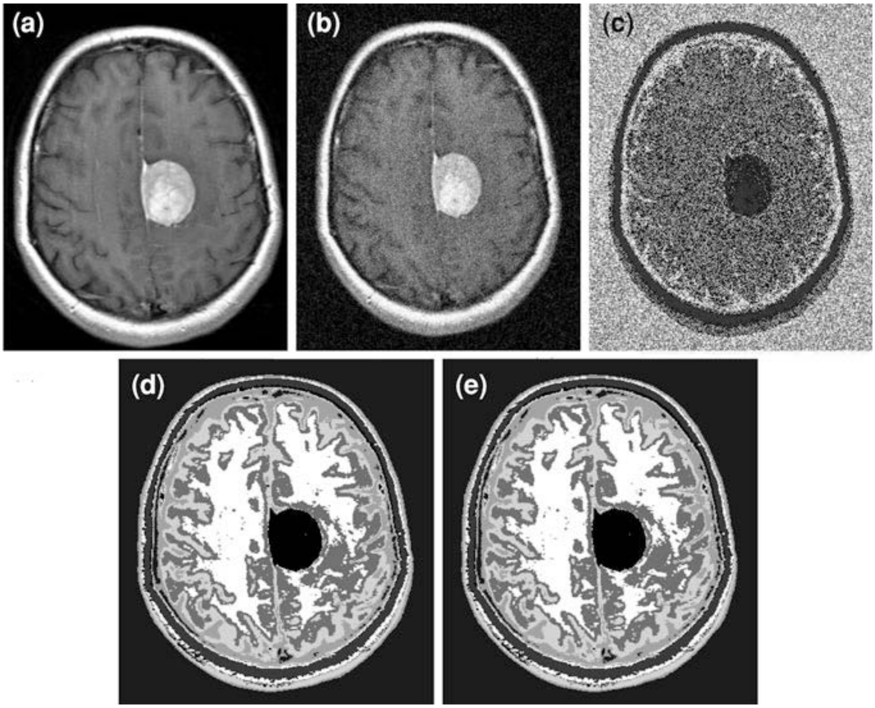


Fig. 26.1 **a** Source image for the study. **b** The image with gauss noise. **c** The segmentation results obtained by using a standard FCM algorithm. **d** The results of the FCMS with the spatial information with ($p = 1,\ 1 = 2$). **e** The result used by the GFCMS method. The experiment shows the segment based on the fuzzy partition are better for the FCMS than the conventional FCM. The result of GFCMS is effective same as the FCMS. However, the GFCMS search the optimum in parallel. The method does not depend on the initial selection and can achieve the global optimum

shows that this algorithm can achieve the good effective result as FCM and better than it, which cannot trap in local optimum.

References

1. Dunn JC (1973) A fuzzy relative of the ISODTA TA process and its use in detecting compact well separated clusters [J]. J Cybern 3(3):32–57
2. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms [M]. Plenum Press, New York
3. Chuang K-S, Tzeng H-L (2006) Fuzzy c-means clustering with spatial information for image segmentation[J]. Comput Med Imaging Graph 30:9–15
4. Maulik U (2009) Medical image segmentation using genetic algorithms [C]. IEEE Trans Inform Technol Biomed 13(2):166–173

5. Pham DL, Prince JL (1999) An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities [J]. Pattern Recogn Lett 20:57–68
6. Maulik U, Bandyopadhyay S (1999) Genetic algorithm-based clustering technique [J]. Pattern Recogn 33(2000):1455–1465
7. Andrey P (1997) Selectionist relaxation: genetic algorithms applied to image segmentation [J]. Imag Vis Comput 17(1999):175–187
8. Scheunders P (1997) A genetic c-means clustering algorithm applied to color image quantization [J]. Pattern Recogn 30(6):859–866
9. Kanungo T, Netanyahu NS, Wu AY (2002) An efficient K-means clustering algorithm: analysis and implementation [J]. IEEE Trans Pattern Anal Mach Intell 7(24):881–893
10. Bezdek JC, Boggavarapu S (1994) Genetic algorithm guided clustering [C]. In: Proceedings of the first IEEE conference on evolutionary computation, pp 34–39
11. Krishna K, Narasimha Murty M (1999) Genetic K-means algorithm [J]. IEEE Trans Syst Man Cybernet Part B 29(3):433–450
12. Coleman GB, Andrews HC (1979) Image segmentation by clustering [J]. Proc IEEE 67:773–791
13. Clausi DA (2002) K-means iterative fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation [J]. Pattern Recogn 35(2002):1959–1972
14. Likas A, Vlassis N, Verbeek JJ (2002) The global k-means clustering algorithm [J]. Pattern Recogn Soc 36(2003):451–461
15. Singh M, Patel P, Khosla D, Kim T (1996) Segmentation of functional MRI by K-means clustering. IEEE Trans Nucl Sci 3(3):2030

Chapter 27
The Real-Time Image Processing System Based on ARM9

Lijun Xu, Guohong Gao, Shitao Yan and Junhui Fu

Abstract With the rapid development of digital image processing technology, the applications of embedded systems are widely used in image processing. This paper first describes the image processing method and process, and accounts for the hardware and software design of ARM9 based real-time image processing with an emphasis. Through testing, the system can overcome the weaknesses of the traditional mode of PC-based image processing system, improve the system's real-time performance and reduce the cost as much as possible.

Keywords Image processing · ARM9 · TMS320C6205 · CPLD · SAA7111

27.1 Introduction

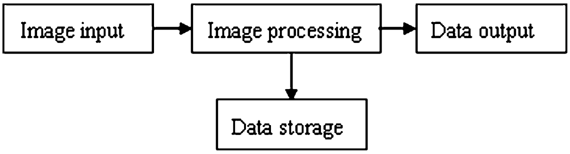
The rapid development of digital image processing technology enables all the problems of image processing to be solved in the form of digital signal processing, which provides a broad space for real-time image processing applications. First of all, there are a large number of mature and fast algorithms in digital signal processing, such as FFT, FHT, etc. These algorithms have been used extensively in image processing. Second, with the rapid development of ultra-large scale

L. Xu (✉)
Institute of Computer and Information Engineering, Xinxiang University,
Xinxiang 453003, Henan, China
e-mail: xljwork@126.com

G. Gao · S. Yan · J. Fu
Institute of Information Engineering, Henan Institute of Science and Technology,
Xinxiang 453003, China
e-mail: 914747841@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_27,

Fig. 27.1 Block diagram of image processing system



integrated circuit, the development of embedded processors provides the possibility for achieving high-speed signal processing. These developments make image processing technology to be applied widely in research, industrial and agricultural production, resources of remote sensing, medical and health, space exploration and other fields [1].

This paper first describes the image processing method and process, and accounts for the hardware and software design of ARM9-based real-time image processing with an emphasis.

27.2 Principles of Image Processing System

A basic image processing system has a specific function for each module, namely, image input, image processing and image storage and other data storage, processing results output as seen in the block diagram in Fig. 27.1.

27.2.1 Module of Image Input

The input module of the image mainly consists of image input and image acquisition. In practice, the image input generally acquires the present pending scene with a CCD camera. Because many camera output signals are analogous, we must use image acquisition cards to achieve the camera and computer interface and make it possible to use a PC.

27.2.2 Data Storage Module

Data storage includes the stored image of the system, the algorithm of image processing and the number of intermediate values of data storage. Image contains a large amount of information, which also requires plenty of storage space for images. Large capacity and fast video memory is essential to the image processing system.

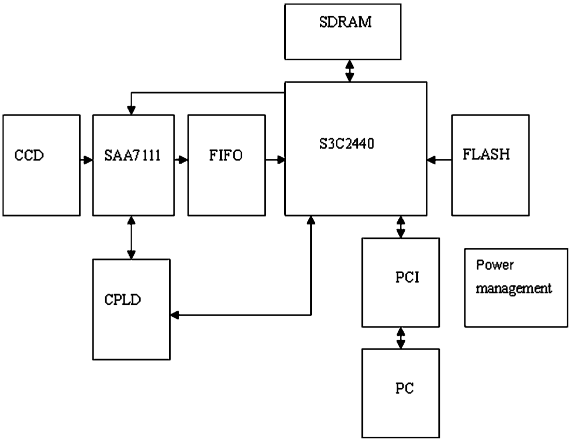


Fig. 27.2 Real-time image processing system structure

27.2.3 The Data Output Module

Image processing systems usually have the image output module and the final result of output image processing. Results can be the final processed image, and can also be other forms of data results of image processing.

27.2.4 Image Processing Module

Algorithms are available in the form of the general description of the image processing and analysis; most algorithms can be realized by software. Of course, the PC can be referenced under the special hardware so as to increase the speed limit or overcome. Especially with the technology and the application of DSP chips, which makes the image processing module performance to increase.

27.3 Hardware Realization of Real-Time Image Processing System

Real-time image processing system is shown in Fig. 27.2, the overall framework. The system consists of two subsystems, one is the image acquisition part, which completes the image of the triggered acquisition. The other is based on S3C2440 ARM9 microprocessor core image processing system, which finishes the interfaces and functions of image processing and PC machine.

27.3.1 Hardware Design of S3C2440 and the FIFO Memory Interface

FIFO memory (FIFO) interface design is more complicated, first, because of the different types of FIFO itself,and secondly because it is varied for the same FIFO in the interface design, and may require additional logic. FIFO can be divided into three types: Synchronous FIFO C Synchronous FIFO), asynchronous FIFOC Asynchronous FIFO) and the trigger FIFO. The FIFO status flag signals are generally provided to control data transmission, including the empty flag/EF (empty flag), full of signs/FF (full flag) and the half-full mark. In addition, most of the FIFO pin can provide specific signals for depth expansion and word extension [2].

27.3.2 DSP Hardware Design and FLASH Interface Partial

Flash memory (FLASH Memory) has been widely used at the end of each application in the recent years. It is a kind of EEPROM, which processes re-write capability online and is more convenient to use. FLASH read operation is identical to the ordinary SRAM, but write is relatively more complicated [3], because you need to write a bunch of command word sequences in the FLASH write. However, from the point of view of interface design, FLASH and SRAM basically are not different. It is somewhat a special attention to space allocation in the interface timing.

S3C2440 is more convenient in the design of asynchronous interface. Users have the flexibility to set different types/speeds of asynchronous direct interfaces to the device. Asynchronous interface signals include four control signals:

/AOE, export permits, valid for the entire time period.
/AWE, written permission of the trigger phase of the write cycle remains in effect.
/ARE, read permission, the trigger phase of the read cycle remains in effect.
/ARDY, ready signal, insert the wait [4].

In practice, a combination of the four control signals is used in order to meet the requirements of different types of asynchronous interfaces, which seamlessly interfaces official EMIF competence. 6205 EMIF asynchronous interface supports 8bit/16bit/32bit ROM (FLASH) access, if the width of the FLASH is less than 32bit, it is a so-called "narrow storage" read the data, EMIF automatically reads the data into one number 32bit value. EMIF working in this case, has the following characteristics: no matter what the width of the memory access, each reading is always carried out by 32bit [4].

EMIF always reads the lower addresses of the data first, which came in LSB, then reads the next byte of data in turn placed at a higher position. This means that regardless of the user who sets the chip's LENDIAN bit of the value, ROM (FLASH) in the data storage must be little endiazi.

The output of the address will automatically shift, which guarantees the narrow memory access operations to provide the correct address. On 16bitROM (FLASH), the address automatically leaves one, on 8bitROM (FLASH) the address automatically leaves two. The higher of the two addresses will be discarded.

27.3.3 S3C2440 and the PCI Interface Hardware Design

S3C2440 chip integrated with a PCI module is convenient to interface design. The PCI module meets the specification requirements PCI2.2, PCI data path width of 32bit, the peak data transfer rate up to 33 M. With this module, it is easy to interface between S3C2440 application system and PCI host. The PCI module chip integrated PCI master/slave bus interface to support the S3C2440 and the direct connection between the PCI hosts [5].

As S3C2440 is a built-PCI module, making use of the system has faster speed and better stability than the PCI bridge chip. In addition, the preparation of its driver is not on the S3C2440 by HPI bus read and write, but is the register directly through the PCI. Built-in PCI interface, including 4 for S3C2440 and host data transfer FIFO, were used as the host write, read by the host processor to write the processor time.

27.4 Design and Implementation of the System Software

Real-time image processing system software design involves the main program, program loading and startup, DMA image transferring and controlling programs.

27.4.1 Flow Description in DSP Main Program Flow Description

The main program throughout the whole system, includes the acquisition module part of the acquisition setup and start SAA7111 chip, interrupt response, moving images collected and processed with the external memory SDRAM, FLASH, and the interface with the PC host, etc. In order to facilitate program implementation, the C implementation is used in the main frame. In the specific algorithm, for some of the key code, the embedded assembly achieves and makes the appropriate software optimization to improve the running efficiency.

After starting the C6205 power-on or reset to complete, the system is loaded and starts to complete system initialization and set of each parameter, that is the system bootstrap, and then it begins to wait for an external interrupt. When the

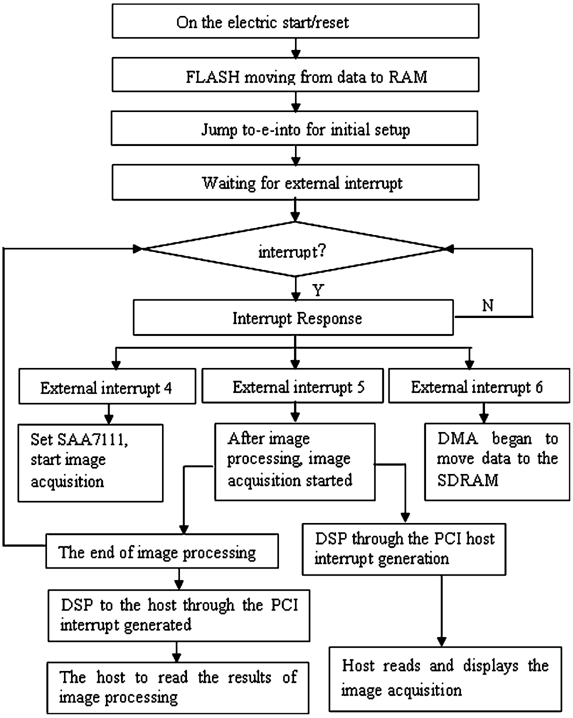

On the electric start/reset
FLASH moving from data to RAM
Jump to-e-into for initial setup
Waiting for external interrupt
interrupt?
Y
N
Interrupt Response
External interrupt 4
External interrupt 5
External interrupt 6
Set SAA7111, start image acquisition
After image processing, image acquisition started
DMA began to move data to the SDRAM
The end of image processing
DSP through the PCI host interrupt generation
DSP to the host through the PCI interrupt generated
The host to read the results of image processing
Host reads and displays the image acquisition

Fig. 27 3 System main program flowchart

trigger external interrupt INT4 images are collected after the election, S3C2440 start SAA7111 capture chip, the image acquisition module is running. When the FIFO half full interrupt generates INT6, S3C2440 starts receiving image data through writing the interrupt service routine. When finished image acquisition occurred after INT5 interrupt by writing the interrupt service routine, DSP image processing begins and the image needs to be sent through the PCI to the host display.

When the image processing is completed, the results need to be sent to PC through the PCI bus master. S3C2440 interacts with the host primarily through three registers to achieve, namely, HSR, HDCR, RSTSRC. RSTSRC can only be visited by S3C2440, HSR can only be visited by the host access.

To achieve S3C2440 interrupt to the host, you first need to host a program on the HSR in the INTAM position 0, and then, in the process of S3C2440 RSTSRC the INTREQ position 1, an interrupt to the host, the host registers by constantly testing and HSR's INTSRC INTAVAL bit to determine whether an interrupt has occurred. S3C2440 achieves the host to interrupt, to write 1 to register DSPINT HDCR interrupt bit to S3C2440, and then write HOSTSW interrupt service routine to respond to the interrupt. S3C2440 host access needs more than write drivers to achieve. The system main program flowchart is shown in Fig. 27.3.

27.4.2 Start Loading in the Program System

Because ultimately image processing runs the main program offline, therefore you first need to program into the rLASH, the selection of the bootstrap ROM way. Using CCS software (Code Composer Studio) to compile the main program, in the C6205 program, you must write cmd file. This file contains the system's storage space allocation, space allocation and the main program as follows:

```
  MEMORY
  {
  RAM-ECTOR:origin = 0 X 00000000,len = 0 X 00000200
         //RAM interrupt vector addresses in the internal
procedures
  RAM-BOOT:origin = 0 X 00000200,len = 0 X 0000FDC0
         //Bootloader in RAM,the address of the internal
  procedures
  SDRAM -CODE:origin = 0 X 00400000,len = 0 X 1000000
         //Address in the main program code in the SDRAM
  ROM -VECTOR:origin = 0 X 1400000,len = 0 X 200
         //Interrupt vector address in the external FLASH
  ROM -BOOT:origin = 0 X 1400200,len = 0xFC
         //Bootloader in the address in the external FLASH
  ROM- CODE:origin = 0 X 80000000,len = 0 X 10000
         //Internal data RAM address
  }
  SECTIONS
  {
  .vectors:load = ROM -VECTOR, run = RAM --VECTOR
         //Interrupt vector segment
  .boot_ load: load = ROM BOOT,run = RAM -BOOT//Block bootstrap
  text:
    load = ROM  -CODE,  run = SDRAM  --CODE//Program  text
  segment
    init: load = ROM -CODE,run = SDRAM_CODE//Initialization
  section
    .const:load = ROM -CODE, run = SDRAM_CODE//Constant section
    .switch: load = ROM -CODE,run = SDRAM_CODE
         //Contains a large switch statement tables above.
  . bss > IDRAM//Uninitialized variable retention storage
  space, mainly refers to global variables and static
  variables.
  .data > IDRAM//The data has been initialized
  . stack > SDRAM --CODE//Distribution system stack space,
the system stack is mainly used for function arguments and
local variables
  .sysmem > SDRAM CODE//Reserve space for dynamic memory
  allocation, dynamic memory mainly refers to the function
  Malloc
  }
```

MEMORY provides the memory field in the ARM at the beginning of the address and length. In the use of ROM bootstrap method, initially all of the program code are stored in rLASII, including the interrupt vector,bootstrap procedures and the main program.

SECTONS assembly code provides for the memory of the field in the field, where the load to burn the code written into the FLASH run is the code that runs in RAM.

First, the plate. out files is compiled by the usage of the above cmd file and main program code, which are then burned ROM code written into the FLASH, it needs to write another cmd file pla seven e. cmd, which reads as follows :

```
plate.out
-a/*Intel hex format*/
-map mainhex.map/*Generate a map file*/
-byte/*Number outpufile locations by bytes*/
/*rather than using target addressing*/
-image/*Specify image mode*/
-memwidth 8/*Define the system memory word width*/
-romwidth 8/*Specify the ROM device width*/
-order L/*Output file is in little endian forma*/
ROMS
{
EPROM:org = 0 X 1400000, len0 X 90000, romwidth8,
    files = {plate. hex}
    }
```

27.5 Conclusion

The ARM9 processor-based real-time image processing system is commonly a set of image trigger, acquisition, processing and transmission on the single board system; it makes full use of ARM9 chips in the high performance image processing, which overcomes the weakness of the traditional PC-based acquisition card and image processing system Real-time, poor treatment and low accuracy, confidentiality and poor scalability, etc. The system takes the main image processing to the ARM9 chip, greatly reducing the load on the computer, making the whole system able to be achieved by others, such as a computer database management, network and so on.

References

1. Liang Hui-jun,Jia Song-hao,Yang Cai, Li Fang-fang (2009) An implementation method of embedded network interface. J Nanyang Norm Univ 12:15–18
2. Song T, Hao X (2009) Research and exploration of Linux-based embedded remote video capture control system. J Huangshi Inst Technol 25:1
3. Zhu W, Wang Q, MA H. (2007) Design of embedded network interface based on ARM9. Microcomput Inf 2007(09):0160–0161
4. Pei-jun L, Jun Z (2009) The video inspect system bases on embedded Linux system. Comput Knowl Technol 5:11
5. Li X, Ji R, Zhang L. (2009) Video collection system based on embedded Linux and ARM9. Electron Meas Technol 32(2):102–104

Chapter 28
A Novel Fuzzy C-Means for Image Segmentation

Cuiyin Liu, Zhang Xiu-Qiong and Xiaoling Li

Abstract In this paper, we present a novel algorithm for fuzzy segmentation of infrared image data using fuzzy clustering. A conventional FCM assigns the data into group, where the data is nearest to the center of group. Although FCM is populated in image segmentation, it still has the following disadvantages: (1) a conventional FCM algorithm does not consider spatial information for clustering. (2) The algorithm is sensitive to noise. In this paper we present a fuzzy-means algorithm that incorporates spatial information and the prior probability of a pixel neighborhood into the membership function for clustering. The modified FCM has a great improvement for noisy image and infrared image segmentation.

Keywords Fuzzy k-means · Spatial information · Clustering · Segmentation

C. Liu (✉) · Z. Xiu-Qiong
College of Computer Science, Sichuan University, Chengdu 610064, China
e-mail: liucuiyin@163.com

Z. Xiu-Qiong
e-mail: zxq_03@tom.com

C. Liu · Z. Xiu-Qiong
State Key Laboratory of Fundamental Science on Synthetic Vision,
Chengdu 610064, China

C. Liu
College of Computer Science, Sichuan Panzhihua University, Panzhihua Sichuan,
Panzhihua 617000, China

Z. Xiu-Qiong
College of Computer Science, Leshan Normal University, Leshan 614000, China

X. Li
School of Information Science and Technology, Chengdu University,
Chengdu 610064, SiChuan, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_28,

28.1 Introduction

Image segmentation is a key process in a variety of applications, such as machine vision, pattern recognition and tracking system and so on. The object of image segmentation is to extract some interesting regions or find the needed structure and consistent or not consistent shape for the next progress. In order to acquire the correct result, variety methodology has been developed to solve this key problem. Clustering is populated for its simplicity and ease of realization. Clustering techniques are unsupervised methods they do not use prior class identifiers.

The hard clustering algorithm assigns a pixel to a group including two situations which are true or not. The fuzzy clustering algorithms assign a data to a group based on the membership. The pattern will not hardly belong to a group. Another presentation membership indicates the possibility of data in group. Fuzzy c-means (FCM) clustering [1, 2] is an unsupervised technique that has been successfully applied to image segmentation. The FCM has been reported by Dunn in [1], and proposed by Bezdek in [2]. The conventional FCM segmentation algorithm only utilizes the gray value of image and does not consider the spatial information, so that the segmentation adaptive range is limited. Chuang in [3] modified the method by altering the membership weighting of each cluster based on spatial information. This scheme greatly reduces the effect of noise classified homogeneous clustering. However, this method cannot completely solve the noisy image segmentation. Ahmed proposed an effective FCM (BFCM) algorithms for bias field estimation and segmentation [4–6]. Although BFCM works well in segmentation of MRI image, it fails to segmentation of infrared image.

To solve the problem of the noise sensitivity and fitting for infrared segmentation, we present in this paper a different approach for fuzzy segmentation of infrared image. Our new method incorporates spatial information, neighborhood prior probability and the membership weighting of each cluster is altered after the cluster distribution in the neighborhood is considered. This scheme greatly improved the incorrect segmentation of noise and remove singular noisy pixel, and also has effective segmentation in infrared image, even with noise. In Sect. 28.2, we reviewed conventional FCM algorithm at first, and propose our new method and describe the steps of MFCM algorithm. In Sect. 28.3, we applied the algorithm to MRI image and infrared image, and then draw comparisons among standard FCM, FCMS, BFCM and MFCM. Finally in Sect. 28.4, we have some conclusion.

28.2 Fuzzy C-Means Clustering Algorithm

28.2.1 Stand Fuzzy C-Means Clustering

FCM is an unsupervised clustering algorithm that has been successfully applied to image segmentation [7]. An image can be represent by $X = \{x_1, x_2, \ldots, x_n\}$ Each $x_i \in R^n$ is a feature vector. The features could be gray value, local character, texture and other characteristic of the image. In this paper, the x_i is the gray value of ith pixel. FCM can assign patterns into each category by minimizing the object function. The algorithms iteratively optimize the value with the minimization of an objective function defined as the equation:

$$J = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m ||x_k - v_i||^2, \quad V = [v_1, v_2, \ldots, v_c], \quad v_i \in R^n, \tag{28.1}$$

$$D_{(k,i)} = ||x_k - v_i||^2 \tag{28.2}$$

where μ_{ik} represents the fuzzy membership of pixel x_k in the cluster of v_i. $D_{(k,i)}$ is the distance from pattern k to the cluster prototype v_i. ||.|| is a metric adopted Euclidean distance in this paper. The exponential m controls the fuzziness of the resulting partition. U is fuzzy matrix and with the constraint is

$$\sum_{k=1}^{c} \mu_{ik} = 1, 1 \le k \le c \tag{28.3}$$

The partition process is an iterative calculation. By minimizing (28.1), using Lagrange multiplier method, the update equations of membership function μ_{ik} and the cluster center v_i are given in Eq. (28.4) and (28.5). The clustering process stops when the maximum number of iterations is reached or when the objective function improvement between two consecutive iterations is less than the minimum amount of improvement specified.

$$\mu_i(k_k) = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{(i,k)}}{D_{(j,k)}} \right)^{1/(m-1)}} \tag{28.4}$$

$$v_j = \frac{\sum_{k=1}^{n} [\mu_i(x_k)]^m x_k}{\sum_{k=1}^{n} [\mu_i(x_k)]^m} \tag{28.5}$$

28.2.2 The Improved FCM Algorithm (MFCM)

The FCM can segment a clean image into correct, which without the noise and homogeneity. In real application, the images for segmentation are corrupted by noise. There are shadow and in-homogeneity field in the image and make difficult for segmentation. The ideal result cannot be achieved by the conventional FCM [8]. In order to improve the adaptability of the algorithms and insensitivity to noise, the spatial information and statistic is an important characteristic to classify pixel will be utilized in the new method.

In order to acquire the good result, a modified FCM algorithm (MFCM) framework is proposed [9, 10]. In modified algorithm, the spatial information and prior probability in neighborhood is used to modify the membership matrix. A modified membership matrix functions are defined as follows:

Step 1. Calculate the prototype of centers, membership matrix and the distance between pattern and the cluster's center. All these process is same as a conventional FCM [11, 12].
Step 2. Use the spatial information and the prior possibility to modify membership value.

At first, we calculate the possibility of pixels belonging to cluster in neighborhood. We defined the result possibility as a vector $P = \{p_1,p_{2,} \ldots, p_c\}$, and the possibility vector is calculated by the following equation:

$$p_i = \frac{n_i}{N} \tag{28.6}$$

where n_i denotes the count of pixel belonging to *i*th cluster, and N is the count of pixels in neighborhood.

Secondly, we calculate the variance of every pixel neighborhood. If the current pixel point with maximum of variance, the membership is updated by Eq. (28.7).

$$\mu_{i,j} = \mu_{i,j} * P \tag{28.7}$$

Otherwise, the membership for current pixel without the maximum of variance is updated by Eq. (28.8).

$$u_{ij}^{'} = \frac{u_{ij}h_{ij}}{\sum_{k=1}^{c} u_{ij}h_{ij}}, \quad h_{ij} = \mu_{i,j} \sum_{k \in \mathrm{NB}(x_j)} \mu_{ik} \tag{28.8}$$

$\mathrm{NB}(x_j)$ represents a square window centered on pixel x_j in the spatial domain. A 5*5 window was often used throughout this work. Just like the membership function, this spatial function h_{ij} represents the degree of pixel x_j belongs to *i*th cluster. This method adopts the method proposed in [2].

The FCM incorporated spatial information and statistic possibility is denoted as MFCM [13, 14]. There are two steps in the scheme. The first is the same as that in conventional FCM to calculate the membership function, prototype and distance matrix. The second is to update membership. There are two situations for the membership updating. One is for the pixel with the maximum of variance the other is for the pixel which is not.

28.3 Experiment and Discussion

28.3.1 Experiment and Discussion

In this section, we describe the application of the MFCM segmentation on MRI images corrupted with multiplicative gain, as well as digital MRI (Figs. 28.1, 28.2).

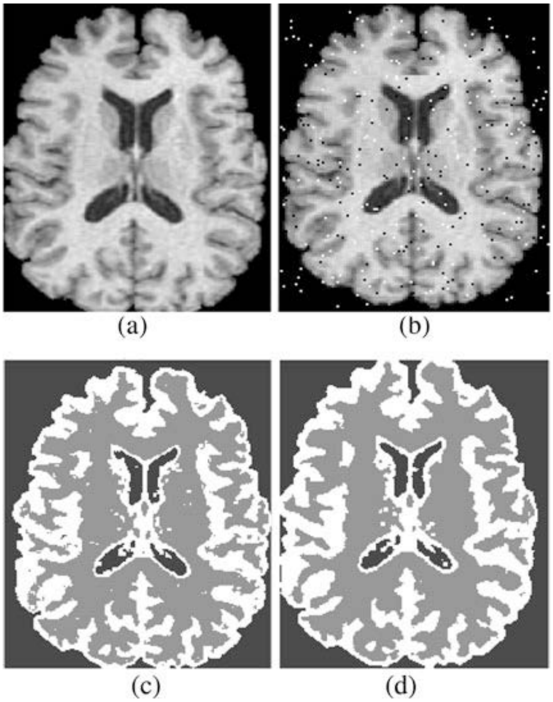


Fig. 28.1 **a** is original image for the study. **b** is the image with pepper noise. **c** shows the segmentation result obtained by using standard FCM algorithm. **d** shows the results of the MFCM

The experiment shows the segment based on the MFCM method are better than the conventional FCM and FCMS. There are some noises in the result image, but the situation in the FCMS segmentation image has been improved. Another FCM clustering was proposed by the Ahmed, which is very useful for bias field estimation and intensity in-homogeneity image segmentation. We have experimented this clustering on infrared image segmentation, which shows this method has no effect in infrared noisy image segmentation.

28.3.2 Cluster Validity Functions

In order to quantitatively assess four methods (FCM SFCM MFCM), the partition coefficient PC and the partition entropy XB are adopted to evaluate the segmentation results. PC and XB are two cluster validity functions which are defined as follows:

$$\mathrm{PC} = \frac{1}{N}\sum\nolimits_{i=1}^{c}\sum\nolimits_{j=1}^{N}(\mu_{ij})^2 \tag{28.9}$$

$$\mathrm{XB} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m||x_j - v_i||^2}{N\min_{i,j}||x_j - v_i||^2} \tag{28.10}$$

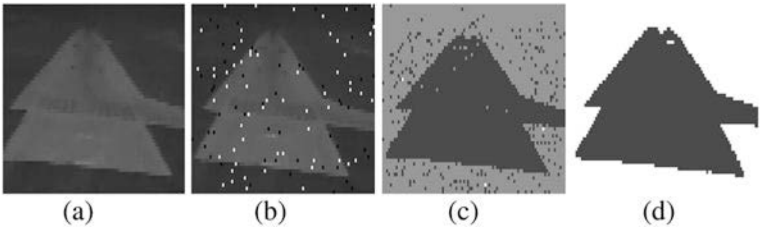


Fig. 28.2 **a** is original infrared image for the study. **b** is the infrared image with pepper noise. **c** shows the segmentation result obtained by using standard FCM algorithm. **d** shows the results of the MFCM

Table 28.1 The clustering results of four images using various FCM techniques

Images	Algorithm	PC	XB
MRI images	FCM	0.4132	125.1098
	SFCM	0.3458	26.9487
	MFCM	0.3674	12.3485
Noisy images	FCM	0.4632	53.4135
	SFCM	0.3345	10.9724
	MFCM	0.3543	13.4555
Infrared image	FCM	0.9162	125.1098
	SFCM	0.9668	26.9487
	MFCM	0.9832	19.3242
Noisy infrared image	FCM	0.8925	53.4135
	SFCM	0.9531	10.9724
	MFCM	0.9342	12.4334

A good clustering result generates the minimum variation within cluster and maximum variation among different groups, it can be measured by the minimum of PC and XB value. Minimizing PC or XB is expected to be a good clustering. (Table 28.1)

28.4 Summary

In order to overcome the sensitivity of FCM to noise, a modified FCM clustering method for segmentation has been presented in this paper. The local spatial information and the prior probability have been used in this method, and experimented on the MRI images and the infrared images. The results show that the standard FCM method is sensitive to noise and the FCMS has a little improvement. The two methods cannot acquire the ideal effectiveness. In application for infrared segmentation, the two methods cannot have the good segmentation. The MFCM exploit the local information and prior probability to reduce the noisy affection and replace the maximum membership with the mean value of neighborhood. The experiment result shows that the MFCM can obtain the satisfying segmentation on noisy MRI images and infrared images.

Acknowledgments The authors would like to thank the anonymous reviewers for their detailed review and constructive comments. This work is supported by the National Natural Science Foundation of China (Grant no. 60736046).

References

1. Dunn JC, Cybernet J (1973) A fuzzy relative of the ISODTA TA process and its use in detecting compact well separated clusters [J] 3(3):32–57
2. Pattern recognition with fuzzy objective function algorithms [M], Plenum Press, New York (1981) pp 121–122
3. Chuang K-S, Tzeng H-L (2006) Fuzzy c-means clustering with spatial information for image segmentation [J]. Comput Med Imaging Graph 30:9–15
4. Singh M, Patel P, Khosla D, Kim T (1996) Segmentation of functional MRI by K-means clustering [J]. IEEE Trans Nucl Sci 43(3):2030–2036
5. Kolen JF, Hutcheson T (2002) Reducing the time complexity of the fuzzy C-means algorithm [J]. IEEE Trans Fuzzy Syst 10(2):263–267
6. Saha S, Bandyyopadhyay S (2007) MRI brain image segmentation by fuzzy symmetry based genetic clustering technique [C]. IEEE Congr Evol Comput 4417–4424
7. Zhang Y, Huang D (2011) Image segmentation using PSO and PCM with Mahalanobis distance [J]. Expert Syst Appl 38:9036–9040
8. Cai W, Chen S (2007) Fast and robust fuzzy c-means clustering algorithms incorporating local information fro image segmentation [J]. Pattern Recogn 40:825–838
9. Bo Yuan, George J. Klir (1995) Evolutionary fuzzy c-means clustering algorithm [C].IEEE 2221–2226
10. Yang J-F, Hao S-S (2002) Color image segmentation using fuzzy C-means and eigenspace projections [J]. Signal Process 82:461–472
11. Zhang D-Q, Chen S-C (2004) A novel kernelized fuzzy C-means algorithm with application in medical image segmentation [J]. Artif Intell Med 32:37–50
12. Zhao F (2011) A novel fuzzy clustering algorithm with non local adaptive spatial constraint for image segmentation [J]. Signal Process 91:988–999
13. Li Ma, Staunton RC (2007) A modified fuzzy C-means image segmentation algorithm for use with uneven illumination patterns [J]. Pattern Recogn 40:3005–3011
14. Tan KS, Isa NAM (2011) Color image segmentation using histogram thresholding-fuzzy C-means hybrid approach [J]. Pattern Recogn 44:1–15

Chapter 29
A New Image Fusion Algorithm for Recognition Capability Enhancements

Zhang Yong-mei, Ma Li and Liu Wen-kai

Abstract Image fusion principles have been widely used in application of imaging remote sensing as an effective means of synergistic information combination. Multi-sensor data fusion can obtain much more and more exact information than single sensor. A sensor often with difficulty realizes target identification, multi-sensors can quite easily identify the goal. A new image fusion algorithm between multi-spectral and panchromatic images based on object recognition is presented, which focuses on the fusion rules of wavelet coefficients and the coherent selection. The algorithm makes full use of different characters of multi-spectral and panchromatic images, which can enhance recognition capability. Experiment data indicate this algorithm improves the fusion quality, and it can benefit the correct target classification. Therefore, the approach proposed is an effective method for target recognition.

Keywords Multi-sensor · Remote sensing · Weighted factor · Target identification

Z. Yong-mei (✉) · M. Li · L. Wen-kai
School of Information Engineering, North China University of Technology,
Beijing 100144, China
e-mail: zhangym@ncut.edu.cn

M. Li
e-mail: mali@ncut.edu.cn

L. Wen-kai
e-mail: 421955907@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_29,

29.1 Introduction

Multi-sensor pixel-level image fusion algorithms mainly include sum fusion, IHS transform and wavelet analysis [1]. Sum fusion algorithm is simple and fast, but both the spatial detail and spectral information characteristics have distortion [2]. The IHS fusion method completely retains the panchromatic image information with rich high-frequency information [3], however, it has some spectral distort [4]. Recently, image fusion algorithm based on wavelet analysis is the main pixel-level image fusion method, which utilizes eyes are more sensitive to local contrast. According to certain rules, it selects the most discriminating features in two or more source images, and preserves these features in the fusion images. But fusion images exist certain ringing artifacts and loss of spatial details.

Aiming at a worse tradeoff between spectrum information and spatial details for present fusion methods, a new algorithm is proposed on the basis of image content self-adaptation fusion rules, which combines with local statistical feature analysis in wavelet domain, focuses on the fusion rules of wavelet coefficients and the coherent selection. In addition, region weighted and consistency checking rules are used to ensure the consistency of high-frequency detail components.

29.2 Multi-Sensor Data Fusion for Goal-Oriented Identification

29.2.1 Self-Adaptation Fusion Rule

In the fusion process, the selection of fusion rules and fusion operators plays a decisive role in fusion quality, and it is one of the difficulties in image fusion. The two-dimension wavelet decomposition is used to decompose the original images to two different parts, namely, the low-frequency part LL and the high-frequency part. The high-frequency part contains the horizontal high-frequency HL, the vertical high-frequency LH and the diagonal high-frequency HH. For the low-frequency part, LL mainly reflects the intensity of the original images. For the high-frequency part, HL, LH, and HH mainly reflect the texture and structure information of the original images, where HL states the horizontal changes, LH shows the vertical changes, and HH represents the diagonal direction changes. For the multi-spectral and panchromatic image fusion, both the spatial details and the spectral information of the fusion image should be taken into account when combing the wavelet coefficients.

The adaptive fusion algorithm for goal-oriented identification selects different fusion rules according to the different frequency domain characteristics. For the high-frequency part, it respectively adopts horizontal, vertical and diagonal Sobel boundary extracting operator in the light of the coefficient direction, selects the largest eigenvalue region of each direction as the combined weight factor for wavelet coefficients. For the low-frequency part, selecting the regional local energy as the weight factor. Finally, adaptively determine the high-frequency and low-frequency coefficients and weight-based method through the selected characters or combination of characters.

After the wavelet transform for panchromatic images, the low-frequency part reflects the major radiation energy. For the multi-spectral images, I component is the main radiation energy after IHS transform. The low-frequency part determines the basic brightness radiation energy, which is the main factor to affect the spectrum preservation for fusion images. For the panchromatic images, it mainly comes from the brightness radiation energy received by panchromatic sensors. In this chapter, calculate the energy statistic of the low-frequency part for multi-spectral and panchromatic images. For the low-frequency component L_M of the multi-spectral images and the low-frequency component L_P of panchromatic images, select the statistical eigenvector as the local area energy E_M and E_P, if the local window is $N{\times}N$, the formulae for the E_M and E_P list as follows:

$$E_M(x,y) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} F(i,j)L_M\left(x-\frac{N}{2}+i, y-\frac{N}{2}+j\right)^2 \tag{29.1}$$

$$E_P(x,y) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} F(i,j)L_P\left(x-\frac{N}{2}+i, y-\frac{N}{2}+j\right)^2 \tag{29.2}$$

In formula (29.1) and (29.2), x and y are respectively the number of row and column in wavelet coefficient plane, $F(i, j)$ is the energy extraction operator, i and j are respectively the number of row and column of the energy extraction operator template. Usually, $N = 3$, $F(i, j)$ is {0,1,0;1,2,1;0,1,0}. The proposed adaptive image fusion algorithm calculates the fusion weights according to the proportion relationship of statistical eigenvector, the formula is as follows:

$$\begin{aligned} C_M(x,y) &= \frac{E_M(x,y)}{E_M(x,y)+E_P(x,y)} \\ C_P(x,y) &= 1 - C_M(x,y) \end{aligned} \tag{29.3}$$

In formula (29.3), E_M and E_P are respectively the local energy value of template statistics, C_M (x, y) is the wavelet coefficients for the multi-spectral images at the point of (x, y), C_P (x, y) is the wavelet coefficients for the panchromatic images at the point of (x, y) and $C_M(x,y) + C_P(x,y) = 1$. Low-frequency wavelet coefficient for the fusion images $\mathrm{LL}_F(x,y)$ can be got in the light of the fusion weights determined by C_M (x, y) and C_P (x, y).

$$\mathrm{LL}_F(x,y) = C_M(x,y) \times \mathrm{LL}_M(x,y) + C_P(x,y) \times \mathrm{LL}_P(x,y) \tag{29.4}$$

In formula (29.4), $\mathrm{LL_M}$ and $\mathrm{LL_P}$ are respectively low-frequency wavelet coefficients of panchromatic images and I component for multi-spectral images before fusion.

For image fusion, the high-frequency part reflects the detail change of the images, so, the high-frequency part needs to enhance the characteristics information of the spatial texture and edge of images, preserve spectral information and prevent from ringing artifacts.

In the chapter, the high-frequency detail components are extracted by selecting the eigenvector as the gradient operators. Considering the orientation of the wavelet coefficients, the gradient operators based on orientation are selected. Using Sobel, extract the horizontal, vertical and diagonal edge features of the wavelet coefficient detail components for multi-spectral and panchromatic images, the wavelet coefficients for three high-frequency sub-bands LH, HL and HH in plane after the wavelet decomposition. Extract the edge characteristic value of local areas by adopting vertical, horizontal and diagonal Sobel operator. The vertical feature descriptor is $V(3,3) = \{1,2,1;0,0,0;-1,-2,-1\}$ which is sensitive to the vertical edges. Similarly, the horizontal feature descriptor is $H(3,3) = \{1,0,-1;2,0,-2;1,0,-1\}$ which is sensitive to the horizontal edges. The diagonal feature descriptor is $D(3,3) = \{-1,0,-1;0,4,0;-1,0,-1\}$ which is sensitive to the diagonal edges. Select the coefficients of largest characteristic value area in each direction as the wavelet coefficients of the location.

Calculate the characteristic statistics according to the above feature descriptors, noted as $G_M(x,y)$ and $G_P(x,y)$, represent the statistical value at the point of (x,y) in multi-spectral and panchromatic images.

$$\begin{cases} G_M(x,y) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P(i,j) Q_M\left(x - \frac{N}{2} + i, y - \frac{N}{2} + j\right) \\ G_P(x,y) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P(i,j) Q_P\left(x - \frac{N}{2} + i, y - \frac{N}{2} + j\right) \end{cases} \tag{29.5}$$

where, $P(i,j)$ is respectively the vertical feature extraction descriptor $V(i,j)$, the horizontal feature extraction descriptor $H(i,j)$ and the diagonal feature extraction descriptor $D(i,j)$ according to the three different directions, $Q(x,y)$ is the largest edge characteristic value in the horizontal, vertical and diagonal components after the wavelet decomposition for multi-spectral or panchromatic images.

Then, calculate the fusion weights based on feature statistics in the following way:

$$\begin{cases} C_M(x,y) = \frac{G_M(x,y)}{G_M(x,y)+G_P(x,y)}, & \text{if } G_M(x,y) \le G_P(x,y) \\ C_M(x,y) = 0, & \text{else} \end{cases} \tag{29.6}$$

$$\begin{cases} C_P(x,y) = \frac{G_P(x,y)}{G_M(x,y)+G_P(x,y)}, & \text{if } G_M(x,y) \le G_P(x,y) \\ C_P(x,y) = 1, & \text{else} \end{cases} \tag{29.7}$$

Similarly, C_M (x,y) means the wavelet coefficient fusion weights at the point of (x,y) in multi-spectral images, C_P (x,y) means the wavelet coefficient fusion weights at the point of (x, y) in panchromatic images, and $C_M(x, y) + C_P(x, y) = 1$. The wavelet coefficients of fusion images can be got according to the above fusion weights. The proposed adaptive fusion rule can select different fusion rules for the high- and low-frequency part after the wavelet transform according to the specific circumstances of each image, and adaptively determine the high- and low-frequency fusion coefficients and weight-based methods by selecting image features or feature combination.

29.2.2 Conformability Choice Criteria

After determining the wavelet coefficients of fusion images, fusion weights for the high-frequency detail components are further regionally weighted to ensure that different wavelet frequency bands with continuous gray changes and consistency spatial detail, thereby, it can improve the continuity and integrality of image content. Considering that the low-frequency components should primarily maintain spectral information of multi-spectral images, the high-frequency components should mainly preserve detail information of panchromatic images, select $k_1 \geq 1, k_2 \leq 1, k_3 \leq 1$. The fusion weights are calculated by $C'_M(x, y) = \min(C_M(x, y) \times k_s, 1)$, $C'_P(x, y) = 1 - C'_M(x, y)$, k_s is respectively denoted as k_1, k_2 and k_3 according to the location of (x, y).

In this chapter, the majority principle is used in consistency checking for the high-frequency detail components of fusion images. If at least six wavelet coefficients in the eight neighborhoods at a location (x, y) come from panchromatic images, then adjust the wavelet coefficients of the location for the wavelet coefficients of panchromatic images.

29.2.3 A Multi-Sensor Fusion Algorithm for Target Identification

In this chapter, a multi-sensor fusion algorithm for target identification is presented to realize multi-spectral and panchromatic image fusion. The specific steps list as follows:

Resample panchromatic and multi-spectral images for the same spatial resolution, then strictly register.

Transform multi-spectral images of low spatial resolution in RGB color spaces to HIS color spaces, achieve the I component characterizing the surface radiation energy, and physical parameters that are H and S components characterizing spectral characteristics.

Respectively transform panchromatic images and I component of multi-spectral images based on wavelet transform.

After the wavelet transform, the high- and low-frequency sub-band wavelet coefficients are weighted fused by adopting image content self-adaptation fusion rule and conformability choice criteria.

Transform the weighted fusion coefficients on the basis of inverse wavelet transform, get a new *I* component, transform the *I* component and *H* and *S* components of the original multi-spectral images based on inverse IHS transform and get the final fusion images.

The spectral information of multi-spectral images is mainly embodied by the *H* component and *S* component, the *I* component reflects the spatial resolution of images, so fuse panchromatic images and the *I* components on the basis of the wavelet decomposition fusion, and the fusion algorithm can improve the spatial resolution and retain spectrum information for multi-spectral images.

29.3 Experiment and Result Analysis

The chapter appraises the fusion images on how to improve the discrimination between objects and background, and the recognition efficiency. When tested whether the fusion images can improve target recognition probability or not, for the specified objectives (depending on the contents of the remote sensing images), the target recognition is respectively observed before and after fusion in order to decide whether to improve the target recognition rate or not.

In Fig. 29.1, the most salient object is a meandering “V”-type road in the mountain, many goals are easily identified in (c). Contrast to rectangular area, namely, the discrimination between object and background is the most significant region. It can be found that (a) is a multi-spectral image, resolution is low, roadside objects in the rectangular area are rather vague and difficult to identify, (b) is a panchromatic image, resolution is high, some space detail information of roadside targets can be identified, however, the information is not enough to identify the targets, (c) is the fusion result between multi-spectral and panchromatic image by adopting the proposed fusion algorithm in this chapter, vegetation texture information is more clear, detail and contour is richer. After fusion image, it can be seen that the spectral information and spatial structure details for the targets organically combine, the color is more natural. It evidently increases the target information, improves the discrimination between object and background and easily identifies the target to be the roadside trees.

We experiment on many multi-spectral and panchromatic images of different terrains including bridges, airports, plants and other objects, the function of the system is tested in a more comprehensive way. The selected experimental remote sensing regions contain many types of objects.

Region segmentation by image fusion is the first step, and the next is feature extraction of target structure in the intersected region. Finally, the target recognition is performed using the structure of the object. The recognition results are shown in Table 29.1. When using a single sensor, the experiment results

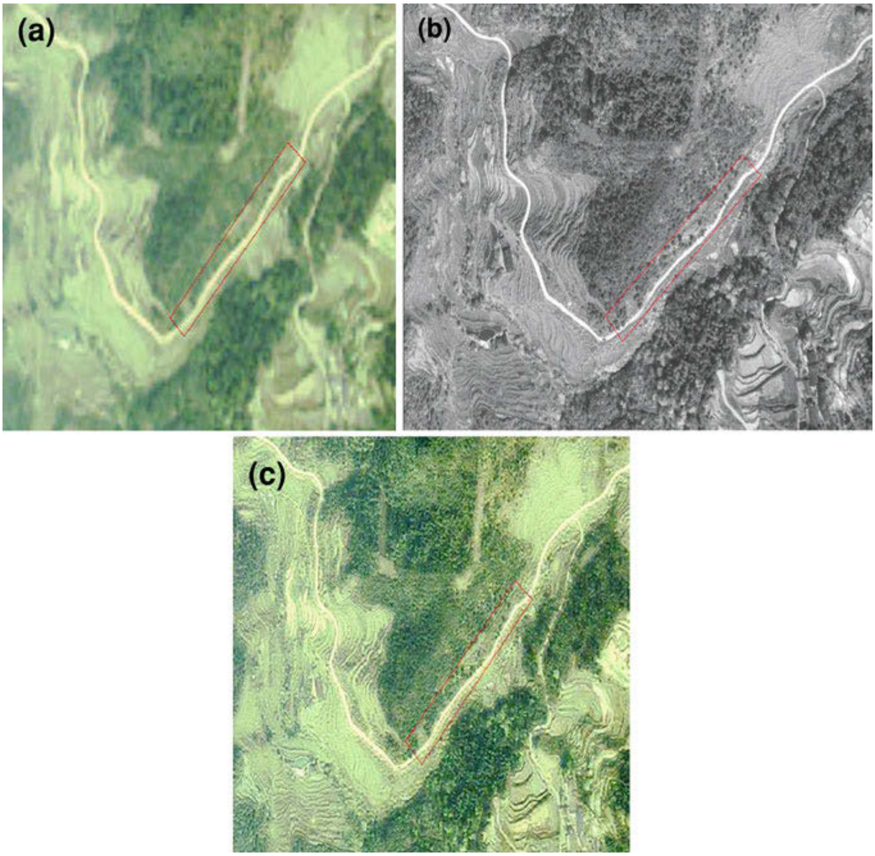


Fig. 29.1 The fusion image between a multi-spectral and panchromatic image. **a** A multi-spectral image. **b** A panchromatic image. **c** The fusion image

Table 29.1 Recognition results

Target type	Correct recognition rate of a single sensor (%)	Correct recognition rate of multi-sensor (%)
Bridges	61.2	89.6
Airports	70.3	94.6
Plants	78.6	97.8

demonstrate that the average recognition rate achieves 70%, false alarm rate is less than 19% and the missing recognition rate is less than 11%. The object recognition system based on fusion can effectively recognize bridges, airports, plants and other objects, the overall recognition rate reaches more than 94%, false alarm rate is less than 4% and the missing rate is less than 2%. Compared with the object recognition method adopting a single sensor, this system has higher recognition rate and reliability, lower false alarm rate and missing recognition rate. It can be seen from Table 29.1, compared with the other goals, bridges have the lower correct recognition rate. The main reason is that the bridge decks or the vegetation on both sides of the bridge can cause interference, which makes the decks be excluded because of not complete conforming to the attributes of artificial buildings.

29.4 Conclusion

In view of spectral distortion in enhancing spatial resolution for current high-resolution remote sensing image fusion, this chapter presents a multi-sensor data fusion algorithm based on goal-oriented identification. Visual and statistical analyses proved that the algorithm can embody effective information for multi-spectral and panchromatic images in fusion images, retain the specific environment, highlight the targets. It clearly shows that multiple data fusion has improved the recognition rate.

Acknowledgment This chapter is supported by Funding Project for National High Technology Research and Development Program (2009AA0101), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR201007121), the Priority Research Project funded by North China University of Technology.

References

1. Wang Z, Ziou D, Armenakis C (2005) A Comparative analysis of image fusion methods. IEEE Trans Geosci Remote Sens 43(6):1391–1402
2. González M, Saleta J, Catalán R (2004) Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Trans Geosci Remote Sens 23(18):1291–1299
3. Wanga L, Sousab WP, Gong P (2004) Comparison of Ikonos and Quickbird images for mapping mangrove species on the Caribbean coast of Panama. Remote Sens Environ 9:432–440
4. Pajares G, de la Cruz JM (2007) A wavelet-based image fusion tutorial. Pattern Recogn 37:1855–1872

Chapter 30
Free Choosability of Outerplanar Graphs

Nana Li, Xianrui Meng and Hongfang Liu

Abstract A graph $G = G(V, E)$ with lists $L(v)$, associated with its vertices $v \in V$, is called L-list colourable, if there is a proper vertex colouring of G in which the colour assigned to a vertex v is chosen from $L(v)$. We say G is k-choosable if there is at least one L-list colouring for every possible list assignment L with $|L(v)| = k\ \forall v \in V(G)$. Now, let an arbitrary vertex v of G be coloured with an arbitrary colour f for $L(v)$. We investigate whether the colouring of v can be continued to an L-list colouring for the whole graph. G is called free k-choosable if such an L-list colouring exists for every list assignment L ($|L(v)| = k\ \forall v \in V(G)$), every vertex v and every colour $f \in L(v)$. We prove that "Every outer plane graph is free $(2, 2)^*$-choosable".

Keywords Outer planar graph · $(L, d)^*$-colouring · Free $(K, d)^*$-choosable

30.1 Introduction

List colourings of graphs are generalisations of usual colourings that were introduced by Vizing [1] and independently by Erdos [2].

A vertex colouring or just colouring of a finite simple graph G is an assignment of a colour to each vertex of G. A colouring is proper if the adjacent vertices

N. Li (✉) · H. Liu
Tangshan College, Tangshan, People's Republic of China
e-mail: lalazinana@126.com

H. Liu
e-mail: lhf_gir@21cn.com

X. Meng
College of Science, Hebei United University,
Tangshan, People's Republic of China
e-mail: xianruimeng@yahoo.com.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_30,

always get different colours. A graph is K-colourable if it has a proper colouring using at most K different colours. A list assignment L to (the vertices of) G is the assignment of a list (set) $L(v)$ of colours to every vertex v of G; and a k-list assignment is a list assignment in which $|L(v)| = k$ for every vertex v of G is present. If L is a list assignment of G, then an L-colouring of G is a proper colouring in which each vertex receives a colour from its own list. The graph G is k -list colourable or k-choosable if there exists an L-colouring for every k-list assignment L of G. The chromatic number $\chi(G)$ of G is the smallest number k such that G is k-colourable. The list chromatic number, or choice number, or choosability ch(G) of G is the smallest number k such that G is k-choosable. C. "Every outer planar graph is $(2,2)^*$-choosable" has been proved and a example of a non-$(1,2)^*$-choosable outer planar graph was given.

In a vertex-coloured graph, the defect def(v) of a vertex v is the number of vertices adjacent to v that have the same colour as v; so a colouring is proper if and only if every vertex has defect 0. A graph G is $(k,d)^*$-colouring if its vertices can be coloured with k colours in such a way that no vertex has defect greater than d. If L is a list assignment of G, then an $(L,d)^*$-colouring is an L-colouring in which no vertex has defect greater than d, and L is $(L,d)^*$-colourable if it has an $(L,d)^*$-colouring. Finally, G is $(k,d)^*$-choosable if it is $(L,d)^*$-colourable whenever L is a k-list assignment. Obviously, $(k,0)^*$-colourable means the same as (properly) k-colourable, and $(k,0)^*$-choosable means the same as k-choosable.

The free list colouring was first raised by Voigt [3] in 1996. Let L be a list assignment of G, where G is called free $(L,d)^*$-colourable if for every vertex $v \in V(G)$ and for every colour $f \in L(v)$ there exists an $(L,d)^*$-colouring $\varphi_{v,f}$ with $\varphi_{v,f}(v) = f$. And G is called free$(k,d)^*$-choosable, if it is free$(L,d)^*$-colourable whenever L is a k-list assignment. When $d = 0$, we call it free k-choosable. In this paper, we shall prove that "Every outer planar graph is free $(2,2)^*$–choosable". In order to prove our main result, we first introduce several useful lemmas and theorems in the next section.

30.2 Some Lemmas and Theorems

Lemma 2.1 (1) *Every free $(k,d)^*$-choosable graph is $(k,d)^*$-choosable*;

(2) *There are $(k,d)^*$-choosable graphs which are not free $(k,d)^*$-choosable.*

Proof (1)⇒(2) It is trivial.

(2) ⇒(1) Consider the complete bipartite graph $K_{2,2}$ and $L(x_1)=\{1,2\}$, $L(x_2)=\{1,3\}$, $L(y_1)=\{2,3\}$, $L(y_2)=\{1,3\}$, where $\{x_1,x_2\}$ and $\{y_1,y_2\}$ are two partite classes of $K_{2,2}$. If we assign color 3 to x_2, then it cannot be continued to a $(L,0)^*$-list colouring of the whole graph.

We have seen that the property "to be free $(k,d)^*$-choosable" is a stricter requirement for a graph than the property "to be $(k,d)^*$-choosable".

Now assume the graph G is $(k,d)^*$-choosable but not free $(k,d)^*$-choosable, this means there exists a k-list assignment L of G, a vertex $v^* \in V(G)$ and a colour $f \in L(v^*)$ such that $\varphi(v^*) \in L(v^*)\backslash\{f\}$ for all L-list colourings of G.

In the following, such a vertex v^* is called a bad vertex and such a colour f is a bad colour.

Lemma 2.2 *Let G be a graph which is $(k,d)^*$-choosable but not* free $(k,d)^*$-*choosable, v^* a bad vertex of G and $F : =\{f_1,f_2,\ldots,f_{k-1}\}$ an arbitrary set of $k-1$ colours. There exists a k-list assignment $L_{v^*,F}$ of G, so that $\varphi(v^*) \in F$ is satisfied for every $(L_{v^*,F},d)^*$-list colouring φ.*

Proof As G is a graph which is $(k,d)^*$-choosable but not free $(k,d)^*$-choosable, we can find a k-list assignment L to G, such that G is $(L,d)^*$-colourable but not free $(L,d)^*$-colourable. Let v^* be a bad vertex, $L(v^*)=\{g_1,g_2,\ldots,g_{k-1},g_k\}$ and g_k is a bad colour. We use the known list assignment L with $\varphi(v^*) \in L(v^*)\backslash\{g_k\}$ (for all $(L,d)^*$-list colouring φ of G) and rename the colours in a suitable way: denote colour set $T_1 = \cup_{v\in V(G)} L(v)=\{g_1,g_2,\ldots,g_{k-1},g_k,g_{k+1},\ldots,g_n\}$, and let colour set $T_2=\{f_1,f_2,\ldots,f_{k-1},f_k,\ldots,f_n\}$, where $\{f_k,\ldots,f_n\}\cap F=\Phi$. Define an injection $\psi : \psi(g_i)=f_i, i=F1,2,\ldots,n$, and let the resulting list assignment be L^*. Because for every $(L,d)^*$-colouring φ, $\varphi(v^*) \in L(v^*)\backslash\{g_k\}$, it is easy to see that for every $(L^*,d)^*$-colouring φ^*, $\varphi^*(v^*) \ \in L^*(v^*)\backslash\{f_k\}=F=\{f_1,f_2,\ldots,f_{k-1}\}$.

Lemma 2.3 ([4]) *There are outer planar graphs which are not $(1,2)^*$-choosable.*

30.3 Main Result

Theorem 3.1 *The following results are equivalent: Every outer planar graph is $(2,2)^*$-choosable.*

Every planar graph is free $(2,2)^*$-choosable.

Proof (2) ⇒(1): Trivial.

(1) ⇒(2): Assume a planar graph G' is $(2,2)^*$-choosable, but $G\prime$ is not free $(2,2)^*$-choosable. In the following, using $G\prime$ we will construct a outer planar graph G^*which is not $(2,2)^*$-choosable. This contradicts to (1).

Let $G^{(1,2)}$ be a outer planar graph which is not $(1,2)^*$-choosable with an m-element vertex set $V\left(G^{(1,2)}\right)=\{v_1,v_2,\ldots,v_m\}$ for some m by Lemma 2.3. Let $L^{(1,2)}$ be a 1-list assignment of $G^{(1,2)}$ and $G^{(1,2)}$ is not $\left(L^{(1,2)},1\right)^*$-list colourable.

Choose a bad vertex v^* of G' and let G' be embedded in the plane in such a way that v^*belongs to the boundary of the exterior face.

Take m copies $\left\{G'_1,G'_2,\ldots,G'_m\right\}$ of this graph G' with the bad vertices $v^*_1,v^*_2,\ldots,v^*_m$ respectively.

Define $G^*(V^*, E^*)$, where $V^* \ : = \cup_{\mathrm{i=1}}^{\mathrm{m}} V\left(G_i'\right)$, $E^* \ : = \cup_{\mathrm{i=1}}^{\mathrm{m}} E\left(G_i'\right) \cup \left\{ \left(v_i^*, v_j^*\right) \right.$ $\left| \left(v_i, v_j\right) \in E\left(G^{(1,2)}\right) \right\}$

Construct a 2-list assignment L^* of G^* in the following way: for $v_j^*(j = 1, 2, \ldots, m)$, take $L^*\left(v_j^*\right) = L^{(1,2)}\left(v_j\right) \cup \{f_k\}$, where f_k is bad colour of bad vertex v^* in G'. And for each one of the other vertices in G^*, take the same colour list as it is in G'.

Since G^* is outer planar, then G^* is $(2, 2)^*$-choosable by (1). Hence there exists an $(L^*, 1)^*$-list colouring φ^* of G^* with $\varphi^*\left(v_j^*\right) \in L^{(1,2)}\left(v_j\right)$ for all $j = 1, 2, \ldots, m$, by Lemma 2.2.

Therefore $G^{(1,2)})$ is $\left(L^{(1,2)}, 2\right)^*$-list colourable, which contradicts to the assumption.

Because "Every outer plane graph is $(2, 2)^*$-choosable" has been proved, so "Every outer plane graph is free $(2, 2)^*$-choosable".

References

1. Vizing VG (1976) Coloring the vertices of a graph in prescribed colors (in Russian). Diskret Anal 29:3–10
2. Erdos P, Rubin AL, Taylor H (1979) Choosability in graphs. Congr Numer 26:125–157
3. Voigt M (1996) Choosability of planar graphs. Discrete Math 150:457–460
4. Woodall DR (2002) Defective choosability results for outerplanar and related graphs. Discrete Math 258:215–223

Chapter 31
Exploring Relation Algebra Division Based on Images Set

Qi Zhong and Wenxiao He

Abstract Introducing theory of images set, modifying two contrary definition of relation algebra in the literatures, a general definition of relation algebra division is proposed and described formally. A general algebra division algorithm based on imaging set theory and a general template of realizing division with select sentence in SQL are presented. Finally, divisions in factual occasions are validated with SQL Sever 2005.

Keywords Relations · Division · Images set · Attribute

31.1 Introduction

In the course of database principle or applications, relational algebra's operations appear frequently. These operations include union, intersection, difference, and division and so on. The basic operations of relation database (union, intersection and difference) are formally described and realized in SQL in Refs. [1] and [2]. In SQL, there is a special operator union, intersection can be realized by where-clause, difference can be realized by existential quantifier (Exists) indirectly. Incomplete Formal description was provided in Ref. [2]. It is not enough to accomplish query that there are division theories without relevant realizable statement. Examples are querying student nos. who take as all elective courses, querying student nos. who take as

Q. Zhong (✉) · W. He
Computer Science College, Neijiang Normal University,
Neijiang 641112, Sichuan, People's Republic of China
e-mail: hdacong@163.com

W. He
e-mail: xiaohe@njtc.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_31,

two elective courses (c1 and c2) and querying student nos. who take as more elective courses than student whose no. is ‘95002’. Simple SELECT-statements cannot realize these queries. The algebra query expressions of these queries reflect that relational algebra division operation is the essence of queries. Definition of relational algebra division is presented in Refs. [2] and [3]. But the definition is special, called illiberal definition. This Chapter extends division operation’s condition and prompts a formal general definition of relational algebra division basing on the definition in Refs. [2] and [3]. This Chapter provides a general algorithm, a general template of SELECT-statement realized division in SQL and practical applications.

31.2 Introduction of Symbols

Symbols are introduced according the formal definition of division operation in Ref. [2].

R, t ∈ *R, t* [*Ai*]. R (A1, A2,…, An) is a relational schema, R named relation. t∈R means t is a tuple of relation R. t[Ai] is a component of attribute Ai in tuple t.

A, t[*A*], *D*[*Yi*]. If A = {Ai1, Ai2,…, Aik}(Ai1, Ai2,…, Aik are part of A1, A2,…, An), then A is called attribute or domain. t[A] = (t[Ai1], t[Ai2],…, t[Aik]) means a component collection of tuple t’s attribute A. D[Yi] means the range of Yi’s value.

Images set Zx. Given a relation R (X, Z), X and Z are the attribute group. When t[X] = x, the images set in R of x is Zx (Zx = {t[Z]|t ∈ R,t[X] = x}). x is a value of attribute set X, Zx is the opposite value set of the Z component.

31.3 Definition of Division Operation

Definition 1 [3] The degree of relation R is (m + n). The degree of relation S is n. R divided by S. The result of this division is a relation which degree is m. There are two preconditions for the division. First, some attributes of R are attributes of S. Second, some attributes of R are not attributes of S, shown in Fig. 31.1.

Let T = R ÷ S, then T is a relation too. The attributes of T are some attributes of R, furthermore these attributes are not attributes of S. The tuples of T consist of the value of attributes opposite tuples of S.

There is formal description of previous depiction.

The degree of relation R(X) is (m + n). The degree of relation S(Z) is n. The sufficient condition of the division between R and S is described below:

$$Z \subseteq X \text{ and } Ai \in R \wedge Ai \in S.$$

Definition 2 [2] Given relations R (X, Y) and S (Y, Z), where X, Y, Z are the attributes groups. The attribute Ys can have different attribute names in R and S individually, but the domain sets must be same. The result of division between R

R

1	2	3	4
a1	b1	c1	d1
a1	b1	c2	d2
a1	b1	c3	d3
a2	b2	c1	d1
a2	b2	c2	d2
a3	b3	c1	d1

S

1	2
c1	d1
c2	d2

T

1	2
a1	b1
a2	b2

Fig. 31.1 Example of division operation

and S is a new relation P(X). The tuples of P include projection of attribute X in relation R, furthermore the images set Yx (Yx is the images set of component x, x is the component value of X) consists of the projection of attribute Y in relation S.

$R \div S = \{tr\ [X] \mid tr \in R \wedge \pi Y(S) \subseteq Yx\}$, Yx is the images set of component x, x = tr [X] [4].

According to the above two definitions, there is an obvious contradiction. The first division definition must meet Z $\subseteq$ X, where Z is all attributes of S, X is all attributes of R. The second division definitions just to meet the R $\cap$ S $\neq \varphi$ [5]. The two contrary definitions have some connections in essence. Weakening conditions of the first definition, it can be a special case of the second definition. Then, the second definition is a generalized definition of division. After weakening conditions, the first definition became the third.

Definition 3 The degree of relation R(X) is (m + n). The degree of relation S(Z) is n. The sufficient condition of R$\div$S is Z $\subseteq$ X and Ai $\in$ R$\wedge$Ai $\notin$ S.

According to Definition 3 and 2, a more general division definition of relational algebra can be got as follows.

Definition 4 [6] Given relations R (X, Y) and S (Y, Z), where X, Y, Z are the attributes groups. X = {X1, X2,…, Xn}, Y = {Y1, Y2,…, Yn}, Yi $\in$ R, Yj $\in$ S, D[Yi] = D[Yj] (the domain of Yi and Yj must be same). Yi and Yj can be named differently. Attributes group Z can consists of zero or several attributes.

(R$\div$S) Yi$\wedge$…. $\wedge$Yj represents R divided by S with one or more same attributes.

(RYi$\div$SYj) represents R divided by S with one or more different attributes.

All same attributes in two relations are operated in division of relation. Maybe there is one same attribute or there are several same attributes.

31.4 Realization of Division

According to the definition of images set and the forth definition of division, the algorithm realization of division and query statement in SQL Server are given latter.

Algorithm realization [7]. Given relations R (X, Y) and S (Y, Z), where X, Y, Z are the attributes groups. X = {X1, X2,…, Xn}, Y = {Y1, Y2,…, Yn}, Attributes group Z can consists of zero or several attributes. The description of division's algorithm is given below.

Step1. Get the component value t[Xi] of X in relation R.
Step2. Get ZXi (the images set of Xi).
Step3. Get projection set on attribute Y in relation S, πY(S).
Step4. Get the set of Xi in ZXi which composite to components of πY(S), the set is the result of R ÷ S.

SELECT-query implementation. Union operation of relational algebra can be realized by SQL statement directly. Intersection, difference and division cannot be realized by SQL statement directly. Intersection and difference can be realized by SQL statement indirectly easily. The implementation of division is more difficult. Division can be implemented by sub-query.

The SELECT-statement of upper algorithm as follows.

```
SELECT * FROM R as r1
WHERE NOT EXISTS
(SELECT * FROM S
WHERE NOT EXISTS
(SELECT * FROM R as r2
WHERE r1.X1 = r2.X1 […and r1.Xn = r2.Xn] and S.Y1 = r2.Y1 […S.Yn = r2.Yn]))
```

Example is given as follows.

SELECT-statements of R's dividing by S with attributes B. The result of (R ÷ S)B is shown in Fig. 31.2d.

```
select distinct A from R as x
where not exists
(select * from S
where not exists
(select * from R as y
where y.a = x.a and s.B = y.B))
```

SELECT-statements of R's dividing by S with attributes B and C. The result of (R ÷ S) B ∧ C is shown in Fig. 31.2e.

```
select distinct A from R as x
where not exists
(select * from S
where not exists
(select * from R as y
where y.a = x.a and s.B = y.B and s.C = y.C))
```

SELECT-statements of T with attribute E dividing by S with attribute B. The result of (RE ÷ SB) is shown in Fig. 31.2d.

	A	B	C
1	a1	b1	c2
2	a1	b1	c3
3	a1	b2	c2
4	a2	b1	c2
5	a2	b1	c3
6	a2	b2	c3
7	a3	b2	c3
8	a3	b1	c3

(a)

	A	E	F
1	a1	b1	c2
2	a1	b1	c3
3	a1	b2	c2
4	a2	b1	c2
5	a2	b1	c3
6	a2	b2	c3
7	a3	b2	c3
8	a3	b1	c3

(b)

	D	B	C
1	d1	b1	c2
2	d1	b1	c3
3	d2	b1	c2

(c)

	A
1	a1
2	a2
3	a3

(d)

	A
1	a1
2	a2

(e)

Fig. 31.2 Example of relational division. **a** Relation R. **b** Relation T. **c** Relation S. **d** $(R \div S)_R$ **e** $(R \div S)_{B \wedge C}$

```
select distinct A from T as x
where not exists
(select * from S
where not exists
(select * from T as y
where y.a = x.a and s.B = y.E))
```

31.5 Conclusions

This Chapter studies some definitions of division, provides a formal definition of division which can be general applied in theory and practice and presents a general algorithm in theory and a general template implemented by SELECT-statement.

The algorithm can realize the narrow division operation and the general division operation. The general template can be applied in typical division query without comprehension on the formal definition. Last, some examples are presented. All examples in this Chapter can be implemented in SQL Server 2000 or SQL Server 2005.

References

1. Codd EF (1970) A relational model of data for large shared data banks. CACM 13(6):377–387
2. Wang S, Sa S (2007) Introduction to database system. Higher Education Press, Beijing, pp 243–247
3. Li H (2007) Database principles and applications. Higher Education Press, Beijing
4. Ullman J (1982) Principles of database systems, 2nd edn. Computer Science Press, New York
5. Hoffer JA, Prescott M, McFadden F (2006) Modern database management, 8th edn. Prentice Hall, New Jersey
6. Zhang Z (2003) How to express division operation with fundamental operations in relational algebra, Shaanxi. J Shaanxi Norm Univ (Natural Science Edition) 31:1–3
7. Lu Z (2005) The realization of 'division' in relation algebra by SQL query, Hebei. J Hebei Inst Technol 27(3):81–84

Chapter 32
Image Interpolation via Graph Cut

Bo Li, Dianxuan Gong and Qifeng Zhang

Abstract This chapter introduces a novel variational method for image interpolation via graph cut. Image interpolation can be seen as a special case of image inpainting when the inpainting mask is chosen as the pixels which will be interpolated. In this paper, an energy variational function is proposed, and the optimization problem is regarded as a labeling problem via solving a minimum cut of a certain graph. Experimental results show that the algorithm can save the computer time and improve the staircase effect occurred in some classical interpolation methods.

Keywords Image interpolation · Graph cut · Total variation

32.1 Introduction

A generic image interpolation takes a picture as input and provides a picture of greater size preserving as much as possible the information content of the original image as ouput. It has played a very important role in many digital image processing operations, such as translation, scaling, rotation and geometric correction. In the recent days, this old topic gains more and more attentions with the rapid development in internet videos, mobile photos, high definition TV, etc.

B. Li (✉) · Q. Zhang
College of Mathematics and Information Science, Nanchang Hangkong University, Nanchang, China
e-mail: bolimath@gmail.com

D. Gong
College of Sciences, Hebei United Universtiy, Tangshan 063009, China
e-mail: dxgong@heut.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_32,

A large class of image interpolation techniques are achieved by means of some simple mathematical interpolation algorithms, such as pixel replication [1], bilinear [2, 3], bicubic [4, 5] or spline interpolation [6, 7]. Unfortunately, these methods, while preserving the low frequencies content of the source image, are not equally able to enhance high frequencies, suffering from unacceptable effects (e.g. blurring, blockiness), especially at edge areas. Adaptive interpolation algorithms were developed to yield better results, usually the local edge orientation was firstly found via edge-map, and then the interpolation was performed along that direction [8, 9], these adaptive and edge-oriented algorithms [10] are classified as the fourth category, and they usually depend on the gradient features [11] or statistical information of images [12].

Image interpolation can also be seen as a special case of image inpainting. Therefore, some classical inpainting methods can be applied to image interpolation. The growing impact of variational techniques in image processing is mainly due to their capability and flexibility in controlling geometrical features of images. For example, total variation minimization, which leads to a curvature term, can retain sharp edges in image processing. In this chapter, we select total variation minimization to facilitate the zooming process so that the unknown pixels after magnification can be filled into faithfully preserve geometric image feature. Some classical variation minimization methods, such as steepest descent method, will produce the staircase effect and only obtain the local optimal solution. So we present total variation image zooming algorithm based on graph cut. In this algorithm, the minimum of the total variation image zooming energy function was transformed to a minimum cut of a graph. Then, some maximum flow/minimum cut algorithms could solve this problem, and get the global minimum of the total variation function.

This chapter is organized as follows. In Sect. 32.2, we give total variation zooming model. Sect. 32.3 presents graph cut algorithm for solving minimum energy function.

32.2 Graph Cut-Based Image Interpolation Algorithm

Many of the problems that arise in early vision can be naturally expressed in terms of energy minimization. In the last few years, a new approach has been developed based on graph cuts. The basic technique is to construct a specialized graph for the energy function to be minimized such that the minimum cut on the graph also minimizes the energy. The minimum cut can compute very efficiently by max flow algorithms. In this section, we consider graph cut algorithm, solving total variation zooming energy function (32.1).

$$E \approx \lambda \sum_{(x,y)\in\Omega} \left[\left| u_{x,y} - u_{x+1,y} \right| + \left| u_{x,y} - u_{x,y+1} \right| \right] + \sum_{(x,y)\in S} \left(u_{x,y} - u^{0}_{x,y} \right)^{2} \qquad (32.1)$$

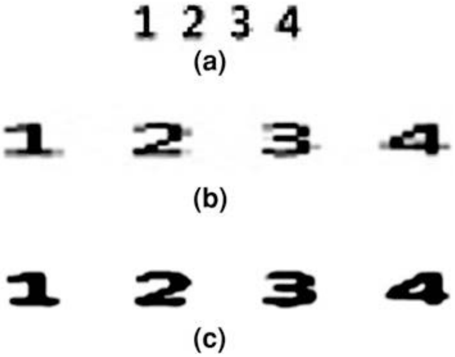


Fig. 32.1 The 15 times interpolation results of Arabic numbers **a** Original image **b** Bicubic interpolation **c** Proposed method

Firstly, image zooming problem can be regarded as the pixel-labeling problem. Every pixel $p \in \mathrm{P}$ must be assigned a label in some finite set L which can be luminance set. The goal is to find a labeling f that assigns each pixel $p \in \mathrm{P}$ a label $f_p \in L$, where f is both piecewise smooth and consistent with the observed data and minimizes the energy function. We rewrite the above model (32.1) as:

$$E = \lambda \sum_{(p,q)\in N} \left| l_{px,py} - l_{qx,qy} \right| + \sum_{p\in P} \left(l_{px,py} - u^0_{px,py} \right)^2 \tag{32.2}$$

where u_p^0 presents luminance value of initial image in the pixel p, l_p presents luminance value of zooming image in the pixel p, N is a set of all pairs of neighboring pixels. N is defined as:

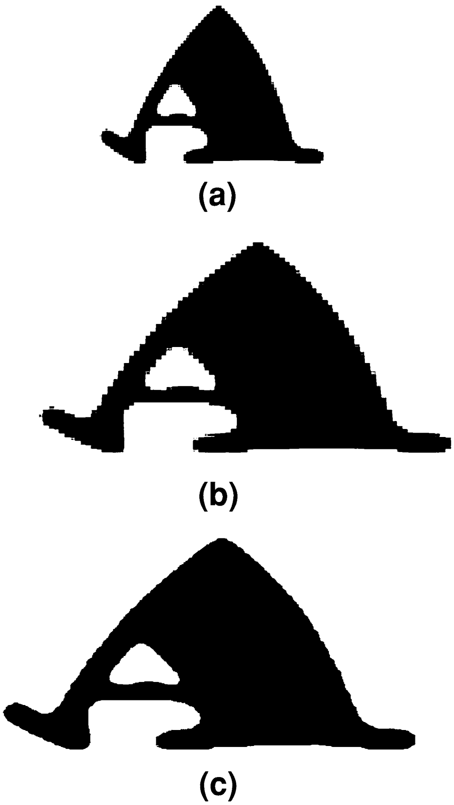
$$N = \left\{ (p,q)/p_x = q_x - 1, p_y = q_y, \mathrm{or}, p_x = q_x, p_y = q_y - 1 \right\} \tag{32.3}$$

where p_x *and* p_y respectively denote horizontal coordinate and vertical coordinate of the pixel p.

Secondly, suppose $G = (V, \mathrm{E})$ is a directed graph with non-negative edge weights that have two special vertices (terminals), namely, the source s and the sink t. An s-t-cut C = S. T is a partition of the vertices in V into two disjoint set S and T such that $s \in S$ and $t \in T$. The cost of the cut is the sum of costs of all edges that go from S to T, being defined as $C(S,T) = \sum_{u\in S, v\in T, (u,v)\in E} C(u,v)$. The minimum s-t-cut problem is to find a cut C with the smallest cost. Due to the theorem of Ford and Fulkerson [12], this is equivalent to computing the maximum flow from the source to sink.

Thirdly, minimizing an energy function via graph cut remains a technically difficult problem. Each paper constructs its own graph specifically for its individual energy function and, in some of these cases, the construction is fairly complex. Boykov presents fast approximate energy minimization via graph cut, namely, movable space algorithm [13]. In our chapter, we apply this algorithm to get minimum cut.

Fig. 32.2 The 5 times interpolation results of English letters **a** Original image **b** Bicubic interpolation **c** Proposed method



Finally, total variation image zooming algorithm based on graph cut is described as follow:

```
Discrete total variation model, obtaining (32.1)
According to local feature, set regularization parameter $\lambda$
Map energy function minimization problem to movable space labeling problem, and solve it.
Initialize labeling $f$
Set success = 0
For each pair of labels $\{\alpha, \beta\} \subset L$
Find $\hat{f} = \arg\min E(f')$ among $f'$ within one $\alpha - \beta$ swap of $f$
if $E(\hat{f}) < E(f)$, set $f = \hat{f}$ and success = 1
If success = 1 goto 2
Return $f$
```

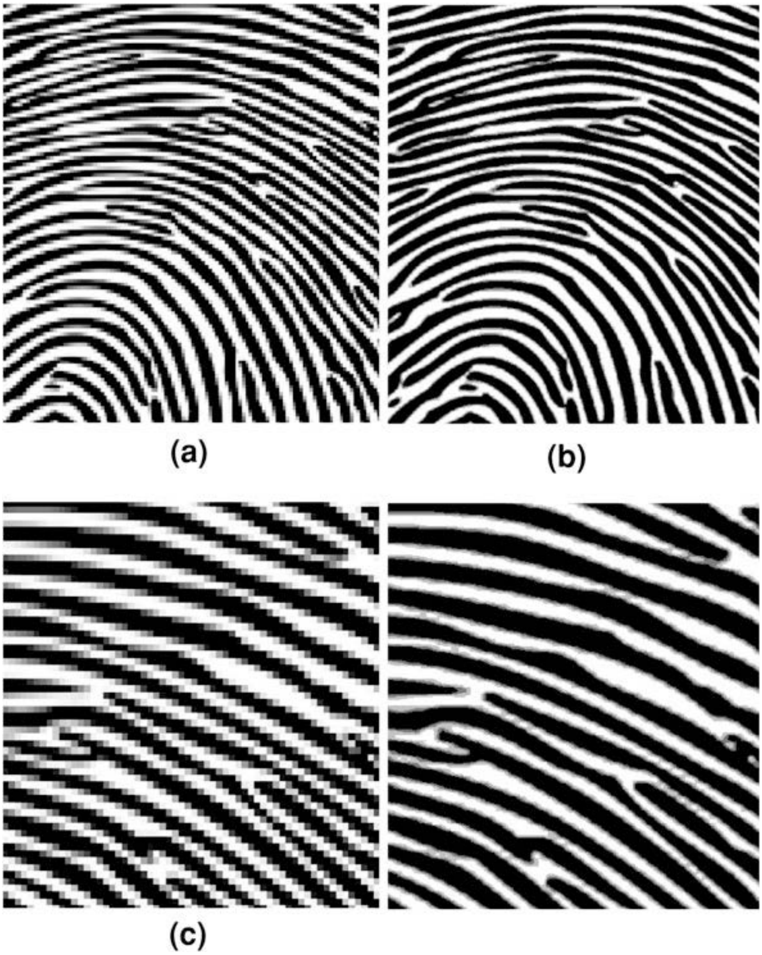


Fig. 32.3 The 10 times interpolation results of fingerprint **a** Bicubic **b** Proposed method **c** Local magnified of two methods

32.3 Numerical Implementation

This algorithm is performed by Matlab 2010a on a notebook computer with Intel Pentium IV CPU 2.0G. In our experiments, we adopt Arabic numbers, English letters and fingerprint images for testing. The results are shown in Figs. 32.1, 32.2 and 32.3, respectively.

From the results of the experiments, we can see that the proposed method can avoid the staircase effectively with 10 or larger times interpolation.

Acknowledgments This work is supported by the Project Sponsored by the Scientific Research Foundation (Nanchang Hangkong University).

References

1. Gonzalez RC, Woods RE (1992) Digital image processing. Addision-Wesley, Reading
2. Maeland E (1988) On the comparison of interpolation methods. IEEE Trans Med Imag 7(3):213–217
3. Parker JA, Kenyon RV, Troxel DE (1983) Comparison of interpolating methods for image resampling. IEEE Trans Med Imag MI-2(1):31–39
4. Keys R (1978) Cubic convolution interpolation for digital image processing. IEEE Trans Acoust Speech, Signal Process ASSP-26(6):508–517
5. Hou HS, Hou HS, Andrews HC (1978) Cubic splines for image interpolation and digital filtering. IEEE Trans Acoust Speech, Signal Process ASSP-26(6):508–517
6. Parker JA, Kenyon RV, Troxel DE (1983) Comparison of interpolation methods for image resampling. IEEE Trans Med Imag 2(1):31–39
7. Dodgson NA (1997) Quadratic interpolation for image resampling. IEEE Trans Image Process 6(9):1322–1326
8. Allebach J, Wong PW (1996) Edge-directed interpolation proceedings of ICIP-96. IEEE Press, Lausanne CH, Vol.III, pp.707–710
9. Biancard, Lombardi L, Pacaccio V (1997) Improvements to image magnification. In: Proceedings of ICIAP 97 Vol 2, pp 141–149
10. Rudin L, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Physica D 60:259–268
11. Chan TT, Shen J (2001) Mathematical models for local non-texture inpaintings. SIMA J Appl Math 62(3):1019–1043
12. Ford L, Fulkerson D (1962) Flows in networks. Princeton University Press, New Jersey
13. Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cut. IEEE Trans Pattern Analysis Mach Intell 23(1):1222–1239

Chapter 33
Image Inpainting Under Local Coordinate System

Bo Li, Xiuping Liu, Dianxuan Gong and Qifeng Zhang

Abstract This chapter focus on the regular texture inpainting problem under local coordinate system. General variational image inpainting models perform well for cartoon images, but poor for textures. In this paper, a novel local inpainting model is proposed by combining the total variation and OABE algorithm. Firstly, the local direction of texture is obtained according to the neighborhood of damaged region, the local coordinate is set up via the local texture direction and its normal direction; then the local variational inpainting model is proposed in this coordinate. We give the discrete Gauss–Seidel algorithm for this model and numerical experiments. The results show that our algorithm perform well for the regular texture images, even for textures like "Y".

Keywords Inpainting · Nonlocal total variation (NLTV) · Discrete cosine transform

B. Li (✉) · Q. Zhang
College of Mathematics and Information Science, Nanchang Hangkong University,
Nanchang, China
e-mail: bolimath@gmail.com

X. Liu
School of Mathematical Science, Dalian University of Technology, Dalian, China
e-mail: xpliu@comgi.com

D. Gong
College of Sciences, Hebei United Universtiy, Tangshan 063009, China
e-mail: dxgong@heut.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_33,

33.1 Introduction

Image inpainting is to fill in some lost or damaged region with available information from their surroundings in a certain rule, so that the restored image approach the original image. As early as the renaissance in Europe, artist have begun to manually restore medieval artwork. With the development of digital image processing technology, digital inpainting technology has been widely applied in many areas, for example damaged photographs, video restoration, image characters or objects removal (such as publication date, microphone), texture filling,and so on. Compared with traditional manual methods, with its fast, effective, automated, without destroying the original, the digital inpainting approach has aroused the concern of many scholars. By the end of twelfth century the concept being proposed, many successful approaches have been proposed. Partial differential equations (PDE) methods and texture synthesis methods are the presently important methods in image inpainting technology.

According to the inpainting mechanism, the methods based on PDE can be divided into two types: microscopic and macroscopic. The methods which are based on diffusion, interpolation and isophotes, belong to microscopic inpainting mechanism, for example, BSCB [1], CDD [2] (Curvature-driven diffusion), horizontal interpolation, and so on. BSCB algorithm was put forward by Bertalmio to be applied to restore damaged photographs. It fills in the damaged region by transmitting the information outside the region along the isophotes direction to the inside region by anisotropic diffusion. The fill-in is done in such a way that isophotes lines arriving at the regions' boundaries when the algorithm is converged. The methods based on variational system, such as TV method [2], flexible repair method [3], Mumford-Shah method [4], as well as the Mumford-Shah-Euler method [4], belong to macroscopic inpainting approach. Inspired by the well performance of total variation methods for image denoising, Chan [2] proposed to use TV method to solve the problem of image inpainting, One of the advantage of this method is that it can serve the discontinuity preserving while repair the image. In order to solve the shortcoming of TV model, such as parameter sensitivity and large amount of computing, Shao [8] proposed an improved version. This method can effectively improve the robustness and increase the computing speed.

Experiments show that the above methods perform well for cartoon images, which is rich in geometrical information. But they cannot solve the inpainting problem for texture images, which is rich in details. Texture image is an important characteristic but it is difficult to describe. Customarily texture is the characteristic of local irregularity and global regularity. Zh [6] proposed an inpainting method based on Markov random field. It has a good effect on the random texture, but the speed is very slow. Criminis proposed a texture synthesis method based on isophotes priority [7]. The algorithm first computes the priorities on the border region, then selects the greatest priority of texture element. It performs well for regular texture, but the inpainting structure information is limited. Yan Niu and TimPoston proposed OABE method [9] which could be applied to inpainting mosaic texture missing after decode. The idea is exploring the known information of the texture direction to construct the elliptic equations for restoratio. But the approach only deals with a square area and is invalid for the "Y"-shaped domain missing.

Images to be inpainted usually have both structural information and texture information. Only using structure or texture inpainting method, the results are not satisfactory. The basic idea is to first decompose the image into the sum of two components with different basic characteristics, and then repair each one of these components separately with structure and texture filling-in algorithms.

By the above analysis, the total variation methods perform poor for texture images, while the OABE algorithm can deal with some types of textures. Inspired by the idea of OABE, in this paper we propose a novel variational texture image inpainting methods under local coordinate system. The main contribution of this algorithm is that it can deal with the texture image via the local coordinate system similar to the OABE, and it also serves the discontinuity preserving by total variation.

33.2 The Proposed Local Variational Inpainting Model

One of the key step of this algorithm is to get the local texture orientation. Because the repaired regions do not have available information, so we will get the approximate local texture orientation from the surrounding information.

We assume that $V_\eta = (m, n), m, n \in Z$ is the texture direction, $V_\zeta = V_\eta^\perp = (-n, m)$, and (V_η, V_ζ) constitute an orthogonal coordinate system, known as the local texture coordinate system, as shown in Fig. 33.1. $\Lambda = \{U, D, L, R\}$ is four neighbor points of the point $O(i,j)$ concerning the vector V_η.

Point LU, LD, RU and RD are respectively the upper left, lower left, upper right and lower right neighbor point of V_η. If V_η is greatly large, for preventing the neighbor points of O into another repaired areas, we should approximate V_η. For example, if $V_\eta = (-10, 16)$, V_η can be approximated as (−2, 3). Experiments have show that the approximation has not effect on the inpainting results.

We modify the TV model to be a structured texture inpainting model, as follows:

$$-\tilde{\nabla}\cdot\left(\frac{\underset{u}{\tilde{\nabla}}}{\left|\underset{u}{\tilde{\nabla}}\right|}\right)+\lambda_e\left(u-u^0\right)=0,$$

$$\lambda_e=\begin{cases}\lambda, & (x,y)\in E\\ 0, & (x,y)\in D\end{cases}$$

$\tilde{\nabla} u = (\frac{\partial_u}{\partial_\eta}, \frac{\partial_u}{\partial_\varsigma})$ is the gradient at point (V_η, V_ς), corresponding to $\nabla u = (\frac{\partial_u}{\partial_x}, \frac{\partial_u}{\partial_y})$, $\tilde{\nabla}\cdot = \frac{\partial}{\partial_\eta} + \frac{\partial}{\partial_\varsigma}$ is the divergence at point (V_η, V_ς), corresponding to $\nabla\cdot = \frac{\partial}{\partial_x} + \frac{\partial}{\partial_y}$.

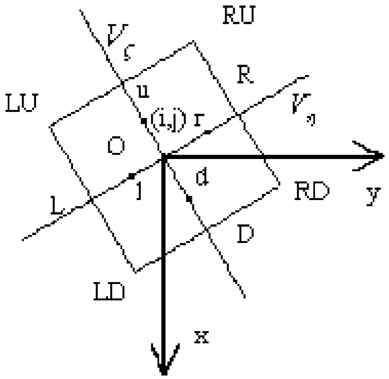


Fig. 33.1 Sketch map of local texture coordinate

In the local texture coordinate system, the Euler–Lagrange equation is equivalent to search minimization of energy function in the form of

$$\tilde{J}_\lambda[u] = \int_{E\cup D} \left|\tilde{\nabla} u\right| d\eta d\varsigma + \frac{\lambda}{2}\int_{E} \left|u - u^0\right|^2 d\eta d\varsigma \tag{33.1}$$

Steepest descent equation is

$$\frac{\partial_u}{\partial_t} = \tilde{\nabla}\cdot\left(\frac{\tilde{\nabla} u}{\left|\tilde{\nabla} u\right|}\right) + \lambda_e\left(u^0 - u\right). \tag{33.2}$$

Minimization of energy $J_\lambda[u]$ is equivalent to minimization of energy $\tilde{J}_\lambda[u]$. Write$\left|\tilde{\nabla} u_\alpha\right| = \sqrt{\left|\tilde{\nabla} u\right|^2 + \alpha^2}$. Similarly,

$$-\tilde{\nabla}\cdot\left(\frac{\tilde{\nabla} u}{\left|\tilde{\nabla} u_\alpha\right|}\right) + \lambda_e\left(u - u^0\right) = 0 \tag{33.3}$$

33.3 The Implementation of Proposed Algorithm

In actual calculation, selection of parameter α has influence on the results, in the term of literature [5]. In the early iterations, we select a larger α. With the process of iterations, we can shrink the value α. Summing up the above, this algorithm for image restoration can be described as follows:

Read into the repaired image and mask image which is used for marking the repaired region.

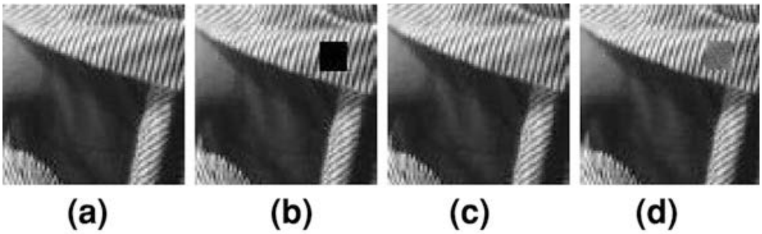


Fig. 33.2 Inpainted result of Barbara cutted image. **a** Original image. **b** Damaged image. **c** Inpainted result of proposed method. **d** Inpainted result of TV

For a fixed restoration area, the OABE algorithm can be used for the calculation of local texture direction.

For assigning a initial value to the parameter, we adopt formula (3) to update the damaged image.

If the number of iteration achieves the largest or the image difference between a certain step and its previous iteration is smaller than a given threshold, out of the circulation. Then, end the inpainting. Otherwise, switch to (3) into next iteration.

33.4 Experimental Results and Evaluation

33.4.1 Experimental Results

In this paper, we consider the standard Barbara cutting image of which repaired region has not noise pollution. TV algorithm can not inpaint texture, so that the inpainting results are poor. In this paper, the algorithm of structured texture inpainting is satisfactory, see Fig. 33.2, 33.3, 33.4, and the PSNR evaluation results are shown in Table 33.1. Besides, it can inpaint "Y" and even "X"area which are not restored by literature [5]. Our approach can also directly restore images which have simultaneously structure and texture, without decomposing.

33.5 Conclusion

In this chapter, we proposed a new image inpainting approach which can inpaint structure and have a good result for inpainting structured texture.In the process of experiments,the search for local texture direction spends a lot of time. So looking for a more effective approach for texture direction is one of our works. In addition, in order to inpaint large curvature texture and large scale texture missing, the adaptive change of texture direction is also focused on.

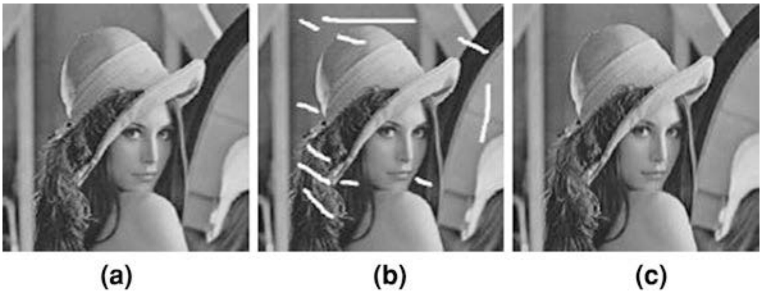


Fig. 33.3 Inpainted result of dirty lena. **a** Original image. **b** Damaged image. **c** Inpainted result of proposed method

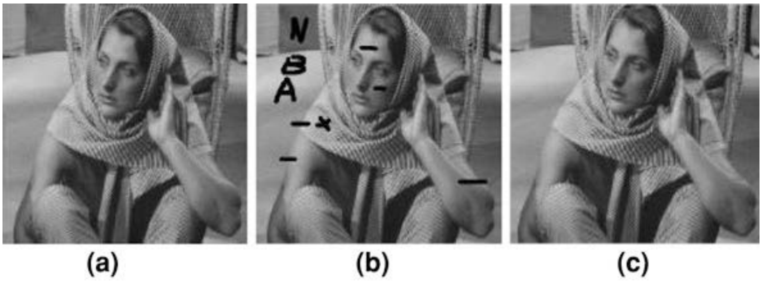


Fig. 33.4 Inpainted result of dirty Barbra. **a** Original image. **b** Damaged image. **c** Inpainted result of proposed method

Table 33.1 PSNR evaluation on the inpainting algorithm

Picture	Damage rate (%)	Number of iterations	Total repair time (s)	Repair before the PSNR	After repair PSNR
Fig. 2	3.51	150	32.8280	40.4802	43.5822
Fig. 3	3.51	150	32.1720	37.617	57.2649
Fig. 4	3.43	150	32.1090	40.28	59.1773

Acknowledgments This work is supported by the Project Sponsored by the Scientific Research Foundation (Nanchang Hangkong University), National Natural Science Foundation of China (No. 60873181).

References

1. Bertalmio M, Sapiro G, Caselles V, Ballester C (2000) Imageinpainting. Computer Graphics (SIGGRAPH July 2000)
2. Chan TF, Shen J (2001) Mathematical models forlocal non¡texture inpaintings. SIAM J Appl Math 62(3):1019–1043

3. Chan TF, Shen J (2002) Euler's elastica and curvature based in paint. SIAM J Appl Math 63(2):564–592
4. Criminisi A, Perez P, Toyama K (2003) Object removal by exemplar-based inpainting. IEEE Proc CVPR 2:721–728
5. Esedoglu S, Shen J (2002) Digital inpainting based on the Mumford-Shah-Euler image model. Eur J Appl Math 13:353–370
6. Bertalmio M, Vese L, Sapiro G, Osher S (2003) Simultaneous structure and texture image inpainting. IEEE Trans Image Process 12(8):882–889
7. Niu Y, Poston T (2005) Using an oriented PDE to repair image textures. VLSM, LNCS 3752, pp 61–72
8. Shao X, Liu Z, Song B (2004) An adaptive image inpainting approach based on TV model. J Circuits Syst 9:113–117
9. Yasuda M, Ohkubo J, Tanaka K (2005) Digital image inprinting based on Markov random field. IEEE CIMCA-IAWTIC, pp 245–249

Chapter 34
A SIFT-Based Approach for Image Registration

Lintao Zheng and Guiping Qian

Abstract Over the past several decades, image registration has emerged as one of the key technologies in medical image computing with applications ranging from computer assisted diagnosis to computer aided therapy and surgery. In this paper, we present a new method for medical image registration, which is based on the Scale-invariant feature transform (SIFT) and TPS. Our experimental results show that the proposed method could achieve greater competitive performance than TPS-based image registration technique.

Keywords SIFT · MLS · Non rigid registration · Medical image

34.1 Introduction

Medical images are increasingly widely used in health care and biomedical research; Medical imaging technologies are altering the nature of many medical processions today. In medical applications, images of similar or differing modalities often need to be aligned as a preprocessing step for many planning, navigation, data-fusion, and visualization tasks. Image registration refers to the process

L. Zheng (✉) · G. Qian
Department of Computer Science and Engineering,
Zhejiang University, Hangzhou, China
e-mail: zhenglintao@zju.edu.cn

G. Qian
e-mail: qianguiping@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_34,

of overlaying two or more images of the same scene taken at different times, under different lighting conditions, from different viewpoints, and/or by different sensors. Image registration is a very common problem in medical image processing. Registration of medical images has been an active topic of research for over the past three decades.

Medical image registration has a wide range of potential applications. These include [1]:

Combining information from multiple imaging modalities, for example, when relating functional information from nuclear medicine images to anatomy delineated in high-resolution MR images.

Monitoring changes in size, shape, or image intensity over time intervals that might range from a few seconds in dynamic perfusion studies to several months or even years in the study of neuronal loss in dementia.

Relating preoperative images and surgical plans to the physical reality of the patient in the operating room during image-guided surgery or in the treatment suite during radiotherapy.

Relating an individual's anatomy to a standardized atlas.

Several good comprehensive surveys of medical image registration methods are extensively reported in the literature [1–4]. In general, the medical image registration methods can be divided into two main categories: feature-based techniques and intensity-based techniques. Feature-based techniques require some preprocessing, prior to registration, to extract relevant information, such as anatomical landmarks, edges, or shapes. In feature-based methods, registration involves the determination of the coordinates of corresponding features in different images such as landmark points, ridges, or surfaces, and the estimation of a geometrical transformation using these corresponding features [8–10]. In contrast to feature-based techniques, intensity-based measures get by without prior preprocessing. Thus images can be registered right after image acquisition. Intensity-based measures use the full raw image information for image alignment. Here, we adopt the former approach.

This paper focuses on a new method using the Moving Least Squares (MLS) transformation and Scale-invariant feature transform (SIFT) feature in medical image registrations. To register two images, SIFT features are selected from the images and correspondence is established between them. In the following sections, we will introduce our method in more detail, and then apply it to medical image registration.

This remainder of this paper is organized as follows. We first introduce the moving least square algorithm used as deformation method in Sect. 34.2. Then we introduce SIFT feature used as the registration criterion of two images in Sect. 34.3.Our experiments and discussion is in Sect. 34.5.

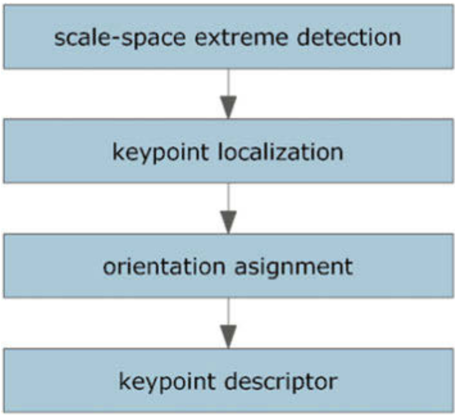


Fig. 34.1 Flowchart of SIFT Algorithm

34.2 Moving Least Squares

First we present an image registration method based on MLS deformation algorithm [5].

After selecting a set of corresponding control points on the source and target images, the MLS deformation technique aims to compute the transformation $l_v(x)$ that best minimizes the least squares error

$$\sum_i |l_v(p_i) - q_i|^2 \tag{34.1}$$

where p_i and q_i are the set of corresponding control points in the source and target images respectively. However, this transformation produces a single affine transformation of the entire image as there is no control over the scaling or shearing in the image. A weighting function included to this least squares error fixes this problem and thus produces a different transformation function for each point of the image.

$$\sum_i w_i |l_v(p_i) - q_i|^2 \tag{34.2}$$

The weighting functions w_i have the form

$$w_i = \frac{1}{|p_i - v|^{2\alpha}}. \tag{34.3}$$

where v is the point of evaluation in the image and α is a parameter of the weighting function whose value decides whether the weights computed are small or large. Because the weighting function w_i is dependent on the point of evaluation, the method is called Moving Least Squares minimization. From the above equation we can observe that as v approaches p_i, the weight w_i approaches infinity and the transformation function interpolates.

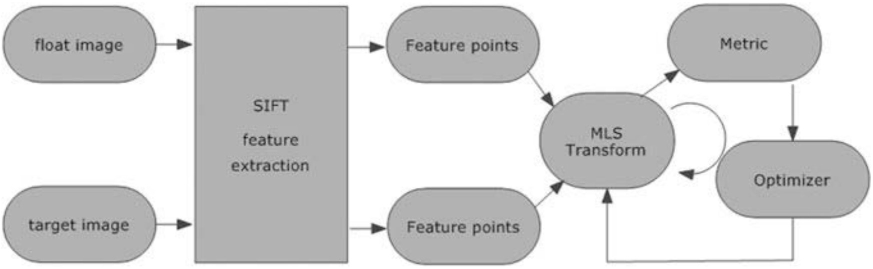


Fig. 34.2 Framework of our proposed method

The transformation function $l_v(x)$ is made up of a simple linear transformation matrix M and a translation vector T as

$$l_v(x) = xM + T \tag{34.4}$$

Here, matrix M can be regarded as a general transformation including effects of scaling shearing and rotating. By removing these components, we can obtain the affine, similarity, and rigid transformation functions. The details can be found in the literature [6]. The translation component can be easily computed by

$$T = q_* - p_* M \tag{34.5}$$

where p_* and q_* are the weighted centroids of the control points given by

$$p_* = \frac{\sum_i w_i p_i}{\sum_i w_i} \quad q_* = \frac{\sum_i w_i q_i}{\sum_i w_i} \tag{34.6}$$

Then the transformation function can now be calculated as

$$l_v(x) = (x - p_*)M + q_* \tag{34.7}$$

The least squares problem can be written as

$$\sum_i w_i |\hat{p}_i M - \hat{q}_i|^2 \tag{34.8}$$

where $\hat{p}_i = p_i - p_*$ and $\hat{q}_i = q_i - q_*$.

Note that Moving Least Squares is very general in that the matrix M does not have to be a fully affine transformation. In the literature [14], there are detailed derivations to M under different constraints. Here, we select rigid deformation. The following results are derivable from the constraints of rigid deformation.

$$M = \frac{1}{\mu_s} \sum_i w_i \begin{pmatrix} \hat{p}_i \\ -\hat{p}_i^{\perp} \end{pmatrix} \left(\hat{q}_i - \hat{q}_i^{\perp T} \right) \tag{34.9}$$

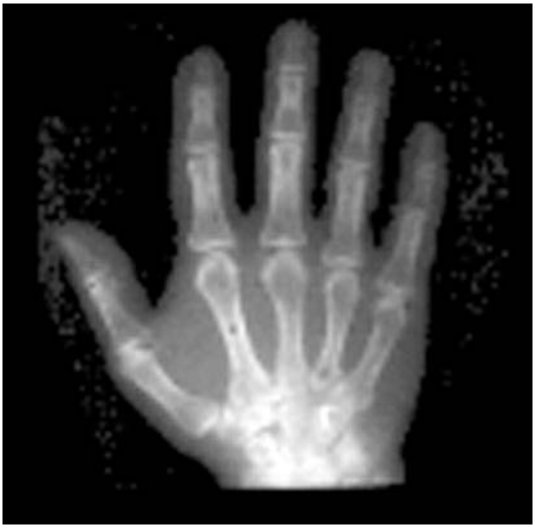


Fig. 34.3 Floating image

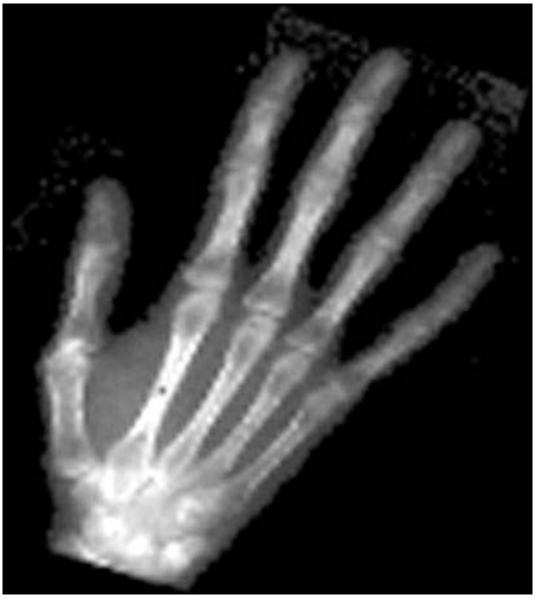


Fig. 34.4 Target image

where $\mu_r = \sqrt{\left(\sum_i w_i \hat{q}_i \hat{p}_i^T\right)^2 + \left(\sum_i w_i \hat{q}_i \hat{p}_i^{\perp T}\right)^2}$. The detailed derivation for the rigid transformations can be seen in [5].

Fig. 34.5 Extract SIFT feature in floating image

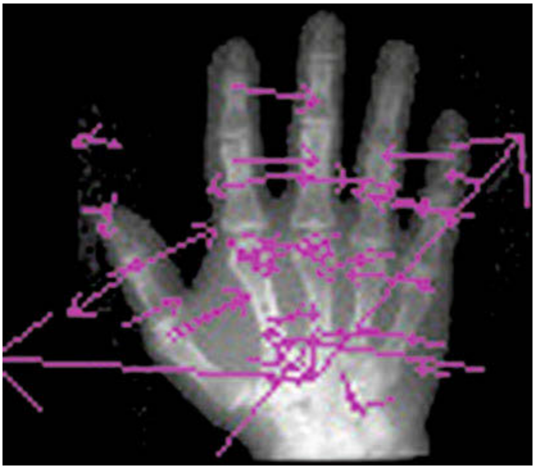
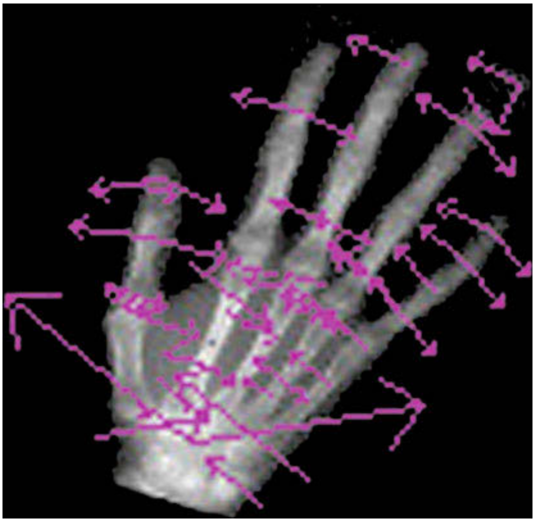


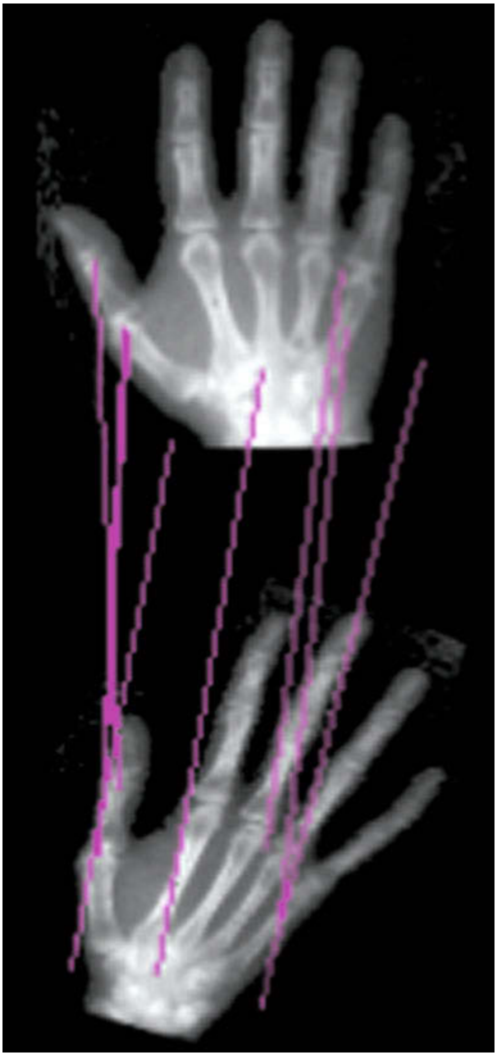
Fig. 34.6 Extract SIFT feature in target image



34.3 Scale-Invariant Feature Transform Feature Algorithm

Scale-invariant feature transform (SIFT) is an algorithm in computer vision to detect and describe local features in images. The algorithm was published by David Lowe in 1999 [6]. It has been successfully applied to a variety of computer vision problems based on feature matching including object recognition, pose estimation, image retrieval, and many others. SIFT, as described in [7], consists of four major stages:

Fig. 34.7 Initial possible corresponding point (including error point)



In the first stage, searches over scale space using a Difference of Gaussian function to identify potential interest points that are invariant to scale and orientation (Fig. 34.1).

In the second stage, the location and scale of each candidate point is determined and key points are selected based on measures of stability.

The third identifies the dominant orientations for each key point based on its local image patch. The assigned orientation, scale, and location for each key point enables SIFT to construct a canonical view for the key point that is invariant to similarity transforms.

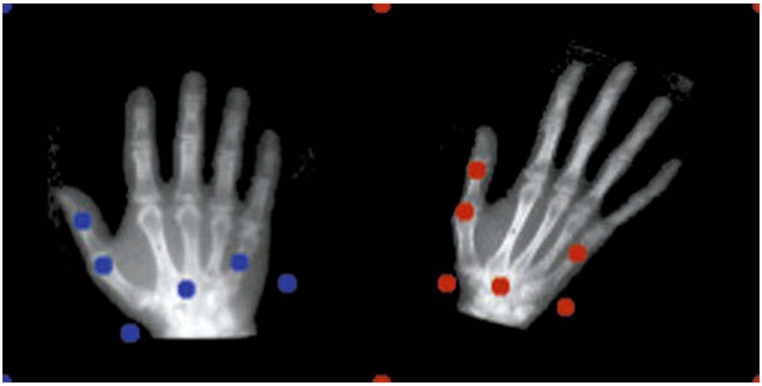


Fig. 34.8 Final corresponding point

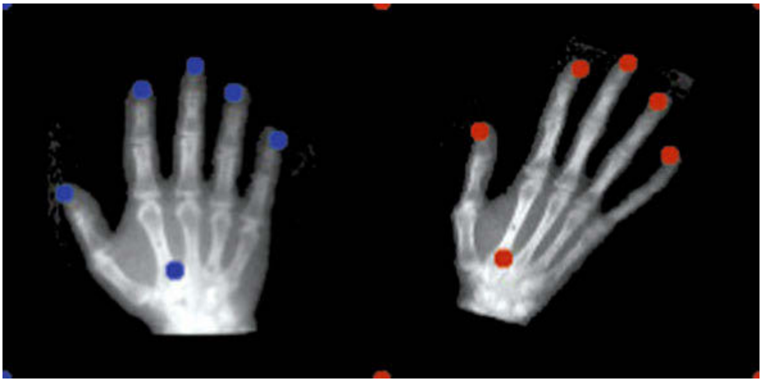


Fig. 34.9 Select manually equal number of points

The final stage builds a local image descriptor for each key point, based upon the image gradients in its local neighborhood.

The four stages will not be discussed further in this paper since our work makes no contributions to those areas. We only use SIFT algorithm to improve the rationality of selecting point. We need not manually select landmark as it is selected automatically by using SIFT.

34.4 Similarity Measure

Till date, a large majority of registration measures in the literature have been presented. In this paper, we use peak signal-to-noise ratio (PSNR) and mean squared error (MSE) as similarity measure to evaluate the performance of our registration method (Fig. 34.2).

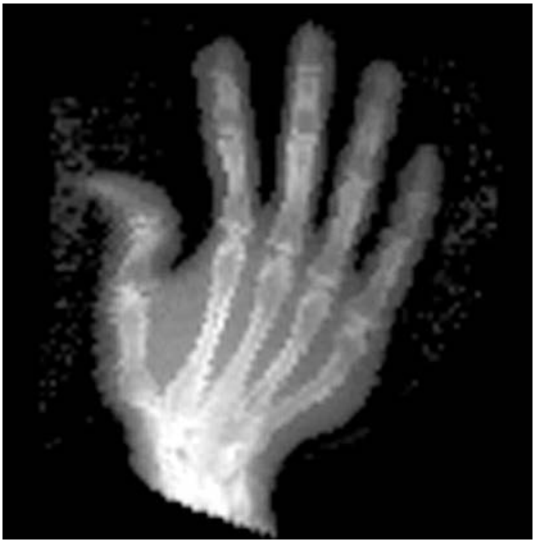


Fig. 34.10 Registered result using our method

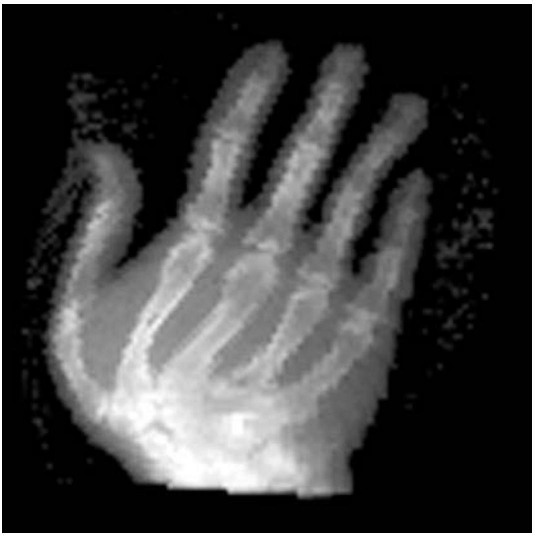


Fig. 34.11 Registered result using manual selection method

Table 34.1 Quantitative analysis results

	PSNR	MSE
OUR method	17.5183	1.1515e + 003
Original MLS method	15.1919	1.9674e + 003

34.4.1 Image Registration Model

34.4.2 Experimental Results

Here, we perform image registration experiments with medical image data to evaluate the performance of the proposed technique. Moreover, we also perform registration with MLS in which the landmarks are selected manually for comparison. The experiment is implemented in C++, and tested on Intel Core 2 Quad CPU Q6600, 2.40 GHz, and 4G RAM. Figures 34.3, 34.4, 34.5, 34.6, 34.7, 34.8, and 34.9

Figures (a–h) show the implementation process of our algorithm. Figure 34.10 is the registered result using our algorithm. Figure 34.11 is the registered result using MLS transformation with the same number of landmarks manually selected. Figures 34.10 and 34.11 show that the proposed technique is better than the original MLS technique. In addition, quantitative analysis results of PSNR and MSE also show that the proposed technique is better than the original MLS technique (Table 34.1).

34.5 Conclusion

In this chapter we have presented a new method for the registration of medical images which is based on the combination of MLS and SIFT technique. We need not manually select landmark as it is selected automatically by using SIFT. Experimental results show that the proposed technique is better than the original MLS technique. According to the experiment, we can conclude that the method proposed is more effective than the original MLS technique.

Acknowledgment This chapter was supported by Zhejiang Provincial Natural Science Foundation of China (Grant No. Y1100018).

References

1. Hajnal JV, Hawkes DJ, Hill DLG (2011) Medical image registration. CRC, London
2. Brown L (1992) A survey of image registration technique. ACM Comput Surv 24:325–376
3. Zitová B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21:977–1000
4. Shams R, Sadeghi P, Kennedy R, Hartley R (2010) A survey of medical image registration on multicore and the GPU. Signal Process Mag IEEE 27(2):50–60
5. Schaefer S, Mcphail T, Warren J (2006) Image deformation Using moving least squares. ACM Trans Graph 25(3):533–540 Jul

6. Lowe DG (1999) Object recognition from local scale-invariant features. Proc Int Conf Comput Vis 2:1150–1157
7. Lowe DavidG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
8. Koshani R, Hub M, Balter J et al (2009) Objective assessment of deformable image registration in radiotherapy: a multi-institution study. Med Phys 35(12):5944–5953
9. Klein A, Andersson J, Ardekani BA et al (2009) Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46(3):786–802
10. Dandekar O, Shekhar R (2007) FPGA-accelerated deformable image registration for improved target-delineation during CT-guided interventions. IEEE Trans Biomed Circuits Syst 1(2):116–127

Chapter 35
Denoising Using Laplacian Mixture Model with Local Parameters in Shearlet Domain

Wei Tian, Hanwen Cao and Chengzhi Deng

Abstract An adaptive Bayesian estimator for image denoising in shearlet domain is presented, where a mixture of Laplace distributions are used as the prior model of shearlet coefficients of images. The mixture of Laplacian probability density function has a large peak at zero and its tails fall significantly slowly than a single Laplacian pdf and the Laplacian mixture model can model shearlet coefficients distribution better. Under this prior, a Bayesian shearlet estimator is derived by using the maximum a posterior (MAP) rule. Simulations with images contaminated by additive white Gaussian noise are carried out to show that the performance in shearlet domain substantially surpasses that in wavelet domain, both visual effect and peak signal-to-noise ratio (PSNR).

Keywords Image denoising · Laplacian mixture model · Bayesian estimator · Shearlet transform

W. Tian (✉) · C. Deng
School of Information Engineering,
Nanchang Institute of Technology, Nanchang, China
e-mail: tw_0930@163.com

C. Deng
e-mail: dengchengzhi@126.com

H. Cao
Department of Science, Nanchang Institute of Technology,
Nanchang, China
e-mail: chwhappy@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_35,

35.1 Introduction

The denoising of natural image corrupted by Gaussian noise is a classic problem in signal processing. One of the most well-known methods in a Bayesian framework is soft threshold shrinkage proposed by Donoho [1]. A single Laplacian pdf is assumed in the soft threshold rule. However, a single Laplacian pdf is a weak model for wavelet coefficients of natural images because it is not fitted to the empirical histogram very well. A mixture of Laplacian pdfs is presented in [2, 3], because Laplacian mixture model has a large peak at zero and its tails fall significantly slower than a single Laplacian pdf. In other words, a mixture of Laplacian pdfs can improve modeling of wavelet coefficients distribution [4].

On the other hand, the shearlet transform has emerged as an exciting new tool for image, and it breaks the limitation of the wavelet transform and provides sparse representation for the objects.

In this chapter, a mixture of Laplacian pdfs with local parameters is used as a prior to capture the processing sparseness of the shearlet coefficients. For exploiting this prior in a Bayesian framework, we designed a MAP estimator to find the denoised image.

The simulation results for image denoising show that our algorithm in the shearlet domain achieves better performance visually and in terms of peak signal-to-noise ratio (PSNR) in comparison with the algorithms in the wavelet domain.

35.2 Shearlet Transform

Shearlet transform recently introduced by Guo and labate in [5], by taking advantage of the theory of composite wavelets, exhibits the following properties.

Shearlets satisfy parabolic scaling. Each element $\hat{\psi}_{j,k,\ell}$ is supported on a pair of trapezoids of approximate size $2^{2j} \times 2^{j}$, oriented along the lines of slope $\ell\, 2^{-j}$ (see Fig. 35.1a). Their supports become increasingly thin as $j \to \infty$, so shearlets are well localized.

An illustration of this frequency tiling is shown in Fig. 35.1b. Shearlets exhibit highly directional sensitivity. The number of orientations doubles at each finer scale.

Shearlets are spatially localized. For any fixed scale and orientation, the shearlets are obtained by translations.

Shearlets are optimally sparse.

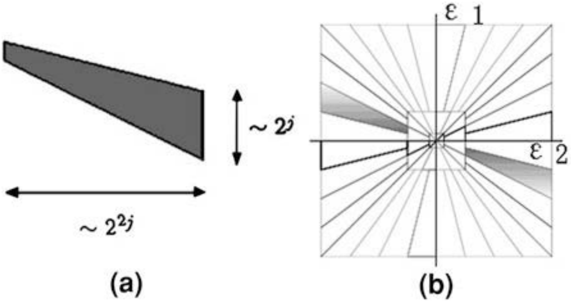


Fig. 35.1 **a** The frequency support of a shearlet. **b** The tiling of the spatial-frequency plane induced by the shearlets

35.3 Bayesian Denoising-Based Bivariate Model

In this section, the denoising of an image corrupted by white Gaussian noise will be considered, i.e. $g = x + n$, where n is independent Gaussian noise. We observe g (a noisy image), and wish to estimate the desired signal x as accurately as possible according to some criteria. In shearlet domain, the problem can be formulated as $y = w + n$, where y is the noisy shearlet coefficient, w is the original noise-free shearlet coefficient and n is noise, which is yet independent Gaussian.

How to get w from y? This is a classical problem in estimation theory. The standard MAP estimator for w given the corrupted observation y is $\hat{w}(y) = \arg\max\limits_{w} p_{w|y}(w|y)$.

After some manipulations, this equation can be written as

$$\hat{w}(y) = \arg\max_{w}[\log p_n(y - w) + \log p_w(w)]. \tag{35.1}$$

Suppose that the pdf of each shearlet coefficient is different from the other coefficients. In this case, we have $y(k) = w(k) + n(k)$, where $k = 1, 2, \ldots, N$ and N is the number of coefficients. Thus we can obtain the MAP estimation for $w(k)$ as

$$\hat{w}(k) = \arg\max_{w(k)}[\log p_n(y(k) - w(k)) + \log p_{w(k)}(w(k))]. \tag{35.2}$$

Assuming the noise is i.i.d. white Gaussian, the noise pdf can be written as

$$p_n(n(k)) = \frac{1}{\sqrt{2\pi}\sigma_n} \cdot \exp\left(-\frac{n^2(k)}{2\sigma_n^2}\right). \tag{35.3}$$

Replacing Eq. 35.3 in Eq. 35.2, it yields

$$\hat{w}(k) = \arg\max_{w(k)}\left[-\frac{(y(k) - w(k))^2}{2\sigma_n^2} + \log p_{w(k)}(w(k))\right]. \tag{35.4}$$

Therefore, we can obtain the MAP estimate for $w(k)$ by setting the derivative to zero with respect to $\hat{w}(k)$. That gives the following equation to solve for $\hat{w}(k)$:

$$\hat{w}(k) = y(k) + \sigma_n^2 \frac{d\log(p_{w(k)}(w(k)))}{dw\,(k)}. \tag{35.5}$$

Laplacian mixture model. Now we need a model $P_w(w)$ for the distribution of noise-free shearlet coefficients. If it is Laplacian and also local,

$$p_{w(k)}(w(k)) = \frac{1}{\sqrt{2}\sigma(k)}\exp\left(-\frac{\sqrt{2}|w(k)|}{\sigma(k)}\right) \tag{35.6}$$

then the estimator is the classical soft threshold shrinkage function,

$$\hat{w}(k) = \mathrm{soft}(y(k), \frac{\sqrt{2}\sigma_n^2}{\sigma(k)}) \tag{35.7}$$

where, $\mathrm{soft}(g(k), \tau(k)) = \mathrm{sign}(g(k))\cdot(|g(k)| - \tau(k))_+, (g)_+ = \max(g, 0)$.

Here, we consider the Laplacian mixture model with local parameters proposed by Rabbani and Vafadust [4], and then the pdf can be written as

$$\begin{aligned} p_{w(k)}(w(k)) &= a(k)\mathrm{Laplace}(w(k), \sigma_1(k)) + (1 - a(k))\mathrm{Laplace}(w(k), \sigma_2(k)) \\ &:= \frac{a(k)}{\sqrt{2}\sigma_1(k)}\exp\left(-\frac{\sqrt{2}}{\sigma_1(k)}|w(k)|\right) + \frac{1 - a(k)}{\sqrt{2}\sigma_2(k)}\exp\left(-\frac{\sqrt{2}}{\sigma_2(k)}|w(k)|\right) \end{aligned} \tag{35.8}$$

Figure 35.2 illustrates the observed and the Laplacian mixture marginal densities in log scale of the shearlet coefficients for several natural images.

As one can easily notice, this model is an acceptable approximation to the empirical histogram illustrated in Fig. 35.2, and the later experimental results also prove this point.

Bayesian estimator and parameters estimation. Solving (35.5) with (35.8), the MAP estimator of $w(k)$ is derived to be

$$\hat{w}(k) = \frac{\mathrm{soft}(\mathrm{y(k)}, \frac{\sqrt{2}\sigma_\mathrm{n}^2}{\sigma_1\mathrm{(k)}}) + R\mathrm{soft}(\mathrm{y(k)}, \frac{\sqrt{2}\sigma_\mathrm{n}^2}{\sigma_2\mathrm{(k)}})}{1 + R} \tag{35.9}$$

where $R = \dfrac{\dfrac{1 - a(k)}{\sigma_2(k)}\left[\mathrm{erfcx}\left(\dfrac{\sigma_\mathrm{n}}{\sigma_2\mathrm{(k)}} - \dfrac{\mathrm{y(k)}}{\sqrt{2}\sigma_\mathrm{n}}\right) + \mathrm{erfcx}\left(\dfrac{\sigma_\mathrm{n}}{\sigma_2\mathrm{(k)}} + \dfrac{\mathrm{y(k)}}{\sqrt{2}\sigma_\mathrm{n}}\right)\right]}{\dfrac{a(k)}{\sigma_1(k)}\left[\mathrm{erfcx}\left(\dfrac{\sigma_n}{\sigma_1(k)} - \dfrac{y(k)}{\sqrt{2}\sigma_n}\right) + \mathrm{erfcx}\left(\dfrac{\sigma_n}{\sigma_1(k)} + \dfrac{y(k)}{\sqrt{2}\sigma_n}\right)\right]}$,

$\mathrm{erfc}x(x) = \exp(x^2)\mathrm{erfc}(x^2)$, $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$ and $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt.$

Equation 35.9 above can be interpreted as Laplacian mixture shrinkage function.

As we can see, this estimator requires the prior knowledge of the noise variance σ_n^2 and the three parameters $\sigma_1(k)$, $\sigma_2(k)$, $a(k)$ for each shearlet coefficient.

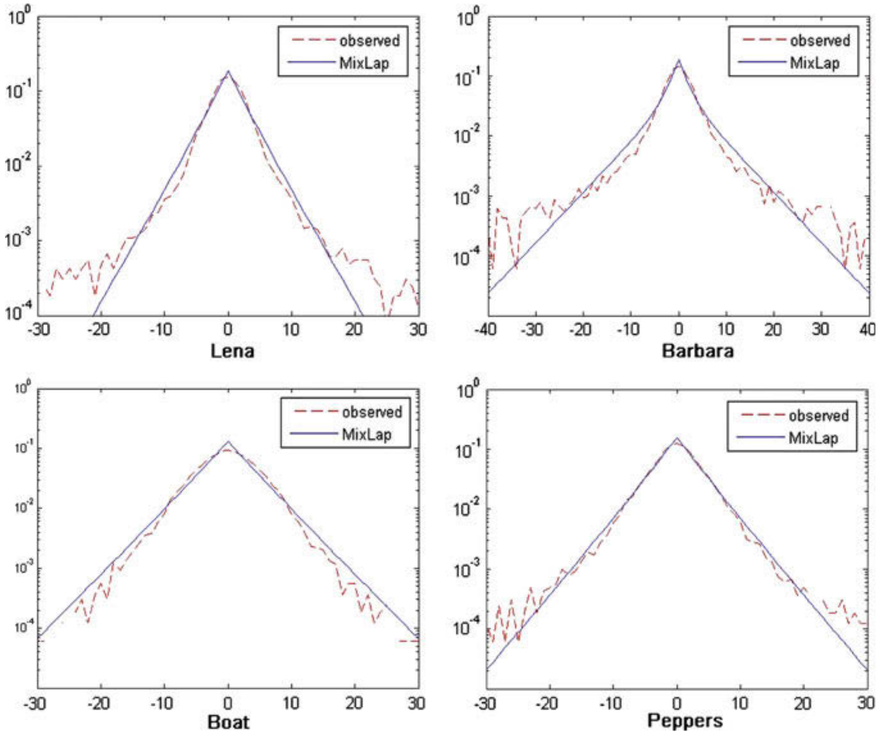


Fig. 35.2 Empirical histogram computed from several natural images with a *dashed line*. A Laplacian mixture pdf is fitted to the empirical histogram with a *solid line*

To estimate the noise variance σ_n^2 from the noisy shearlet coefficients, Monte Carlo method in [7] is adapted.

To estimate$\sigma_1(k)$, $\sigma_2(k)$ and $a(k)$ for each shearlet coefficient $w(k)$, a square window $N(k)$ centered at $w(k)$ is considered. The corresponding Expectation–Maximization algorithm in [4] is adapted.

The E-step calculates the responsibility factors:

$$r_1(k) \leftarrow \frac{a(k)\mathrm{Laplace}(\mathrm{w}(\mathrm{k}), \sigma_1(\mathrm{k}))}{a(k)\mathrm{Laplace}(w(k), \sigma_1(k)) + (1 - a(k))\mathrm{Laplace}(w(k), \sigma_2(k))},$$

$$r_2(k) \leftarrow \frac{(1 - a(k))\mathrm{Laplace}(\mathrm{w}(\mathrm{k}), \sigma_2(\mathrm{k}))}{a(k)\mathrm{Laplace}(w(k), \sigma_1(k)) + (1 - a(k))\mathrm{Laplace}(w(k), \sigma_2(k))},$$

where $k = 1, 2, \ldots, N$ and N is the number of shearlet coefficients.

Table 35.1 Comparison of PSNR values [dB] for Laplace and Laplacian mixture method in wavelet and shearlet domain

Image	Denoising scheme	Wavelet–Lap (7 × 7)	Wavelet–LapMix (9 × 9)	Shearle t-lap (7 × 7)	Shearlet–LapMix (9 × 9)
Lena	$\sigma_n = 10$	34.15	35.35	34.46	35.41
	$\sigma_n = 10$	30.94	32.31	31.28	32.17
	$\sigma_n = 20$	29.14	30.58	29.56	30.64
Barbara	$\sigma_n = 10$	31.90	33.76	33.08	34.53
	$\sigma_n = 20$	28.06	29.90	29.17	30.77
	$\sigma_n = 30$	25.98	27.73	26.92	28.53
Boat	$\sigma_n = 10$	31.80	33.28	32.36	33.43
	$\sigma_n = 20$	28.48	29.88	28.95	29.93
	$\sigma_n = 30$	26.60	28.03	27.09	28.15

The M-step updates the parameters
$a(k), \sigma_1(k), \sigma_2(k), a(k) \leftarrow \frac{1}{M}\sum_{i\in N(k)} r_1(i), \sigma_1(k) \leftarrow \sqrt{2}\frac{\sum_{i\in N(k)} r_1(i)|w(k)|}{}$
$\sum_{i\in N(k)} r_1(i),\ \sigma_2 \leftarrow \sqrt{2}\frac{\sum_{i\in N(k)} r_2(i)|w(k)|}{\sum_{i\in N(k)} r_2(i)},$
where M is the number of coefficients in the square window N(k) centered at $w(k)$.

35.4 Experimental Results

This section gives simulation results to show the efficiency of our method. We use three 512 × 512 grayscale images, namely Lena, Barbara and Boat, as test images. I.d.d. Gaussian noise at different variances levels is generated and imposed to them. For the shearlet transform [6], 5 levels of decomposition, 9 directions in both horizontal and vertical cone are adapted, and the quadrature mirror filter employed is Daubechies's symmlet with 8 vanishing moments.

There are four methods that we compare, and the results are shown in Table 35.1. The Wavelet–Lap method refers to the Bayesian method based on a Laplace model in the wavelet domain, and Wavelet–LapMix method refers to the method based on a Laplacian Mixture model in wavelet domain. The Shearlet–Lap and Shearlet–LapMix methods are similar but in shearlet domain. Table 35.1 shows PSNR values (in dB) of the denoised images obtained by the four methods above with three different noise standard deviations, 10, 20 and 30. PSNR values in orthogonal wavelet domain are taken from paper [4]. Lena, Boat and Barbara images are used for this purpose.

From Table 35.1 it can be found that in the same domain, using mixture model can achieve an improvement of 1 dB approximately. On the other hand, methods based on shearlet domain perform a little better than those on wavelet. The influence of correlation also depends on the content of an image. For example,

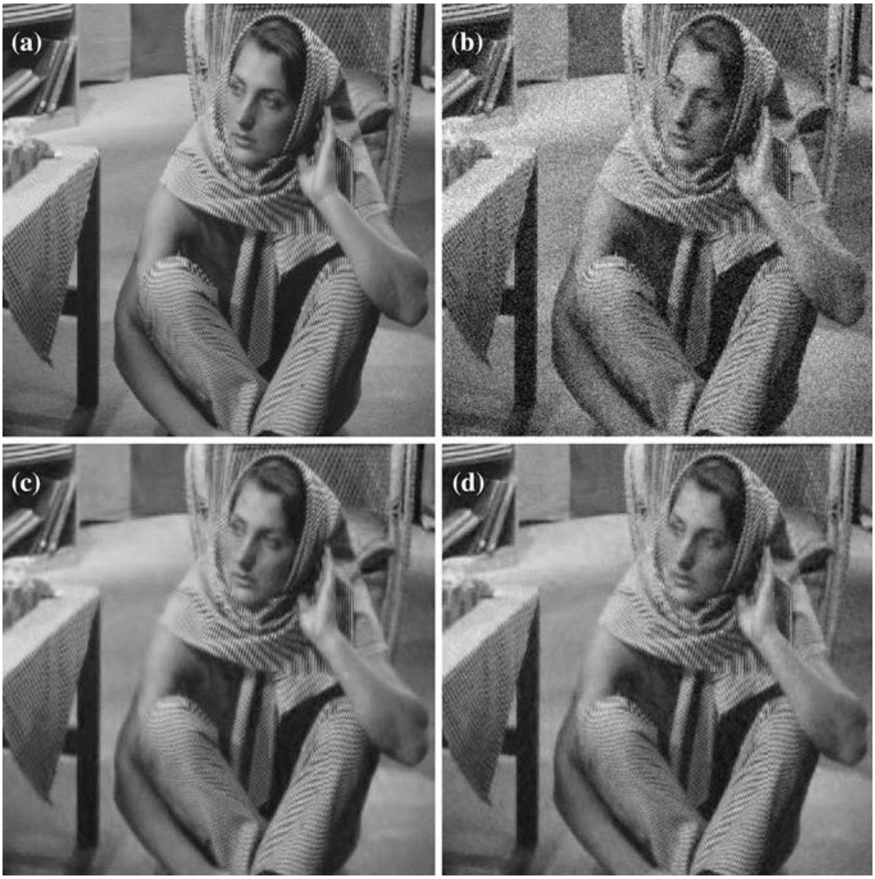


Fig. 35.3 Visual comparison of various denoising method on test image Barbara (at noise level 20). **a** Original; **b** noisy image, PSNR = 22.10 dB; **c** Wavelet–LapMix, PSNR = 29.90 dB; **d** shearlet–LapMix, PSNR = 30.77 dB

"Barbara" image has a large area of textures, which can be well captured by shearlet transform, and therefore be of most benefit than the other two images.

To compare the visual effect in a different transform domain, Fig. 35.3 shows the estimated images in wavelet and shearlet domain for Barbara with noise level 20. From the results we can find that the new proposed method yields better denoising results.

35.5 Conclusion and Future Works

We introduced a new statistical representation for shearlet coefficients, based on Laplacian mixture probability densities which has a large peak at zero and tails fall significantly slowly. The Laplacian mixture model can model shearlet coefficients distribution in each subband very well. Using it as the prior, the MAP estimator has been derived. The experimental results have shown that the performances in shearlet domain is superior to those in the wavelet domian in terms of the PSNR as well as visual quality.

In practice, the variance of the shearlet coefficients of natural images is quite different from scale to scale. We will make further research to improve the denoising performance by considering marginal variances. It might also be possible to use shearlet transform to derive the corresponding estimator.

Acknowledgment This is supported by Youth Science Research Project of Education Department of Jiangxi province in China (GJJ10269 and GJJ11250).

References

1. Donoho DL, Johnstone IM (1995) De-noising by soft-thresholding. IEEE Trans Inform Theory 41:613–627
2. Raghavendra BS, Subbanna Bhat P (2006) Shift-invariant image denoising using mixture of Laplace distributions in wavelet-domain. Lect Notes Comput Sci 3851:180–188
3. Rabbani H, Vafadoost M (2006) Wavelet based image denoising with mixed Laplace model. In: Proceedings of the 11th international computer society of Iran computer conference (CSICC 2006), Tehran, pp 21–26
4. Rabbani H, Vafadust M (2008) Image/video denoising based on a mixture of Laplace distributions with local parameters in multidimensional complex wavelet domain. IEEE Trans Signal Process 88:158–173
5. Guo K, labate D (2007) Optimally sparse multidimensional representation using Shearlets. SIAM J Math Anal 39:298–318
6. Information on http://www.shearlab.org/index_software.html
7. Crouse MS, Nowak RD, Baraniuk RG (1998) Wavelet-based statistical signal processing using hidden Markov models. IEEE Trans Signal Process 46:886–902

Chapter 36
The Region of Interest for Image Reconstruction Methods Based on Feature and Color

He Yan and Xiufeng Wang

Abstract The Region of Interest (ROI) Image reconstruction methods based on feature and color would be useful for the non-contact detection of defects in composite, metallic, and hybrid composite/metallic structures. An improved adaptive method of processing image data in multi-objective optimization has been developed to enable automated, real-time reconstruction of possibly engineering design, parameter estimation, and image reconstruction. There are three approaches for this purpose: Firstly, ROI with a gradient-based method improve quality enhancements and moderate convergence efficiency, and continue to develop the gradient evaluations "smart" imager cells method for image reconstruction. Secondly, the multi-objective framework will integrate the analysis image reconstruction for feature and color, instead of relying on one image codes to perform the analysis for all disciplines, and develop artificial intelligence algorithms for image classification based on the "smart" imager cells approach for simplifying multi-objective optimization. Lastly, the ROI model develop artificial intelligence algorithms for image classification based on the "smart" imager cells approach for simplifying multi-objective optimization, implementation of image reconstruction to optimize the ROI model in which lieu to the computationally expensive functions.

Keywords Image reconstruction · Region of interest · Multi-objective optimization · Imaging technology

H. Yan (✉) · X. Wang
School of Materials Science and Engineering,
Shaanxi University of Science and Technology, Xi'an, China
e-mail: yanhe@yahoo.cn

X. Wang
e-mail: exw@sust.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_36,

36.1 Introduction

Image reconstruction method has been shown to be useful for the non-contact detection of defects in composite, metallic, and hybrid composite/metallic structures. An improved adaptive method of processing image data in multi-objective optimization has been developed to enable automated, real-time reconstruction of possibly engineering design, parameter estimation, and image reconstruction. For optimization problems associated with Region of Interest (ROI) low accuracy function and gradient values are frequently much less expensive to obtain than high accuracy values [1]. When high accuracy evaluations are unavailable or prohibitively expensive, the ROI image reconstruction methods were based on feature and color. The optimization involves a ROI of two prior image reconstruction methods: one method based on adaptive detection of shape features; other methods based on adaptive color segmentation. The optimization of image reconstruction methods based on feature and color approach is flexible in application and extremely reliable, providing optimal results for all optimization problems attempted. This considerably slows the entire production and maintenance process [2, 3]. There are three approaches for this purpose: Firstly, ROI with a gradient-based method to improve quality enhancements and moderate convergence efficiency. Secondly, the multi-objective framework will integrate the analysis image reconstruction for feature and color, instead of relying on one image codes to perform the analysis for all disciplines. Lastly, implementation of image reconstruction to optimize the ROI model in which lieu to the computationally expensive functions [4–6].

36.2 Object Recognition and Multi-Objective Optimization

Object recognition. The ROI image reconstruction methods are based on feature and color. This optimal adaptive reconstruction of the ROI involves interaction between a shape-feature-based and a color-segmentation-based method in a cyclic algorithm performance. Using shape adaptive features and color adaptive features from the previous cycle life, ROI containing the object are identified in the present image by means of feature detection and color segmentation. The ROI is then used for sampling data to adapt a new shape and color features for the image during the next cycle life (Fig. 36.1). The methods can be used with any ROI image reconstruction application where 2D images are taken as slices of a larger object. These could include machines, materials for inspection, geological objects, or human scanning.

According to Fig. 36.1, several uncertainty models, the trust region method could readily be implemented in integrated circuitry to make a compact, real-time object-recognition system. It has been proposed to demonstrate the feasibility of such a system by integrating a 256-by-256 active pixel sensor with adaptive principal component analysis and adaptive color segmentation. All of methods are

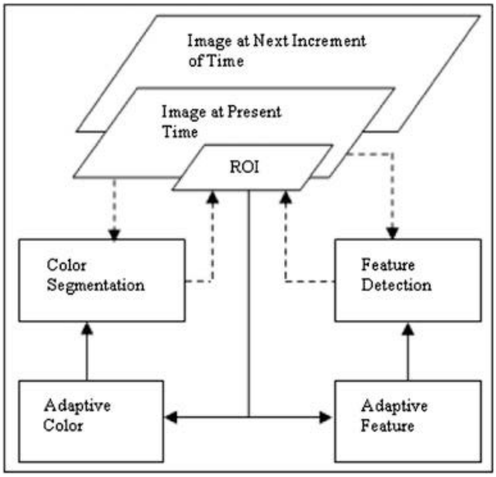


Fig. 36.1 Optimal reconstruction of the ROI involves interaction between shape-feature-based and a color-segmentation-based method in a cyclic algorithm performance

made to interact with each other in a cyclic life system to obtain an optimal solution of the object-recognition problem in image reconstruction environment.

A possible result is to provide a minimized adaptive step that is used to obtain by the two component methods when changes of color and apparent shape occur. Another possible results of the interaction is to increase, beyond the accuracy of the determination of a ROI within an image which contains an object that one seeks to recognize. The effect of an adaptive learning sequence enables to the multi-objective framework to update its recognition output and improve its recognition capability [7].

Multi-objective optimization. The multi-objective optimization can be used to solve Image reconstruction problems. In the case of the "ROI" for a multi-objective optimization problem is typically a range or a set of solutions, which represent trade-offs in objective space. In the optimization process, conflicts might arise among the various objective functions, i.e., the optimal values of each individual objective, in general, will not occur for the same decision variable vector. As a result of multiple local minima problems, multi-objective optimizations are able to find the global optimum results while pareto optimal solutions may converge to the local optimum value. The pareto front, it requires optimization problems with only two objectives, which must be curves in two dimensions. The multi-objective provides a large convergence efficiency enhancement for problems with non-convoluted pareto fronts and degradation in efficiency for problems with convoluted pareto fronts [8, 9].

36.3 Image Reconstruction and Foveal Vision

Image reconstruction. Image Reconstruction Technology is an emerging discipline of image capture and image-data processing that offers the prospect of greatly increased capabilities for real-time processing of large, high-resolution images for such purposes as automated recognition and tracking of ROI. Image Reconstruction Technology offers a solution to the image-data processing problem. In order to identify and track the shape adaptive features and color adaptive features without the means of dynamic adaptation to be afforded by Image Reconstruction Technology, it would be necessary to post-process data from an image-data space consisting of t-bytes of data.

Foveal vision focus on ROI. A foveal-vision image sensor is designed to offer higher resolution in a small ROI within its field of view. Exploiting these constraints systematically in conjunction with spatial shape characteristics resulted in increased image sequence processing efficiency by orders of magnitude. Foveal vision reduces the amount of unwanted information that must be transferred from the image sensor to external image-data-processing circuitry. Active pixel integrated-circuit image sensors that can be programed in real time to effect foveal artificial vision on demand are one such example [10]. The ROI image reconstruction methods control both a shape-feature-based and a color-segmentation-based of image regions. In limited search regions how to parallel processors and the extractions of features by special algorithms depending on the situation encountered. There is geared to object recognition for which corresponding generic knowledge is represented in ‘object processor groups’. It leads to efficient Image Reconstruction Technology and to modular object recognition based on feature and color.

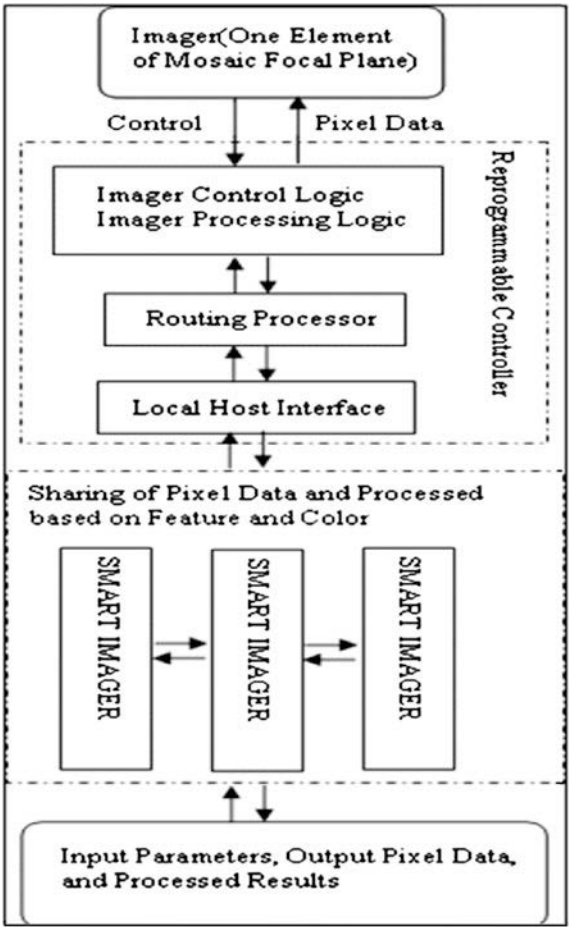
Figure 36.2 describes a mesh-connected Image Reconstruction Technology architecture as applied to a focal-plane built from “smart” image cells, each of which would contain adaptive principal component analysis and adaptive color segmentation. The multi-objective framework provides a networked autonomous array of reprogrammable controllers with “smart” imager cells processing of image data from individual image sensors. Based on experience with real-time processing for feature detection and processing, the image sensors can also have multiple pixel data outputs where each output has dedicated processing circuitry in its associated controller to achieve high function and gradient evaluations.

Each controller includes a routing processor to implement the network protocol and define the network topology for real-time transfer of raw pixel data and processed results between controllers. The processing and networking capabilities of the controllers will enable real-time access to data from multiple image sensors-“smart” image cells, each of which would contain adaptive principal component analysis and adaptive color segmentation. The application-level control of one or more ROI sharing of detected data features among smart cells.

36.4 Algorithm Performance on Image Classification

The “smart” imager cells of the Image reconstruction method’s powerful application of the principle of minimum complexity offered great promise for high performance restoration of conventionally compressed images. As has been

Fig. 36.2 Built from "smart" imager cells, each of which would contain adaptive principal component analysis and adaptive color segmentation


Imager(One Element of Mosaic Focal Plane)
Control
Pixel Data
Imager Control Logic
Imager Processing Logic
Routing Processor
Local Host Interface
Reprogrammable Controller
Sharing of Pixel Data and Processed based on Feature and Color
SMART IMAGER
SMART IMAGER
SMART IMAGER
Input Parameters, Output Pixel Data, and Processed Results

demonstrated in "smart" imager cells image reconstruction, the "smart" imager cells method is capable of correctly deducing the properties of structure finer than the diffraction limit. This essentially means the "smart" imager cells method correctly reproduces spatial pareto optimal solutions that are not present in the data. It is capable of reproducing feature and color because a minimum complexity model that correctly matches the ROI in a cyclic algorithm performance. In other words, the "smart" imager cells method should be able to deduce correct features that have not been recorded in compressed images. The image reconstruction of this program was to develop "smart" imager cells software to demonstrate this capability in a practical manner [11].

This study proposed three goals: (1) to continue to develop the gradient evaluations "smart" imager cells method for image reconstruction, (2) to develop image compression techniques based on the "smart" imager cells method, (3) to develop artificial intelligence algorithms for image classification based on the "smart" imager cells approach for simplifying multi-objective optimization. It was decided to investigate the ability of the "smart" imager cells method to provide superior restorations of images compressed with standard image compression schemes, specifically shape-feature-based and a color-segmentation-based method in a cyclic algorithm performance trust RIO algorithm and gradient evaluations results and conclusion.

36.5 Results and Conclusion

In this chapter, the ROI image reconstruction methods are based on feature and color. The optimization involves a ROI of two prior image reconstruction methods: one method based on adaptive detection of shape features; other method based on adaptive color segmentation. There are three approaches for this purpose: Firstly, ROI with a gradient-based method improves quality enhancements and moderate convergence efficiency, and continue to develop the gradient evaluations "smart" imager cells method for image reconstruction. Secondly, the multi-objective framework will integrate the analysis image reconstruction for feature and color, instead of relying on one image codes to perform the analysis for all disciplines, and develop artificial intelligence algorithms for image classification based on the "smart" imager cells approach for simplifying multi-objective optimization. Lastly, the ROI model develop artificial intelligence algorithms for image classification based on the "smart" imager cells approach for simplifying multi-objective optimization.

The ROI Image reconstruction methods based on feature and color would be useful for the non-contact detection of defects in composite, metallic, and hybrid composite/metallic structures.

References

1. Vitali R, Haftka RT, Sankar BV (2002) Multi-fidelity design of stiffened composite panel with a crack. Struct Optim 23(5):347–356
2. Viana FAC, Kotinda GI, Rade DA, Steffen V Jr (2007) Tuning dynamic vibration absorbers by using ant colony optimization. Comput Struct. doi:10.1016/j.compstruc.2007.05.009
3. Jin C, Qingze X, Jianfeng Y, Yuan L (2010) Numerical analysis based on the pressure bulkhead of the multi-objective optimization, 3rd international conference on information management. Innov Manag Ind Eng 2:187–190. doi:10.1109/ICIII.2010.209
4. Holst TL, Pulliam TH (2003) Evaluation of genetic algorithm concepts using model problems, NASA/TM–212813
5. Heikkila J (1997) Accurate camera calibration and feature based 3-D reconstruction frommonocular image sequences. Infotech Oulu and Department of Electrical Engineering, University of Oulu, ActaUniv. Oul. C 108

6. Goel T, Vaidyanathan R, Haftka RT, Shyy W (2004) Response surface approximation of pareto optimal from in multi-objective optimization. AIAA paper 4501
7. Duong T, Duong V, Stubberud A (2008) Object recognition using feature-and color-based methods. NASA Tech Br 10:32–33
8. Crespo LG, Kenny SP (2005) Reliability-based control design for uncertain systems, AIAA J Guidance, Control Dyn 28(4):15–30
9. Marduel X, Tribes C, Trépanier JY (2006) Variable-fidelity optimization—efficiency and robustness. Optim Eng 7:479–500. doi:10.1007/s11081-006-0351-3
10. Hoenk M, Monacos S, Nikzad S (2009) Synthetic foveal imaging technology. NASA Tech Br 9:12–13
11. Puetter R, Yahil A (2002) The Pixon method for data compression image classification, and image reconstruction. Goddard Space Flight Center. Document ID: 20020052633; Report number: UCSD-20-5313/CSS8397/28397A

Part IV
Education and Informatics

Chapter 37
The Application of Network Blog in College English Teaching

Wenlong Wan, Wenxian Xiao, Zhen Liu and Yulan Li

Abstract In recent years, more and more scholars have begun to pay attention and study network Blog. Under the guidance of constructivism theory, the significant assistance value of network Blog in the teaching of college English is first expounded, and then the approaches of using network Blog to assist the teaching of college English are discussed, at last, the limitations and their corresponding countermeasures are pointed out.

Keywords Network blog · Applications · Countermeasures

37.1 Introduction

At present, because of the conflicts between the resources of university education and the scale of education, most institutions teach English in the way of large classes. The traditional teaching model still dominates the foreign language teaching, so it is difficult to fully mobilize the enthusiasm of the students, unable to effectively cultivate students' practical skills of using language. Therefore, to build a secondary English teaching environment and use it to provide a rich input language, reconstruct two-way teacher-student interaction, support students' active language acquisition and a positive output to make up for the lack of classroom teaching, is an important issue college English teaching reform must settle.

With the development of information technology, network Blog technology has become more and more popular [1]. Introducing network Blog technology into the

W. Wan (✉) · W. Xiao · Z. Liu · Y. Li
Henan Institute of Science and Technology, Henan Xinxiang, 453003, China
e-mail: wanwenlong@hist.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_37,

teaching of college English to establish the college English teaching mode of "class-based, supplemented by network Blog", a new teaching environment can be created. This model expands and extends the "single classroom teaching"; enable teaching and learning activities flexible beyond the control of time and space; to maximize the student's language learning opportunities in experience, practice, participation, cooperation and exchange; has good applicability on making up for the lack of classroom language input.

37.2 Introduction of Network Blog

Network Blog is the fourth way network exchange following email, BBS and ICQ, which is the personal "Reader's Digest" in internet age, an online diary with the hyperlink as a weapon, represents a new way of life and new ways of working, also represents a new way of learning. The pioneer studying network Blog in China is Xianghui Mao. In 2003, he founded the Chinese education network Blog, which caused great concern in the education sector. In recent years, more and more researchers began to pay attention to and study Blog, such as Husan Hua, Wang Xiaodong, Xingfu Kun and so on. Compared to other teaching methods in online teaching, network Blog has unparalleled advantages [2]. For example, the establishment of personal forums generally needs to pay, and users must be registered in order to express his personal point of view; but network Blog is free with more space, the management of a single web page need not be registered to express his personal point of view, chatting media QQ ICQ, MSN can transmit information by text, voice and video chat, while communication model is single; but the resources can be shared on network Blog, by releasing personal point of view to achieve multi-dimensional interaction. Web applications generally need to buy space, learn to make web software knowledge; the network Blog is free, interactive and powerful, simple operation, no need to learn software knowledge, but also maintain a permanent exchange of records, easy to manage, with strong personality and the operation is also relatively simple [3].

37.3 The Theoretical Basis of the Blog Used in English Teaching

Constructivism believes that: knowledge is not got by the teaching of teachers, but in certain social and cultural situations learners achieve it through the manner of the construction of meaning with the help of other people (including teachers and learning partners) and the necessary information and means of learning. Since learning is the construction process of meaning realized through collaborative activities among people in a certain socio-cultural context with the help of others. Constructivism advocates learner-centered learning under the guidance of teachers, that is to say, we have stressed the role of the learner's cognitive subject, without ignoring the guiding role of teachers, in which teachers are the helper, facilitators of construction of meaning, rather than imparting knowledge and instilling those. Students are the main information processor and the active constructer of meaning, rather than passive recipients of external stimuli, and the object taught. “Situation”, “cooperation”, “exchange” and “construction of meaning” are the four elements of their environment. Network Blog can provide a learning environment for learners, becoming an important way to build collaborative learning through the interaction of the network Blog conversation and exchange; Blog can also organize problem-based learning, case-based learning and discussion, cooperation and personalized learning, reflective learning for learners and ultimately achieve the overall increase through the learner’s construction of meaning.

37.4 The Assistance of Network Blog in the College English Teaching Network

1. Network Blog can greatly improve students’ ability to actually use the language. Network Blog increases the amount of training in listening, speaking, reading, writing and translation, thereby improving students’ ability of actually using the language. Blog is a network tool setting listening, speaking, reading, writing and translating as one [4]. A large number of English articles in network provide students with the most abundant reading materials, providing very good support for students to take the initiative to acquire language, vocabulary consolidation, migration, knowledge points and constitute the new icon. Writing is almost the student’s entire production throughout and recovery process of network Blog. It is a more sophisticated cognitive activity. Students will use the grammar and vocabulary they have learned in the article during the writing process on purpose, which can effectively improve their actual ability of the use of language. Furthermore, links or network Blog can add sound and video files as listening training resources and diversity of media dramatically improve the students’ interest and quality in listening. Meanwhile back function in the network Blog also provides students with a semi-colloquial and non-real-time communication environment [5].
2. Network Blog provides a harmonious and equal interaction platform for teachers and students.

With the continuous enrollment of college, most institutions teach English by way of large classes. In the traditional foreign language teaching mode, because of the lack of discussion and communication between teachers and students, in most of the cases, students passively accept what the teachers teach, so it is difficult to mobilize students’ enthusiasm of initiatively discussing problems. But the network Blog technology as a supplementary teaching tool can solve this problem quite properly. Equal and harmonious relationship between teachers and students largely determines the quality of college English teaching, while the equal and harmonious relationship between teachers and students is to communicate not only affected by the communication of teachers and students in classroom, but also heavily influenced by the exchange of extra-curricular. The network Blog not only offers a broad platform for the exchange of teachers and students in class, but also provides a new platform to promote exchange between students in and out of the class [6]. In the network Blog communication, teachers and students can use written expression to conduct a deeper, more comprehensive exchange, which will help to make up for the lack of classroom interaction, effectively enhancing students' motivation, and can greatly improve their learning enthusiasm [7].

3. Network Blog helps to create a better learning atmosphere and enhance the students' self-learning ability.

Students have a very broad interest, and are easily interested in new things, so network Blog as a new way of teaching is very popular in the students. Network Blog can construct an ideal language information environment for students. Not same as the classroom teaching, in the Blog, students may choose their own learning resources, which is suitable for themselves according to their specific circumstances. In the teaching with the aid of network Blog, the students truly become the subject of learning, while teachers have become an information provider, facilitator. Network Blog can help students choose certain topics to research under the guidance of teachers, producing some original ideas, meanwhile, able to provide a good interaction mechanism through the network Blog to improve their own ability of self-learning by cooperation.

4. Network Blog as a platform for teachers' teaching reflection

Teachers use network Blog record the important plot and story in the process of teaching, carry out teaching reflection on "teaching problems" and "teaching conflict" occurred in the process of the teaching, and draw teaching summary. Through network Blog entries, teachers teaching the same subject can enhance the communication and collaboration, and remote training can be done to teachers, in which teachers can share their own teaching experience with other teachers and do research on teaching methods, and thus conduct conscious teaching reform, improve their teaching methods and forms of organization, which is helpful for teachers' development [8].

37.5 The Methods and Procedures Using Network Blog to Assist the Teaching of College English

1. To plan and construct the teaching network Blog site. Under the guidance of the constructivist learning theory, through the awareness and understanding of the concept of network Blog, plan to set up the teaching network Blog site to provide an interactive platform for teaching and learning among teachers and students, enabling network Blog to fully play its functions as network Blog can

improve students' actual use of language, an interactive platform for teachers and students, the creation of a better learning environment, platform for teaching reflection, try to achieve the goals of higher education reform in the use of modern information technology.

2. Establish teacher network Blog. Through the organization and management of teachers, carry out teaching activities using network Blog to achieve the interactive online teaching and learning, in which teacher's main job is to lay out the teaching task, provide learning resources, answer inquiries, organize the discussion as well as the expressing of personal opinion. Teachers can divide its own network Blog column into multiple functions, including home page, blog (log), photo albums, music and friends and so on. Logging the content associated with teaching activities on the network Blog in the form of text, which can be associated with classroom learning, such as previewing and reviewing the contents of the text, background text, exercises; can also be associated with extra-curricular learning. Image content can be placed inside the album, and the video data can be linked URL. In the links section, teachers can help students to link on the website, such as, listening online, onestopenglish, online dictionary, cocoa listening network, PubMed gas stations and other domestic and foreign English language learning website [9]. Teachers put lesson plans and reference materials to curriculum into their teaching network Blog beforehand, and ask students to preview before the class. After class, students ask questions through teachers' network Blog and QQ, and teachers use network Blog's comments function and restore function to tutor for students. Teachers' requirements and assignments are made directly on the network Blog, after completing assignments on time; students tell the teachers the location of their assignments, teachers click to see students' work. While greatly enhancing the efficiency of these duties, which also greatly reduces the labor intensity of teachers.
3. Encourage students to develop a class network Blog and personal network Blog. The class network Blog is a collective network Blog, and class members have permission to add content to the network Blog. This model can offer help for the building of learning resources, learning communication and collaboration of foreign language learning. The class network Blog as a place for students to learn and communicate will help foster a good learning atmosphere of the class; and gradually form a group network Blog of learning and communicating, and to further guide the students to self form network learning organization based on a common of professional direction. Personal network Blog is for students' independent learning by means of the review of the problem, the results show, the accumulation of resources, the building of groups, etc.
4. Through relevant network Blog cultural seminars to enhance the student's network Blog knowledge, enable them to know and understand the concept of network Blog, network Blog culture and spirit, to consciously apply the network Blog to learn experience, share and grow. Increasing teachers' awareness of the network Blog, using network Blog for teaching reflection, teaching reform, and the common exchange between peers and improve their own quality and take the professional growth path [10].

37.6 The Constraints and Solutions Using Network Blog to Assist the Teaching of College English

Although network Blog has many advantages over other assistance tool in college English teaching, it really has some constraints in the process of applying it in college English teaching, which are mainly reflected in: first, many universities lack network classrooms with slower speed of network, without a good learning environment, unable to meet students' needs of the online implementation of individual, independent learning. For the security of network information, universities limit the opening of Internet cafes; therefore it is very difficult for students' access. Bars inside and outside the school tend to be noisy, if students lack good control ability, it is difficult to concentrate on their studies, and listening training is particularly difficult; second, many college English teachers' ability of operating the network multimedia technology is limited, and the ability of using software to present course content needs to be improved. As time and energy, teachers invested are limited, the resources on teachers' Blog and the recommended resources are not sufficient, and the fewer guidance and language amendments for students' Blog affected students' learning result; students' information literacy and independent learning ability are lower, which influenced students' effect and level of Blog production; network speed can not meet the need, leading some contents can not be normally listened to, which wasted students' time and affected the enthusiasm of students to use the Blog; lack of effective teaching resources for educational Blog to support the teaching of English language.

As to the above problems, the achievement of the application of network Blog in English language teaching also needs to take the following measures: first, the need to seek the support in hardware and software. Universities must construct the hardware platform network Blog supporting college English teaching, which mainly includes: the establishment of autonomous language learning center, adhere to the Foreign Language Learning Center serves foreign language teaching. Established suitable computer networks, making more efforts to build the education Blog site based on internet technology. Specially-assigned person is responsible for monitoring online behavior of students to ensure that students' learning effect in a network environment, eliminate the negative impact in the process of network learning. To construct software teaching platform including system software and application software for network Blog's aid to college English teaching, provide for teachers and students with tools and services. In the process of construction, a principle should be paid attention to, namely, that is simple and universal. Thus facilitate students' access to learning resources. Second, use of network Blog should have some information literacy and technology literacy, form the awareness on the management and application of knowledge, otherwise, it is difficult to complete the construct of a good network Blog Group. So on one hand, teachers should change the concept of teaching, recognizing the importance of network Blog's assistance to teaching English, on the other hand, universities should increase the support of teachers in teaching and research work from the mechanisms and funding to mobilize teachers to renew ideas in the process of building interactive teaching model, learn new teaching ideas and teaching methods. Meanwhile should increase the intensity of the training of teachers to help English teachers master the application of multimedia network technology to improve the quality of college English teachers, professional standards and the ability to present course content with software. Third, the network Blog supporting English language teaching should pay attention to cultivate student' ability to access online teaching resources, master the methods network Blog supporting college English learning, and master the basic knowledge and skills of online learning, including: the basics of using a computer, basic operation of windows, internet browsing web pages, the attendance of the BBS discussion, able to download the file, send and receive e-mail and broadcast multimedia courseware. Using such platform as Youth Academic Forums of Department of Foreign Languages to strengthen publicity of network Blog culture in the form of lectures, Blogs, salon, etc., and gradually improve their information literacy and independent learning ability, improve their production level of the network Blog [11]. Fourth, to seek the support of the network center to improve students' learning environment of network, improve network speed in the use of broadband services, which is conducive in building second class of English on network Blog, is conductive in the promotion and use of networking tools to enable them to more effectively serve the teaching and research. Fifth, to construct teaching resources on network Blog supporting the English teaching. The construction of teaching resources not only requires a lot of capital investment, but also requires a lot of manpower, particularly the need for the talents who have both teaching experience and know software development. This kind of people have been very scarce and universities must conduct conventional teaching at the same time, so manpower can be put in this area is limited, therefore the independent development of all new teaching resources is very difficult to afford, and it is not necessary. Thus, the current universities may better and more economically build teaching resources on network Blog supporting the college English teaching through such ways as the collection, transformation, integration of traditional teaching resources, collect shared resources, purchase or exchange of multimedia network resources, and develop lesson plans and online courseware.

37.7 Conclusion

E-Learning researchers Xianghui Mao once said: with in-depth of educational information, the network Blog tool or the application of network Blog ideas will be generally accepted in basic education, higher education and the working environment. Use of network Blog technology to assist college English teaching is an effective way of improving college English teaching and an important method of innovative teaching. Network Blog technology as a means of network technology, reflects the direction of contemporary development, will deem to play an important role in college English teaching.

References

1. Lixia Ma (2004) Blog in the organization of information exchange [J]. Inf Sci 7:881–883
2. Xiangdong Chen, Jiping Zhang (2003) Blog culture and modern education technology [J]. E Educ Res 3:17–21
3. Mao Isaac (2003) Blog will become an important tool in education [J]. China Distance Educ 1:73–76
4. Fang Xingdong, Liu Shuanggui (2003) Blog (blog) the application of technology in the field of education research [EB/OL]. http://mail. Nhu. Edu. Tw/ ~ Society/e2j/36/36214. Htm
5. Kekang (2005) The theory and method on the deep-level integration of information technology and curriculum. E Educ Res 1:8–9
6. Stephen Downes (2004) Educational blogging educause review, 39(5): (Sep/Octo)
7. Huang Hua (2007) English writing based on blog [J]. Wuxi Inst Commer Technol 3:89–91
8. Ying-hong Chai, cited Qin Gang, Li Cui (2006) Blog supporting English language teaching research [J]. Foreign Lang Educ 10:46–48
9. Lei Dong (2004) Education blog break new ground [J]. Equip Mod Educ China 12:54–57
10. Wei Qi (2004) The application of blog in education [J]. Inf Technol Educ 2:11–13
11. Hua Husan, Xiaodong Wang (2004) The application of blog in education teaching [J]. Distance Educ J 1:10–12

Chapter 38
The University Computer Foundation Educational Reform Searches Analyzes

Ning Li, Zhen Liu, Wenxian Xiao and Hui Ye

Abstract On the basis of analysis of the status quo of computer basic teaching,this text points out the necessity for the current curriculum reform in basic computer and put forward specific ideas for reform. And it emphasizes teaching should be scientifically classified into different levels throughout the basic university computer teaching and theory and practice go hand in hand accompanied with teaching models of basic computer with professional and distinctive features to meet the demand for basic computer education in the university.

Keywords University · Computer base · Courses · Teaching · Reform

38.1 Introduction

Computer basic proficiency is not only an important and indispensable part in the knowledge structure of college graduates, but also an important condition to measure the standard of graduates for the employers [1]. However, basic computer courses are very practical, technical updates are very fast, as the first course of the general students on computer courses, making it face the different levels of computer object. In this case, how to innovate to achieve a sub-professional, sub-levels of computer experiment teaching environment [2], and to meet a better basic computer classroom instruction, enabling students to access the computer skills future career requires and it increases students' strength and confidence in employment, it is extremely necessary and urgent.

N. Li (✉) · Z. Liu · W. Xiao · H. Ye
Institute of Information Engineering, Henan Institute of Science and Technology,
Xinxiang 453003, China
e-mail: 30240547@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_38,

38.2 The Current Status of Teaching of Basic Computer in University

38.2.1 The Students Different Starting Points, Level Uneven

With the development of computer network technology, at present, China [3] has already begun in the primary and secondary spread of information technology education, but because of regional differences, primary and secondary level of computer literacy is inconsistent, uneven level of students in computer applications. Such an objective starting point for differences in the level of the students due to regional differences in the teaching process requires us to co-ordinate arrangements.

38.2.2 Different Subjects Require Different Levels of Basic Computer Applications for the Students

Information society continued to develop in depth, all walks of life continue to accelerate the process of information; the integration of computer technology and a number of professional teaching greatly enriched the content of professional courses, this integration has become a new trend of technological development; the requirements of professional ability of computer applications are increasingly strong and have diversified characteristics. Discipline covered in paper, history, philosophy, economics, management, science, engineering, agriculture, medicine and many other professionals, and each subject requires the students should have not exactly the same computer knowledge [4]. We used in the teaching process a uniform curriculum and uniform teaching plan, the final exam is taken in the form of school exams. This often just completes imparting knowledge to the course, but not combining the characteristics of their disciplines, which to some extent, makes students present blindness, whom do not know why to learn these things and for what purpose. This led to the teaching process that is comprehensive, in fact, the knowledge truly grasped and applied is very little.

In view of the different levels of universities and on the basis of the different needs of the computer professional curriculum, unified teaching of basic computer courses no longer meets the needs of new situation, therefore, according to teaching contents, areas and levels of computer basic courses, reforming the teaching model is very inevitable.

38.2.3 The Check-up System is Not Flexible Enough to Achieve the Goal of Promoting Learning by Examines

Performance evaluation is an important link in the process of teaching, which offers guidelines to students' learning behavior and learning methods. And unreasonable performance appraisal will hinder the improvement of students' overall development and comprehensive makings. The entire school implements the united examination on the assessment of Computer Foundation [5–6]. It can reflect a student's testing level to a certain degree, but cannot comprehensively reflect his command of curriculum knowledge.

38.3 Second, the Location and Characteristics of Basic Computer Education in College

The perspective of basic computer education in college is from top to bottom, from the surface to point, which is the prominent and intrinsic difference from other courses and belongs to its own feature that is decided by the blending of theory and implication as well as the coherent combination of multi-science [7].

As far as I am concerned, the aim of college's basic computer education has the following three points:

(1) It enables students to have a correct understanding of its working process essentially
(2) It enables students to have the corresponding specialized operation skills in computer application
(3) It gives students computer knowledge good enough to adapt them to the demand for professional development and the ability of re-learning and implication.

The realizations of three teaching purposes will make the student have a correct understanding of computer, and attain the level of comprehension. Learning computer is the process which requires one to be a good master of a comprehensive study, and the understanding in its theory is also stratified. University education of basic computer not only gives the human by "fish", the more important is that it teaches him "how to fish". The aim of basic computer education in college not only to impart students with "fish" education but also to teach a person "fishing". Whether the introduction to the theory is a good one, has a direct influence upon the study and application in computer. Calculator among them, the first solution to be resolved is the primary entry to theory and operating implication, and the second is the improvement of ability in both of the foundations.

Such cases not only occur in non-computer professional learning, but also in the professional. The meaning of introduction to theory of computer foundation teaching lies not only in the guidance to learning and application, but also in the further deepening of learning and application. The theoretical introduction fails to resolve the difficulties in computer learning, and falls into the marsh of bind touching on elephants. The theoretical threshold in basic computer education lays great stress on the mastery and know-how logically in mind to the complete computer principle by way of learning and with his present knowledge capacity available. It not only stresses on the teaching in theoretical details but also the first level in the standing of the overall computer theory.

The underestimation for theoretical entry in non-computer specialized computer foundation teaching is actually caused by the deviation in the understanding of teaching aims and the superficial thinking. The two above-mentioned awkward nesses are just the reflection of the instructional weakness and poor efficiency in theoretical education of efficiency of expression. Hence, the theoretical instruction in computer foundation learning is the key to the realization of computer foundation teaching.

The classification of computer basic education scientifically sheds lights on the computer-teaching rules, but in the current teaching practice, it is the undesired crux not to be fully aware of its characters and to apply them to hierarchical teaching. For another, such analog defects still exist as its inefficiency and inability to cope with the "the theory and the application", "the depth and the breadth", "the pursuit of advanced level" and "the relative stabilization in teaching" approved by professor Haoqiang Tan.

For some reason, the university computer education should combine with its features, closely around the three teaching objectives, readjust its teaching levels, and be divided into two levels at the first stage of learning, namely:

(1) basic computer knowledge;
(2) computer learning combined with profession.

38.4 Implementation Methods of the Reform Based on the Computer Basic Teaching

38.4.1 Depending on the Different Professional Needs of the Computer, the University Can Classify the Students, Deep Professional Needs with Flexible Organization of Contents

According to our school, students can be divided into three categories: science and engineering, liberal arts and sports art. According to professional settings for different types of courses, curriculum development, integration and optimization of teaching content, from the macro level. We should meet the teaching needs of different students. That is, according to the different disciplines, not only to set a different course, but also the same course in the teaching process should be distinguishly treated. For students of economics and management, in the process of the course, you can highlight statistics on excel applications; for the art (especially fashion design and interior design category) professional students, with mastering the basic foundation of knowledge, teachers should give full play to the professional nature of teaching objects, enabling them to design artistic-looking documents, animation and to minimize theoretical explanations, thereby enhancing student learning and mastery of computer interest and confidence.

38.4.2 On the Basis of the Classification, According to the Situation in the Basic Computer Knowledge and Operating Skills to Master for a Freshman, We Stratified Education

Different levels teaching of basic computer is guided by constructivism, drawing CBE teaching philosophy, according to the level of knowledge and ability of students will be divided into different levels of students, supplemented by different instructional design, while the use of appropriate teaching strategies based on students specialty adjust the teaching content, so that different levels of students in learning ability, learning speed can achieve the same teaching objectives.

For the freshmen, after organizing hierarchical test, based on examination results the students are divided into A, B two classes. Students with test scores greater than or equal to 50 classified into A; students with test scores less than 50 min into B. For students with more than 85 sub-test scores may be exempted from taking "computer based" courses, but asked to complete part of the "Computer Basics" course of experiments and exercises, and required to choose their elective course about computer information technology, or to learn more a level computer courses as attendant student.

Through the layers, we establish reasonable teaching programs, in ensuring the basic requirements of teaching, based on the students, appropriately adjust the course content and organize teaching with a appropriate depth, breadth, focus, teaching methods, reflects the people-oriented, individualized teaching philosophy.

38.4.3 Student-Centered, Reform Teaching and Methods of Assessment

(1) The establishment of network teaching platform

With the rapid development of computer technology and its integration with other cross-disciplinary, the content of basic computer course teaching continues to increase, but the trend of higher education reform program is to reduce a large number of class hours while giving students more autonomy to study and a broader space of development. Now it is very common that the students have computers. In order to solve the conflicts with dramaticly increased lectures, basic practice teaching of university computer should conduct online learning, online jobs, online discussions and testing and extra-curricular activities with network teaching platform. Therefore, we need to establish and develop a network teaching platform, such as to set the shared areas of teaching resources in our computing center web site to achieve the resources shared; to arrange areas for the exchange of basic computer course, and to arrange teachers responsible for special instruction; for some of more concentrated problems, they can discuss at the meeting of the department to meet the needs of students and to timely solve the problems students encounter. Using the campus network, students can download and watch teachers teaching courseware, instruction manuals and other experiments, conduct self-study and intensive training, extend learning time and space and enrich extra-curricular learning activities of students.

(2) Apply what they learn, and actively build a practical platform for students

The main purpose of computer education is to enable students to apply to the computer as an intelligent tool for future study, work and life. In order to create a good cultural atmosphere called “Learning” on campus, provide a stage to enable the students to play personal skills and demonstrate wisdom and talents, enrich the cultural life of students after school and enhance students hands-on practical ability, colleges and universities should actively carry out a variety of extracurricular activities.

(3) Focus on assessment of students hands-on skills

Examination is an important part in teaching and an important means of checking the effect of teaching, consolidating knowledge, improving teaching and achieving educational goals. But the traditional rote-based written examinations clearly can not reach that goal. To achieve the training objectives regarding application as the main line, we must establish proficiency test-centered full test mode, emphasize the practical ability of students, improve hands-oriented performance and implement the new evaluation system, that is, course score $=$ (work $+$ attendance) $\times$ 10% $+$ (practical skills $+$ learning performance) $\times$ 20% $+$ integrated design of experiments. Through the above three aspects of the assessment results, we can check whether the teaching activities at different levels have achieved the desired results; if there are problems, we can easily find out the reasons and to solve it so as to ensure the quality of computer basic teaching and learning activities.

(4) To strengthen teaching staff, improve teaching quality

With teaching basic computer course at different levels, the role of teachers ranges from the initiator to the director and organizer, whom play an important role in the implementation of the guidance, learning management and assessment.

Teachers should pay attention to the original knowledge structure of students, that is, different people, different style of teaching and guidance, so that collective teaching has a good combination with individual teaching. At the same time, teachers should change the “indoctrination” type of teaching methods and adopt a variety of teaching methods to fully mobilize the active participation of students.

38.5 The Conclusion

Computer basic education is an important part of higher education, for many non-computer majors, involving almost every profession, so its reform needs to be considered more carefully. Through describing the reform of teaching contents and improvement of testing method, this text explores the multi-level teaching mode of computer basic course. Basic computer teaching school must win the attention and support of leaders from relevant departments. Only through common interest and deepening the reform of basic computer teaching, and constantly perfecting the curriculum system of basic computer education can we develop competitive compound talents.

References

1. Shibing Zhou (2008) University teaching of computer basic course [J]. Jiangnan Univ: Educ Sci 28(2):64–67
2. Jiang Li, Xingfen Li, Zhengke Huang (2009) Sub-sub-level professional computer based experimental teaching and research [J]. China Power Educ 138:143–145
3. Qi Li (2008) Basic computer curriculum system [J]. Comput Knowl Technol 13:769–771
4. Tao Zhang, Aidong Wang (2009) Of computer teaching reform on the basis of [J]. Taizhou Univ 6:80–84
5. Xiane Cun, Runqin Cai, Yufeng Pu (2009) Universities basic computer non-computer professional teaching and research [J]. Comput Educ 24:13–15
6. Ding Jianmin,WangYuQin,Liu Wei (2006) In computer basic course teaching innovation teaching mode and teaching methods were discussed [C]//University computer foundation course report BBS program committee. Basic computer course report BBS university of proceedings: 2005. Higher education press, Beijing, pp 385–388
7. Kun Liu (2006) Universities based the computer specialized computer explore new teaching mode [J]. Shaanxi Educ (Theory Version) 12:124–125

Chapter 39
Overview of Several String Pattern Matching Algorithms in Data Structure Teaching

Shukun Liu, Meiling Cai and Hao Peng

Abstract Data structure is one of the most important base courses of computer profession, and string is the most important content of data structure. This paper discusses the definition of string and explains several string pattern matching algorithms detailed which will be helpful for students to study the course. The main differences of three kinds of string matching algorithms which can be used to help students understand the essence of the string matching are described in this paper. In this way, students can master the methods of string matching quickly and teachers can achieve the purpose of teaching easily.

Keywords String · Pattern matching · Algorithm

39.1 Introduction

Data structure is a base course of computer major and a main course of computer science. In order to master all kinds of data structures which are often used and improve the ability using data structure of solving problem, the students must study the course hardly. String is the more important knowledge in the data structure teaching [1, 2]. Especially most algorithms of matching are very

S. Liu (✉) · M. Cai · H. Peng
Department of Computer Science and Technology,
Hunan International Economics University, Chang sha, China
e-mail: Liu_shukun@163.com

M. Cai
e-mail: caimeiling418@sina.com.cn

H. Peng
e-mail: peng_hao1978@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_39,

important in solving the practical problems [3]. For example, in many operations of software, the process of finding is an application of string matching.

In general, the string is a finite sequence which is composed of zero or more characters. The sequence is usually denoted by S = ‘$a_1a_2a_3$… a_n’, where S is the string name of the sequence of characters enclosed in quotation which denotes the value of string [4, 5]. In a string, a_i (1 <= i <= n) can be letter, number, underscores or other character. The number of characters contained in the string is called string length. String of zero-length which is called empty string does not contain any characters. Usually if a string is just composed of zero blank or many blanks then it is called blank string. If there are consecutive characters in any string then the consecutive characters can be called as subsequence of the string, on the other hand the string that contains the corresponding substring is called the main string. The serial number (or location) of substring in the main string is defined as the position of the first character of the substring appearing in the main string usually.

For example, suppose P is a string and P = ‘This is a string’, [6]Q is a string, and Q = ‘is’. Then we call Q as a substring of P, and P is the main string. We have found that the string Q appeared twice in the string P, but the first time when the string Q appears in the string P with the position 3. So we call that the serial number of Q is three. Especially empty string is a substring of any string. Any string is a substring of its own.

39.2 The Definition of Abstract Data Type of String

The ADT of string is as follows [7, 8]:

ADT String {

Data object: D={a_i |a_i∈ CharacterSet, i = 1,2,…,n, n≥0}

Data relation: R1={< a_{i-1}, a_i > | a_{i-1}, a_i ∈ D, i = 2,…,n}

Base operations:

StrAssign (&T, chars)

Initial condition: chars is a const of string.

Result: the value of chars is the value of T.

StrCopy (&T, S)

Initial condition: the string of S is exist.

Result: the value of S is assigned to T.

//Attention: the operation of strcopy can not achieved in the way T=S but in the //way of function call.

StrLength (S)

Initial condition: S is a string and S is exist.

Result: the return number is called the length of the string.

StrEmpty (S)

Initial condition: S is a string and S is exist.

Result: if the string is empty, the result is true else the result is false.

StrCompare (S, T)

Initial condition: the string of S and T are exist

Result: if S>T, the return value is positive number; if S<T, the return value is negative number; if S==T, the return value is 0.

Concat (&T, S1, S2)

Initial condition: the string of S1 and s2 are exist

Result: the return value is a new string which is composed with s1 and s2.

SubString (&Sub, S, pos, len)

Initial condition: S is a string and S is exist. 1≤pos≤StrLength(S) and 0≤len≤StrLength(S)-pos+1.

Result:The value of sub will be a substring from the position of pos of string S.

Index (S, T, pos)

Initial condition: The variable S and T are string, They are not empty.1≤pos≤ StrLength(S).

Result: If there is a substring in the string S with the same value of string T, then the first position after the position pos of the main string is returned. Otherwise, the return value is zero.

Replace (&S, T, V)

Initial condition: The variable S and T are string. They are not empty.

Result: Replace the string T in the sting S with the same value of string T with string V.

StrInsert (&S, pos, T)

Initial condition: The string of S and T are exist, 1≤pos≤StrLength(S)+1.

Result: Insert string T at the posth position of string S.

StrDelete (&S, pos, len)

Initial condition: The string of S is exist, 1≤pos≤StrLength(S)-len+1.

Result: Delete len characters at the string S from the pos position.

DestroyString (&S)

Initial condition: The string of S is exist.

Result:String S is destroyed.

ClearString (&S)

Initial condition:The string of S is exist.

Result:String S will be empty.

} ADT String

The min sub operation collection is composed with StrAssign (&T, chars), StrCopy (&T, S), StrCompare (S, T), StrLength (S), Concat (&T, S1, S2) and SubString (&Sub, S, pos, len) in above operations. That is those operations are not accomplished by other operations. Otherwise other operations such as ClearString (&S) and DestroyString (&S) can be accomplished by these operations.

39.3 Several String Pattern Matching Algorithms

The logic structure of string is similar to the liner table. The difference between them is that the data object of the string is constrained in character subset. But the basic operations of the string are very different with liner table. The single element is considered as the data object in linear table but in string the whole string is considered as the data object. There are some string pattern matching algorithms as follows.

39.3.1 Simple Algorithms

For example, returning the position that the substring T in the main string S from the pos position [9]. If the main string S does not contain the sub string T, then the return value is zero, else return the concrete position. The algorithm is as follows:

```
int Index(SString S, SString T, int pos)
//return the position of substring T first appear in the string S after the position
//pos of the main string S.
{
//if does not exist the return value if 0. T is not empty,1≤pos≤StrLength(S).
i=pos;
j=1;
while (i<=S[0]&&j<=T[0])
 {
      if (S[i]==T[j])
             {
                ++i;
                ++j;
             }
//continue compare the next characters
           else
         {
            i=i-j+2;
            j = 1;
         }
//the pointer withdraw and begin to retrieve again
         }
         if (j>T[0])
         return i-T[0];
         else
         return 0;
}
```

39.3.2 Head–Tail Matching Algorithms

The question which will be resolved is the same as Sect. 3.1, but the matching method is not according to the sequence of from head to tail, but first compare the head character to the last one,at last compare the second character with the n-1 character. For example, main string M=‘ababcabcaaabcbaabc’, pattern string N=‘abcba’. The compare process is that first compare the first character of M and the first character of N, if they are the same then continue comparing the last character of N with the xth character of the main string (x is the length of the pattern string) if they are not the same then exit the compare process. Then begin to compare from the position of y(y is the sum of the length of the pattern and the first position of the main string). The algorithm is as follows [10]

```
int Index_FL(SString S, SString T, int pos)
//then the first position of the substring T after the position pos of the main string
//S is returned. If there is
//no string T in the string S,then the return value is zero.
//String T is not an empty string.
{
  1≤pos≤StrLength(S).
  sLength=S[0];
  tLength=T[0];
  i=pos;
  patStartChar=T[1];
  patEndChar =T[tLength];
  while (i<=sLength-tLength+1)
{
    if (S[i]!=patStartChar)
           ++i;
    //finding the matching point again
    else if (S[i+tLength-1]!= patEndChar)
            ++i;
  //the last character of the pattern string does not match with the original
  //character.
else
   {
      k=1;
      j=2;
  //checking the situation of the interval characters matching
  while(j<tLength && S[i+k]=T[j])
     {
         ++k;
          ++j;
     }
  if (j==tLength)
      return i;
  else
     ++i;
  //restart to the next matching process
    }
}
return 0;
}
```

39.3.3 The Algorithm of KMP(D.E. Knuth, V.R. Pratt, J.H. Morris)

When S[i]!=T[j], there is the result: S[i..i+j-2]==T[1..j-1]. If T[1..k-1]== T[j-k+1..j-1],then S[i-k+1]==T[1..k-1]. The algorithm is as follows:

```
int Index_KMP(SString S, SString T, int pos)
  {
  //An algorithm named KMP which can compute the
//first position of string T in the string S afer the position pos using the pattern
//function. T is not an empty string.
//1≤pos≤StrLength(S).
  i=pos;
  j=1;
while(i<=S[0]&&j<=T[0])
   {
      if(j==0||S[i]==T[j])
          {++i;++j;}
      //continue to compare the next character
        else
         j=next[j];
     //move the pattern string right.
   }
   If(j>T[0])
     return i-T[0];
   //matching successfully
   else
      return 0;
   }
```

The definition of function next of the pattern string:

When j=1, next[j]=0;

When ‘$p_1p_2\ldots p_{k-1}$‘=’$p_{j-k+1}\ldots p_{j-1}$’

next[j]=max{k|1<k<j}; other situation next[j]=1;

The computation of function next is a process of recursion. The concrete process is as follows:

When next[1] = 0;

Suppose: next[j] = k; and T[j] = T[k]

Then next[j+1] = k+1

If T[j]!= T[k] then we will return to before. The essence of checking process is comparing, the difference is that the main string and the pattern string is the same. The compute algorithm of next function is as follows:

```
      void get_next(SString &T, int &next[])
        {
  //compute the value of next function of pattern string,then store the value into
  //the array with the name next.
```

```
        i=1;
        next[1]=0;
        j=0;
      while(i<T[0])
       {
        if (j==0||T[i]==T[j])
         {
           ++i;
           ++j;next[i]=j;
         }
       else
          j=next[j];
      }
  }//get_next
```

There is a special situation we must consider: S=‘ bbbbbbbbbbbbbbbbbbba’, T=‘bbbba’. According to the computation method of next function which is described above, the value of next[j] is 01234. There will be many compare times which are not necessary in this situation. For example, when the last character b is compared with a, we will find that they are not equal, so the function next will be computed. But the value of the next function will be 4 if the old method is used. That is the fourth character of T will be compared with the fifth character of string S, but we will find that the character a which will be compared with character b is the same to the character a which has been compared. So the character in the position may not be equal to the character a. The process of this comparison is not necessary. So in this situation the computation of next function will be repaired. If the next value of the current character is equal to the value of next of the former character, the value of next will be the current character’s next value. Otherwise the computation method of the next function will be according to the old way. The value of next function of the pattern string which has been repaired is nextval[j]=00004. The repaired method that is computation algorithm is as follows:

```
  void get_nextval(SString &T, int &nextval[])
     //compuate the repaired value of next function of pattern string T and store the
     //value into the arraywith the name nextval.
  {
   i=1;
   nextval[1]=0;
   j=0;
  while(i<T[0])
  {
   if(j==0||T[i]==T[j])
        {
         ++i;
           ++j;
       }
     if (T[i]!=T[j])
```

```
            next[i]=j;
        else
          nextval[i]=nextval[j];
}
   else
    j=nextval[j];
     }
   }
```

39.4 Conclusions

There are three different algorithms to solve the same problem, but the time complexity of them is different. The first algorithm that is the simple algorithm with the idea compares each character according to the sequence of from head to tail. Only if all of the characters are same, the conclusion of matching can be drawn. Its time complexity is O(m × n). The second method that is head–tail method can reduce the time of dealing question, but not at all the situations the compared times can be reduced. Its worst time complexity is O(m × n). So the time complexity of the head–tail algorithm is same as the simple algorithm. The compare think of KMP algorithm is more complicated than the other algorithm, it is achieved using the way of next function, with the smaller time complicity O(m + n). So in the teaching process of string, we can tell the compare algorithm sequence: simple algorithm, head–tail algorithm and KMP algorithm. The effect of teaching can be better than before.

Acknowledgments The authors would like to thank the anonymous reviewers for their constructive comments and suggestions on improving the quality of the paper. This work is supported by Scientific Research Fund of Hunan International Economical University (Project name: The research of program design of contract based on Java modeling language, No. 4), Hunan Provincial Natural Science Foundation of China (No. 10JJ6092).

References

1. Yan W, Wu W (2002) Data structure (C program language edition)[M]. Tsinghua University Press, Beijing
2. Yan W, www.freekaoyan.com data structure teaching notes[EB/OL]
3. Zhou H, Chen H (2010) Reflection on teaching of data structure course [J]. Theory Pract Edu 30(6):62–63
4. Geng X (2007) Research and exploration of data structure teaching[J]. J Changchun Norm Univ (Nat Sci) 26(3):104–105
5. Fan J, Chen W, Xu X, Yu X, Hou Z (2010) Suggestions of teaching and learning data structures and algorithms on the college[J]. Comput Edu 16:17–20
6. Walter S (2003) JAVA-an introduction to computer science and programming[M]. Higher Education Press, Beijing

7. Bruce E (2004) Think in Java[M], 3rd edn. Higher Education Press, Beijing
8. Zhang F, Huang Z, Wang Y (2010) The research of teaching reform of data structure[J]. J Tianzhong 25(5):87–88
9. Wang W, Zhang L, Shi Y, Zhang H (2010) Exploration about the practice teaching of data structure[J]. Comput Education 13:155–157
10. Li W, Zhang M, Ji Z (2010) Study and practice of methodology in teaching of the introduction section of data structure[J]. Comput Edu 16:75–78

Chapter 40
Design and Research Intelligent Answering and Voter-Timing Machine Based on AT89C51 MCU

Xiaokan Wang, Zhongliang Sun and Lei Wang

Abstract An Intelligent Answering and Voter-timing Machine Based on AT89C51 MCU is designed. Priority encoder circuit, the latch and the decoding circuit are designed, respectively, and in the same the input signal of participating team is separately diplayed on the monitor. The alarm circuit is started by the control circuit and the host switch, timing circuit and decoding circuit can generate second-pulse output signal to achieve timing function. After the experiment analog and simulation, the results shows that the answering has characteristics with quick, multi-functional and strong practical features.

Keywords AT89C51 MCU · Answering · Timing · Alarm · Sequence control

40.1 Introduction

Regardless of the school, the factory, the army or the educational television program, they will host a variety of intellectual competition, then they will be used in the occasion [1, 2]. At present there are various intelligent competitive answering machines in the market, but the majority products are early designed by analog circuits, digital circuits and the combination of analog circuits and digital circuits.

X. Wang (✉) · Z. Sun · L. Wang
Henan Mechanical and Electrical Vocational Group,
Zhengzhou, Henan, China
e-mail: sunzhl2008@126.com

Z. Sun
e-mail: wxkbbg@163.com

L. Wang
e-mail: wangxiaokan@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_40,

These answering machines are already quite mature, but the function of the circuit is relatively more complex, high cost, many break downs and simple display (some do not even have display circuit), so it cannot judge the behavior of pressing the button early and is not easy to upgrade the circuit [3]. With the rapid development of science and technology in the recent years, the applications of the microcontroller are going deeper which simultaneously leads the traditional control test to change every day and the answering machine has a breakthrough development.

40.2 System Schematic Design

40.2.1 Analysis of Answering Machine System Diagram

Answering machine is a priority decision circuit designed for answering first of the intellectual competitors [4, 5]. The intellectual competitors can be divided into several groups (eight groups), each group is controlled by an answer switch which respectively is SW0, SW2, …, SW7. The competitor of each group needs to judge and answer the questions presented by the director in a short time, and then press first answering button and answer the question. The group number will display in the monitor when the first person presses the button, and the circuits of other groups will be blocked at the same time. After answering questions, the director will restore all the buttons and the system will be ready to begin the next answering round. The system has two ways to remind the first answering: first, the speaker will produce the “Music” warning when the first person presses the button; the other, encoding and decoding circuit will display the number of the corresponding group when the output pin brights the LED [6]. The system circuit is composed of input switch array, answering channel, voting channel, discriminant group control, sound tips and LED display. The answering machine system schematic is shown in Fig. 40.1.

40.2.2 Design of Overall System Circuit

The overall system circuit is shown in the Fig. 40.2 which mainly including display module, answering module and voting and timing module.

40.2.3 Answering Channel

In order to achieve the eight answering functions we could use input pins of two interrupt requests in the microcontroller and the pulse input pins of its timer/counter T0 and T1 may extend input pins of two interrupt requests. By this way we can fully

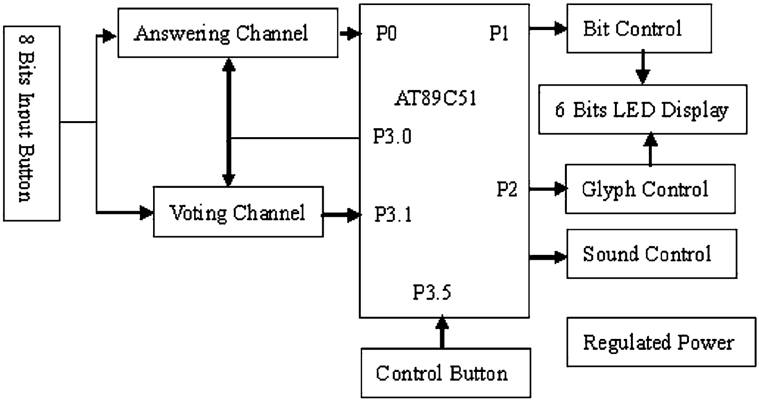


Fig. 40.1 Answering machine schematic system

use interrupt characteristics to respond to answering signal in time and reduce the error caused by the small difference time of two groups. If the answering machine groups are more than four groups in the system, we need to add a logic gate circuit before the input pins of four interrupt requests. Realization of this design is through by SW0, SW2, …, SW7.

40.2.4 Voting Channels

Turn on the power system, the concentrator and voter reset and self-checking. After the initialization of voter, it will be in the communication receiver state (sign/confirm button beside the LED begin flashing [7]). After the initialization of concentrator, it is transferred to the line communications testing of voter and address frame is sent directly to check the voter. Address frame format is the address number of each voter which uses the address bits (the nineth bit is 1) way.

The voter will send the interrupt when it receives frame signal to determine whether it is compliant with the local address, if not, the information will be discarded, otherwise it continues to receive frame signal; if they meet, then turn to interrupt handling and sending "local normal" information, the frame format will be the slave address. After the concentrator receives the information and determines whether it is correct to store the correct "voting normal" information; then the concentrator begins to check the next voter. If the received information is not correct, concentrator will send another time after delaying for a period of time. If the two times' information are not correct, then this voter is breakdown. The breakdown state of voter machine will be expressed with a null byte. After the turnaround of the inquiry, the voter not asked by checking the results of earlier inquiries is finally confirmed for the first time. That is, analysis which just saved the voter status word to determine the voter is questioned.

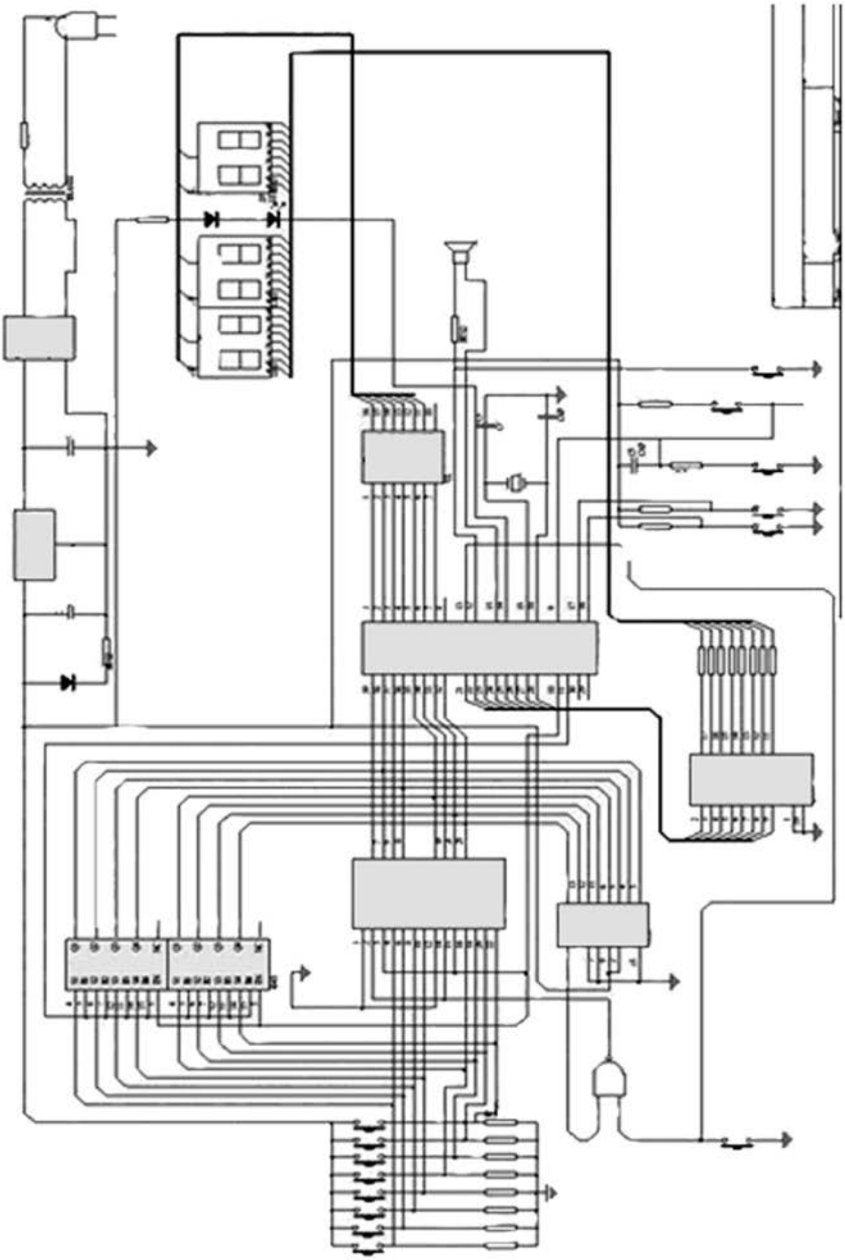


Fig. 40.2 The overall system circuit

40.2.5 Latch Section

Character display could use the output latch of serial-parallel 8-bit 74HC541 in the latch section [8]. Firstly, we must convert the characters into the corresponding shape code and send it to 74HC541 by the serial port, then 74HC541 would change the data received from the serial port into parallel output to the data tube (to provide the driver of a-dp). A common anode eight digital displays are applied in the display module.

40.2.6 Sequential Control Circuit

Sequential control circuit is the key of answering machine design which will mainly complete the following three functions.

When the director turns the control switch to "start" position, the speaker has sound and the first answering circuit and the timing circuit enters to the working condition of normal first answering.

When the players press the answering key, the speakers will sound, then the first answering circuit and the timing circuit will stop working.

The speakers will sound if nobody answers in the setting time, the first answering circuit and the timing circuit will stop working at the same time.

40.2.7 Bits Control

If P1 port outputs the shaped code, the digital tube will display the first answering group number. In order to improve the signal's driving capability, we would use the one-way driver 74LS244 to drive digital tube.

40.2.8 LED Display Module

The LED display screen can display the change of the numbers, text, graphics and video which not only can be used for indoor environment and but can also be used in outdoor environment. It has the incomparable advantages of projector, TV wall and LCD screen [9]. So the development prospect of LED is extremely broad, it is moving in the direction of higher brightness, higher resistance climate, higher luminous density, higher luminous uniformity, reliability and full color direction. This paper selects the LED because of its characteristics of high brightness, low working voltage, low power, miniaturization, long life, impact resistance and stable performance. In Fig. 40.2 when pressing the button, the data will be sent into AT89C51 through two-way I/O of P1, encoding and transforming the data into the corresponding binary code. Then the data will be transformed and locked

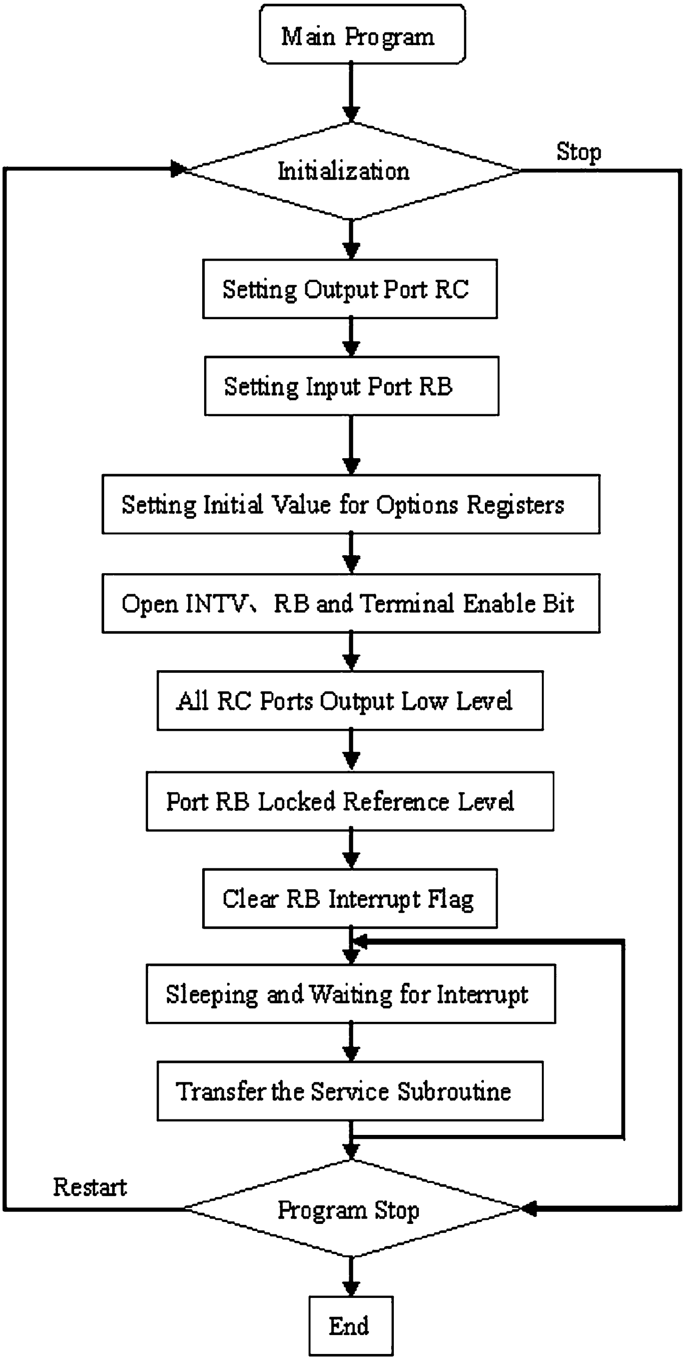

Main Program
Initialization
Stop
Setting Output Port RC
Setting Input Port RB
Setting Initial Value for Options Registers
Open INTV、 RB and Terminal Enable Bit
All RC Ports Output Low Level
Port RB Locked Reference Level
Clear RB Interrupt Flag
Sleeping and Waiting for Interrupt
Transfer the Service Subroutine
Restart
Program Stop
End

Fig. 40.3 The main program flow chart

parallel to 8-bit 74HC541 by the serial port P3.0 and P3. It simultaneously carries on decoding the binary code of the data of input digital display screen in the parallel form by the output port of O0-O6. Here, the hundredth position of digital display screen is controlled by P3.5, the tenth position is controlled by P3.4 and the unit of it is controlled by P3.3.

40.3 Software Design

The main flow chart of answering machine system software is shown in Fig. 40.3. It also added a pure 10 ms delay subroutine for preventing jitter.

40.4 Conclusion

This design uses AT89C51 microcontroller as the core of the logic control and signal generation which makes its race really just, fair, open and objective by taking full advantages of microcontroller. Even if the difference of several groups' answering time is a few microseconds, the system may also tell which group is in priority.

Acknowledgments This work was supported by the Innovation Fund for Small Technology-based Firms of China (No. 09C26214101775, No. 2009GJD00027), the Natural Science Foundation of Henan (102300410240), Innovation Scientists and Technicians Troop Construction Projects of Zhengzhou City (10CXTD160).

References

1. Li WY, Xie WC (2008) Design and realization of competitive answer machines based on EDA[J]. Sci Technol Eng (11):70–72, 76
2. Luo YX (2010) The design and simulation of the smart responder based on multisim 9[J]. Microcomput Inf 25(09):183–184, 200
3. Wang DW, Zhang JQ (2009) Design and realization of eight-way contest device based on singlechip[J]. J Jiamusi Univ (Nat Sci Edn) (03):32–34
4. Cheng QM, Chang L, Wang MM, Wang YF (2010) Design and implementation of competition answer based on the freescale 16-bit single chip microcomputer[J]. J Shanghai Univ Electr Power (03):71–75, 84
5. http://www.21IC.com
6. Cui BL, Zhang L (2009) Analysis and design of eight-bit SPC-QL electronic circuit[J]. Mod Electron Tech (20):206–207, 210
7. Qiao RF (2009) Design of the rush-answer machine based on MCU and PC[J]. Electron Sci Technol (08):46–49, 54
8. Zhang HR (2005) Electronic circuits and applications[M]. Tsinghua University Press, Beijing
9. Li J, Chen JB, Zhang LL (2010) Probability density function estimation of stochastic processes[J]. Chin J Appl Mech (03):53–57, 211–212

Chapter 41
The Research on E-Learning Interactive Technology Based on Discuz! Software

Yanshuang Zhou, Hong Li and Na Wang

Abstract Based on the existing achievement analysis of the current E-learning, this dissertation introduces the interactive function of the Discuz! software, realizes the E-learning interactive study system based on the discuz! software and elaborates the function of the E-learning module.

Keywords Discuz! Software · E-learning · Interaction · System design

41.1 Introduction

One of the main features of the information era is digitization. In the information era, the study is closely related with the development of the information techniques, which has multimedia and network technology as its core. The information technology is focused on digitization [1]. Due to the application of information technology to the education, the environment, resources and method of study are

Y. Zhou (✉)
School of Information Engineering, HanDan College,
HanDan City 056005, China
e-mail: shuangshuang223@126.com

H. Li
Electronic and Information Engineering Experiment and Training Center,
HanDan College, HanDan City, China
e-mail: lilihonghong@163.com

N. Wang
Modern Educational Technology Center, Henan Polytechnic University,
JiaoZuo City, China
e-mail: wangna@hpu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_41,

moving toward digitization, forming a digitized study environment, study resources and study method.

The CEO Forum on Educational Technology of America (abbreviated as ET-CEO forum) held the third annual meeting in June 2000 with the theme of the strength of digital study: the integration of the digital content [2]. This method of combing the digital technology with the teaching content is called digital study (E-learning). It presents the concept of E-learning [3]. Through the combination of digital technique and curriculum, it can create a digital study environment, and bring the digital resources and learning approach into the study of students. As a result, the school can make use of the information techniques efficiently in order to realize the full application of the information techniques. Then the students can study according to cooperation and creativity in order to achieve the purpose of training the creativity spirit and ability of the students.

The platforms of the current the worldwide popular E-learning are: WEB CT, Virtual-U, WISH, WEB Course in a Box, BlackBoard CourseInfo, LUVIT and Learning Space [4, 5]. These platforms require more expenses and advanced techniques, so the cost will be too high and reliance on IT will be more. Therefore, the author introduces an E-learning interactive learning system based on the technique of discuz!, aiming to overcome these difficulties.

41.2 The Analysis of the Key Techniques

41.2.1 Introduction of the E-Learning

E-learning is digital learning, which means studying in the environment of information techniques of network, communication, computer, artificial intelligence and multimedia [6]. It regards the information techniques as the tool of teachers and students through the networking education environment, digital education content, and intelligent learning tool based on the advanced education theory. All of these are aiming to realize brand-new methods for students and teachers so that the efficiency of education will be increased.

The design of E-learning teaching system is a comprehensive project for realizing the electronic teaching and study. It has the feature of networking teaching and pays more attention in controlling and researching the study process. This forms special networking education.

41.2.2 Openness

In the system of E-learning, both the teaching resources and objects are open. Some schools call this openness “3A”, that is, “Anytime”, “Anywhere”, “Anybody” [7].

41.2.3 Individuation

The design of E-learning system focuses on the students so that all of the study resources and content are related with the learners. They can choose their own learning methods and make the study plan based on their own situation and demands. At last, the students can start their own special study. In the E-learning system, the autonomous and individual learning can be reflected fully.

41.2.4 Interaction

The learners can report their progress freely by using the information techniques and network techniques; they can ask questions and apply courses. The teachers can be an expert group formed in the virtual environment, communicating with learners in an equal way. The teachers can teach students according to their aptitude and give them guidance and advices. As for learners, they can also do cooperative learning under the guidance of the teachers. This is the personalized feature. As a result, the learning progress could either be the learning behavior of individuals, teachers or the E-learning system, or it also could be an interactive study progress of a group.

41.2.5 The Introduction of Software in the Discuz! Forum

Crossday Discuz! Board abbreviated as discuz! 7.0 is a universal software system of community forum [8]. It is an open source software designed by Comsenz Inc. The users can build a forum with perfect function and strong load capacity by simply installing it. The infrastructure of discuz! adopts the most popular PHP + MySQL, it is an efficient solution of forum system which has perfect design and can be used in various environment of the server. This free software can be downloaded from www.discuz.com, it takes only 20 s to finish the whole installing.

41.3 The Realization of Interactive Function in E-Learning System Based on the Discuz!

41.3.1 The Introduction of the Interactive Study

Interactive study is the most important part of the E-learning system, providing a wider study space for learners.

Interactive study will break the traditional time concept. This breakthrough can be explained from two aspects: on the one hand, it shows the subjectivity and autonomy of the students. The students can carry the interactive study with teachers and classmates at any time; on the other hand, this study method and skill of network communication can satisfy the demands of lifelong learning.

Interactive learning will get unlimited expansion in the space. The distance-learning students can realize the interactive learning through the E-learning system only if they have access to Internet.

The study method will have essential change. In the interactive study, teachers still play an important role. They lead, stimulate and help students through E-learning system. The traditional relationships between students and teachers will be replaced by the democratic and equal friendship. The interactive study will provide more communication ways for teachers and students through establishing interactive network study environment. Through the system, the students can make a statement and ask questions in real-time, discuss the problems online as well as do some other non-real-time communication and discussing. There is no any united thoughts, languages and movements, instead it encourages personality, accepts multi-thoughts, forgives the mistakes and cultivates encourage. All of these aim to make the education a quality-oriented education that can cultivate scientific spirit and creativity.

The evaluation method of interactive study will be benefit for reforming the traditional evaluation model. In the interactive learning, the examination is not the only standard to judge the students' performance. The learning effect will be measured by the activity of students, their knowledge and the ability of solving problems in the virtual environment. Research and analysis: E-learning system, which is based on the discuz! software can provide efficient communication for students, teachers and schools; it plays an important role in increasing the activity of students and ensuring the study quality.

41.3.2 The Realization of Interactive Function in E-Learning System Based on the Discuz!

The core subject of the system construction is interactive communication, which is designed for Internet interactive communication. The users can discuss the problems of writing online through the function of SMS and message replying in the discuz!. Users can issue their own works by using the function of posting. The teachers can modify the exercises through replying and other users can express their own opinions on it. In addition, the students can modify their own exercising at any time. From this, we can see that the E-learning system which is based on the discuz! can remove the distance of time and space between students and teachers so that they can communicate with each other at anytime and anyplace. It contains the whole process of publishing, discussing, resources sharing, communicating, coaching and modifications and saves it on the Internet. The relationship between students and teachers or students and students is both way interactive. In the interactive study of E-learning, all of the posts are spectacular because both the teachers and students are in a multi-interactive network study. The people can communicate with each other directly. This interactive process may be recessive (identified by the number of visit) or dominant (identified by the response).

41.4 The Structure Construction of the E-Learning Based on the Discuz!

41.4.1 System Construction Based on the Discuz!

E-leaning system requires installing the discuz! first for establishing the forum of the E-learning system [9]. It will identify the authorities through the logging. Non-registered users can only browse the information, but registered users can issue, reply, upload and download except for the authority of non-registered user.

Registered users of discuz! have following features:

Choose interested discussion group.
Select some interested content to read.
Issue some own opinions in the discussion group.
Allow to publish various notice or send some news and information.
Express own opinions on some subjects.
Material sharing (software, document, materials, etc.).
Send and receive internal E-mail.
Real-time chatting in the website.

Upload and download resource. Because the registered users can discuss openly, and upload and down load resources, it realizes the resource sharing in the E-learning system or non-real-time communication.

Take the courses of educational technology as an example, it designs an E-learning system based the discuz! (See Fig. 41.1).

41.4.2 Function Introduction of System Module

In order to ensure the safety of teacher management module, the password of administrators should be changed and a backup of the data of the forum should be made at regular intervals. The teachers can set the forum moderator and let the moderator to maintain their own pages in order to reduce the teacher's workload. Meanwhile, the teachers have to publish Tehran teaching resources and manage the questions of students. The courseware-released module provides a back-stage

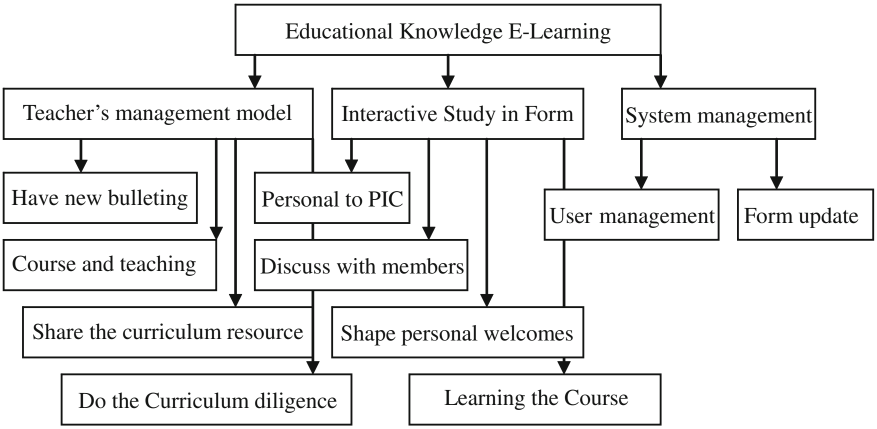


Fig. 41.1 Structure chart: E-learning functional module of the education technique courses based on the discuz! Forum software

management for teachers. Even the teaching do not understand the network techniques, they can still display their teaching plan, key to the exercises and multimedia courseware on the website so that the students can study at anytime. As for the students' management module, it is used to record the problems of study. The teachers can answer these problems in classroom or on the website. Resource management module provides guidance material, teaching cases, software of learning tools, the speech record and achievement. The resources are recommended by teachers or collected by learners. Through the module, the teachers can modify, delete, edit the resources online and upload attachments. The students can browse, comment, download, upload and search the related teaching resources.

Interactive learning module is the core area for the learners to learn knowledge, exchange ideas, answer questions and construct knowledge, which mainly includes three submodules, question and answer online area, discussion area and study and evaluation. In the question and answer online area, teachers will organize students to participate in real-time question and answer regularly. And according to the characteristics of the mechanical courses, the module mainly aims to realize the real-time browse of mechanical graphics online, edit, remote network assistance feature on the basis of the text and voice communication. Forum is a mature platform for open asynchronous communication with key functions to set subforum, post topics, reply to topics, combine topics, upload attachments, edit, delete and search. While question setting and discussion sub-module is mainly used for synchronous communication, where both real-time discussion of knowledge and regular the network meetings organized by teachers or team leaders are available. Moreover, the speech content of this module will be tracking module and marking module. Learning process tracking sub-module records the times of the learners' and teams' access to the system, use time, the amount of resource access, the number of times to participate in discussion and exchange, resource uploading and the number of documents. While marking submodule's major functions include self-assessment of learners, peer assessment of team members, assessment of members beyond the team, teachers' assessment, and evaluation team, providing the evaluation methods of marking and making comments. Each learner's final scores include the individual scores drawn from self-assessment, peer assessment of team members, assessment of members beyond the team, and teachers' assessment and pulsing the weighted scores derived from peer assessment of team members, assessment of members beyond the team, and teachers' assessment, etc.

The system administration module provides the user management and system's use help. The User administration module provides different user the different registration and the debarkation jurisdiction management. The identities of login interface are different for different status's user. This is recorded by the system for later browsing. The study and evaluation of submodule mainly includes learning process-system divides the user into three types: teachers, students, and administrators, and each of them have different privileges, in which the administrators have the highest authority. The administrators could carry on the management of the system and can participate in the knowledge library's construction. The teachers are mainly involved in knowledge library's construction activities, and should answer the questions raised by students and carry on the explanation. The student user may study each item of resources provided by website, and may also discuss mutually or inquire to the teachers. The help instruction module provides system's use help, the study strategy instruction issue, the browsing retrieval, the instruction of help seeking and the real-time early warning. The help information could be established into different instruction scope, including the public help information, the group help information and individual help information, which presents for different user.

41.5 Conclusion

The requirements of the E-learning system based on the discuz! is low. It will bring good interactivity, safety and stability to online education by making full use of the advantages of discuz!. Moreover, it can also realize the communication of students and teachers or students and students, exploit the huge potential of exploration learning and cooperative learning and improve the education and result of e environment.

References

1. Discuz! Home: http://www.comsenz.com
2. Wang L, Zhang X (2010) The study of the length approach in college english based on discuz!7.0[J]. Crazy Engl Teach 1:013–014

3. Zhang J, Xingjian S (2010) The design of medical informatics BBS based on discuz! [J]. Sci Technol Inf 15:145–146
4. Di A (2008) The study of real-time interactive about the network education platform based on discuz! [J]. Pioneer Sci Technol Mon 12:37–38
5. Minghui B, Yongliang W, Wang L (2009) Design and implementation of E-learning system based on Web service [J]. Comput Appl Softw 5:226–231
6. Yang W, Rong Y (2002) The research and design of intelligent E-learning [J]. Comput Eng Appl 38:078–083
7. Tao X, Wang X, Wang L (2008) Design and implement of E-learning system of mechanical teaching. J TianJin Polytech Univ 27:44–48
8. Xu W (2009) Web service oriented e-learning system [J]. J Shenyang Norm Univ (Nat Sci) 7(3):53–56
9. Payne M, Stephenson JE, Morris WB, Tempest HG, Mileham A, Griffin DK (2009) The use of an e-learning constructivist solution in workplace learning [J]. Int J Ind Ergon 39:548–553

Chapter 42
Analysis of Drop-Out Due to Diseases in Key Primary and Secondary Schools of Lubei District of Tangshan City From 2003 to 2008

Zhou Lei, Wang Xiaohong, Li Huilan, Zhang Guobin and Wu Jianhui

Abstract This Chapter aims to calculate the students' drop-out rates caused by various diseases and make the comparison between the drop-out rates of primary and secondary schools ($\chi2$ test) and to understand the current situation of drop-out due to diseases in key primary and secondary schools of Lubei district of Tangshan city from 2003 to 2008. The results show that mental illness led to the lowest drop-out rate of students, while cardiovascular disease led to the highest. And drop-out rates caused by various diseases was lower than 5‰ in total. There were no significant differences between the drop-out rates of primary and secondary schools. The disease spectrum in children and adolescents has changed, and cardiovascular disease has replaced infectious disease to be the main disease, causing students to drop-out. So it is imperative to strengthen the prevention and treatment of cardiovascular disease.

Keywords Disease · Drop-out · Analysis

42.1 Introduction

Drop-out due to diseases is one of the reasons affecting the normal life and learning in primary and secondary school students. Better understanding of students' drop-out due to diseases is necessary to explore and to take appropriate preventive measures to reduce primary and secondary students' drop-out.

Z. Lei (✉) · L. Huilan · Z. Guobin · W. Jianhui
ebei United University, Tangshan 063000, Hebei, China
e-mail: ncmcyjsb@163.com

W. Xiaohong
Tangshan Centers for disease control and prevention,
Tangshan 063000, Hebei, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_42,

The current situation about drop-out due to diseases in key primary and secondary schools of Lubei district of Tangshan city from 2003 to 2008 was deeply analyzed.

42.2 Subjects and Methods

42.2.1 Subjects

Students in key primary and secondary schools of Lubei district of Tangshan city from 2003 to 2008

42.2.2 Materials

According to the 3rd item of table 26 in health statistics, questionnaires on drop-out due to diseases in key primary and secondary schools of Lubei district of Tangshan city from 2003 to 2008 were filled in, which included some reasons leading to students' drop-out, such as mental illness, infectious disease, cardiovascular disease and other diseases.

42.2.3 Diagnostic Basis

The diseases leading to students' drop-out were approved by relative hospitals.

42.2.4 Statistical Methods

All data were put into computer to calculate the students' drop-out rate. The comparison between the diseases leading to students' drop-out was carried out (Fisher exact test).

42.3 Results

42.3.1 Situation on Drop-Out Due to Diseases in Key Primary and Secondary Schools of Lubei District of Tangshang City from 2003 to 2008

Table 42.1 showed that the total drop-out rate of primary and secondary school students was close to 4‰. Mental illness led to the lowest drop-out rate of students, followed by infectious disease, while cardiovascular disease and other diseases led to higher.

Table 42.1 Situation on drop-out due to diseases in key primary and secondary schools of Lubei district of Tangshang city from 2003 to 2008

Study stage	Year	Number of schools	Number of students	Mental illness n(‰)	Infectious disease n(‰)	Cardiovascular disease n(‰)	Other diseases n(‰)	Total amount rate n(‰)
Primary	2003	3	2,625	0(0.00)	1(0.38)	6(2.29)	5(1.90)	12(4.57)
	2004	1	637	0(0.00)	0(0.00)	5(7.85)	4(6.28)	9(14.13)
	2005	2	1,317	1(0.76)	2(1.52)	2(1.52)	2(1.52)	7(5.32)
	2006	2	2,084	0(0.00)	1(0.49)	1(0.49)	2(0.98)	4(1.92)
	2007	2	2,064	0(0.00)	2(0.97)	1(0.48)	0(0.00)	3(1.45)
	2008	2	1,697	0(0.00)	2(1.18)	2(1.18)	0(0.00)	4(2.36)
	Total	12	10,424	1(0.10)	8(0.77)	17(1.63)	13(1.25)	39(3.74)
Secondary	2003	3	3,736	0(0.00)	1(0.27)	3(0.80)	7(1.87)	11(2.94)
	2004	1	1,210	0(0.00)	1(0.83)	2(1.65)	8(6.61)	11(9.09)
	2005	2	2,649	2(0.76)	0(0.00)	6(2.27)	7(2.64)	15(5.66)
	2006	2	3,044	1(0.33)	0(0.00)	5(1.64)	6(1.97)	12(3.94)
	2007	2	3,024	0(0.00)	2(0.66)	6(1.98)	0(0.00)	8(2.64)
	2008	2	3,647	0(0.00)	2(0.55)	5(1.37)	1(0.27)	8(2.19)
	Total	12	17,310	3(0.17)	6(0.35)	27(1.56)	29(1.67)	65(3.76)

42.3.2 *Comparison Between the Diseases Leading to Students' Drop-out*

Table 42.2 showed that Fisher exact test indicated that there were no significant differences between the drop-out rates caused by the above mentioned four diseases and the total drop-out rate in primary and secondary school students ($P > 0.05$).

Table 42.2 Comparison between the diseases leading to students' drop-out

	Drop-out rate caused by mental illness (‰)	Drop-out rate caused by infectious disease(‰)	Drop-out rate caused by cardiovascular disease (‰)	Drop-out rate caused by other diseases (‰)	Total drop-out rate (‰)
Primary school	0.10	0.77	1.63	1.25	3.74
Secondary school	0.17	0.35	1.56	1.67	3.76
P	0.517	0.110	0.504	0.236	0.534

Table 42.3 Sequence of diseases causing students' drop-out

Disease	Number of drop-out students	Percentage (%)
Mental illness	4	3.85
Infectious disease	14	13.46
Cardiovascular disease	44	42.31
Other diseases	42	40.38
Total	104	100.00

42.3.3 Analysis to Sequence Diseases Causing Students' Drop-out

Table 42.3 showed that cardiovascular disease led to the highest drop-out rate of students, followed by other diseases, infectious diseases and mental illness.

42.4 Discussion

With the rapid development of society and economy, health levels and people's living standards have been greatly improved, and the healthcare measures of children have gradually perfected. And the disease spectrum of children and adolescents has changed. It is reported that in Futian district from 1998 to 2001, among the diseases causing primary and secondary school students' drop-out, infectious disease ranked first, followed by cardiovascular disease [1]. study results on primary, secondary and college students from 1996 to 1999 in Shanghai also showed that infectious disease was the most serious disease [2]. Our results showed that cardiovascular disease has replaced infectious disease as the main disease causing students' drop-out, which was in consistent with some reports [3]. The main reason lied in China's more emphasis on the prevention and research of infectious disease so as to reduce the prevalence of some infectious diseases such as tuberculosis and hepatitis in the past 10 years. Medical research has found that cardiovascular disease has great relation with genetic causes, and obesity and dietary factors are its main incentive. Research showed that obesity rates increased year by year in primary and secondary schools [4, 5], so it is necessary to prevent obesity in primary and secondary school students. Obesity is mainly caused by the lack of knowledge on a reasonable nutrition, by too much calorie intake, especially sweets intake, by less time devoted to the physical activities which reduce the heat consumption; and even by general food intake. Prevention and treatment to obesity includes launches nutrition education and reasonable arrangements for meals, encouraging more participation in sports activities, advocating more reasonable and balanced diet and so on.

Other diseases mainly included accidental injury, skin diseases, urinary diseases, disorders and other facial features. Our results showed that other diseases (ranking No. 2 in proportion), cannot be ignored. We should strengthen education on safety and health knowledge to enhance self-protection capability of the primary and secondary school students.

Although infectious disease ranked No.2 from the bottom, it still accounted for higher percent (13.46%). This study mainly defined urban primary and secondary school students as subjects. It was reported that [6], the situation of drop-out due to infectious disease in rural primary and secondary school students was more serious. So it is suggested that we should continue to strengthen the prevention and research and control of infectious diseases, particularly in rural primary and secondary school students.

References

1. Wang X, Kai Z, Liang H et al (2003) Analysis of school drop-out due to sickness, of death in Futian district from 1998 to 2001. Doctor 17(1):53–54
2. Gendi G, Peng N, Zhou Y et al (2000) Analysis of school drop-out due to sickness and of death in Shanghai. Shanghai J Prev Med 12(11):514–515
3. Huang K, Xu L et al (2003) Analysis of high school students' drop-out due to illness. Chin Sch Health 24(3):284
4. Sheng L, Yang B, He J (2000) Dynamic analysis of common diseases in students in Henan province from 1991 to 1999. J Prev Med 11(6):347–348
5. Zhou S, Li Y (2002) Analysis to monitoring results of common diseases in students of Dezhou city. Chin J Health Insp 9(3):156–157, 154
6. Song H (1999) Analysis to drop-out due to sickness in Hubei province in1998. Hubei J Prev Med 10(6):35

Chapter 43
Design and Development of University Educational Administration Information System Based on J2EE

Lanqing Liu and Sanyou Ji

Abstract With the rapid development of education, the informationization process of university educational administration is growing, which accelerates the reformation of education methods, teaching process and management mode. To cater to the new requirements of this time, the university educational administration information system is studied in this chapter. The demand of the system construction, the main technical problems and solutions during its development process are analyzed. By using the J2EE-based platform, a reasonable system structure is designed, and a teaching prototype system is devised and realized finally.

Keywords University educational administration information system · Information technology · J2EE

43.1 Introduction

These years, with the promotion of higher education and the consequently expanded enrollment, a lot of problems arise in the university educational administration system, such as the increasing management workload, the shortage

L. Liu (✉) · S. Ji
College of Logistics Engineering,
Wuhan University of Technology, Wuhan, Hubei, China
e-mail: bingheku6@sina.com.cn;hslgllq@163.com

S. Ji
e-mail: jisanyou@126.com

L. Liu
Academic Affairs Office, Huangshi Institute of Technology,
Huangshi, Hupei, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_43,

of teaching resource and the poor teaching quality [1]. Thus educational administration faces an unprecedented challenge. As an important tool of educational administration, educational administration information system plays a significant role not only in improving the accuracy and promptness of educational administration, but also in realizing of scientific management, green management and digital management.

According to the universal law of university educational administration, the design and implementation of university educational administration information system are discussed in this Chapter. The aim is to achieve centralized management, decentralized operations and information sharing, to make the educational administration digital, intelligent and comprehensive, to reduce workload and management costs and to improve speed and accuracy of information handling, office efficiency and management level, and thus achieve the modernization of educational administration.

43.2 Key Technologies

The users of educational administration information system vary from educational administrator to teachers and students. Each role has different requirements. For instance, in the system, teachers need to view their teaching work, calculate teaching workload, record students scores and so on; students require to inquire their scores, choose courses, evaluate teaching level and so on. The specific requirement of Educational Administration Information System is designed and developed using J2EE technology in this chapter [2].

J2EE defined a wealth of technical standards for realizing distributed application. It provides development tools which meet these standards and API to support the development enterprise application. These techniques contain several aspects including database access, distributed communication, security and so on. Application based on component/container is the heart of J2EE technology [3]. Component represents units which can be developed and distributed in multiple languages. Container represents a software entity which is running on server and used to manage specific type of components.

J2EE provides four different types of containers to manage the corresponding components:

Application container: to manage independent Java applications.
Applet container: to provide an execution environment for Applet.
Web container: to manage Web components, such as Servlet and JSP.
EJBcontainer: to manage EJB components, such as conversation Bean and entity Bean.

Through these components and containers, J2EE architecture can not only provide independence while developing and deploying but also provide portability among different types of middle-tier server.

43.2.1 Servlet

Servlet is some Web components which can generate dynamic content. It is used to develop server applications and extend server functionality. It provides an effective mechanism which is used for server-based business logic and for interactions between Web-based clients. Servlet can also provide a light and manageable alternative to common CGI scripts.

43.2.2 JSP

JSP is another type of J2EE Web components. It is composed of HTML page and the embedded Java code. It can receive request from client and generate HTML response page dynamically. In addition to the function of packaging and extending Servlet, JSP provides a flexible way of writing that combined front-end static HTML with back-end program.

43.3 System Design

43.3.1 Overall System Design

According to the routine tasks, staff and data involved, business process flow and personal work experience, the overall system design is shown in Fig. 43.1:

43.3.2 Function Modules Design

The function modules design is illustrated in detail by taking planning management module for example [4].

Planning management module. This module is the precondition to realize sectors including student's course selection, scheduling and so on. Student's training plan is transformed into semester plan through the module, so that the teaching plan can really become the guidance documents during whole teaching procedure. The teaching task of each semester can be generated in accordance with the teaching plan. Through this, the relevant departments can draw up course offering plan, teaching venues request, teachers and choosing teaching classes.

Courses database management. The course database contains every course and its public properties involved in the educational system. It is the most important data which runs through the whole procedure of system management, including teaching plan, course scheduling, course selection, test, scores and graduation audit management.

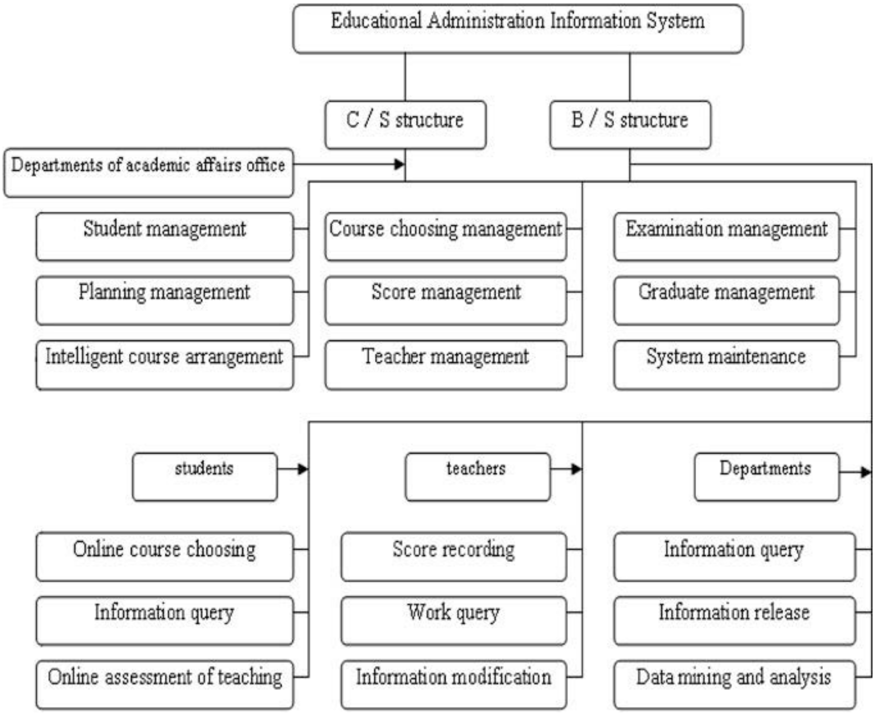


Fig. 43.1 Framework diagram of Educational Administration Information System

Professional planning management. According to the teaching plan and provisions of the school's student management, this module can determine the various types of credits needed in each learning stage for students of different professionals and grades, individual learning plans can be made flexible in keeping with individuals. Besides, this module can monitor status of students learning for ensuring implement of the credit system.

Teaching task management. Set up teaching task to meet the requirements of every major's teaching plan for each semester. Then the teaching task required can be implemented by the relevant departments, and classes can be generated and teachers can decide to provide resources for student's course choosing (as is shown in Fig. 43.2).

43.3.3 System Implementation

The business layer provides "data transfer" function. It can send back the required data which is obtained from the data layer to client layer in response to its order. Besides, the business layer also provides business rule checking for the data

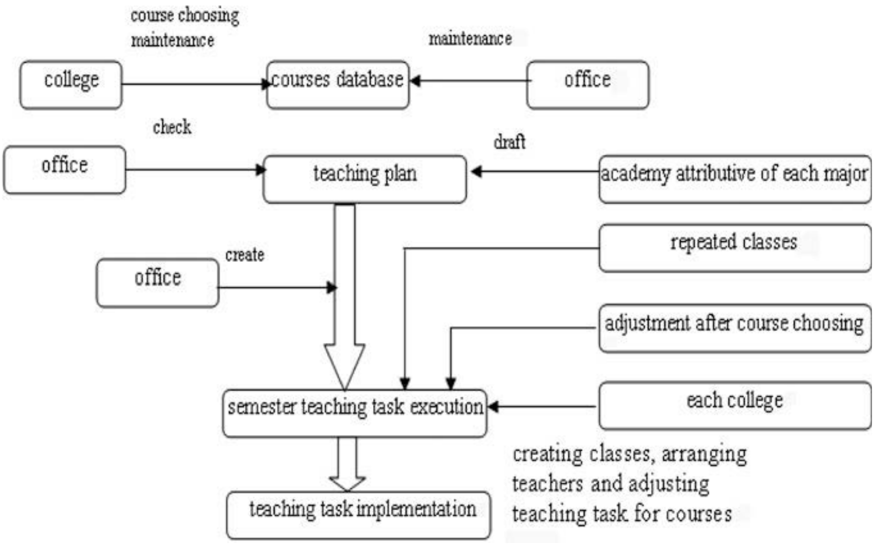

course choosing maintenance
maintenance
college
courses database
office
check
draft
office
teaching plan
academy attributive of each major
repeated classes
office
create
adjustment after course choosing
semester teaching task execution
each college
creating classes, arranging teachers and adjusting teaching task for courses
teaching task implementation

Fig. 43.2 Flow chart of the teaching task implements

submitted by client layer, so that the data can be submitted to the data layer for storage only if it is consistent with the rules. Therefore, business layer can be considered as a logical bridge between client layer and data layer.

In J2EE framework, Servlet and JavaBean can be used to realize the middle logical layer, it is popular in the small- and medium-enterprises system. In this way, JavaBean is used to encapsulate the logical operations of database, Servlet is used to control business processes and the intermediate operations are implemented by the using of JavaBean (code DAO and DTO). The program is modularized through this way which makes the program simple and clear.

The first page of the system is login interface. Users should fill the user name and password, select user type and then submit to the system. System would search in the login table corresponding to the user type. If the information is correct, the user will be allowed to login, and if not, the user will be required to log in again with an error warning. After successful login, the user permission will be recorded in the system for the verification in each of the following interface. User permissions are determined by user ID which is stored in the database in advance. User permissions include: student user, teacher user, teaching secretary user, educational administrator user and system administrator user.

Some of the key code is as follows.

```
// Servlet to control login processes
public class LoginAction extends javax.servlet.http.HttpServlet implements ActionInterface,javax.servlet.Servlet {
protected void doGet(HttpServletRequest request, HttpServletResponse response) throws ServletException, IOException {
        process(request,response);
  }
  protected void doPost(HttpServletRequest request, HttpServletResponse
response) throws ServletException, IOException {
  process(request,response);
  }
  // Request of client browser is encapsulated as a HttpServletRequest object
which contains all the information including address and parameters of the request,
data submitted, files uploaded, client ip and the client operating system.
  public String process(HttpServletRequest request, HttpServletResponse
response) {
            // Create user session object
  HttpSession session=request.getSession();
            Map map=new HashMap();
            try {
                    boolean isLive=false;
                    // Get the inputted user ID and make simple verify
                    String id=request.getParameter(“p_id”);
                    if(id==null||id==““){
                            return “index.html”;
                    }
                    int p_id=Integer.parseInt(id);
                    // Get the inputted password and make simple verify
                    String password=request.getParameter(“password”);
                    if(password==null||password==““){
                                    return “index.html”;
                    }
                    // Get the inputted validation code and make simple verify
                    String random=request.getParameter(“randomcode”);
                    String randomcode=(String) session.getAttribute
(“randomcode”);
                    if (!randomcode.equalsIgnoreCase(random)) {
                            return “index.html”;
                    }
       // create a database connection
                    Connection con=
DataSourceFactory.getDataSourceFactory().getConnection();
                    // create DAO for manipulating the login information table of
database
                    LoginDAO logindao=
DAOSourceFactory.getLoginDAO(con);
                    // Traversing the login information table, comparing with the
inputted user ID and password to verify if it is legitimate user
       List list=logindao.login(p_id, password);
                    Iterator its=list.iterator();
```

```
                while(its.hasNext()){
                      LoginDTO dto=(LoginDTO)its.next();
        if(dto.getP_id()==p_id&&dto.getP_password().endsWith(password)){
                                 isLive=true;
                       }
                }
  // If the user authentication failed, remained in the home
                if(!isLive){
                      return "index.html";
                }else{
  // If the user authentication passed, the user role and permission will be
acquired from the relevant tables for the verification in each of the following
interface.
                        List listusermanage=
logindao.findusermanage(p_id);
                        map.put(p_id, listusermanage);
                        session.setAttribute("p_id", p_id);
                        session.setAttribute("managemap", map);
                }
         } catch (SQLException e) {
                // TODO Auto-generated catch block
                e.printStackTrace();
         }
  // user authentication succeed, jump to another interface
         return "first.html";
         }
  }
```

43.4 Summary

A deep-going study on university educational administration information system is made in this chapter. The system is designed on the basis of characteristics and rules of educational administration and implemented by the use of J2EE. Along with the rapid development of higher education, the exploitation and application of educational administration information system is becoming more and more important in the daily work of educational administration. Only through the combination of modern computer technology, network technology, communication technology and educational administration, the pace of informationization of university educational administration can be accelerated and the quality of higher education can be improved consequently.

References

1. Zhou X (2008) Web information systems [M]. VLDB Database School, Beijing, China
2. Couch J (2002) J2EE Bible (translated by Ma L). Publishing House of Electronic Industry, Beijing, China
3. Zhao Z (2005) The research and realization of educational administration system of the academic years credit system. Guangdong University of Technology, Guangdong, China
4. Peng D, Minglan S (2001) Plan and research of the information management system of teachers. J Archit Educ Inst High Learn 5(13):56–58

Chapter 44
Analysis and Evaluation of Order Based on Rough Sets

Zhiyuan Zhao, Shujuan Li, Mingshun Yang and Qilong Yuan

Abstract Focus on the evaluation index system for customer orders are too complicated and the evolution weights are too subjective, this paper presents a method. On the basis of the original data, the indexes are reduced based on the rough set theory according to the importance of the indexes, objective weights of reduced the indexes are, the priority sequence of multiorders is assessed by constructing an object close to the ideal point of degree programs on a number of priority assessments. A case study shows that this method is effective to assessment orders.

Keywords Order assessment · Rough sets · Index weight · Close-degree

44.1 Introduction

With the technology developing, global competition and market environment changing rapidly and customer's demand for diversified and personalized, the manufacturing business is becoming increasingly complex. More and more

Z. Zhao (✉) · S. Li · M. Yang · Q. Yuan
The School of Mechanical and Precision Instrument Engineering,
Xi'an University of Technology, Xi'an, China
e-mail: ymz@163.com

S. Li
e-mail: shujuanli@xaut.edu.cn

M. Yang
e-mail: msyang@xaut.edu.cn

Q. Yuan
e-mail: qlyuan@xaut.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_44,

enterprises will adopt the way make-to-order to respond to these changes in order to gain a competitive advantage [1].

According to the way make-to-order, priority decision of the order becomes the core of manufacturing demand management, and is most important issue to consider when the production plan is made. Based on the manufacturing capacity of enterprise, the sequence of order to arrive and the due time of each order, enterprises must determine the process sequence of the orders and meet customer needs as much as possible within its ability. These will make the enterprises to better achieve fast and accurate response to customer demands, shorten delivery time and quickly occupy the market, improve the degree of information integration and enhance their ability to respond to the market [2–4]. Traditional evaluation methods for orders which include comprehensive scoring method, standardized scoring method and the efficiency coefficient method [5] and modern evaluation methods for orders which include principal component analysis, factor analysis, hierarchical analysis and the concept of entropy-based evaluation for small volume orders etc. [3, 4], but these evaluation methods have some deficiencies such as the index is too complicated, and redundancy.

In a comprehensive evaluation of customer orders, evaluation index system is extremely complicated due to the diversification of customer information, and there are often some redundant indicators which not only increases the cost of research and statistics, but also easy to make some important indicators are weakening; thus, these indicators will lead to evaluation of unreasonable results. In this paper, the manufacturing enterprises with the way make-to-order as the background applies rough sets reduction method to evaluate index system for customer orders reduction; focus on the shortcomings that the weights of each indicator to determine is too subjective in the practical application, the objective weight of each indicator will be determined by the importance of indicators then the method which is close to the the ideal point is used to comprehensively evaluate customer orders.

44.2 Rough-Set-Based Attribute Reduction

44.2.1 Overview of Rough Set

The rough set theory is used to study imprecise, inconsistent, incomplete and other incomplete expression of knowledge and information, learning, theory and method of induction. It does not require prior knowledge and only uses the information provided by the data itself, while the premise of retaining critical information is reduction of the data and to obtain the minimum expression of knowledge. It can also identify and assess the dependence between the data, reveal a simple conceptual model and obtain knowledge which is easily confirmed from the empirical data [6, 7]. Therefore, the priorities of orders from customers will be evaluated based on rough set in this study.

Suppose there are information systems $\{U, C, V, F\}$, where: $U = \{x_1, x_2, \ldots, x_n\}$ is non-empty finite set called the domain; $C = \{c_1, c_2, \ldots, c_m\}$ is non-empty, finite set of attributes; V is attribute value, for $c \in C$, V_c is the range of attributes c, the attribute value for objects i under the condition j is $v_{ij}(i = 1, 2, \ldots, n, j = 1, 2, \ldots, m)$; $f : U \times C \to V$ is an information function that is denotes for every x; as the attributes value V_c is a continuous real value, this information system is a real property information systems. As for this real property information systems, the definition for no distinguish relation $\operatorname{ind}(P)$ of attribute set P is $\operatorname{ind}(P) = \{(x_s, x_t) \in U \times U | \forall c \in P, f(x_s, c) = f(x_t, c)\}$. If$(x_s, x_t) \in \operatorname{ind}(P)$, there is no distinguish between x_s and x_t. For any $X \subseteq U, \underline{R}(X) = \left\{x \in U | [x]_R \subseteq X\right\}$ is a lower approximation of x, $\overline{R}(X) = \left\{x \in X | [x]_R \cap X \neq \phi\right\}$ is a upper approximation of x, $\operatorname{pos}_R(X) = \underline{R}(X)$ is known as r positive region of x.

44.2.2 Property Value in Processing Information Systems

As dimensions are of different range for the order assessment system, then value range interval is also different, in order to attribute value into a dimensionless value, and so that all property values fall within the range from 0 to 1, and is handled as follows:

$$v'_{ij} = \frac{v_{ij} - \min\limits_{1 \le i \le n} v_{ij}}{\max\limits_{1 \le i \le n} v_{ij} - \min\limits_{1 \le i \le n} v_{ij}}, \quad x_i, c_j \in C \tag{44.1}$$

where $\max\limits_{1 \le i \le n} v_{ij}$ denotes the maximum value of j property range, $\min\limits_{1 \le i \le n} v_{ij}$ denotes the minimum value of j property range. On this basis, the value of the property is discrete. Order $\mathrm{STEP} = \max\limits_{1 \le i \le n} v'_{ij} - \min\limits_{1 \le i \le n} v'_{ij} \Big/ 3$, property value of each attribute is divided into five grades, $\max\limits_{1 \le i \le n} v'_{ij}$ is the highest level 5, $\left(\max\limits_{1 \le i \le n} v'_{ij}, \min\limits_{1 \le i \le n} v'_{ij} + 2 \times\right.$ STEP.] is the higher level 4, $\left(\min\limits_{1 \le i \le n} v'_{ij} + 2 \times \mathrm{STEP}, \min\limits_{1 \le i \le n} v'_{ij} + \mathrm{STEP}\right]$ is the middle level 3, $\left(\min\limits_{1 \le i \le n} v'_{ij} + \mathrm{STEP}, \min\limits_{1 \le i \le \mathrm{n}} \mathrm{v}'_{\mathrm{ij}}\right)$ is the lower level 2, $\min\limits_{1 \le i \le n} v'_{ij}$ is the lowest level 1.

44.2.3 Attribute Reduction

For the information system S = (U, C, V, F), $c \in C$, if $\operatorname{ind}(C - \{c\}) = \operatorname{ind}(C)$, it will explain c is unnecessary in C, else c is necessary in C. Unnecessary attributes in the information system is redundant, remove from the information system will not change the capabilities of classification of information system; Conversely, if a necessary attribute is removed from the information system, it will change the capabilities of classification of information system.

If $\forall c \in C$ is necessary in C, attributes set C is independent, otherwise C is relevant. For the associated attributes, set which contains extra attributes can be reduced. All the necessary attributes which constitute attributes set C is called core(C). For $P \subseteq C$, if P is independent and ind(P) = ind(C), P is a reduction of C, the relationship of core and reduction is as follows:

(1) If C is independent, $P \subseteq C$, then P is independent;
(2) $\mathrm{core}(C) = \cap \mathrm{red}(C)$, $\mathrm{red}(C)$ express all reduction of C.

44.3 Determine the Index Weight Based on the Importance of Information System Attribute

Set knowledge P export in domain U is X, $X = \{X_1, X_2, \ldots, X_n\}$, probability distribution of σ which composed by P in the subset of U is $(X|P) = \begin{pmatrix} X_1, X_2, \ldots, X_n \\ p(X_1) \cdots p(X_n) \end{pmatrix}$, $p(X_i) = \frac{|X_i|}{|U|}, i = 1, 2, \ldots, n$, $|\bullet|$ means the number of elements contain in set, the information entropy $H(P)$ of properties set P is $H(P) = -\sum_{i=1}^{n} p(X_i) \log_2^{p(X_i)}$ (When $p(X_i) = 0$, rule $0 * \log_2^0 = 0$). For the previous information system, the importance of attributes $c \in C$ in A are:

$$\mathrm{SGF}_{(C-\{c\})}(c) = H(C) - H(C - \{c\}) \tag{44.2}$$

It can be seen that the importance of property $c \in C$ in C is measured by the exchange of information entropy after removed $\{c\}$ in C. Then weight $c_i \in C$ is defined as:

$$\omega(c_i) = \frac{\mathrm{SGF}_{(C-\{c_i\})}}{\sum_{i=1}^{m} \mathrm{SGF}_{(C-\{c_i\})}(c_i)} \tag{44.3}$$

44.4 Evaluation Method of Attribute Importance and Closeness of the Target

Based on decision matrix and the target weight received, this paper adopts the following goals based on assessment methods of attribute weights and the close-degree to determine the priority of order.

For a multi-objective decision problem, suppose there are n assessments program, each assessment program has m evaluation index, evaluation matrix is $R = \left(r_{ij}\right)_{m\times n}$, suppose their relative ideal solution is $G = (g_1, g_2, \ldots, g_m)^T$, which $g_i = \max\limits_{1\leq j\leq n} r_{ij}, \mathrm{i} = (1, 2, \ldots, m)$. The weighted hamming distance of program x_j and G defined as the close-degree $\mathrm{d}\left(x_j, G\right)$ the program for the ideal solution that is to be ideal point G, and according to the size of closeness determine rank of program strengths and weaknesses.

$$\mathrm{d}(x_j, G) = \sum_{i=1}^{n} \left(\omega(C_i)\omega(c_{ij})\left|g_i - r_{ij}\right|\right) \quad j = 1, 2, \ldots, n \tag{44.4}$$

44.5 Case Study

44.5.1 Index Reduction

The evolution order for standard parts of an automobile factory is taken as example, which illustrates the previous method. Through in-depth understanding of the enterprise, the order assessments index system is established and evaluation of primitive data is done and are shown in Table 44.1.

Formula (44.1) is used to process raw data and discretization,and the results are shown in Table 44.2.

Through the observation of Table 44.2, in properties c_{11} and c_{13}, the evaluation of the corresponding object property values are the same. That is, for these five objects, properties c_{11} and c_{13} have the same resolving power, so just keep one property; here keep c_{11}. Similarly, for c_{22} and c_{23}, c_{31} and c_{32} and c_{33} and c_{41} the corresponding attribute values of evaluation object are the same. Considering to retain c_{22}, c_{31} and c_{33}, remove property c_{13}, c_{23}, c_{32} and c_{41} corresponding to the column in Table 44.2, and obtain the preliminary simplified information system. According to non-identified relation, we can obtain:

$$U/\mathrm{ind}(R) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8\}, \{x_9\}\}$$

$$U/\mathrm{ind}(R - \{c_{11}\}) = \{\{x_1, x_4\}, \{x_2\}, \{x_3\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8\}, \{x_9\}\}$$

$$U/\mathrm{ind}(R - \{c_{12}\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8\}, \{x_9\}\}$$

$$U/\mathrm{ind}(R - \{c_{21}\}) = \{\{x_1\}, \{x_2\}, \{x_3, x_6\}, \{x_4\}, \{x_5\}, \{x_7\}, \{x_8\}, \{x_9\}\}$$

$$U/\mathrm{ind}(R - \{c_{22}\}) = \{\{x_1\}, \{x_2, x_7\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_8\}, \{x_9\}\}$$

$$U/\mathrm{ind}(R - \{c_{31}\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8\}, \{x_9\}\}$$

$$U/\mathrm{ind}(R - \{c_{33}\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5, x_8\}, \{x_6\}, \{x_7\}, \{x_9\}\}$$

Table 44.1 The data for order evaluation indexes

Indexes		Scheme								
		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Customers' information C_1	Social effect c_{11}	1.5	1.65	1.7	1.3	1.1	1.7	1.6	1.1	1.34
	Honest c_{12}	52	25	12	52	37	12	24	36	52
	Development essential c_{13} (%)	60	75	80	40	23	80	68	23	51
Orders information C_2	Emergency degree of orders c_{21}	23	52	13	23	40	31	52	40	21
	Number of order(thousands) c_{22}	2	1.8	1.6	1.8	1.5	1.55	1.7	1.5	1.8
	Order importance c_{23} (%)	50	40	25	45	20	26	35	20	44
Manufacturing capacity C_3	Design ability c_{31}	5	4	4.5	3.8	4	4.5	3.9	3.9	3.8
	Manufacturing capacity c_{32}	5	4.5	4.8	4.3	4.5	4.8	4.85	4.85	4.3
	Adaptive rate of manufacturing resource c_{33} (%)	75	83	85	75	77	85	81	85	75
Economic indicators C_4	Other fees c_{41}(%)	6	10	12	6	7	12	10	12	6
	Product profit rate c_{42} (%)	43	47	50	35	39	50	47	42.5	50

Table 44.2 Discrete the data

Object U	Attributes C										
	c_{11}	c_{12}	c_{13}	c_{21}	c_{22}	c_{23}	c_{31}	c_{32}	c_{33}	c_{41}	c_{42}
x_1	4	5	4	2	4	4	5	5	1	1	3
x_2	4	3	4	5	5	5	2	2	4	4	4
x_3	5	1	5	1	2	2	4	4	5	5	5
x_4	3	5	3	2	4	4	1	1	1	1	3
x_5	1	4	1	4	1	1	2	2	2	2	3
x_6	5	1	5	4	2	2	4	4	5	5	5
x_7	4	3	4	5	3	3	2	2	4	4	4
x_8	1	4	1	4	1	1	2	2	5	5	3
x_9	3	5	3	2	4	4	1	1	1	1	5

$$U/\text{ind}(R - \{c_{42}\}) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4, x_9\}, \{x_5\}, \{x_6\}, \{x_7\}, \{x_8\}\}$$

Attributes of c_{12} and c_{13} are unnecessary, attributes of c_{11}, c_{21}, c_{22} and c_{33} are necessary. Hence core of information system is $\text{core}(C) = \{c_{11}, c_{21}, c_{22}, c_{33}, c_{42}\}$, there are two reduction $R_1 = \{c_{11}, c_{12}, c_{21}, c_{22}, c_{33}, c_{42}\}$ and $R_2 = \{c_{11}, c_{21}, c_{22}, c_{31}, c_{33}, c_{42}\}$, consider R_1 only, then the corresponding index set of first and secondary indicators is: $\{C_1, C_2, C_3, C_4\} = \{\{c_{11}, c_{12}\}, \{c_{21}, c_{22}\}, \{c_{33}\}\{c_{42}\}\}$.

44.5.2 Calculation of Weights

$$U/\text{ind}(C_1) = \{\{x_1\},\{x_2,x_7\},\{x_3,x_6\},\{x_4,x_9\},\{x_5,x_8\}\},$$

$$U/\text{ind}(C_1 - \{c_{11}\}) = \{\{x_1,x_4,x_9\},\{x_2,x_7\},\{x_3,x_6\},\{x_5,x_8\}\},$$

$$U/\text{ind}(C_1 - \{c_{12}\}) = \{\{x_1,x_2,x_7\},\{x_3x_6\},\{x_4,x_9\},\{x_5,x_8\}\}$$

$$H(C_1) = -\left(4\times\frac{2}{9}\log_2\frac{2}{9}+\frac{1}{9}\log_2\frac{1}{9}\right),$$

$$H(C_1 - \{c_{11}\}) = -\left(3\times\frac{2}{9}\log_2\frac{2}{9}+\frac{3}{9}\log_2\frac{3}{9}\right),$$

$$H(C_1 - \{c_{12}\}) = -\left(2\times\frac{2}{9}\log_2\frac{2}{9}+\frac{3}{9}\log_2\frac{3}{9}\right),$$

$$\text{SGF}_{(C_1-\{c_{11}\})}(c_{11}) = H(C_1) - H(C_1 - \{c_{11}\}) = \frac{1}{3}\log_2\frac{1}{3} - \frac{2}{9}\log_2\frac{2}{9} - \frac{1}{9}\log_2\frac{1}{9},$$

$$\text{SGF}_{(C_1-\{c_{12}\})}(c_{12}) = H(C_1) - H(C_1 - \{c_{12}\}) = \frac{1}{3}\log_2\frac{1}{3} - \frac{4}{9}\log_2\frac{4}{9} - \frac{1}{9}\log_2\frac{1}{9},$$

$$\omega(c_{11}) = 0.28,\ \omega(c_{12}) = 0.72,$$

similarly, the first weight indicators are $\omega(C_1) = 0.26$, $\omega(C_2) = 0.12$, $\omega(C_3) = 0.26$, $\omega(C_4) = 0.26$; the second weight indicators $\omega(c_{21}) = 0.57, \omega(c_{22}) = 0.43, \omega(c_{33}) = 1, \omega(c_{42}) = 1$.

44.5.3 Comprehensive Assessment

When assessing nine of the given object to be evaluated, in the index matrix there are no experts in the case of the subjective weight, target weight can be obtained directly from the original data. First, the ideal solution is to determine the relative ideal point: $G = (g_1, g_2, \ldots, g_m)^T = (5,5,5,5,5,5)^T$. Nine programs obtained by the formula (44.2) is relatively close to the ideal degree programs and are:

1.8896,0.9672,1.1772,1.9624,2.0532,0.9720,1.0704,1.2732,1.4424.

The smaller program shows the priority order: $x_5, x_4, x_1, x_9, x_8, x_3, x_7, x_6, x_2$.

44.6 Summary

Since the influence factors on orders is large and complex the evaluation index system is too complicated, and there are many redundant indicators, making some important indicators weak. For this problem, information system is established;

provide reduction method based on the attribute of rough set. This method starts from raw data processed by discretization using the rough set reduction theory, solve reduction of index system, by solving Objective weight of the indicators after reduction, the relative closeness with the ideal point methods is adopted for optimization of customer orders, and for order production order priority assessment provides a feasible approach.

References

1. Tao LH, Feng WD, Xu B (2008) The evaluation model of production order under networked manufacturing. Mach Tool Autom Manuf Tech 10:101–105
2. Xu XJ, Sun YM (2001) Study on determination method of priority of production order[J]. Ind Eng Manag 6(3):36–42
3. Liu GX, Lv XL (2004) Study on the evaluation method of low-volume orders based on the concept of entropy [J]. Mech Electr Prod Dev Innov 17(02):22–24
4. Chen ZX, Qiu J (2007) Evaluation manufacturing demand order based on neural network[J]. Wuhan Univ Technol (Inf Manag Eng) 29(6):81–84
5. Wang PH (1999) Research project evaluation methods[J]. Res Manag Rev 20(3):18–24
6. Shao LS, Qiu YF (2008) Rough set-based supplier evaluation and selection[J]. J Liaoning Tech Univ 27(04):591–596
7. Zhang WX, Wu WZ (2001) Rough set theory and method[M]. Science Press, Beijing

Chapter 45
Professional Management Capability of Sports Competition in Universities and Colleges

Dinghong Mou, Qingshan Ma and Xiaobing Fan

Abstract Management of universities and colleges sports competition needs the sufficient integration of all required resources. To integrate and coordinate different kinds of work, besides the necessary professional knowledge, there must also be some managerial capabilities and techniques. Although sports competitions are somehow different from other competitions, many capabilities are almost the same, which are actually the basic capabilities to manage successful competitions. With the methods of the literature review, expert interview, logical analysis, etc., a study is made on the professional ability of management universities and colleges sports competitions, and some theoretical references are provided to realize the aim of universities and colleges sports competition management.

Keywords Universities and colleges · Sports competition · Professional capability

45.1 Introduction

Universities and colleges sports competition is one of the main contents in universities and colleges sports work and in universities and colleges students' campus life. It is the most active and bright scene on universities and colleges

D. Mou (✉) · Q. Ma · X. Fan
Faculty of P.E, East China Institute of Technology, Fuzhou, China
e-mail: Moudinghong330@163.com

Q. Ma
e-mail: qingshanma22@163.com

X. Fan
e-mail: Fanxiaobing12@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_45,

campus. It is very meaningful to promote the carrying out of Universities and colleges sports activities. The triune model of universities and colleges physical education (PE) teaching, sports training and research can be fully shown in sports competitions [1]. Competitions are a powerful support to sports training and performance. Such factors as forming universities and colleges sports competition systems, methods, content selection, as well as the scale of the competition, measures for rewarding, etc., will certainly guide the development of many sports. Contradictions between the current universities and colleges sports competition mode and the new idea of management of these competitions require physical workers to stick to the law of conducting universities and colleges sports competition, to improve relevant managerial capabilities, in order to push forward universities and colleges sports competition [2].

45.2 Study Object and Study Methods

45.2.1 Study Object

Physical workers and university students who participate in universities and colleges sports competitions.

45.2.2 Study Methods

Literature review. Look up relevant documents and materials about PE teaching and universities and colleges sports competition. According to the requirements, look them up in the school library and in many journals. Compare and analyze involved research methods and materials. Provide references for the study, design and conception here.

Expert interview. Interview some experts about relevant problems, to get their help and guidance.

Logical analysis. During the process of analyzing the literature and investigating the materials, the following logic methods are used: classification and comparison, induction and deduction and analysis and synthesized.

45.3 Results and Analysis

Sports competition management refers to the fact that those who host the sports competition should carry out the managerial role of planning, organizing, implementing, controlling and reasonably using and allocating the investment of financial, material, human resources and the use of information technology; to effectively produce the products of sports competition and relevant service; to employ all resources into sports items, so as to achieve the settled goal and satisfy different kinds of needs. The specific manifestation of this management is shown in the effective and systematic managerial activities such as planning, organizing, conducting, controlling and coordinating during the whole process of the competition, thus to accomplish its goal. [3].

Sports competition management is a complex process, and is restricted by both the subjective and the objective conditions. For instance, the manager's quality, its goal and strategy, professional ability of communicating and coordinating, personnel enrollment, ability of sports training and managing, knowledge of managing special sports competitions, logistic service, ability of managing information system and level of understanding. Man is the most important factor in the management. His managerial skills and his own guiding theory and ideology decide the managerial efficiency. To improve this efficiency, it is necessary to make clear what the professional ability of management sports competitions is and then study it. As far as the professional capabilities of management universities and colleges sports competition is concerned, they include the following.

45.3.1 Capabilities to Communicate and Coordinate

Division of work in sports competition is the same as in other activities. It needs the full-timers to take charge of marketing, gyms, public relations, meals, rooms, reception and entertainment. Sports competition managers are very important to link all these parts. Communication is a tool to create a coordinated management. In more detailed words, communication is a process during which people convey their ideas and transfer information. Through this, people can share mutual feelings and knowledge, eliminate misunderstanding and enhance mutual understanding. Effective communication and coordination can readily solve the problems existing in different departments, so as to guarantee the smooth development of sports competition.

45.3.2 Capabilities to Recruit and Train Staff

To successfully hold a sports competition, only one able manager is far from enough. The biggest trial lies in recruiting, training and managing the staff. The professional training will, guide them to change perspective of observing things, continually stimulate their willingness to serve, and help them to form good occupational behaviors and habits. Those in charge of the sports competition should have the integral professional capabilities in the following three aspects: to make sure the competition situation is safe for all; to be certain that the working staff should have enough ability and experience; to guarantee a cooperative team.

45.3.3 Managerial Knowledge to Hold Special Sports Competitions

Sports competition managements should have more and more relevant managerial knowledge, which can make them more confident and more believable. Rich professional knowledge can bring others' trust and appreciation, which is also very necessary in social communications. Without required knowledge about special sports competitions, it is possible to neglect some important segments during the planning process because of their cultural difference or other special reasons. For example, there are many items in the track-and-field competition, ranging from the sign in, allocation of booking, timing, recording to adjudicate and risk management. It is very likely to be different from holding an international swim competition. To hold a high-qualitied sports competition, the managements must have very rich relevant knowledge.

45.3.4 Techniques to Management Sports Competitions

During the process of managing the sports training, managements play a very critical role. To accomplish the training task and to achieve the planned aim, they are necessarily required to grasp better managerial theories, to master managerial techniques, and to coordinate different kinds of relationships. Managerial techniques place the emphasis on such factors in the managerial function as plan, organization, personnel, supervision and control. According to the different competitions' particularity and specialization, make a sufficient integration of organizational resources and employ them in all kinds of big- or small-scale sports activities. This professional ability also includes coordinating organizational conflict, doing away with unexpected events in and out, possessing knowledge about sports law and carrying out leadership skills.

45.3.5 To Improve Logistic Management Capability

The task of the logistic management is to use all kinds of management methods, to organize, guide and coordinate the activities of logistic staff, so as to fulfill the logistic task effectively and high-qualifiedly, as well as guarantee the smooth development of all work. Among all the management work, logistic management is very important, while it is often neglected. With the other conditions being the same, whether a logistic unit is good or bad mainly depends on the level of management. The logistic department is supposed to provide material security to its belonged institution and it should serve the functional activities. The logistics stand for a series of logistic activities, such as transportation, storage, service and information. From the perspective of working contents, the logistic management covers a lot. For a sports competition manager, the logistic service system reflects his ability to control the logistic service system of sports competitions. Business, big or small, such as eating, wearing, living, transportation, security, gyms, ticketing, performance and medical care are all parts of the logistic service [4]. The logistic management should offer reliable material safeguard, and improve the utilization ratio of personnel, financing and properties, thus to promote the working efficiency. Combine the personnel, financing and properties in the best possible way and use them most sufficiently and effectively. This can stimulate people's subjective initiative to bring each one's ability into full play; this can also increase the financial utilization effect to put every coin into actual use; what is more is, this can display the best possible potential of materials and equipments to make the best use of things.

45.3.6 Capabilities to Market and Attract Funds

Marketing is an integral part of sports competitions because sports competitions are producing invisible products, which are strongly characterized by the invisible nature. The nature of providing the consumers, including the audience, with service consumption decides the fact that marketing is the certain content and task of sports competitions. Any plan in operating the competition and any change in the management process, will definitely influence the consumers' feeling and experience. The changes may be shown in gyms, facilities and agenda. From the eyes of consumers and benefit-oriented competition participants, the marketing is all. Therefore, how to meet the consumers' need, and how to draw up satisfactory marketing strategy are of critical importance. The idea to develop the sports competition products should be reflected in the promotion, which can in turn guarantee the successful marketing of ticket price, relay broadcasting and funding reward. Meanwhile, attach great importance to publicize and reward the sponsors, and maintain a good cooperative relationship with them. Besides, the organizers should not only work hard to realize the telecast, but also provide the broadcasters with excellent service, and improve the broadcasting quality. This helps to achieve the double goal of publicizing the sports event and enhancing the effect of the sponsors' publicity [5].

45.3.7 Capabilities to Manage Information System

IT application in management of sports competition is guided with advanced managerial ideas. The managers should on one hand advance with times, and on the other hand be equipped with professional knowledge in this field and be ready to bring the last science and technology into the sports competition. This is in fact a working process of using automatic instruments to survey, demonstrate, evaluate and control. To collect, process and apply sports information should be integrated into the working process, and as a result form a complete and organic whole. Through delivery of information, realize the automation of the work. During the process of management, the information technology should be widely used. Consequently, it can dominate resources such as personnel, material, finance and information more scientifically, more reasonably and more effectively.

This is helpful to better reach the goal of the sports competition. With the ever-developing science and technology, the managers must equip themselves with basic ability and knowledge of using computers. Low cost, quickness and convenience brought by the network make the managers communicate with people from different places and get their immediate feedback. Based on the present sports management system, through the accumulation of real-time data in the working process, depending on the data warehouse, dig a large number of historical data, construct an effective sports management model, to further guide and control the work. It is the latest developing trend at present.

45.4 Conclusion and Suggestions

To realize the expected aim raised by the universities and colleges sports competitions, it is necessary to resort to effective management. The managerial capabilities include capabilities to communicate and coordinate, capabilities to recruit and train staff, knowledge to hold special sports competitions, techniques to manage sports competitions, improve logistic management ability, capabilities to market and attract funds, as well as capabilities to manage the information system. Attract more and more universities and colleges students to participate in the universities and colleges sports competitions, to make them feel the joy brought by sports. To satisfy their various needs should be the guiding idea to promote the progress of universities and colleges sports competition reform, and it also calls for the improvement of relevant management capabilities.

References

1. Kesenne S (2006) Competitive balance in team sports and the impact of revenue sharing. J Sport Manag (20):39–51
2. Humpreys B (2002) Alternative measures of competitive balance. J Sports Econ 3(2):133–148
3. Bennett G, Lachowetz T (2004) Marketing to lifestyles: action sports and generation Y. Sport Mark Q (13):239–243
4. Clark JM, Cornwell TB, Pruitt SW (2005) The relationship between major-league sports' official sponsorship announcements and the stock prices of sponsoring firms. J Acad Mark Sci 33(4):401–412
5. Milczarek M (2001) Where the kids are: grassroots events allow companies to reach teens on their own turf and cut through the clutter. Mark Mag 106(31):13

Chapter 46
Design and Research on Teaching Platform of Stage Task Using JavaEE

Jun Ou, Yun Pei, Min Chen, Qingxiu Wu and Shizhuan Li

Abstract Nowadays, the teaching platform applied to network teaching did not meet the learning needs of students mastering the knowledge and skill gradually. Combined with task-teaching pattern, it presented a new teaching mode of phased tasks, designing a teaching platform based on JavaEE. The platform enabled the students to learn courses independently, look for problems and find answers, according to learning tasks of each stage. The platform improved the students' enthusiasm, initiative and creativity.

Keywords JavaEE · Stage task · Teaching platform

46.1 Introduction

With the rapid development of computer and network technology, information technology gained more and more popularity in the field of education. Network teaching mode gradually became the principal means of modern teaching. It not only takes the teacher as leading and the student as chief body but also achieves interactive teaching ideas. Through the consolidation and management of teaching resources, it can optimize teaching resource, by improving learning efficiency.

J. Ou (✉) · Y. Pei · M. Chen · Q. Wu · S. Li
Department of Network Engineering, Hainan College of Software Technology,
Qionghai 571400, Hainan, China
e-mail: xhogh@hotmail.com

Y. Pei · M. Chen
University of Electronic Science and Technology of China,
Chengdu 610054, Sichuan, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_46,

However, most of the network teaching platforms are beneficial to teach with the resources on the publishing server and unable to meet the learning needs of students mastering the knowledge and skill whose characteristics are gradual and orderly in progress.

At present, other teachers have understood the advantage of task teaching and abandoned the simple traditional classroom teaching mode, and combined classroom teaching with the project practice. Practice proves that the teaching methods have contributed to improving students' computer skills. But, there is still a developing space in two areas: (1) guiding students step-by-step through a task to acquire professional skills and professional knowledge; (2) improving further students' learning initiative, enthusiasm and creativity. According to the characteristics, design the teaching platform of stage task based on JavaEE.

The traditional network teaching and the implemented ways of task teaching are integrated into the method adapting to the characteristic of task teaching. Its stage task teaching improved the efficiency of the students learning and understanding the knowledge and skills. This platform made teaching quality and management very efficient. It made students better grasp the professional theoretical knowledge and the practical skills with accomplished tasks [1].

46.2 Teaching Mode of Stage Task

It presented a new teaching model: teaching mode of stage task. Teaching process described below.

First of all, give the first stage of a case. When students completed the first stage, and submitted the works to the teacher. They can enter tasks in the next stage after the check of teachers. Students can study reference (documentation and videos) provided by the teaching platform to resolve the problems encountered in the process of completing the task.JSP programing in computer courses achieves a new publishing system project. The project can be broken down into many cases, and one of the stories for the news release system adds new features. It includes tasks: landing page design; landing validation feature and implementation of verification code function. Cases are done step-by-step. Students must finish the task ahead, and then do the next task. Three functions are completed. News publishing system realizes the feature of adding news. Students grasp the relevant technology and related knowledge through implementing the entire function [2, 3].

Periodic task teaching is useful for students to acquire knowledge and skills: (1) through the learning process, students understand the knowledge required for completing the task; (2) through a cooperative process, students further understand the knowledge; (3) by completing the tasks of this practice, students master the skills and understand the knowledge of completing task. Learning maps is shown in Fig. 46.1:

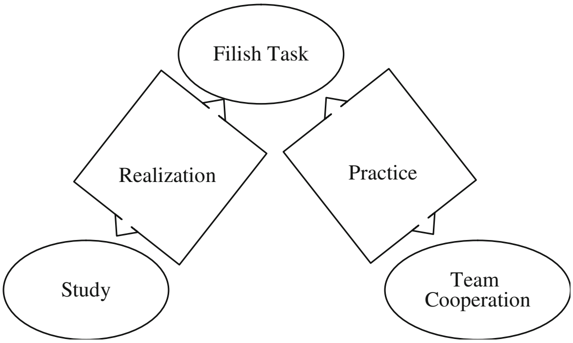


Fig. 46.1 The structure of study

46.3 System Design Idea

Teaching platform of stage task should be based on the practical needs of teaching, and designed according to the following principles.

Principles of functional integrity. Take advantage of existing infrastructure and technology on a large scale for different levels through a combination of information flow. Enable the data to be unified and efficient management.

Modularization principles. System solutions follow “modular” principle, structure flexible, on demand. Arbitrary combination of all basic modules and the business modules can run, which are omnidirectional to meet present and future demands. Using the principle of object-oriented data packaging, can improve platform scalability and stability.

Friendly principles. User friendly interface is easy to use, which enhances the efficiency of work and study.

Principle of openness. All procedures and interface are with uniform standards, making the system with excellent portability. Open interface, ease of application extension, with portability.

Security. Has a higher operating efficiency of network security and the security system uses multilevel protection system, including network-level, database-level and application-level, with a high degree of security and confidentiality.

High performance. Guarantee fast response speed, and stable system. On the premise of ensuring system reliability and stability, it improved performance. Use multithreaded data connection pools and load balancing technology, to make good parallel performance, ensuring more user interaction when large amounts of data happen.

Scalability principle. Database structure should be designed to fully consider the needs of development needs, transplants, with good extensibility, scalability and a moderate amount of redundancy [4].

46.4 Use Case Analysis of Task Management

Teachers role in task information management includes the following function: project classification management, task management, reference management; task classification information management includes: adding, viewing, modifying and deleting task classification information, and task classification information includes classification number, name and belongs Professional; task management includes: adding, viewing, modifying and deleting task information,and task information is includes task stage number, and name; reference management includes: adding, viewing, modifying and deleting reference information, and reference information includes task name, classification, belongs project and added date.

46.5 Main Function Design

System module is most important in the overall system design. Modular refers to solving a complex problem top-down point-by-layer process so that software system is divided into several modules. Each module can perform a specific function. According to the certain method, all modules were organized as a whole to complete the function required by the system. The system is divided into a number of modules in order to reduce the complexity of software systems, readability and maintainability, but the division cannot be arbitrary and should try to keep their independence. That is to say, each module realized independent subfunction of the complete system requirements, and linked with other modules. and the interface is simple, that is, high cohesion and low coupling is possible, improving the independence of the module, laying the foundation for designing high quality software architecture.

It uses a structured design in the summary design of the system. Structured design is based on the data flow graph of requirements analysis phase. According to certain steps, it mapped into a software structure. Under demand analysis of task teaching platform, it can be designed in detail. The entire system can be designed as several module or subsystem. Each subsystem is relatively independent and interrelated to function module. In subsystem, it has different entity function module to support relevant function and service. First, entire system is divided into several small problems, small modules. It has project information management, task information management and personnel information management. Then, further breakdown of the module, add details. Basic personnel information management includes departmental information management, position information management, user information management, authority information management and project information management including adding, modifying, deleting and viewing project information.

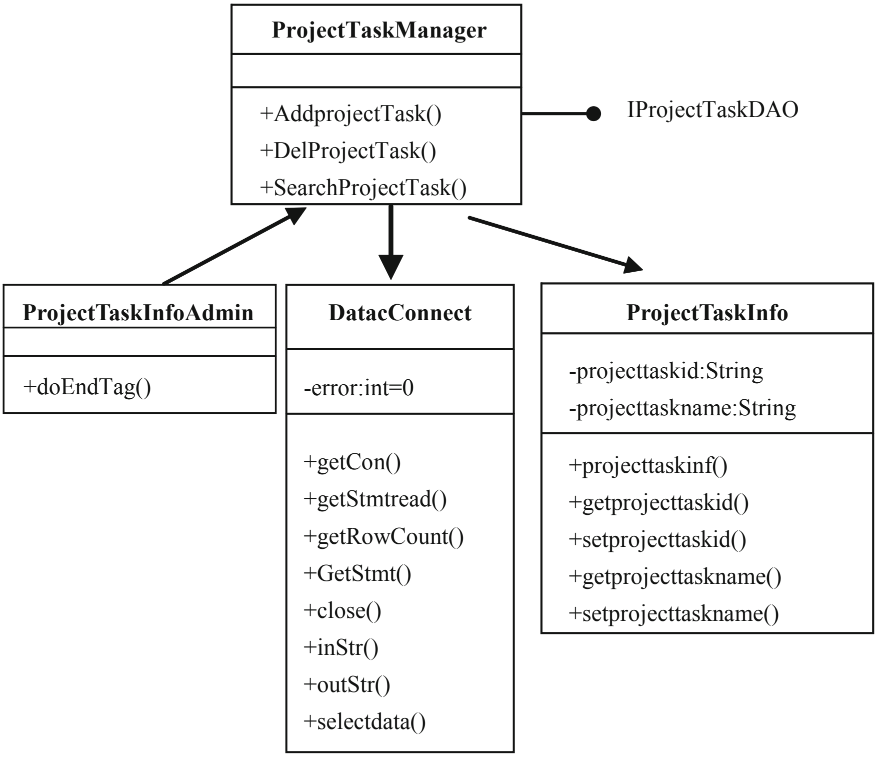


Fig. 46.2 Class map of task management

46.6 Stage Task Module Design

Task classification management is including adding, viewing, modifying and deleting task classification information. Specific feature of the operation is that users access to login interface of the teaching platform. After the authentication and authorization, user issues a request to the server-side to perform a functional operation. Based on the resource configuration file of Web.XML, server is parsing the request accepted. According to the different request, the processing component will be called to execute the request. ProjectTaskInfoSave, ProjectTaskInfoUpdate, ProjectTask-Admin, ProjectTaskUser are accepted Widget request, which are corresponding to each request of adding, modifying and viewing task classification information; IProjectTaskDAO is a interface of session layer in task classification information management; ProjectTaskManager implements the interface; ProjectTaskInfo is an entity of task classification information. Data Connect is all database access layer. Class diagram of task classification information management is shown in Fig. 46.2:

It is followed that analysis of the executing process is done by adding task of classification information. User logins the category page of adding tasks.

In accordance with the requirements of the system adding a task to a category information, it fills in the necessary, entire and accurate information. Client-side JavaScript inputs validation function check (), and verifies the information submitted. After authentication, it will submit requests to the server-side. Server is parsing to accepted request based on the resource configuration file of Web.XML.

```
<servlet>
<description>To Save ProjectTask information</description>
<display-name>ProjectTask Info Save</display-name>
<servlet-name>ProjectTaskInfoSave</servlet-name>
<servlet-class>com.logistic.servlet.ProjectTaskInfoSave</servlet-class>
</servlet>
<servlet-mapping>
<servlet-name>ProjectTaskInfoSave</servlet-name>
<url-pattern>/ProjectTaskinfosave</url-pattern>
</servlet-mapping>
```

According to the information in the configuration file as above, the request submitted is adding a task category information, and then calls the component com.logistic.servlet.ProjectTaskInfoSave executes the request. Public void doPost (HttpServletRequest request,HttpServletResponse response) throws ServletException. IOException method gets client submitted all information in widget ProjectTaskInfoSave and creates task classification session Bean ProjectTaskManager for implementing task classification information management session layer interface. IProjectTaskDAO calls public int AddProjectTask (String ProjectTaskid,String ProjectTaskname) method to implement adding task classification operation; eventually updating on database calls public int Updata (String SQLs) method in database access layer Data Connect to implement.

46.7 Result

The task teaching mode of network teaching platform was on analysis. On the basis of UML modeling technology, the teaching platform of stage task was designed based on JavaEE. Using structure of B/S, JavaEE architecture platform and SQL data storage technologies, combined with popular strut+Spring+Hibernate framework technology, full support for all aspects of teaching and provide task teaching with a variety of integrated services. The platform brought a great deal of convenience in testing, easy maintenance and many other aspects. It enabled students to master the knowledge and skill characteristics gradually. System platform in the pilot phase was recognized by the vast number of teachers and students. It effectively improved the quality and efficiency of teaching task, with high practical value.

Acknowledgment This work is supported by the Foundation of office of Education of Hainan Province under Grant Hjkj2010-55 (Colleges Scientific Research Projects of Hainan Province: Based on J2EE Project-teaching Interactive System of “2+1” Training Mode) and Hjsk2011-92 (Research on system maintenance talent training mode of ability-oriented).

References

1. Pajares A, Guerri JC, Belda A et a1 (2002) JMFMoD: a new system formedia on demand presentations. In: Proceedings of 28th Euromicro conference, 4–6 Sept 2002, pp 160–167
2. Zhang Q (2008) A development application design and implementation of teaching case based on JavaWeb [u]. Comput Edu (13):98–100
3. Guo H, Chen S-Q (2008) A design and implementation of a Java learning platform based on J2EE [u]. Comput Inf Technol (7):32–36
4. Li Z, Zhu Q et al (2008) A development and construction of experimental teaching platform based on JavaEE. Comput Edu (2):110–112

Chapter 47
The Research and Application of Methods in Mathematics Analysis Course

Bin Ran

Abstract There are many problems in the current Mathematics Analysis courses in science and engineering colleges and universities. People's traditional concepts of Mathematical Analysis courses need to be updated. This article discusses the teaching content and method of Mathematical Analysis and puts forward suggestions for improvement, such as Mathematical culture and numerical experiments to be brought into teaching, etc.

Keywords Mathematics analysis · Numerical experiments · Constructivism theory

47.1 Introduction

Mathematical Analysis is an important lesson in all college Mathematic specialities. It is not only an important basis of the fellow course to students who major in Mathematics but also plays a very important role in cultivating students' mathematic quality, and the deliverance of Mathematical cultural knowledge. At present, the instruction in Mathematical Analysis courses in all science and engineering colleges in our country is not satisfactory [1]; in this computer era, the traditional method is not practical. Hence, we much take the impingement of Mathematical Analysis courses seriously.

B. Ran (✉)
College of Mathematics and Computer Science,
Changjiang Normal University,
Chongqing 408000, China
e-mail: rbfl@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_47,

47.2 The Sense Change to Mathematical Analysis Courses

The traditional sense to Mathematical Analysis is that it is just a tool of mathematic theory, and has no function. Except for technological teaching. It is the same with regard to all Mathematical Analysis courses in science and engineering colleges' mathematic specialities instruction [2]. Therefore, all kinds of mathematic specialities such as information and calculation science, statistics, and other under specialities value the tool and infinitesimal calculus theory much, before the Mathematical Analysis courses' instruction, and the calculation questions and provident questions are also the main question in the exams.

Nowadays people are admitting that mathematics is not only a tool, but also a culture. Therefore, it should have two educational functions, that of technology and culture. Studying this course can improve students' comprehensive characters and cultural accomplishment, make students learn to think and enhance the ability of solving problems. Mathematical Analysis study can offer one a view from a mathematician, enriching one's ways of observing the world; offer a wise brain, help in thinking rationally; offer a curious mind, ensuring a strong desire for knowledge; offer a set of researching patterns, making it as the telescope and microscope for exploring the secrets of the world; offer a new optometry for seeking one's dreamland in cross discipline and use one's diligence and intelligence in invention and creation. Stewart, an academician of the England imperial academy of science, once pointed out that our world is based on a math foundation, math does not avoid from blending into our whole culture.

Hence, colleges of science and engineering math majors offering a course in Mathematics Analysis should lay stress on quality education. We can divide it into two special fields, mathematics analysis culture function and technological function, combined with fostering the goals of polytechnic university math majors and modern math developing trends. Mathematical Analysis education should also be in with the times, thus training undergraduates to fit the demand of the times.

47.3 The Improvement of the Mathematical Analysis Course

The improvement of the course has been the most important part of the course reform for universities. Being the mathematical basic theory, Mathematical Analysis course has been put in the first place. For this we have put forward the following steps for reform.

47.3.1 Mathematical Culture's Bring in

Since Mathematic Analysis has both cultural and technical functions in education, we should not just analyse the concepts, theorems, principles and formulas, as in the past [3]. We also need to tell something about the Mathematical culture, the origin, and the question ‘what is mathematics’ that refers to the attitude about mathematics.

Combined with the teaching of analyzing mathematics, we could introduce some history along with the relevant experts. By doing this, not only would we have wonderful but also necessary teaching. Because ‘the most precious part of a kind of sciences is its history, for the reason that sciences just give us knowledge, however, we could get wisdom from its history.’ This introduces much about those aspects of science which go through the times and those important concepts that inform and develop. At the same time, evaluate them, give them some exportation, and do not touch much more on the detail. In this way, students can be made to feel that the course is productive.

47.3.2 The Alternation of Recent Mathematics Theories

While we lighten the students’ calculating burden, we should also notice that some recent mathematics theories, such as “Recent Algebra”, “General Functions Analysis”, “Tuopu Learning” and similarly some mathematics branches are applying to modern technology day by day. Apart from the above, so do “the Theory of Control” and “Vague Mathematics” and so on. Among the mathematics majors in polytechnic universities, often only the major of “Application Mathematics” establishes the course of “General Functions Analysis”, while only the major of “Information and Calculating Science” establishes “Vague Mathematics”, but the theories of recent mathematics are not introduced coherently in other majors [4]. We use the characteristics of Mathematics Analysis courses having long study hours and long semesters to give some introduction to Mathematics Analysis. It not only makes students know more about the new mathematics instruments, but also makes a basis of studying mathematics further in the future.

Meanwhile, the content of Mathematics Analysis courses should also link up to the exam programs of postgraduate schools, as there are more and more students nowadays.

47.3.3 The Combination of Software “Mathematic”

The content of Mathematics Analysis courses in polytechnic majors should come of age. Nowadays, Mathematics is changing at an unprecedented rate because of the introduction and development of computers. And Mathematics has come to the Computing Mathematics age. However, the content of Mathematics Analysis courses is in great contrast with it, and the content is still almost the same since the last ten years. In fact, in Mathematics Analysis courses, some contents which we think are so important, such as “the Solution of Differentiation”, “Equation the

Calculation of Integration", "Derivative Operation", "the Operation of Unfold the Functions to levels" and "some Formula Calculation" etc. [5] are very easy with Mathematics. Students of mathematics majors should master the application of Mathematics, which can make students get rid of complex calculations and change their perspective of studying mathematics. Besides, it makes students not only study mathematics, but also make use of mathematics to solve practical problems. Because knowledge is power, and the origin of power is the use of knowledge.

47.3.4 Mathematical Analysis and Design of Experiments

In the course of Mathematical Analysis can be mixed mathematical experiments and mathematical modeling concepts together because mathematical experiments can make students more realistically and intuitively understand the Mathematical theory. In the Mathematical analysis by students, hands on experiments in mathematical modeling to further and improve mathematical experiments can solve the actual problem by using Mathematical theory. Here, the actual problem referred to may have no readymade answers, no fixed method and no special mathematical tools and methods. However, students can choose the most suitable tools to set up the mathematical model and find the solution from the knowledge gained. This is a creative work that can cultivate students' innovative spirit and practice ability and also inspire students' enthusiasm about Mathematical Analysis.

At the same time, Mathematical modeling is popular among domestic universities. This point is not special from the American Mathematical modeling contest enrollment situation of the recent years. The lesson of Mathematical Analysis should comply with the characteristics of Sichuan and make students have a good beginning.

47.4 Reform Mathematical Analysis Teaching Methods

The Mathematical Analysis teaching methods' reformation is an ancient and longstanding problem. The keys are, first, most of the teachers who are masters or doctors, teaching Mathematical Analysis in science and engineering colleges and universities often have their own professional researches, and they are usually not teachers of college graduates. What is more, they do not know Mathematical teaching methods. Second, a majority of the Mathematical Analysis teachers think that mathematics students learning maths is the process of "Pleasure comes through toil". The students must have the spirit to endure hardships and overcome difficulties. Thus a Mathematical Analysis teacher needs to know not Mathematical teaching theories only, but to pay great attention to Mathematical teaching methods. Therefore some teachers teach Mathematical Analysis as a book from heaven, whereas the students in class are like "Duck listen to Ray". The results, of course, are that the modern educational concepts and personnel training goes against the objectives. At the same time students also develop the psychological fear of mathematics learning.

Mathematical Analysis is a basic course to majors in Mathematics, hence a better teaching method could arouse students' interests in what they have learned as a profession. This makes them get salutary learning. On the contrary, an unwarrantable Mathematical Analysis teaching method would give students a deep impression that mathematics is dull and dry and learning mathematical expertise is useless. Mr. Wenjun said with deep feeling when he achieved the State Supreme Science and Technology Award, "I was not interested in math when I was a child, and even once when I was a sophistry I lost confidence in math. But a teacher, Mr., whose brilliant lecture changed my perspective on mathematics. And my interest increased up was like the geometric progression. Finally I established mathematics to be my life-long occupation." Thus we can see how important the University of Math Teachers' Teaching Methods are. How to adapt to modern mathematics teaching is the Mathematical teachers' unthinkable duty.

47.4.1 With the Constructivist Theory to Guide the Teaching of Mathematical Analysis

The course which is in terms of variable mathematics is more complex than primary mathematics in middle school. Hence, it should obey the law of education mathematics.

Although you have already imagined mathematics of education to be very complex, the results of mathematics of education is more compiled.

To want to teach the course of Mathematic Analysis, you should know the theory of mathematic education and students should know how to study. The contemporary theory of studying states: "mathematic object is in terms of method of thinking, mathematic activity is in term of ideological activity". Thus, studying new knowledge of mathematic is a kind of studying course to make up typically. The theory of constitution is a common education theory. It believes that students should not be imitative to constitute, and not accept all knowledge that teachers give. One knowledge of mathematics is based on how to handle the experience, communicate and introspect, all of which make great progress in mathematic. It also believes that teachers and students should make up a study community. In this community, the students have the main part while teachers are like the coach as in sports activities.

To teach by building constructing theory directed by Mathematical analysis, first we should know clearly what the students have mastered and then we can start our Mathematical study according to the knowledge that the students have mastered. For instance, the sum and minus changing into product formula of triangle function is not required anymore in the university entrance exam. But in the college Mathematical studies, the product changing into sum and minus formula is used much more frequently, so it is necessary for us to review and strengthen the basic knowledge before we give instructions. The second the teacher ought to be the promoter of the building constructing activity, sort out the knowledge so that the students can get it easily; make the knowledge like the apples on the tree which the students can touch but need to take some effort to get it. For example, before using the polar coordinate in Mathematical analysis we need to introduce the polar coordinate according to the basis of panel right angle coordinate system knowledge. Thus students will master their own way of studying. Then accumulation of knowledge is like the pyramid and connected together.

47.4.2 Use the Modern Teaching Method

Using the modern teaching method has become a common phenomenon, so Mathematical analysis teaching method is facing the opportunity and challenge. Some teachers go against using the multimedia teaching method to instruct Mathematical Analysis as they believe that it is normal and completely different from the old teaching system. But what we cannot deny is that the multimedia teaching method has many advantages, such as more visuals and more information. It has no diction and fault and it depends on how people use and operate it. So we should make full use of the multimedia, for instance we can save time if we use the multimedia to review our knowledge, the students can visually realize the process of the continual summarization by the multimedia, the trim integral part is more visual which make the students ensure the integral limit of accumulation integral more accurately and fast, the guarantee of the Mathematical experiments insertion and the photos which we use to introduce the mathematician, can also be shown to the students easily, and so on.

The multimedia educational method should make the process of Mathematical Analysis richer and more various. Therefore, there is a higher request for Mathematical Analysis teachers. First, a good knowledge of producing multimedia tasks and operating multimedia is necessary. Second, adding your educational thought and method of this text into text and in the performing test. Finally, according to the class reaction, the teacher should always correct this test.

47.5 Outlook for Mathematical Analysis Education

A good teacher is not only able to make his students understand what Mathematical Analysis is, but also to make a difference in the life of students; to make math as a life-long career when a math-major student strides forward into the university. What in his curious eyes exposes a confutation about the major-course study and a worry about the future? Mathematical Analysis is an important course which students face at the beginning. So what and how can we improve the students' math accomplishment as well as rising interests and confidence in the major-course study which is a question which our teacher should think about. There is a heavy burden and a long road in the reforming of Mathematical Analysis education. Let us strive together to make this event a smooth one for the country with the talented-person-training well done.

Acknowledgments This chapter is supported by the Educational Project of College of Light Industry, Hebei Polytechnic University. Project supported by Educational Commission of Hebei Province of China (No. 2009448).

References

1. Zhang S (2005) Mathematics education and mathematics culture [J]. Math Bulletin 44(1):4–9
2. Stockhausen G, Keitel C, Kilpatrick J (1992) Mathematics curriculum development [M]. Shanghai Education Publishing House, Shanghai
3. Tu R (2001) Mathematical instructional theory in the light of constructivism [J]. J Nanjing Norm Univ (Soc Sci) 3(2):77–82
4. Xu L (2009) Exploration and practice on the reform in mathematics course in normal Universities [J]. J Math Edu 9(2):1–6
5. Ji F (2007) Higher mathematics curriculum suggestions for improvement [J]. J Tonghua Teach Coll 28(8):91–93

Chapter 48
Research on the Risk of College Stadiums BOT Project and Countermeasures

Li Hongchang and Li Hongyu

Abstract In this chapter, the build-operate-transfer (BOT) model in the investment and financing system of the universities in China is researched. The necessity and feasibility of applying the BOT model and the notable problems are posed, and then the risk factors in financing process are analyzed and a series of risk reduction strategies are proposed.

Keywords BOT model · University physical education · Financing · Risk

48.1 Introduction

The build-operate-transfer (BOT) concession model has been a major trend during the recent two decades in the privatization of public sector infrastructure projects. BOT is an approach the private sector utilizes to obtain a granted concession for completing a specific project independently. First, the government department passes the franchise agreement. Second, the promoters carry out financing, designing, constructing, managing and maintenance of this project and operate facilities stated in the concession contract and recover its investment, operating and maintenance expenses from the project. Finally, the ownership of the project has to be returned to the public sector once it is entirely completed [1, 2]. To carry out a BOT project, both sectors take advantage of risk sharing from each other.

L. Hongchang (✉) · L. Hongyu
College of Physical Education, Beihua University,
Jilin 132021, China
e-mail: guomingli95@163.com

L. Hongyu
e-mail: lhc2100@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_48,

Along with the development of education in China from the stage of elite education to the stage of mass education, physical education in universities face various constraints, among which, lack of advanced technology and lack of infrastructure are two major drawbacks. Physical education businesses need a lot of funds. Due to high cost and lack of funding in building new campuses, physical education in universities cannot only depend on the government and schools which will bring enormous pressure to school finance. Thus physical education in universities should expend financing channels relying on the market. To overcome or alleviate these constraints, the country is encouraging local and foreign private sector involvement in the provision of infrastructure projects or services.

BOT project as a new model of financing is a meaningful attempt of expending physical education financing channels. BOT project is widely utilized by the construction of public infrastructures because of its efficient investment structure and significant social benefits and is also applied in the construction of new campues successfully.

The main method in adopting a BOT model in constructing college stadiums is as follows: the school provides the land-use rights and the investors provide the funds from the co-financing department. They formulate the stadium programs and construction requirements and entrust the construction tasks to the contractors by signing the concession contract. After the completion of the stadium, the ownership of the project attribute to the school and the management rights will be ceded to the private sector. The private sector must satisfy the classroom teaching and obtain the cost of investment by operating independently. At the end of the agreement, the management rights will be returned to the school. In this way, the school can achieve the objective of promoting school sports facilities.

In this chapter, we analyse the necessity and feasibility of applying the BOT model and pose the notable problems, and then analyse the risk factors in financing process and propose a series of risk-reduction strategies.

48.2 Practical Significance of College Stadiums BOT Financing

48.2.1 Be Beneficial to Lighten the Public Educational Financial Burden and Promote the Cause of Physical Training

In recent years, physical education has developed rapidly in our country. But with the number of students and sports events between universities continually increasing, college stadiums cannot adapt to the requirement of modern development. Construction of physical facilities requires a heavy outlay. So BOT is helpful to lighten the public educational financial burden and promote the cause of physical training.

48.2.2 Be Beneficial to Reduce the Economic Costs and Improve the Quality of Stadiums

Upon successful construction of the project, the actual concession period can last for 10–50 years depending on the type of the project. The private sector must take a long view in such contract negotiation, because of the long duration and high capital costs of infrastructure projects and changing priority of the host governments. The private sector can consider the quality. So BOT is beneficial to reduce the economic costs and improve the quality of stadiums.

48.2.3 Be Beneficial to Improve the Economic Benefit

If college stadiums are used only in physical education teaching, this is a costly waste. But this cannot only rely on the strength of the school. The private sector of BOT project can make full use of physical stadium resources. So BOT is beneficial for improving the economic benefit.

48.2.4 Be Beneficial to Promote Nationwide Health-building Drive

The BOT project can prove to be beneficial to the public and cut down sports consumption. It will remarkably promote the construction of gymnasium and stadium facilities and professional sports teams of the host city, increase the sports population of the city, expand the operation experience of holding comprehensive sports events, and improve the managerial level of such activities and sports culture of the host city, and promote the development of the sports industry of the host city, thus promoting the sports competitiveness of the host city in an all-round way. So BOT is beneficial to promote the nationwide health-building drive.

48.3 Problems Requiring Attention in College Stadiums BOT Financing

48.3.1 Perfect BOT Financing Legislation, Strengthen Supervision and Inspection

During the process of BOT project's implementation, there exist many difficulty and problems caused by the deficiency of legal norm, sequential stipulation, feasibility, whole content and complete set law etc. Our country should draw up

College Stadiums BOT Financing management systems that not only accord with the situation of our country but also conform to common international practices. The government should formulate College Stadiums BOT Financing policies and legislations that can be used to regulate investors and define the right and incumbency of members and bring College Stadiums BOT management under a system of laws and standards.

48.3.2 Establish the Common Government Administrative Setup

BOT project executive process is a long and complex process. This involves multiple agencies such as the financial, taxation, municipal and traffic control departments and involves many companies such as management, loan bank and insurance companies. The government should set up a special agency to exercise a system of unified administration. The State shall formulate preferential policies to support the College Stadiums BOT project.

48.3.3 Establish a Risk Prevention Mechanism

The school and private sectors may face many uncertain factors associated with BOT projects during the planning, construction and operation periods. As for the varied uncertainty factors such as land acquisition delay, completion delay in construction, concession period change, interest rate, toll regulation, political change, some may become risk factors and some will remain uncertain when BOT projects are being undertaken. The aim of a BOT concession agreement therefore is to reduce those mentioned risks produced through contract negotiation. However, it is a critical step for two parties to assess risks among many uncertain factors, and then to determine the primary and secondary risk factors of concern to individual parties before the concession negotiation. Thus the school and private sectors should establish a risk prevention mechanism for assuring that the BOT project can be held successfully.

48.3.4 Make Adequate Preparations for the Return of the BOT Project

BOT project implementation experiences at home and abroad reveal that the private sectors may lose project management and project maintenance toward the end of the agreement. So the school should have a periodical check on the BOT project. The school should set about making preparations for withdrawing stadiums two years before the contract ends.

Extract home and abroad successful experiences.

So far, many colleges have carried out BOT projects with some successful cases. We should learn from both their successes and mistakes, and learn to profit from our own experience as well as that of other countries.

48.4 The Controllable Risk of College Stadiums BOT Project and its Countermeasures

Since the BOT project has the special advantages of reducing the financial pressure of government and study, expand financial support and improve management, it plays a more and more important role in reformation. However, it also brings up a more complicated environment and more risk, which make the managers pay more attention to risk analysis.

There are various risks associated with BOT projects, such as social and political risks, environmental risks, technical risks as well as economic risks[3]. The process of the whole BOT project is very long and needs a multitude of participants. So the risks may emerge at different stages of the project life cycle. Social and political risks include internal resistance, nationalization, labor resistance, political influence, uncertainty of government policy and instability of government, corruption including bribery, unfair process of selection of private investors, changes in laws and regulations, inefficient legal process and legal barriers. Economic risks include devaluation risk, foreign exchange risk, inconvertibility of local currency, inflation risk, interest risk and small capital market demand and supply risk, incapable investors, too small number of interested investors, general liability risk, management risk and price escalation.

Although Finnerty [4] defined nine risk factors in BOT projects, including completing, technical, material supply, economic, financial, currency, political, environmental, and force majeure risk, BOT risks are classified as controllable risks and uncontrollable risks from risk factors. Uncontrollable risks are the chance that a loss of this kind may occur beyond the scope of the project the company can control. Such risks relate to the conducive environment of the micro-market, including political risk, policy and legislative risk, political risk, commercial risk, inflation risk, force majeure risk, etc. Controllable risks are the chance that a loss of this kind may occur in the scope of the project that the company can control in the process of stadium construction, operation and management, including capital risk, contractor risk, material supply risk, market risk, operational risk, etc. The division of these two kinds of risks is neither relative nor absolute. In this section, we focus on how to countermeasure some controllable risks of college stadium BOT projects.

48.4.1 Capital Risk and its Countermeasures

Capital risk mainly includes two aspects: first, risks that credits or financiers of the project cannot provide funds according to the contract stipulation; second, risks caused by unreasonable capital structure after BOT project financing. In BOT projects, in order to avoid such risk, the private investors should arrange properly the fund proportions in the construction of stadiums, choose the proper fund forms and improve the utilizing effect of stadiums. In the course of college stadiums BOT financing, the investors' fundamental concern is the ability to pay the debt. The investors may assess the risk by using debt ratio

$$\mathrm{CR} = \frac{\mathrm{DCF}}{\mathrm{D}} \tag{48.1}$$

where DCF denotes that the discounted cash flow and D denotes that the loans. Generally, CR should be limited to 1.3–1.5.

48.4.2 Contractor Risk and its Countermeasures

Contractor risk includes design contractor, construction contractor and operation contractor. This risk relates to all processes of the project, from design, construction to operation. From this point of view, contractor risk is an all-process risk of BOT project, mainly including the technique risk, management risk, intelligence risk, credit risk, cost control risk, project delay risk and so on. In the real operations of BOT project, it mainly uses the contractor risk protection to avoid the risk since the main causes of this kind of risk are the credit and strength of the contractor. Currently, the departments with high credit and strength, mainly country-owned companies, are less risky. The operations are mainly done by BOT professional project companies, which will be a little risky unless some unpredictable change in government policies are encountered.

For contractor risk, schools can require payment, maintenance, performance, advance payment, and completion guarantees from the contractor. The bond from commercial guarantee companies, on the one hand, can transfer and reduce the contractor risk to project company, on the other hand, it can oblige the contractor to do their best to finish the project following the project design in the time limit by the effect of guarantee companies. Because the contractor must be monitored by the guarantee companies, the risk about finishing project in time limit and the risk about quality are reduced. Some districts of China have brought out several district-wide laws to operate project guarantee systems, which provide good preconditions for BOT project investor to avoid risk, and make a great effort to let our national construction management to meet the international demands.

48.4.3 Operational Risk and its Countermeasures

When facing the corruption of officials of local governments, the operational cost would be very high. Most of the time and money would be spent on coordination, but not on the operational improvement and service improvement, which will lead to great increase in operational cost, great decrease of operational profit, and finally non-profitable project. Currently, facing the market economy in China, the profit fluctuates with the market and the sponsors cannot rely too much on the government's protection . Hence they should take care of market risk.

References

1. Hwang YL (1995) Project and policy analysis of build-operate-transfer infrastructure development, Ph.D. Dissertation, Department of Civil Engineering, University of California, Berkeley
2. Walker C, Smith AJ (1996) Privatized infrastructure:the build operate transfer. Thomas Telford Publications, London
3. David AK (1996) Risk modeling in energy contracts between host utilities and BOT plant investors. IEEE Trans Energy Convers 11(2):359–366
4. Finnerty J (1996) Project financing. Wiley, New York

Part V
Enabling Technologies

Chapter 49
Research on Awareness Driven Schedule for Sensors in Web-Based IOT

Haoming Guo, Shilong Ma and Feng Liang

Abstract In smart IOT, sensor resources are organized to monitor events in real world for all time that may produce huge number of data. Due to difference of resources' awareness of event, the valuable data may be flooded by dull data if task processes without discrimination. This paper introduced an approach, called Awareness Driven Schedule (ADS), which enables involved resources that provide differentiated data service on their capability of awareness. Upon ADS, a middleware, called SSRDSs is built for CAE's SPON. As dull data could be banned out, applications in IOT may be more efficient.

Keywords Sensor · Resource schedule · IOT · Web service

49.1 Introduction

In smart IOT, sensors are primary resource objects to be organized [1]. Sensor is to monitor events in real world [2]. The primary need is not functionality but sensors' awareness. The target events are unexpected, ambiguous and dynamically developing. Consequently, sensor's awareness is changing. As a result, the huge data collected by sensor should be processed by task with discrimination. The sensor resource schedule and data organization policy are different from the conventional one. For example, in earthquake application: Seismological Precursory Observation (SPO), SPO's goal is to find out exceptional vibration's intensity and location

H. Guo (✉) · S. Ma · F. Liang
Computer School, Beihang University,
New Main building G1134, Xueyuan street 37#,
Haidian district, Beijing, China
e-mail: guohm@nlsde.buaa.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_49,

exactly and assess affected area in real world. Various sensor resources are spread all over the area. These sensors are automatically responsible to collect data, monitor exceptional vibration's occurrence, be manipulated by tasks, etc. As shown by above example, sensor resource in SPO's world has two distinguished features: (1) Data's quantity is huge: Sensor is to watch real world constantly. It generates data as it works continuously. (2) Data quality is not static: data's accuracy is affected by its working conditions around and signal intensity it receives. As target event changes, data accuracy may change either. Some data may be highly valuable while others may by dull to application. As a result, resource's service quality to task is differentiated.

In IOT's world, the goal of sensor resources schedule is to enable differentiated continuous services in accordance with resource's data quality and its awareness. In this schedule approach, task publishes data requirements. All resources are involved in accordance with whether they could provide required data. Higher accurate the resource's data is, better service it provides. Through this awareness driven schedule, resources are organized in accordance with their leveled data so that unnecessary huge data transfer may be reduced while guarantee credible and continuous data service for tasks.

49.2 Related Work

Derived from the traditional wireless sensor network, Sensor Web is now widely used in fields such as Bio-complexity mapping of the environment, Military applications, flood detection, traffic management and, etc [3].

Conforming to the SWE standard, NICTA Open Sensor Web Architecture (NOSA) [4] is a software infrastructure aimed at harnessing massive computation power of grid computing in sensor web. However, the quality of data is not considered as a standard way for result processing.

Earth Observing 1 (EO-1) [5] is to provide a sensor web for assisting transient science events observation. It allows autonomous resource optimization for mission resources and autonomous event triggered image by cooperating sensors on the ground together with the satellites and other assets. Although EO-1 considers the location of the sensors, the quality and accuracy of the data from the source are still processed indiscriminately.

According to the above mentioned projects, it can be seen most of the information-driven middleware implements the scheduling by processing the data indiscriminately, this scheduling mechanism is suitable for information browsing activities, but inefficient for emergencies which involve variety of parties and monitors distributed dynamic event sources because of longer processing time and more resource. Therefore a sophisticated and effective mechanism is required for data filtering and scheduling.

49.3 Awareness Driven Schedule

49.3.1 Whole View of ADS

As mentioned above, ADS's goal is to enable differentiated continuous services and dynamic resource involvement in task in accordance with resource's data quality and its awareness. In ADS, tasks register awareness requirement (AR) to Awareness Requirement Registration (ARR) and Create Task Awareness Schedule (TS) in Task Awareness Scheduler Manager (TSM). In TS, data channels are defined and built for data differentiation and service forwarding. In the AR, information type and data value definition are listed. TSR searches all resources that can provide same data as defined by AR and invoke. Resources create task object handler (TOH) in local task object handler pool (THOP). Once a resource is aware of target event, it collects data and check data channels' definition from related TS. Data transfer frequencies are listed in TS's data channel definition. Resource retrieves the frequency information by which it transfers data. In TS's data channel definition, data process services are defined. All data from one channel is about to forward to the specific service. If resource's data value shift from one range to another, TS may link related data channel to resource and reassign data transfer job. If resource lost awareness of event, it cut off link to TS. If a resource finds the event, it checks TSR with the event information and retrieve related AR upon which links to TS are built.

49.3.2 Awareness Requirement Registration

As mentioned above, ARR is to organize task's (AR) through which resource could involve into task's working group. ARR's definition is shown as: ARR = {ARi|AR1, AR2,…ARn}. AR is task's awareness requirement definition. In AR, task lists what requirements of resource to pool as working group and task's data channel definitions. Its definition is shown as below:

$$AR = \{ID,\ taskID,\ resPTable,\ dataChannelList\}$$

taskID is the task's identity through which data channels forward messages to right destination. resPTable is to define what kind of resource task needs. It consists of a table of property. resPTable = $\{p_i|p_1, p_2,\ldots p_m\}$. p = {name type, value}. While looking for resources from resource registration at the beginning, if AR's all property requirements matche one resource's properties, the resource may be included in task's initial working group. dataChannelList is to create data channels for task. Data channel is to create link between resource and related data process in accordance with resource's awareness or its data accuracy.dataChannelList = $\{dcf_j|dc_1, dcf_2,\ldots dcf_k\}$.

$$\mathrm{dcf} = \{\mathrm{ID, taskID, proc, maxValue, minValue, frequency, transMod, cacheSize}\}$$

In dcf, proc is object in task to process the data with required accuracy. maxValue and minValue are to define range of the channel. During data transferring, resource sends data on the way which is defined by transMod. cacheSize is to apply data cache in resource's data pool.

Once task's AR is registered, its Task Awareness Scheduler (TS) will be created in TSM. TSM consists of a group of TS as: TSM = {$\mathrm{TS_i}$ | $\mathrm{TS_1}$, $\mathrm{TS_2}$,…$\mathrm{TS_n}$}. TS is to keep contact with resources, receive and forward data to right data process object in task and schedule resource's service. Its definition is as: TS = {ARID, dcs} ARID is corresponding with AR'ID in ARR. One AR has one TS created. dcs is data channel list in TS. It consists of a group of data channel: dcs = {$\mathrm{dc_j}$| $\mathrm{dc_1}$, $\mathrm{dc_2}$,…$\mathrm{dc_m}$};

$$\mathrm{dc} = \{\mathrm{ID,\ dcf\ ID,\ RHOPool}\}$$

dcfID is data channel's definition identity through which data channel may retrieve information from corresponding AR. RHOPool is pool of resource handler object (RHO). RHO is persisted for task to receive data from corresponding resource. Once a resource is invoked, a RHO will be created and pooled in related data channel's RHOPool. RHO is defined as below:

$$\mathrm{RHO} = \{\mathrm{ID,\ taskID,\ resBinding,\ dataCache}\}$$

taskID is to reserve RHO's hosted task identity. The resBinding is used to keep the binding information. Through the information, RHO may redirect messages to right resource. dataCache is used to cache data collected by resource by time order. After AR's TS is created, ARR searches for all resources. The result resources of the search are organized as initial working group of task. All resources in the working group would be invoked to create link with AR's TS. In invocation, a task handler object (THO) will be created for the request. THO's definition is shown as below:

$$\mathrm{THO}\ = \{\mathrm{ID,\ taskID,\ RHOID,\ dataCache,\ dcf}\}$$

The whole AR registration is shown in Figs. 49.1 and 49.2.

49.3.3 Data Linkage for Resource and Task Process

After data channel connection, the resource needs to get relevant data channel's information. By the information, resource retrieves data channel's definition dcf from ARR. In dcf, data channel's data range is defined and the resource transfer data by rules of dcf while data is within the range. Resource's THO return data to task's data channel first. Data channel forward the data to related RHO. RHO looks

Fig. 49.1 Communication for AR registration $RegProc = Task.\overline{regAR} < AR > .ARR.\tau.\overline{createTS} < AR > TSM.\tau.createTS(TSID) ARR.\overline{lookFor} < resPTable@AR > RR.lookFor(resList)ARR.invokProc.0 invokeProc = ARR.\overline{ivkTHO} < TSID > THOPool.\tau.\overline{createCon} < TSID > THO.\overline{getTS} < TSID > TSM\tau.getTS(TS).createConnProc.0$

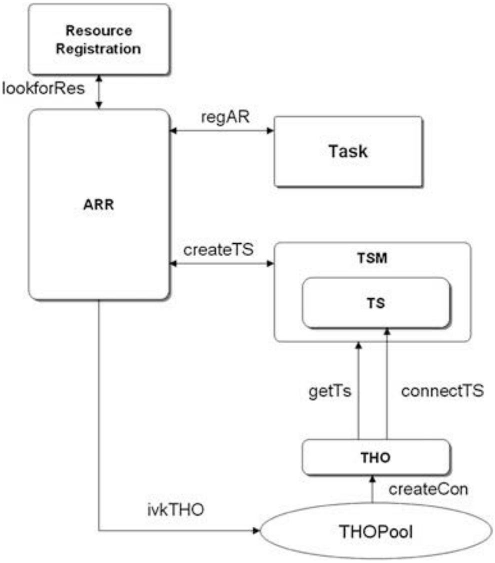
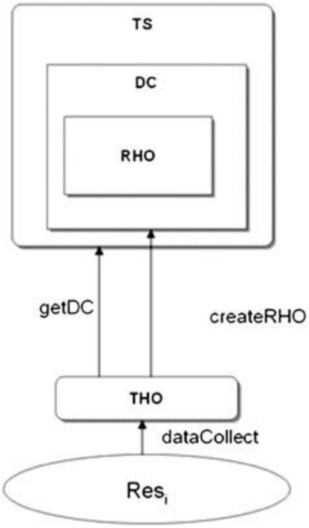


Fig. 49.2 Communication for TS registration $createConnProc = THO.\overline{getDC} < dataCache > TS.getDC(DC).\overline{createRHO} < resBinding > DC.createRHO(RHOID).0$



for data channel's definition from ARR and retrieve process object of task which is persisted in dcf. Then RHO transfer the data to the process object (Fig. 49.3).

In the in task's awareness requirement, task's process object is persisted. During implementation, THO transfers data back to RHO through it working data channel. RHO looks for the process object linked to hosted data channel's

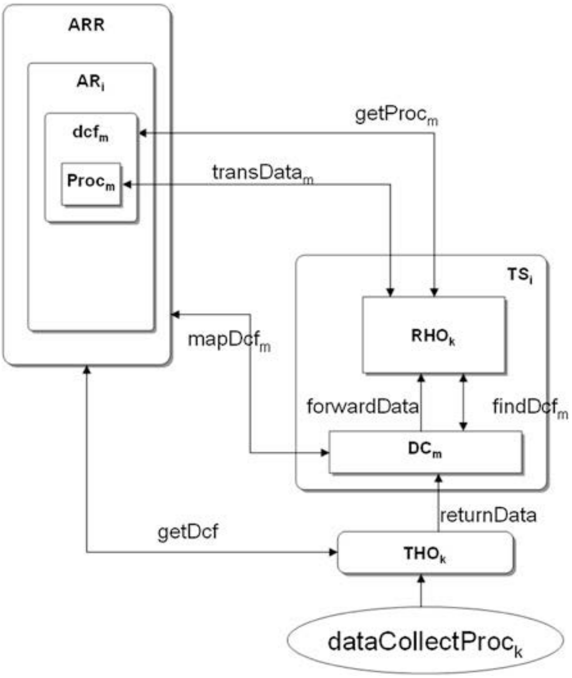


Fig. 49.3 Communication for data collection task $serProc_k = THO_k\overline{getDcf}<dcfID@DC> ARR.getDcf(dcf).dataCollectProc_k.\ THO_K.\overline{returnData_k}\ <dataCache_k> DC.forwardData_k<dataCache_k> RHO_k.TaskDataProc_k.0$ $TaskDataProc_k = RHO_k.\overline{findDcf_m}\ <dcfID>.DC_m\overline{.mapDcf_m}\ <dcfID> ARR\overline{.mapDcf_m}(dcf_m)$ $\overline{findDcf_m}\ <dcf_m>.RHO_k.\overline{getProc_m}<NULL> dcf_m.getProc_m(Proc_m).RHO_k\ \overline{transData_m}<dataCache_k>.0$

definition and forward data to it. In ADS, resource may provide data service constantly by this approach.

Resources are to monitor real world's event and collect data. In ADS, resource's data collection job is ruled by its linked data channel. In data channel's definition, maxValue and minValue are to define current data channel's range. If one resource's collected is within the range, it keeps data collection for current data channel. Otherwise, resource looks for new channel in current TS and collect data by the new one's rule.

49.3.4 Resource Awareness Orientation

Once a resource's collected data is out of current data channel's range definition. It may check ARR for new oriented data channel and shift-related RHO from old hosted data channel to the new one. If resource lost awareness of the event, it will be removed from data channel. During this process, the ROH shift request message is defined as:

$$\text{changeDCReq} = \{\text{ID},\ \text{RHOID},\ \text{resBinding},\ \text{taskID},\ \text{oldDCID},\ \text{newDCID}\}$$

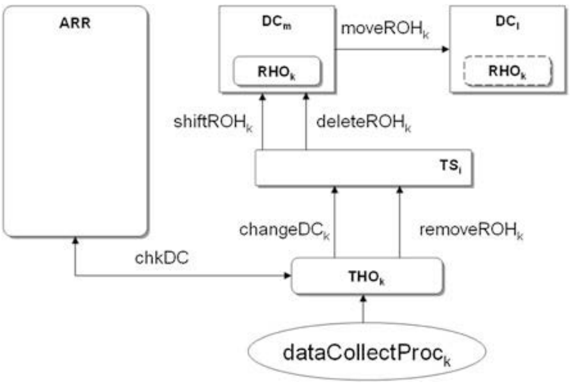


Fig. 49.4 Communication for awareness shift $AwChangeProc_k = THO_K \overline{chkDC} <dataCache > ARR.chkDC(chkResult). ((chkResult = NULL). \overline{removeROH_k} <removeROHReq_k > .TS_j. \overline{deleteROH_k} <RHO@removeROHReq_k > . \tau.0 + (chkResult = dcf_l). \overline{changeDC_k} <changeDCReq_k > .TS_j. \overline{shiftROH_k} <changeDCReq_k > DC_m.moveRHO_k <RHO_k > .DC_l\tau.0)$

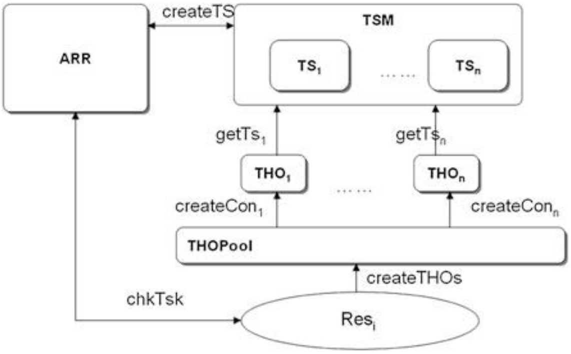


Fig. 49.5 Communication for service shift $chkNewTsk = dataCollectProc. \overline{chkTsk} <chkTaskReq > ARR.chkTsk(tskResp). \overline{createTHOs} <tskResp > .THOPool.\tau. (createCon_1 <TSID_1@tskResp > THO_1 \overline{getTS_1} <TSID_1@tskResp > TSM\tau. getTS_1 (TS_1).createConnProc_1 .0|createCon_2 <TSID_2@tskResp > THO_2 \overline{getTS_2} <TSID_2@tskResp > TSM\tau. getTS_2(TS_2). createConnProc_2.0| \ldots.createCon_n <TSID_n@tskResp > THO_n\overline{getTS_n} <TSID_n@tskResp > TSM\tau. getTS_n(TS_n). createConnProc_n.0)$

In the request message, RHOID is related RHO identity. TS retrieves the object through the identity. resBinding is information about resource. taskID is current task's identity through which locates related TS. oldDCID is current linked data channel's ID and newDCID is the new data channel to link.The RHO remove request message is defined as:

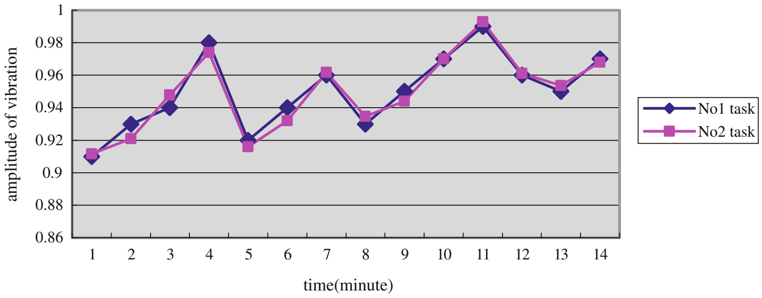


Fig. 49.6 Comparion of ADS and conventional

$$\text{removeRHOReq} = \{\ \text{ID},\ \text{RHOID},\ \text{resBinding},\ \text{taskID},\ \text{oldDCID},\}$$

The process is shown as below:

If a resource begins to sense the event, it checks ARR with the data and its own property for tasks which require the data from ARR. ARR may return a list of available tasks' AR. The resource create connection with the tasks' TS and begin to provide data service. The check message is defined as below:

$$\text{chkTaskReq} = \{\text{ID},\ \text{dataCache},\ \text{pTable}\};$$

the ARR returned message is defined as:

$$\text{tskResp} = \{\text{TSID}_1,\ \text{TSID}_2, \ldots\ldots \text{TSID}_n\}$$

the process is shown in Figs. 49.4 and 49.5.

49.4 Application and Test

Upon ADS, a Seismological Sensor Resource Data Service System (SSRDSs) is built for CEA's Seismological General Scientific Data Platform (SGSDP) built for SPON. In test, 120 resources are deployed to simulate application environment. The resource's data collection working frequency is about 60Hz which means one resource may produce 60 data per second. All 120 resources may produce 360 data per second. The amplitude of simultaneous vibration is about 1 mm. All resources could sense the vibration. According to their distance to the simultaneous vibration, two tasks were executed for comparison. No.1 collected all data directly from resource without discrimination. No. 2 task created 4 data channels and resources provided differentiated data services. During implementation, 120 resources transferred data back to No. 1 task at about 360 data per second. In No. 2 task, resources transferred at about 168 data per second. For No. 2 task, data load was

47% of No. 1 task. Data lost may lead to certain accuracy lost. Figure 49.6 shows two task's data aggregation curve. No. 2 task's result's accuracy is lower than No. 1 task's. However, it was in application's accuracy requirement.

The test above shows effectiveness of ADS for data concentrated applications of smart IOT.

49.5 Conclusion

This paper introduced an approach called ADS. Through task's requirement, ADS organize all involved resource and enable them to provide differentiated data service to task. Higher a resource's awareness is, more detailed data it should collect and transfer. As a result, low awareness of resources only need to provide limited and periodic data service. Resources' data service are differentiated with their awareness. Dull data are banned out.

Acknowledgment This research work was supported by the Special Program for Seism-Scientific Research in Public Interest "Research in Online Processing Technologies for Seismological Precursory Network Dynamic Monitoring and Products" (NO. 201008002).

References

1. Jensen F (2006) Introduction to computational chemistry. Wiley, New York
2. Chaouchi H (2010) The internet of things: connecting objects. Wiley-ISTE, New York
3. Hu W, Bulusu N, Chou CT, Jha S, Taylor A, Tran VN (2009) Design and evaluation of a hybrid sensor network for cane toad monitoring. ACM Trans Sen Netw ACM 5:1–28
4. http://eo1.gsfc.nasa.gov/new/validationReport/index.html
5. Fernandez-Baca D (1989) Allocating modules to processors in a distributed system[J]. IEEE Trans Softw Eng 15:1427–1436

Chapter 50
Enabling Sensor Resource to Provide Constant, Consistent and Continuous Service for Web-based IOT

Haoming Guo, Feng Liang and YunZhen Liu

Abstract Sensor resources are required to provide constant, consistent and continuous service for tasks in IOT. However, due to conventional web technology's communication limitations, these sensor resources' applications are confined while expanding upon internet. This paper introduced a message exchange approach, called ICDME, to enable the resources to keep constant contact with tasks, intervene task's work initiatively and organize continuous data flow for web-based smart IOT. In this approach, the handler objects are used to cache data and create communication between resource and task. Through ICDME, task may retrieve data from sensors continuously and constantly without building long connection repeatedly.

Keywords Message exchange · IOT · Web service

50.1 Introduction

Web-based smart IOT is a result of combination of smart IOT and service centric web [1, 2]. In traditional web fabrics, the primary objects are warped by web service interfaces. Through SOAP message exchange, people or programs may access those resources. One service implements a specific data process or other works [3]. However, instead of resources organized by conventional service centric web, in web-based smart IOT, the primary objects are smart sensors or

H. Guo (✉) · F. Liang · Y. Liu
Computer School, Beihang University, New main building G1134,
Xueyuan street 37#, Haidian district, Beijing, China
e-mail: guohm@nlsde.buaa.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_50,

devices [4]. Those sensors or devices are borders between the real world and computing environment, and tools to monitor and change the physical environment. Their coordination policies and message exchange needs are different.

As in Seismological General Scientific Data Platform (SGSDP), there are different types of sensors. They can work under different settings so that they can protect themselves from being overloaded and collect real-time data more accurately. The sensors monitor and collect specific data around continuously. While seismological exception vibration takes place, the information is to be transferred to Precursory-Watching-Application (PWA). PWA is to process the data collected. Data consistency and timeliness are essential. In PWA, continuous data flow collected by sensors should be provided tasks for processing and analyzing. Working status information of sensors will be used as reference for task to process data. When sensors' working statuses change, the corresponding information in PWA task should be updated to guarantee credibility of data reference. The application of the example above shows different resource communication needs of web-based smart IOT. Once connected, resources constantly provide data collection services for task. Resources adjust working status autonomously and this change should be refreshed in tasks. Tasks process data collected from these resources with reference to the working status. Resources could initiatively intervene the task's work with notification of it working status information. They are organized for application tasks with longer connection, more initiatives and for providing continuous services. The message exchange requirement between resource and application is called Initiative and Continuous Duplex Message Exchange (ICDME) which is different from traditional service centric web in which communication between resource and task occurs in request–response way periodically.

50.2 Related Works

As service centric web technology began to merge with IOT, researchers' attentions have been attracted for the issue. In conventional web applications, Constant communication between resource and task is realized by frequent connection. However, the frequent connection establishment between request and response end may cost unnecessary burden for system. The time delay and data lost between invocations are worse and unacceptable for applications such as PWA, mentioned above. Consequently, engineers have worked for stateless web service's communication to enable more complex message exchange. The WS-Notification family has been introduced for that purpose. In IOT's world, however, sensors will be connected temporally for tasks. Unexpected message should be exchanged and processed between task and resources so that it is hard to develop IOT's application, such as in PWA, upon WS-Notification. Moreover, in WS-Notification, the whole system's stability is heavily relied on Subscribe/Notification center that makes it choke point of the whole system.

As researchers realize the different needs of data exchange in IOT web, new middleware are introduced to organize tasks and resources. In the paper [4], data collection and service organizing have been listed as prior problems for sensor webs' development. In the paper, an architecture named, IrisNet, is introduced. In IrisNet, distributed databases are built for sensors as buffer. Through query tool, system provides data access service to tasks. On the other hand, based on SWE some researchers has introduced approaches to build sensor web. The paper introduced Message bus [5] pattern and a publish/subscribe communication infrastructure to realize the goal. However, data and changed metadata are exchanged separately by SWE service which makes the data inconsistent.

The works introduced above, try to answer the resource and data organization for tasks in web-based smart IOT. However, the key requirements of constant, consistent and continuous service are not well addressed.

50.3 ICDME Approach

50.3.1 Whole View of ICDME

In application process, task creates resource handler object (RHO) for resources it connects. The RHO is to receive the resource invocation request message, add invocation information to the message, dispatch the message and return the result message to request. Once the RHO is created, it will be cached in resource handler object pool (RHOP) which is used to cache RHOs and enable task or resource access specific RHO. At the end of task, RHOP is responsible to dispose RHOs it caches. In resource invocation, RHO receives request from task. The message will be sent to the relevant resource service through invocation agent interface (IAI). Once the resource service is invoked, it may retrieve task handler object (THO) from task handler object pool (THOP). If there is no such object exists in the pool, a new one will be created by resource service and cached in THOP for future use. For every task, resource service creates one unique THO. The THO will be used to implement resource works, cache and dispatch working status information, organize continuous data flow and exchange message with its mapping RHO at the task end. The THO will be cached in THOP tile task's end. During resource invocation, THO has resource to work and return invocation result message back to task. Through resource handler object access service (RHOAS), THO and RHO exchange messages. While the task is going on, THO continues to manage resource's work and collect data. When working status changed, resource working status maintenance (RWSM) collects the information and delivers it to THO, THO redirects the message to RHO through RHOAS. Consequently, RHO may notify task react to the change. Data collected by resource will be cached by THO. By certain time configuration, the cached data may be sent back to RHO by THO through RHOAS. The RHO is responsible to organize received data in time order.

Through this approach, there will be no lost of data between task requests. As a result, duplex message exchange and continuous data flow between resource and task will be realized for web-based smart IOT.

50.3.2 Definition of ICDME

As introduced in Sect. 50.3.1 RHO is created uniquely by task for resource it will access. It is responsible to invoke resource, redirect work status changing information and organize data flow. It is composed of five parts:

$$\text{RHO} = \{\text{ID}, \text{ appID}, \text{ taskID}, \text{ resBinding}, \text{ dataCache}, \text{ paraArr}\}$$

ID is the RHO's identity. Through ID, RHOP seek and retrieve the object. appID is application's identity. taskID is to reserve RHO's hosted task identity. Through taskID, RHOP organize RHOs table. In task, the RHO may be connected to a specific resource, the resBinding is used to keep the binding information. Through the information, RHO may redirect messages to the right resource and THO. dataCache is used to cache data collected by resource by time order. It consists of a list of dataCell object by time order, As dataCache = {dataCell i | i = 1, 2,…n}. dataCell is defined as:

$$\text{dataCell} = \{\text{ID, RHOID, THOID, timeStamp, message, resBinding, pSnapShot}\}$$

In dataCell RHOID is used to identify to which RHO the dataCell belongs to. THOID is to identify from which THO the dataCell is collected. timestamp is to tell when the data is collected. By timestamp, dataCell object may be ordered in the dataCache. Message is the data collected from resource in the period of timestamp; resBinding is to identify resource's address which collects the data; pSnapShot is a table of resource working status parameters stored during data collection. It consists of parameters such as: pSnapShot = {pk | j = 1, 2,…m}; p is a parameter consisting of two parts: p = {name, value}; name is the parameter's name, value is the parameter's data. In RHO, paraArr consists of list of working status parameters as: paraArr = {pj | j = 1, 2,…m}

As counterpart of RHO, THO is created in resource end by resource service. THO is to be cached in local THOP. It is responsible to implement resource work process, collect data, redirect working status message, etc. THO consists of five parts

$$\text{THO} = \{\text{ID}, \text{ taskID}, \text{ RHOID}, \text{ RHOAddr}, \text{ dataCache}, \text{ paraArr}\}$$

ID is THO's identity, through ID, THOP seeks and retrieves the object. taskID is to reserve THO's related task identity. Through taskID, THOP organizes THO table. RHOID is the RHO's identity for the THO. RHOAddr is related to the THO's address. dataCache is used to cache data collected by resource by time order. Its definition is same as RHO's. paraArr is a list of working status parameters. Its definition is same as RHO's. RHOP is hosted in task end. It caches and manages RHOs for tasks. It consists of resource bind table, as:

$$\text{RHOP} = \{\text{resTbli}|\ \text{i} = 1, 2\ldots\text{m}\};$$

resTbl is table for specific resource. All task objects connecting to the resource are cached in it. It is defined as:

$$\text{resTbl} = \{\text{ID, resBinding, tskObjs}\};$$

ID is resTbl 's identity. ResBinding is resource's address information through which resTables are sorted by RHOP. TskObjs is a list of object tables for tasks of the application. It is defined as:

$$\text{tskObjs} = \{\text{taskID, RHOArr}\}$$

taskID is task's identity. RHOArr is objects array for the task. It consists of RHOs as:

$$\text{RHOArr} = \{\text{RHOk}|\ \text{k} = 1, 2\ldots\text{n}\}$$

THOP is hosted in resource end. It caches and manages THOs for resources. It consists of application table, as:

$$\text{THOP} = \{\text{taskID, THOArr}\};$$

taskID is task's identity. THOArr is the objects' array for the resource. It consists of THOs as: RHOArr = {THOk I k = 1, 2…n}

50.3.3 Constant Resource Contact

While task is going to connect a resource, it sends out message first. The resource invocation message is defined as:

$$\text{resIvkReq} = \{\text{ID, appID, taskID, resBinding, msg}\};$$

ID is request's identity. appID is application's identity; taskID is task's identity; resBinding is resource binding information. It includes the resource address and the interface the request is going to access. It is defined as: resBinding = {resAddr, interface}. msg is request SOAP message content. At the beginning, RHO which corresponding to the resource binded will be retrieved from RHOP for the request. If there is no such object, a new one will be created and cached in RHOP. The RHO retrieves message is defined as:

$$\text{RHOGetMsg} = \{\text{appID, taskID, resBinding}\}$$

Through appID, taskID and resBinding, RHOP retrieves related RHO and return to request. The RHO invocation process is defined as:

$$RHOIvk = \overline{RoReg} < RHOgetMsg > ROHP.\tau.RoReg(RHO).\overline{ResIvk} < resIvkReq > RHO. \\ \overline{ToIvk} < resIvkReq > IAIIvk.ToIvk(resIvkResp).\tau.ResIvk(resIvkResp).0 \quad]]$$

The invocation request message from RHO will be transferred to resource through invocation agent interface. By the invocation request message's taskID value, THO will be retrieved from THOP. If there is no such object exists in THOP, a new one will be created and cached by THOP and returned. The IAI invocation process is defined as:

$$IAIIvk = ToIvk(resIvkReq).\overline{ToReg} < taskID@resIvkReq > THOP.ToReg(THO). \\ \overline{Ivk} < resIvkReq > THO.Ivk(resIvkResp).\overline{ToIvk} < resIvkResp > RHO.0$$

The invocation request is transferred to THO which is responsible to start working implementation and returns response message back. The process is shown as:

$$THOIvk = \overline{resImp} < msg@resIvkReq > workProc.resImp(resIvkResp). \\ \overline{Ivk} < resIvkResp > IAIIvk.0$$

The whole resource invocation process is organized as shown in Fig. 50.1.

50.3.4 Consistent State Reference

In task, resources are connected and invoked. During its implementation, their working status may change. The change information should be transferred back to task which will take the information as reference will handle data collected from the resources. The message which contains resource's working status change information is defined as:

$$\text{resWsMsg} = \{\text{ID},\ \text{timestamp},\ \text{resBinding},\ \text{pSnapShot}\}$$

timestamp is to identify the message's time information. resBinding is resource's bind information that is used by RHUP to deliver the message to the related RHO. pSnapShot is resource's working status parameter array. Its definition is shown as in Sect. 50.3.1.

In resource end, once the working status changes, the process is shown as:

$$THONotf \ = \overline{WSM} < resWsMsg > THOP.\tau.\overline{WSMNotf} < resWsMsg > RHOAI.0$$

The message will be transferred from THOP to RHOAI to realize working status information reference maintenance in task.

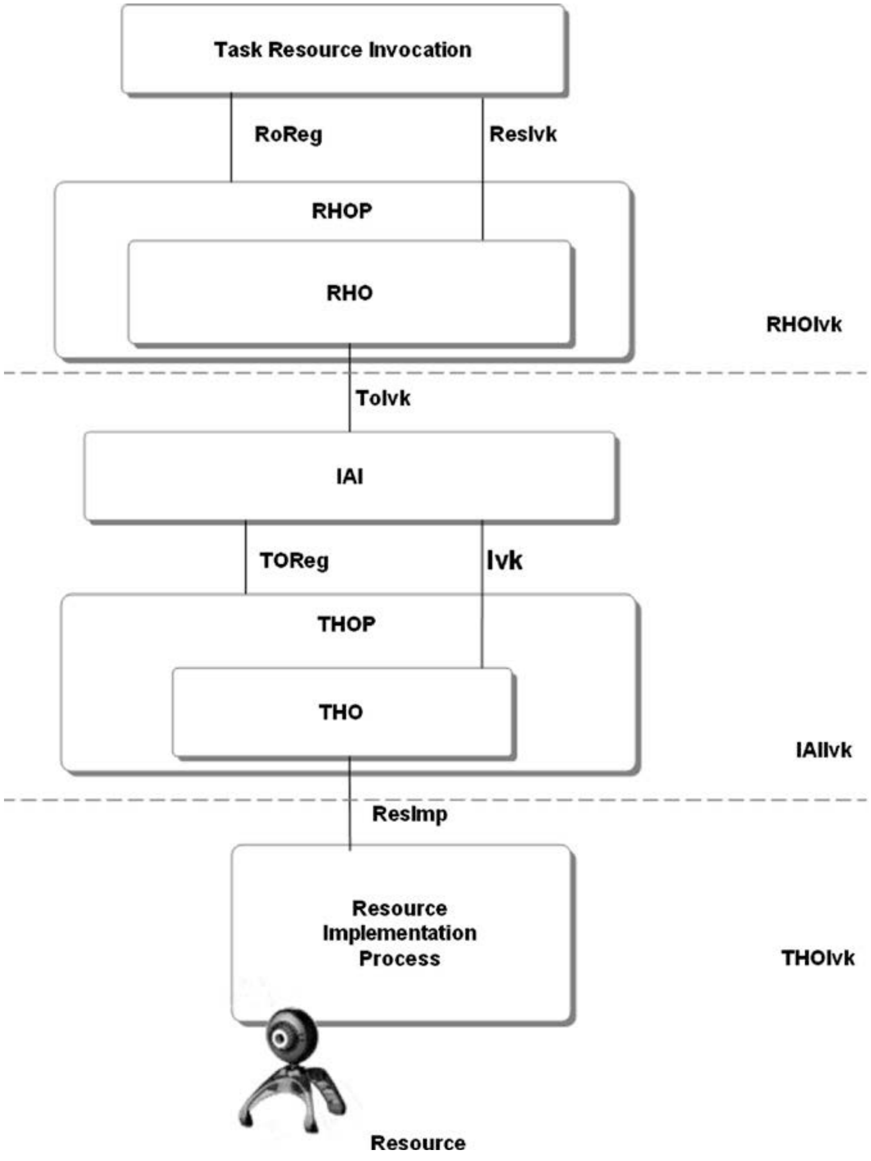


Fig. 50.1 Resource invocation process

Once RHOP receives the message, it retrieves resource table related to resBinding information of the message. Consequently, RHOs in the table will be notified and parameters will be refreshed. Tasks hosting the RHOs will task further actions for the changes if necessary. The process is defined as:

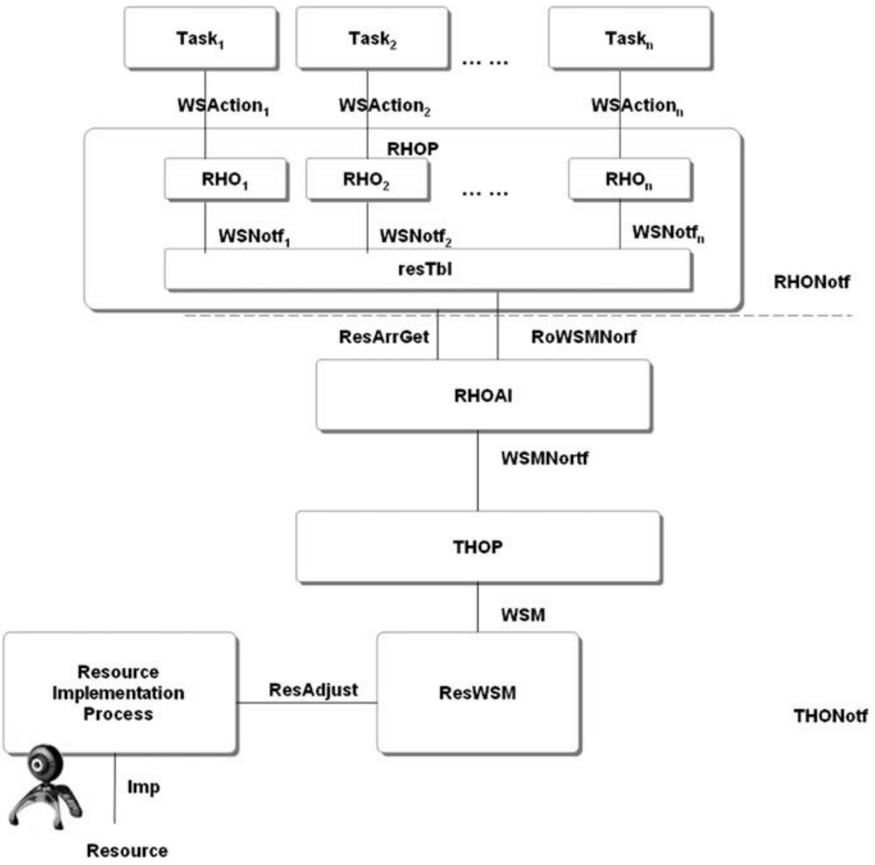


Fig. 50.2 Working status consistency process

$$
\begin{aligned}
&RHONotf = ResArrGet(resBinding@resWsMsg)RHOP.\tau.\overline{ResArrGet}<resTbl>.\\
&\overline{RoWsMNotf}<resWsMsg> resTbl., RHOsinthetablewill\\
&((\overline{WsNotf_1}<resWsMsg> RHO_1.\overline{WSAction_1} < pSnapsot@resWsMsg > .Task_1.0)|\\
&(\overline{WsNotf_2}<resWsMsg> RHO_2.\overline{WSAction_2} < pSnapsot@resWsMsg > .Task_2.0)|\ldots\ldots\\
&(\overline{WsNotf_n}<resWsMsg> RHO_n.\overline{WSAction_n}< pSnapsot@resWsMsg > .Task_n.0)).0
\end{aligned}
$$

The whole process is organized as shown in Fig. 50.2.

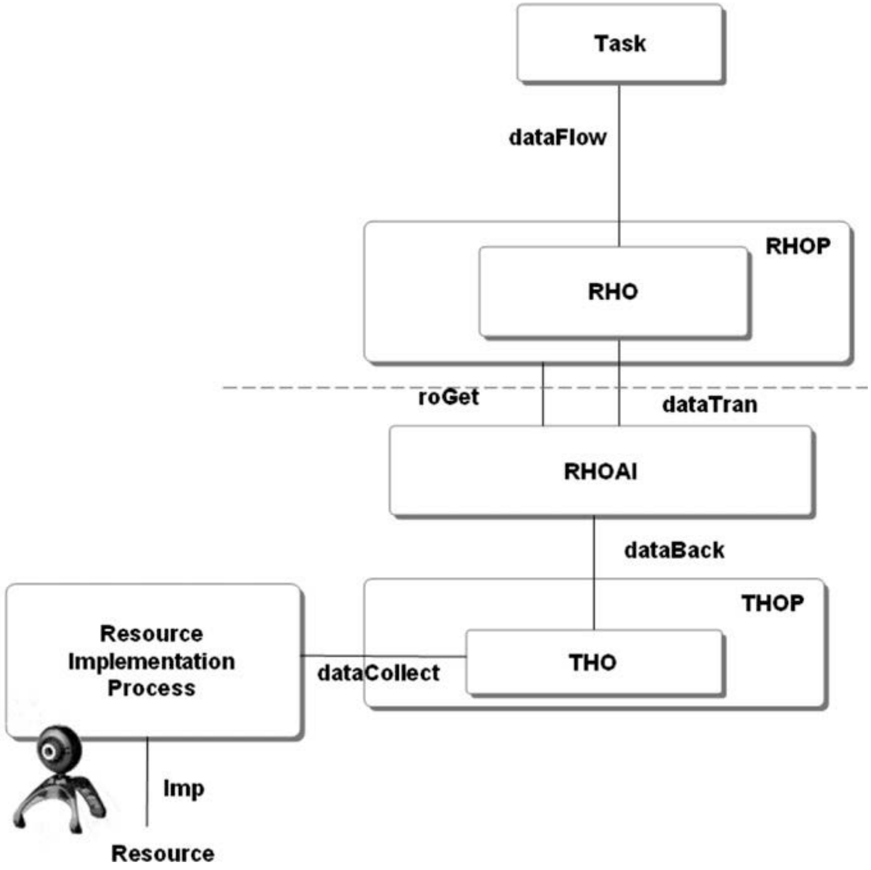


Fig. 50.3 Data flow organization process

50.3.5 Continuous Data Flow

Resource, as introduced in PWA applications, is to collect data for application tasks. Their data collection operation is continuous for messages to be sent back to task. In ICDME, once resource is connected, RHO and THO are created as pair and cached in task end and resource end. Data collected by resource are transferred by THO to task through its paired RHO. The data message is defined as:

$$\text{resDataMsg} = \{\text{ID},\ \text{taskID},\ \text{RHOID},\ \text{resBinding},\ \text{dataCache}\}$$

ID is the message's identity. TaskID is task's identity which is paired with RHO hosted. RHOID is to tell RHOP to which the data message ineeds to be dispatched. resBinding is resource binding information that is to be used by RHOP to retrieve relavent RHO. dataCache is data collected in the period.

Table 50.1 Comparison of ICDME and conventional web message exchange

	Average data lost (%)	Task average net load (%)	Sensor average net load (%)
In ICDME	< 0.1	< 15	< 22
In conventional web (5 s interval)	2.4	10.3	9.2
In conventional web (1 s interval)	1.1	17.5	15.8
In conventional web (0.5 s interval)	0.7	22.1	21.2

$$dataBk = dataCollect(message)THO.\overline{dataBack} < resDataMsg > RHOAI. \overline{roGet} < RHOID@resDataMsg > RHOP.roGet(RHO). \overline{dataTran} < resDataMsg > RHO.\tau.\overline{dataFlow} < message@dataCache@resDataMsg > .0$$

The whole dataflow organization process is shown in Fig. 50.3.

50.4 Application and Test

Based on ICDME, a middleware, called ‘SmartMe’, is built for CEA’ SPON to enable resource provide constant, consistent and continuous service for tasks while running. In the test, 12 sensors are deployed to simulate PWA’s application work. Through SmartME’s ICDME approach, they collect data and provide continuous data flow to the task. Because the message exchange mechanism is duplex, task does not need to invoke resource repeatedly to retrieve data. Sensors send data by certain frequency. The order of sensors to transfer data is scheduled by certain policy. In the test no data lost has occurred. Net load of data transfer for sensors and task are lower. The table shows difference between ICDME and conventional web (Table 50.1).

50.5 Conclusion

The limitation of conventional web technology slow resource’s advance as bridge for human and the world. In this paper, we introduced a mechanism, called ICDME, to enhance IOT’s expansion upon web. Through this mechanism, tasks may create constant connection with resource, enable consistent update of resource working status in task and organize continuous data flow for IOT web’s applications.

Acknowledgements This research work was supported by the Special Program for Seism-Scientific Research in Public Interest “Research in Online Processing Technologies for Seismological Precursory Network Dynamic Monitoring and Products” (NO. 201008002)

References

1. Jensen F (2006) Introduction to computational chemistry. Wiley, New York
2. Lu Y, Yan Z, Laurence TY, Huansheng N (2008) The internet of things: from RFID to the next-generation pervasive networked systems (wireless networks and mobile communications). Auerbach Publications, Boca Raton
3. Newcomer E (2004) Understanding SOA with web services (independent technology guides). Addison-Wesley, Reading
4. Gibbons PB, Karp B, Ke Y, Nath S, Srinivasan S (2003) IrisNet: an architecture for a worldwide sensor web. IEEE Pervasive Comput v2(i4):22–33
5. Hohpe G, Woolf B (2003) Enterprise integration patterns: designing, building, and deploying messaging solutions. Addison-Wesley, Reading

Chapter 51
Frequency Reuse Analysis Using Multigraph Theory

Hui Zhang, Xiaodong Xu, Jingya Li and Haiyuan Liu

Abstract Considering the inter-cell interference coordination in LTE system, many frequency reuse schemes are concluded in this proposal. On this basis, the theoretical basis of frequency reuse is described respectively from graph theory and algebraic analysis principle. In graph theory, the coloring theory in multi-graph and the level interference-limited theory are focused, resulting in an frequency reuse analysis from the optimization problem. In algebraic analysis principle, this proposal gives a quantitative analytic algebra to describe the relationship of frequency reuse factor (FRF) between cell center and cell edge. Then the frequency reuse optimization problem is transformed into two-dimensional coordinate system, which enables to consider taking analytic algebra method to solve it. Moreover, the simulation is taken into compare the system performance with different FRF schemes. The results show that it needs to take into account the size of FRF, both cell-edge and cell-center performance to choose the appropriate inner radius.

Keywords OFDMA · Frequency reuse · Multigraph theory · Algebraic analysis

H. Zhang (✉) · H. Liu
College of Information Technical Science, Nankai University, Tianjin, China
e-mail: bupt2008@gmail.com

H. Zhang · X. Xu
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing, China

J. Li
Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_51,

51.1 Introduction

In LTE and its evolution system, many frequency reuse schemes are derived from soft frequency reuse (SFR) and fractional frequency reuse (FFR). Kim et al. [1] propose an incremental frequency reuse (IFR) scheme that reuses effectively the radio spectrum through systematic segment allocation over a cluster of adjoining cells, which divides the entire frequency spectrum into several spectrum segments. Li et al. [2] give a cooperative frequency reuse scheme for coordinated multi-point transmission. Liang et al. [3] propose a frequency reuse scheme for OFDMA based two-hop relay enhanced cellular networks. Chen et al. [4] give an approach based on large-scale optimization to deal with networks with irregular cell layout. Wamser et al. [5] give some different strategies for user and resource allocation are evaluated along with FFR schemes in the uplink. Assaad et al. [6] give an analysis of the inter-cell interference coordination problem and study the optimal FFR. Imran et al. [7] propose a novel self-organizing framework for adaptive frequency reuse and deployment in future cellular networks, which forms an optimization problem from spectral efficiency, fairness and energy efficiency. Novlan et al. [8] give a comparion of FFR approaches in the OFDMA cellular downlink, which mainly focuses on evaluating the two main types of FFR deployments, respectively Strict FFR and SFR.

However, the existing research mainly major in the form of application under different cellular scenarios, but lack in analyzing the theoretical basis for each frequency reuse scheme. It is important to find some necessary theories to guide the design of frequency reuse scheme. Considering this problem, this paper makes an analysis of frequency reuse theoretical basis from two angles, respectively graph theory and its algebraic analysis principle. By means of the above theoretical tools, we try to summarize the rules of frequency reuse design, giving a way to find the optimal frequency reuse scheme.

51.2 Multigraph Theory

In Ref. [9], the coloring method in graph theory and the collection idea are taken into frequency resue sets, dividing cellular users into diffetent sets and providing a unique frequency reuse strategy to each set, which effciently raise the cellular frequency reuse factor. Hale [10] adopts graph theory to discuss the problem of frequency allocation optimization, proposes a k-level interference-limited theory for the whole frequency sets. Moreover, it establishes an optimization model for frequency allocation. Du et al. [11] analyzes the cellular frequency planning by multigraph T-coloring method. In graph theory, the concept of multigraph means that every pair of points is at most connected with K edges, also whithout no self-loop, such graph is written as K-multigraph. The concept of T-coloring can be defined as: In graph G, color each point in the set $V(G)$ by T classes of colors, and the colors are different among each adjacent point, written as T-coloring in G graph. Reference [12] describes some basic characteristics of frequency

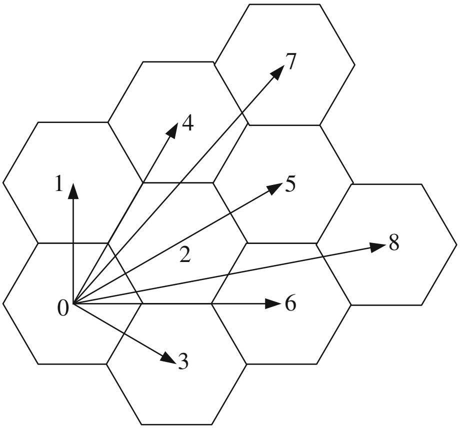


Fig. 51.1 Inter-cell interference level

allocation in T-coloring. In view of graph theory in frequency optimization, we further analyze the frequency reuse optimization problem using multigraph coloring theory and k-level interference-limited theory.

Assume each cellular cell cluster is divided into n limited districts $\{a_1, a_2, \ldots, a_n\}$, and allocate the frequency $f(a_i)$ into each district a_i. For the reuse of co-frequency resources, the level of interference always differs among districts due to the path loss of inter-cell interference. As shown in Fig. 51.1, the co-frequency interference strength of No. 0 cell to adjacent cell is inversely proportional to the path distance. By multigraph theory, the interference level can be mapped into interval according to the strength, which enables to optimize the frequency allocation.

As shown in Fig. 51.2, in order to analyze the inter-cell co-frequency interference, the interference strength in different districts is divided into several intervals by the descending range, respectively $\{[i_1, i_2], [i_2, i_3], \ldots, [i_{m-1}, i_m]\}$, which map into each interference level l: $\{l_1, l_2, \ldots, l_m\}$.

When District a_u and District a_v exist in the interference with the same level, the frequency allocation in this area should satisfy:

$$\{a_u, a_v\} \in E \Rightarrow |f(a_u) - f(a_v)| \notin T(l) \tag{51.1}$$

In particular, when $K = 2$, $T(l_0) = \{0\}$, which means the interference among a_u and a_v are the level l_0. In order to optimize frequency allocation, it is necessary to allocate a different frequency among a_u and a_v. When $K = 2$, $T(l_1) = \{0, 1\}$, it means the interference among a_u and a_v are in the levell_1, so the frequency allocation for such two areas should not only satisfy the difference, but also should not be adjacent.

Furthermore, we consider K different cells $\{G_0, G_1, \ldots, G_{K-1}\}$, and the number of frequency allocation area is $V(G)$ for each cell. Moreover, the co-frequency interference among a_u and a_v are with the same level l. For inter-cell interference level l, we define the taboo collection $T(l)$ for frequency allocation, as follows:

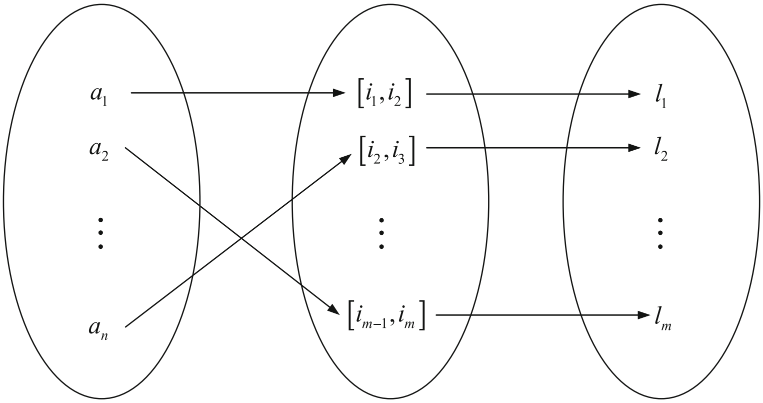


Fig. 51.2 Inter-cell interference level mapping

$$G_0 \supseteq G_1 \supseteq \cdots \supseteq G_{K-1} \tag{51.2}$$

$$T(0) \subseteq T(1) \subseteq \cdots \subseteq T(K-1) \tag{51.3}$$

In the view of graph theory, the frequency reuse aims to allocate frequency into each point in multigraph, finding the allocation function f to make the co-frequency interference minimum. Moreover, it enables to color all the cells $\{G_0, G_1, \ldots, G_{K-1}\}$ by $T(l)$ color. In other words, the formula (51.1) is established under the function f.

51.3 Principles of Algebraic Analysis

Based on the multigraph theory, we propose an algebraic analysis method for frequency reuse, which changes the relationship of cellular frequency reuse factor (FRF) into quantitative algebra analytic formula, taking two-dimensional coordinates to solve this frequency reuse optimization problem.

As shown in Fig. 51.3, considering SFR, taking No.0 Cell as an example, we define the cell-center region 0C (0 Center, simply written as 0C) as variable y, the cell-edge region 0E (0 Edge, simply written as 0E) as variable x. In this way, we take two-dimensional coordinates to analyze the function of x and y. Since 0C is in the cell-center region, its maximum value of FRF is equivalent to 1, so the variable y should satisfy:

$$0 < y \leq 1 \tag{51.4}$$

0E is in the cell-edge region, so its FRF is related with relevant partition way for cell-edge. When its FRF takes 1/k (k = 3, 7,…, n), the variable should satisfy:

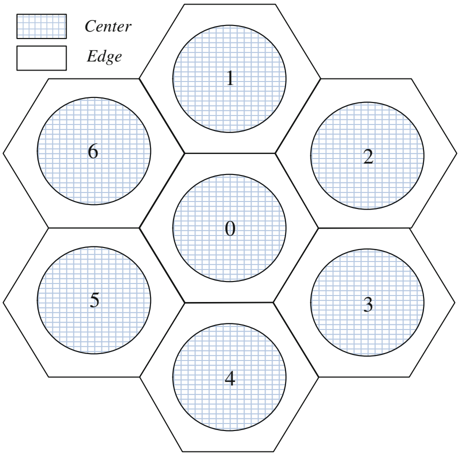


Fig. 51.3 Cellular frequency reuse

$$0 < x \leq \frac{1}{k} \tag{51.5}$$

For 0C and 0E that are still in the same cell, the total sum of FRF for the whole cell should not be more than 1, that is

$$0 < x + y \leq 1 \tag{51.6}$$

At the same time, when it is not divided for cell-center and cell-edge, the total FRF can be expressed as

$$0 < x + y \leq \frac{1}{k} \tag{51.7}$$

Based on the above analysis, as shown in Fig. 51.4, we take the sub-function as follows:

$$\begin{cases} 0 < x + y \leq 1, 0 < x \leq \frac{1}{k}, 0 < y \leq 1 \\ 0 < x + y \leq \frac{1}{k}, 0 < x \leq \frac{1}{k}, 0 < y \leq \frac{1}{k} \end{cases} \tag{51.8}$$

The Algebraic analysis approach opens a new way to the analysis of cellular frequency reuse and optimization. Based on this idea, 0C and its reuse region {1E, 2E, 3E, 4E, 5E, 6E}, 0E and its interference region {1E, 2E, 3E, 4E, 5E, 6E} can be described in the a form of two-dimensional coordinates, which enables to theoretically search the optimal cellular frequency reuse shceme.

51.4 Numerical Results

In order to compare the system performance on the basis of graph theory and algebraic analysis principle, we select SFR scheme and take system simulation with different FRFs. The simulation parameters are given in Table 51.1.

Fig. 51.4 Frequency reuse factor coordinates

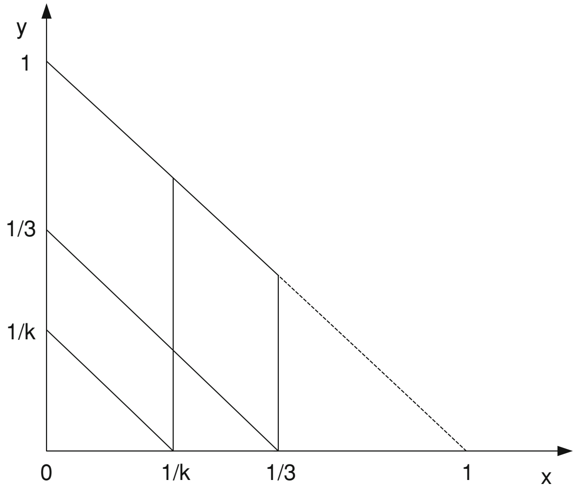


Table 51.1 Simulation parameters

Parameters	Value
BS total Tx power	43 dBm
Inter eNodeB distance	500 m
Carrier frequency	2 GHz
Minimum distance between UE and eNB	30 m
Shadowing standard deviation	8 dB
Shadow fading	Lognormal
Macroscopic pathloss	128:1 + 37:6 log10(R)
Simulation length	500 TTIs
Shadow fading correlation	Inter-site: 0.5, intra-site: 1
Number of simulations	200 per scenario
Thermal noise density	−1174 dBm/Hz
Bs antenna gain	15 dBi

As shown in Fig. 51.5, with the inner radius increases, the number of cell-edge user decreases, making its average rate to increase, improving the performance of cell-edge user. The larger the FRF, the better the performance.

As shown in Fig. 51.6, when the inner radius increases, more users are classified as cell-center users. For each fixed FRF, the frequency resources in cell-center is relatively constant, the increase of cell-center users may lead to insufficient resource allocation and larger co-frequency interference, so the average rate for cell-center user will decrease as the inner radius increases.

Therefore, it needs to take into account the size of FRF, both cell-edge and cell-center performance to choose the appropriate inner radius. With the FRF increases, the inner radius should be increased, making a rational division of cell-center users and cell-edge users.

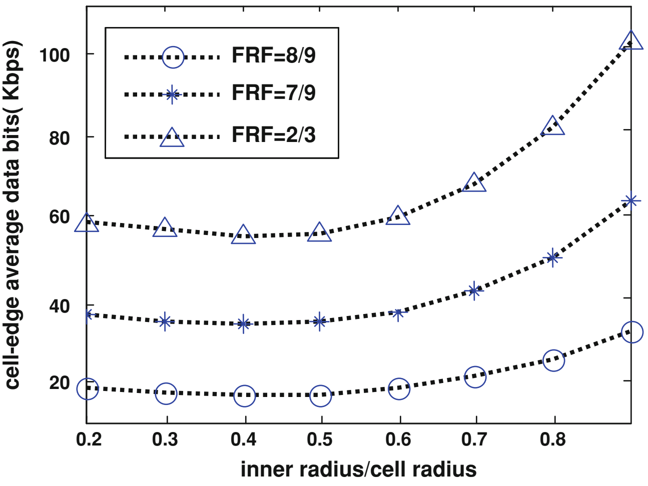


Fig. 51.5 The average rate for cell-edge user (SFR)

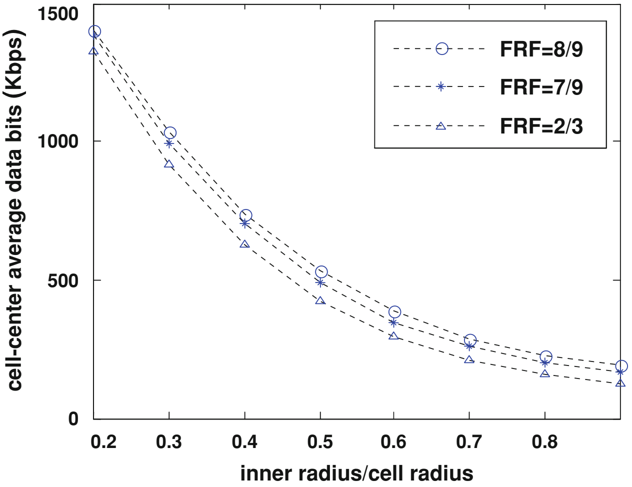


Fig. 51.6 The average rate for cell-center user (SFR)

51.5 Conclusion

This paper gives an analysis on the frequency reuse scheme in inter-cell interference coordination, and tries to find the theoretical basis of frequency reuse from graph theory and algebraic analysis principle. In graph theory, we focus on coloring theory in multigraph and the level interference-limited theory, making an analysis of optimization problem in frequency reuse. In algebraic analysis principle, this proposal gives a quantitative analytic algebra to describe the relationship of FRF between cell center and cell edge. Then the frequency reuse optimization problem is transformed into a two-dimensional coordinate system, which enables to consider analytic algebra method to solve it. Moreover, the simulation is taken to compare the system performance with different FRF schemes, and the results show that it needs to take into account the size of FRF, both cell-edge and cell-center performance to choose the appropriate inner radius. In the future, we would further take the Monte Carlo system simulation to compare different frequency reuse schemes and find the optimal scheme according to the proposed method.

Acknowledgment This research is supported by "the Fundamental Research Funds for the Central Universities", Natural Science Foundation of Tianjin (No.10JCYBJC00100).

References

1. Kim KT, Oh SK (2008) An incremental frequency reuse scheme for an OFDMA cellular system and its performance. IEEE vehicular technology conference-spring, pp 1504–1508
2. Li J, Zhang H, Xu X et al (2010) A novel frequency reuse scheme for coordinated multi-point transmission. IEEE vehicular technology conference-spring, pp 1–5
3. Liang M, Liu F, Chen Z et al (2009) A novel frequency reuse scheme for OFDMA based relay enhanced cellular networks. IEEE vehicular technology conference-spring, pp 1–5
4. Chen L, Yuan D (2010) Generalized frequency reuse schemes for OFDMA networks: optimization and comparison. IEEE vehicular technology conference-spring, pp 1–5
5. Wamser F, Mittelstadt D, Staehle D (2010) Soft frequency reuse in the uplink of an OFDMA network. IEEE vehicular technology conference-spring, pp 1–5
6. Assaad M (2010) Optimal fractional frequency reuse (FFR) in multicellular OFDMA system. IEEE vehicular technology conference-fall, pp 1–5
7. Imran A, Imran MA, Tafazolli R (2010) A novel self organizing framework for adaptive frequency reuse and deployment in future cellular networks. IEEE international symposium on personal indoor and mobile radio communications, pp 2354–2359
8. Novlan T, Andrews JG, Sohn I et al (2010) Comparison of fractional frequency reuse approaches in the OFDMA cellular downlink. IEEE global telecommunications conference, pp 1–5
9. R1-050896 (2005) Description and simulations of interference management technique for OFDMA based E-UTRA downlink evaluation. Qualcomm, 3GPP
10. Hale WK (1980) Frequency assignment: theory and applications. Proc IEEE 68(12):1497–1514
11. Du J, Zhang Y, Zhan S (2006) An algorithm to compute the span of the *T*-colorings of multigraphs. J Hebei Acad Sci 23(3):1–4
12. R1-050738 (2005) Interference mitigation considerations and results on frequency reuse. Siemens, 3GPP

Chapter 52
A New Algorithm of GSM Co-Channel and Adjacent Channel Interference Optimization

Lina Lan, Xuerong Gou and Wenyuan Ke

Abstract Most current methods of GSM frequency planning evaluate interference and assign frequencies based on the measurement reports. The same or adjacent frequencies are assigned to cells close to each other which cause co-channel and adjacent channel interference, and reduce the network performance. The traditional method to check and allocate the new frequencies is by man power which costs much time and the accuracy is not satisfied. This paper proposes a new intelligent algorithm of analysis and optimization on co-channel and adjacent channel interference based on the cells basic configuration information. The algorithm defines an interference evaluation model analyzing various factors such as the base station layer, the azimuth ward relationship, the cell neighborhood relationship, etc. The interference performance of each frequency can be evaluated and the problem frequencies can be optimized. This method is verified by a large number of actual datasets from an in-service GSM network. Contrast with the traditional method, this method demonstrates advantages in intelligence, accuracy, timeliness, and visualization.

Keywords GSM frequency optimization · Co-channel and adjacent channel interference · Interference evaluation model · Base station site layer · Azimuth ward · Neighborhood relationship · Triangulation

L. Lan (✉) · X. Gou · W. Ke
School of Network Education, Beijing University of Posts
and Telecommunications, Beijing 100088, China
e-mail: lindalan2002@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_52,

52.1 Introduction

The current frequency allocation methods always use the measurement data to analyze the interference, indirectly considers the disturbance intensity which is directly related with the distance, azimuth, relative position, and other factors. In the measuring period many problems may happen such as the sampling point's number is insufficient, barrier block, the base station failure or other problems caused by accidental factors. Those problems would affect the accuracy of measurement data, and the same or adjacent channel would exist in frequency allocation results. The same and adjacent channel interference affects the network performance. Therefore, the factors such as distance, azimuth, and relative position of the network should be considered to the adjacent channel interference verification and optimization [1–5].

The Co-channel and adjacent channel interference occurs among the cells in the neighboring area. In different regions, the distance of base stations is different. How to determine the scope of the neighboring area is a problem. The antenna azimuth ward launching rally signal is an important reason causing interference. It is difficult to automatically check if the azimuth is rally launching. To optimize the interference is to assign a group of new frequencies. So how to evaluate the interference of the frequency and choose the best frequency is the key problem.

This paper proposes a new intelligent method of analysis and optimization on co-channel and adjacent channel interference based on the cell basic configuration information. The method defines an interference evaluation model analyzing various factors such as the base station layer, the azimuth ward relationship, the cell neighborhood relationship, etc. The interference performance of each frequency can be evaluated and the problem frequencies can be optimized. The second part in the paper introduces the relevant concept definition of interference. The third part introduces the method and procedures, including interference level division and interference vector model definition. The fourth part analyzes the optimization result. At last review the advantages and characteristics of the approach.

52.2 Check Scope of Cells

Co-channel and adjacent channel frequency interference are mainly caused by the neighboring area [2]. The cells in the coverage area of the serving cell should be checked.

There are large amounts of base stations in the large mobile network. The base station intensity varies much in different regions. In countryside, the distances among base stations are much farther than in city. The cell coverage radius in countryside is much wider than in city. The coverage radius can't be defined as a constant value. Thus, the cell coverage is usually indicated by base station site layer and azimuth award side.

Cells of site layer 1 are in the first circle located in the nearest base station around the serving cell. Cells of site layer 2 are the second circle located in the base station external around the first circle and so on. The cell coverage area is inside of site layer 3.

The forward of the azimuth is the area, 120° around the center line of the azimuth, and the other side means backward. In general, the coverage area of a cell should be in the forward site layer 3 and the backward site layer 1. The area of inside of forward level 1–3 and backward level 1 is the right coverage region. The area of outside of forward level 3 and outside of backward level 1 should not be covered. The cells set C in coverage area of the serving cell should be checked whether or not the co-channel and adjacent channel frequency interference with the serving cell. The cells on the outmost layer those azimuths outwards to the serving cell azimuth are not in the set C, but those azimuths inwards to the serving cell are in the set C.

52.3 Frequency Optimization Algorithm

52.3.1 Interference Level Classification

In the actual frequency optimization, in order to avoid interference with co-channel or adjacent channel, frequency allocation should meet the following requirements [6]:

(a) The cells with the same stations can not be assigned with co-channel or adjacent channel frequency.
(b) The neighboring cell can not be assigned the same frequency, and should avoid adjacent channel.
(c) The azimuth sector can not rally with the co-channel or adjacent channel.
(d) Other neighboring district cells should avoid co-channel or adjacent channel.

According to the principle of verification, the co-channel or adjacent channel interference can be analyzed by kinds of factors such as whether they are in the same station, neighborhood relations, and azimuth sparring and other factors. The interference levels from low to high are divided into 8 levels accordance with the various kinds of interference and its impact, as shown in Table 52.1.

In Table 52.1, the interference levels are classified from level 0 to 7. The interference value is becoming bigger and bigger from 0 to 7. Level 0 indicates no co-channel or adjacent channel. Level 1 indicates existing adjacent channel. Level 2 indicates existing co-channel. Level 3 indicates existing adjacent channel in the neighboring cell. Level 4 indicates existing adjacent channel in the rally azimuth. Level 5 means existing co-channel in rally azimuth. Level 6 indicates existing

Table 52.1 Interference level classification

Level	Is same base station	Is neighboring cell	Is rally azimuth	Is co-channel or adjacent channel
0				N
1	N	N	N	Adjacent channel
2	N	N	N	Co-channel
3	N	Y	N	Adjacent channel
4	N		Y	Adjacent channel
5	N	N	Y	Co-channel
6	N	Y		Co-channel
7	Y			Co-channel or adjacent channel

co-channel in the neighboring cell. Level 7 indicates existing co-channel or adjacent channel in the cells in the same base station. The interference of level 0 is the smallest, and the interference of level 7 is the biggest.

Table 52.2 shows the verification result of the co-channel or adjacent channel of a cell on frequency point 34 and 90.

In frequency point 34, there are 1 of interference level 6, 2 of level 3 and 2 of level 1. In frequency point 90, there are 1 of interference level 5, and 1 of interference level 2. The frequency performance can be evaluated with the number of the interferences. Sum the interferences number in each frequency in a cell, the total performance of a cell can be evaluated. If there is the high level interference, the frequency should be reassigned and make the optimization.

52.3.2 Frequency Interference Vector

Considering the interference level, the site layer of cells, the frequency interference vector is defined as A = (a1, a2, a3, a4, a5, a6). In the vector, element a1 means the main level (e.g. the biggest level) of interference; element a2 means the smallest (e.g. nearest) site layer of the main interference source; element a3 means the number of main interference source; element a4 means the smallest site layer of the interference source; element a5 means the number of interference source in the smallest site layer (e.g. a4); element a6 means the total number of all the interference source.

The performance of each frequency can be indicated with the vector A. For example, there are three frequencies in a cell. The performance vector of each frequency is A1, A2 and A3. The vector values are as following expression:

$$\begin{aligned} \mathrm{A1} &= (4, 2, 1, 2, 1, 4) \\ \mathrm{A2} &= (4, 2, 2, 2, 1, 5) \\ \mathrm{A3} &= (4, 3, 1, 2, 2, 5) \end{aligned}$$

Table 52.2 The verification result of the co-channel or adjacent channel of a cell

Frequency point	Interference level 7	Interference level 6	Interference level 5	Interference level 4	Interference level 3	Interference level 2	Interference level 1
34	0	1	0	0	2	1	2
90	0	0	1	0	0	1	0
Total	0	1	1	0	2	2	2

Compare the performance of A1, A2 and A3 from a1 to a6. The a1 of the A1, A2 and A3 is as the same as 4. Then compare a2. The a2 in A3 is 3, and it is bigger than 2 of a2 in A1 and A2, that means the site layer of main interference source in A3 is farer more than A1 and A2, so the performance of A3 is better than A1 and A2. Then compare a3 of A1 and A2, the number of main interference source of A2 is 2 which is bigger than 1 of a3 of A1, so the performance of A1 is better than A2. Therefore the interference level of them is A2 > A1 > A3, the excellence in performance of frequencies is A3 > A1 > A2. So the best frequency could be chosen to be assigned.

With the interference vector, the frequency can be chosen according to performance values. The serious interference frequency can be found. A group of best frequencies can be selected from the optional frequencies to replace the serious interference frequencies to achieve frequency optimization.

52.3.3 Frequency Optimization Process

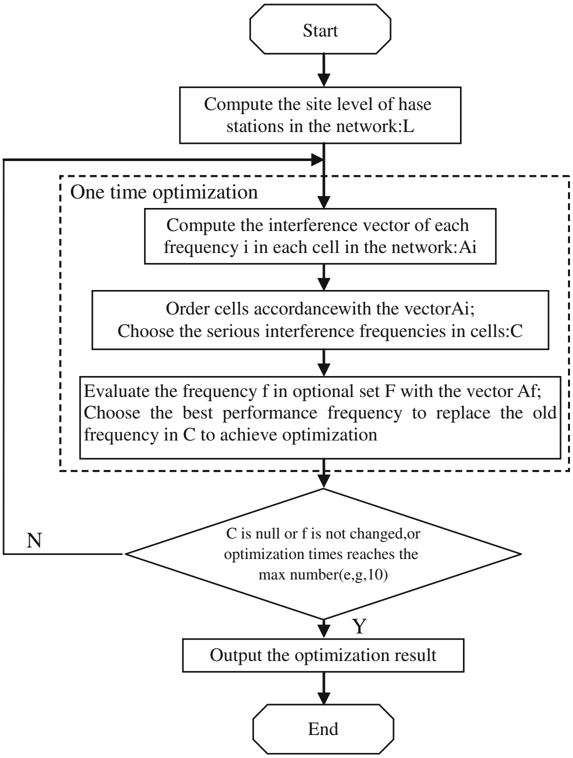
The frequency optimization process is shown in Fig. 52.1.

The process consists of the following steps:

(1) Compute the site layer of base stations in the network and get the matrix L.
(2) Compute the interference vector of each frequency i in each cell in the network and get the vectors Ai.
(3) Order the frequencies in the cells in accordance with the vector Ai, and choose the serious interference cells and put them into the cells set C. C is the scope of cells to be optimized.
(4) Evaluate the frequency f in the optional frequency set F with vector Af, and choose the best performance frequency to replace the old frequency to achieve frequency optimization.
(5) At end of one time of optimization, judge whether the end condition is met. If the C is null or the frequency is not changed, or the optimization times reaches the max number (e.g. 10), all the optimization ends; otherwise go to step (2) to start a new time of optimization.
(6) If the optimization end condition is satisfied, then output the result and the process is ended.

In the process, steps 2, 3, and 4 composed one time frequency optimization. At the beginning of every time of optimization, the interference vector with the modified

Fig. 52.1 Frequency optimization process


Start
Compute the site level of hase stations in the network:L
One time optimization
Compute the interference vector of each frequency i in each cell in the network:Ai
Order cells accordancewith the vectorAi;
Choose the serious interference frequencies in cells:C
Evaluate the frequency f in optional set F with the vector Af;
Choose the best performance frequency to replace the old frequency in C to achieve optimization
C is null or f is not changed,or optimization times reaches the max number(e,g,10)
N
Y
Output the optimization result
End

frequency should be computed again, and the new serious interference cell set C would be found to optimize. The optimization should be executed several times loops to resolve the new interference impact by the new frequency assigned.

The end condition of the process is intelligently designed. If there is no serious interference cell (e.g. the interference level >4), or the optimization times reaches the max number (e.g. 10), or the frequency is not changed (e.g. optimal frequency has been allocated), the optimization will end to output the result. This method avoids the duplication of operations and improves operational efficiency.

52.4 Optimization Result Analysis

The method has been used to do the verification and frequency optimization of co-channel and adjacent channel in a GSM network in a province. The cell number is about 400. The serious interference level is 4, that means if the interference level is bigger than 4, the frequency should be optimized. The max number of optimization loop time is 10. The optimization result analysis is shown in the Table 52.3.

Table 52.3 Statistics of frequency optimization results

Item	Level 7	Level 6	Level 5	Level 4	Level 3	Level 2	Level 1
Interference number of BCCH in old approach	1	2	18	56	124	259	628
Interference number of TCH in old approach	3	70	26	102	449	512	987
Total interference number in old approach	4	72	44	158	573	771	1615
Interference number of BCCH in new approach	0	0	0	0	110	225	555
Interference number of TCH in new approach	0	0	0	0	381	485	966
Total interference number in new approach	0	0	0	0	491	710	1521
Decreased number of interference of BCCH	1	2	18	56	14	34	73
Decreased number of interference of TCH	3	70	26	102	68	27	21
Total decreased number of interference	4	72	44	158	82	61	94
Decreased rate of interference of BCCH (%)	100	100	100	100	11.29	13.13	11.62
Decreased rate of interference of TCH (%)	100	100	100	100	15.14	5.27	2.13
Total decreased rate of interference (%)	100	100	100	100	14.31	7.91	5.82

In Table 52.3, after the optimization, the interferences of level 4–7 are all removed. The co-channel and adjacent channel frequency are all cleared up. The optimal rate is 100%. In general, the higher the interference level, the more obvious is the optimization. For the level outside the scope of the optimization (e.g. level 1–3) of the interference, because its interference impact less on system performance, they can not be completely eliminated, but the interference has also been decreased.

52.5 Conclusions

This paper presented a novel co-channel and adjacent channel frequency interference analysis and optimization method in GSM network. The approach consists of (1) analyzing the basic data of cells which consists of site layer of base station, azimuth ward, and neighboring relationship of cells, (2) classifying the interference level based on the basic data and evaluating the frequency performance with an interference vector model, (3) giving a process of the frequency optimization in a network.

Unlike the previous approaches based on Drive Test (DT), this new approach is based on the basic configuration data collected from OMC, thus the raw data are comprehensive, real-time, and costless.

The approach is evaluated by large amounts of data from a real GSM network. Over 90% co-channel and adjacent channel frequency interference problems can be discovered and the accurate is satisfied. The work efficiency of frequency optimization is greatly improved.

Currently, the interference analysis of this method does not consider the measurement factors. It could not be used directly for frequency planning. In the actual work of network optimization, this method can be used as supplementary of frequency planning to verify and optimize the phenomenon of the co-channel and adjacent channel interference to improve network performance.

References

1. De Pasquale A, Magnani NP, Zanini P (1998) Optimizing frequency planning in the GSM system. IEEE Trans ICUPC'98 1:293–297
2. Neskovic N, Neskovic A, Paunovic D (2001) Automatic frequency planning algorithm in a real land mobile radio system design. In: Proceedings of IEEE transactions on TELSIKS, pp 511–520
3. Deng Y (2005) Wireless cover and network optimization. Shanxi Electr Tech 03:45–46
4. Jin X, Pan Y, Song J (2003) The status and development trend of the mobile communication network optimization technology. Telecom Tech 12:1–3
5. Zhao T, Liu S (2008) GSM wireless network optimization. Sci Consult 03:49
6. Han B (2001) GSM theory and network optimization. Machinery Industry Press, Beijing

Chapter 53
Preliminary Study on Telemetric Vehicle Emission Examination

Pak-kin Wong, Chi-man Vong, Weng-fai Ip and Hang-cheong Wong

Abstract Vehicle engine emissions closely relate to air-ratio called lambda (λ). The current practice of examination of gasoline engine emissions is partially based on λ reading at engine idle speed. For the λ value over a specific vehicle examination standard, it indicates that a significant amount of emissions is produced. The concept of telemetry with traffic lights is proposed that allows lambda data collected from the vehicle on road to be reported in real time at a traffic light. The data can be sent via a radio transmitter in the vehicle to a radio receiver mounted on the traffic light which is connected to a PC. So, the authority can monitor the engine emission data via computer networks while the vehicles stop at red traffic lights and enforce the vehicle owners taking immediate action to fix their vehicle emission problems if necessary.

Keywords Vehicle emissions examination · Telemetry · Wireless networks

53.1 Introduction

Vehicle engine emissions associate with air-ratio called lambda (λ) [1]. References [2, 3] mentioned that if λ is 1% lower than the stoichiometric value, "1", then carbon monoxide and hydrocarbon emissions will be significantly increased. If λ is

P. Wong · C. Vong · W. Ip · H. Wong (✉)
Faculty of Science and Technology, University of Macau, Macau, China
e-mail: hcwong@umac.mo

P. Wong
e-mail: fstpkw@umac.mo

C. Vong
e-mail: cmvong@umac.mo

W. Ip
e-mail: AndyIP@umac.mo

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_53,

1% higher than the stoichiometric value, up to 50% of nitrogen oxides may be produced. Therefore, engine emissions control is a hot research topic other than engine performance improvements [4–8]. Modern automobiles have built-in lambda sensors, which are installed in the upstream and downstream positions of the catalytic converter. The lambda sensors send real-time lambda signals to the electronic control unit (ECU) of the car. Therefore the ECU can monitor and control the conversion efficiency of the catalytic converter, and warn the driver when the converter badly breaks down. Usually, on delivery, the gasoline engine with the catalytic converter can produce very low amount of emissions. However, when the engine ages, significant emissions will be produced. At that time, engine maintenance, particularly the catalytic converter, is required. Unfortunately, the car owners usually are not aware of their engine health before their mandatory vehicle examinations. However, vehicle examination cannot be taken every day for each car; it is only taken annually or after 6–10 years of car registration. The worst case is that the car owners may ignore any engine maintenance even the car produces a warning signal because of lack of self-regulation.

There are a large number of passenger cars equipped with gasoline engine in the world. It can be imagined that a huge amount of engine emissions are produced every year. If the car owners and the governmental authority can be notified of the engine health of the cars through an information system, the car owners can get their engines fixed earlier because the mandatory vehicle emission test is carried out frequently. This act can cause great reduction of the engine emissions every year. Hence a green urban environment can be achieved. With the concept of telemetry with traffic lights, such an information system for this notification service to the car owners and the governmental authority becomes possible.

Telemetry [9] is a technology that allows remote measurement and reporting of information. The telemetry can enable the λ values collected from the vehicle on road to be reported in real time at a fixed location such as a traffic light. The data can be sent via a radio transmitter in the vehicle to a radio receiver mounted on the traffic light. The radio transmitter is connected to the signal line of the in-car lambda sensor and sends out the λ reading along with the unique identity of the transmitter or car wirelessly. The radio receiver is connected to a PC which can behave as a server and share the data with the other authorized computers in relevant government departments via the Internet. Hence, the authority can easily monitor the engine emission data while the vehicle stops at some selected traffic lights. Since traffic light is a central component in the whole traffic system and every car must be stopped at red light, it is the best timing to wirelessly detect the lambda reading of an engine at idle speed. With this concept, all running cars can be examined for their engine emissions continually. This provides the possibility to detect early signs of emission problems and warn the drivers for repairs when the engine λ reading at idle speed exceeds the local vehicle examination standard. Ultimately, the engine emissions can be effectively inspected or even controlled via this information system. In this chapter, the objective mainly studies the implementation issues of this idea and its feasibility.

53.2 Enabling Technologies

The technical issues of the telemetry about radio transmitter, radio receiver, and transmission to server are discussed in the following sections.

53.2.1 Radio Transmitter

The engine λ value is measured by the ECU and transmitted by an in-car telemetric radio. There are two main functions of the radio transmitter. First, it reads the engine λ value from the ECU. The second function is data conversion. Since the engine λ value is a millivolt analog signal which cannot be directly transmitted. A built-in radio modem in the transmitter is used to convert the digital data into a form that can be transmitted by Radio Frequency (RF). This exactly likes the landline modem except that it transmits its data wirelessly instead. The radio receiver converts the transmitted data back to original signal which can be read by a PC.

53.2.2 Radio Receiver

As mentioned before, the traffic light is an important component in urban traffic system. In addition, it is a good place to stop a car and perform wireless inspection for engine emissions. The telemetric examination operates only when it is RED light. When it is GREEN light, the radio receiver stops operation because of several reasons. First, the vehicle emission examination is only measured at idle speed. Second, it is unnecessary to inspect every car continuously. It is already enough to just periodically sample and examine if the λ reading of an engine exceeds the local vehicle examination standard or not. Anyway, a car must eventually stop by a RED light and get examined within a time period, such as a week or several days. It is acceptable to have a weekly telemetric examination. Third, considering that there are several hundreds of traffic lights in a city, lower examination rate can reduce the number of data transmission and the cost of powerful server, making the whole system more practical.

53.2.3 Transmission to Server

Most traffic lights in a city are just connected and controlled for shifting signals which do not provide the capability of data transmission. Once the radio receiver obtains relevant information, the most cost-effective way for data transmission to the back-end server is done by wireless telecommunication technology such as GPRS or 3G. In fact GPRS is already enough because the amount of transmitted data just takes several kilobytes at one time. However, in Macau or Hong Kong, 3G data transmission is so popular that the price is nearly the same as GPRS. For example, an unlimited monthly data transmission provided by one of the Macau Internet service providers requires a subscription fee of 28 USD. Therefore, in the current system design, 3G transmission is recommended to connect the radio receiver and the back-end server.

53.3 System Design

The system design of the whole information system can be depicted in two major—road situation and in-car radio transmitter installation. The road situation considers how a radio receiver at a traffic light communicates with an in-car radio transmitter. As shown in Fig. 53.1, several cars stop in front of a traffic light and their radio transmitters broadcast at the same time. The λ reading along with the ID of a car or transmitter are picked up and sent to the back-end server via wireless telecommunication for further processing. If the λ reading exceeds a specific vehicle examination standard, for example in Hong Kong standard: $\lambda \leqq 1.0 \pm 0.03$, the back-end server will send a message to the car owner for the notification of engine maintenance. The authority may even adopt this information system to control the engine emissions. Any car having failed λ reading does not get fixed within, say, one week may be fined but this has to go through a careful legislation.

For the sake of reliability of λ reading, it is suggested that only the cars closest to the traffic light within a certain distance (e.g., 70 m) are examined. Hence every time, about ten cars are examined at one traffic light so that the back-end server would not be overloaded with the overall examination rate for hundreds of traffic lights.

Another important design is the installation of in-car radio transmitter as shown in Fig. 53.2. There are actually two lambda sensors in an exhaust system; one is settled before the catalytic converter while another one after it. Both of them are used to evaluate if the catalytic converter is functional or not. In order to measure the engine emissions as discussed in the proposed information system, it is only necessary to connect to the lambda sensor in the downstream position of the converter. It can be seen that only a simple wiring is enough to obtain the lambda signal, which is a low voltage (from 0 to 1 V). This signal is then sent to the radio transmitter RFI 9256 whose antenna is installed on the roof of the car for optimal signal transmission position and future easier maintenance. After forwarding to the back-end server through 3G transmission, the lambda signal is interpreted as the real λ reading according to some predefined calibration data for checking any failed cases.

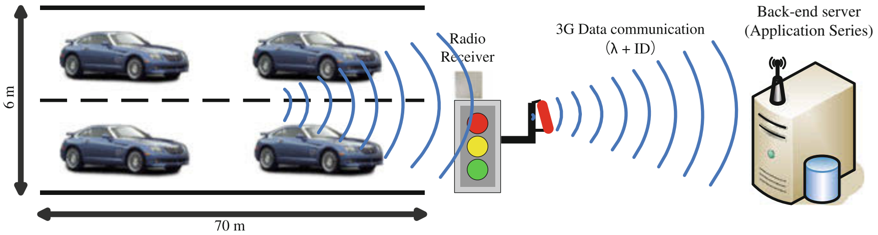

6 m
70 m
Radio Receiver
3G Data communication (λ + ID)
Back-end server (Application Series)

Fig. 53.1 Radio transmittion on road

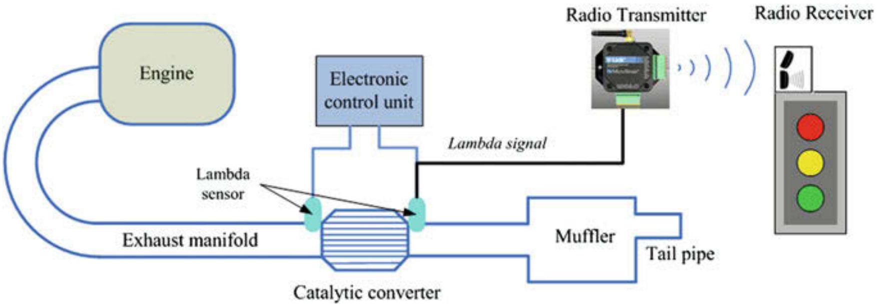


Fig. 53.2 In-car installation of radio transmitter

53.4 Experiments

In order to verify the effectiveness of the proposed system, some experiments have been conducted. In current application, the most critical part is the communication among the traffic light and the cars, while the issue of the data transmission from the radio receiver at the traffic light to the back-end server may be negligible because of the maturity of 3G telecommunication technology. The communication between the traffic light and the cars can be evaluated with two factors: effective distance and reliability. The first factor reflects whether a radio transmitter can be picked up by the radio receiver in a certain distance, while the second one indicates the accuracy of the λ reading.

53.4.1 Experimental Setup

The experiments can be conducted through a simulation of road situation as illustrated in Fig. 53.1. An open area of about 200 square meters is necessary, where a pole of three meters tall simulating the traffic light was set up. The radio receiver and 3G device were installed at the top of the pole. The back-end server

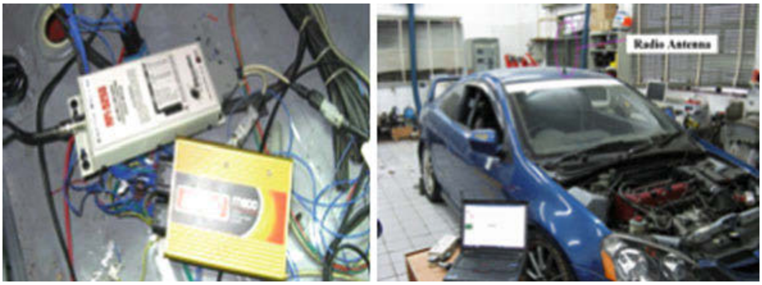
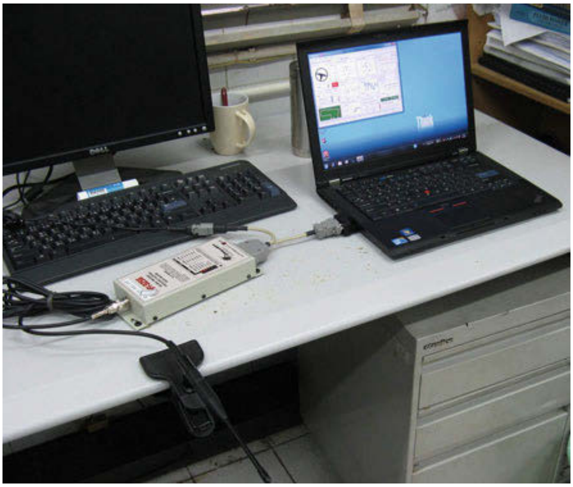


Fig. 53.3 In-car radio transmitter and the location of the radio antenna

Fig. 53.4 Radio receiver connecting to a PC



can be set up in a distant room. The in-car telemetric radio transmitter is RFI9256 is connected to a test car ECU (MoTeC M800) for λ reading and the antenna was installed on the roof of the test car as shown in Fig. 53.3, while the radio receiver was connected to a PC (Fig. 53.4).

53.4.2 Verification of Effective Distance

The prototype vehicle was located at different distance from the pole in order to measure the maximum and minimum possible effective distance for the telemetry. We have tested ten times for each distance at different times. The effective distance can actually be at least 100 m.

53.4.3 Verification of Reliability

Another important experiment test is the reliability of the lambda signal. Those received signals are forwarded to the back-end server and converted to the real λ readings to check if any possible distortion happens. This can be measured by comparing the received λ reading with the true reading at the in-car lambda sensor which can be obtained using MoTeC M800 ECU. Calibration may be necessary if the measured λ reading at the back-end server is significantly different from the true λ reading at the in-car lambda sensor. The experimental results show that a distance exceeding 75 m may have high distortion. Therefore, a distance of 70 m is recommended in the system design for the sake of reliability.

53.5 Implementation Issues

There are several issues for the proposed information system, namely, cost, security, and legislation, which are discussed in the following sections.

53.5.1 System Cost

Cost is a very important issue in implementing any system. In terms of system cost, there are two folds in current application—for car owners and authority. For the car owners, the cost is just the purchase and installation of radio transmitter connected to the lambda sensor. So far, this is only a simulation but the final price of such a simple radio transmitter can be estimated to be less than 150 USD under mass production. This price is recommended to be subsidized by the authorities. For the authority, a set of radio receiver and 3G transmission device is required for a traffic light. Each set takes about 5000 USD. For one hundred sets of these devices, 500 K USD is required. In addition, a back-end server with customized application takes 50 K USD. Other odds, such as maintenance fees, are estimated based on situation. However, most of the hardware is one-time-off investment. In summary, about 700 K USD is necessary to install such an information system in a city with a scale of 100 traffic lights, which is very cost-effective if the vehicle emissions can really be controlled. Therefore the implementation cost for current proposed system will become very feasible in the near future.

53.5.2 Security

Security is usually applied to protect important personal information, especially while transmitting through public area wirelessly. However, the current data transmission comprises only two simple components—vehicle or transmitter ID, and λ reading. None of them can reveal any important information without the back-end server and the database of car owners. Therefore, no threats about security can be visible at current time but it may be left for further studies in the future.

53.5.3 Governance and Legislation

Another important pre-requisite for the proposed information system is governance and legislation. Without legal support, quite a large amount of car owners will not actively install radio transmitter. In addition, these radio transmitters cannot be installed without government support and supervision. However, careful legislation usually takes years and this is the major obstacle to the implementation of the information system.

53.6 Conclusions

In this chapter, the feasibility of an information system under the concept of telemetry under traffic lights for mandatory vehicle emissions examination is studied in a technical viewpoint. With the innovative concept, an information system is proposed to form a wireless connection between traffic lights and gasoline vehicles. Then the fuel combustion indicator, air-ratio (λ), can be received along with the corresponding vehicle or transmitter ID wirelessly. Hence, the engine health can be easily examined easily. Furthermore, this information may be provided to the authority for engine emission control.

In addition, the cost for the proposed information system is comparatively low and acceptable. With the rapid development of technology, the cost will become lower and lower under mass production. Although the cost of the system is low and the service is tempting, several important issues such as governance and legislation must be carefully examined by legal professionals before these services can really be promoted.

References

1. Course WH, Anglin DL (1993) Automotive mechanics, 10th edn. McGraw-Hill, New York
2. Wong PK, Tam LM, Li K, Vong CM (2010) Engine idle-speed system modelling and control optimization using artificial intelligence. Proc IMechE Part D J Automob Eng 224(1):55–72
3. Manzie C, Palaniswami M, Ralph D, Watson H, Yi X (2002) Model predictive control of a fuel injection system with a radial basis function network observer. J Dyn Syst Meas Cont Trans ASME 124(4):648–658

4. Vong CM, Wong PK (2011) Engine ignition signal diagnosis with wavelet packet transform and multi-class least squares support vector machines. Expert Syst Appl :8563–8570. doi: 10.1016/j.eswa.2011.01.058
5. Vong CM, Wong PK (2010) Case-based adaptation for automotive engine electronic control unit calibration. Expert Syst Appl 37(4):3184–3194
6. Vong CM, Wong PK, Ip WF (2010) Case-based classification system with clustering for automotive engine spark ignition diagnosis. In: IEEE/ACIS 9th international conference on computer and information science (ICIS). IEEE Press, New York, pp 17–22
7. Vong CM, Huang H, Wong PK (2010) Engine spark ignition diagnosis with wavelet packet transform and case-based reasoning. In: IEEE international conference on information and automation (ICIA). IEEE Press, New York, pp 565–570
8. Vong CM, Huang H, Wong PK (2009) Case-based reasoning for automotive engine electronic control unit calibration. In: IEEE international conference on information and automation 2009 (ICIA 2009). IEEE Press, New York, pp 1380–1385
9. Vasseur JP, Dunkels A (2010) Interconnecting smart objects with IP: the next internet. Morgan Kaufmann, New York

Chapter 54
The Heuristic Algorithm of Stacking Layer for the Three-Dimensional Packing of Fixed-Size Cargoes

Liu Wang-sheng, Yin Hua-yi and Li Mao-qing

Abstract According to the actual operation of working people, a heuristic algorithm of stacking layer that meets the requirements of stability and convenience for load- and-unload was proposed. First, choose the stacking direction according to the position of the compartment door. Second, optimize the combination of length, width and height along the stacking direction to minimize the remaining space. Finally, optimize each layer's layout. In each layer's layout, adopt long–short edge combination mode for each edge. Considering the flatness and stability of loading and unloading, the number of long–short edge is related. Experiment results show that the algorithm can maintain the requirements of stability and convenience of loading and unloading, and also has nice space utilization.

Keywords Fixed-size cargoes · Three-dimensional packing · Stacking layer method · Heuristic algorithm

54.1 Introduction

Currently, the logistics and distribution services have become important components of Enterprise competitive power, as the quality of freight loading has a direct influence on the efficiency of distribution, and further impacts the working

L. Wang-sheng (✉) · Y. Hua-yi · L. Mao-qing
School of Computer and Information Engineering, Xiamen University,
Xiamen 361005, China
e-mail: kollzok@yahoo.com.cn

L. Wang-sheng
Research Centre of Modern Logistics, Jimei University,
Xiamen 361021, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_54,

efficiency of the distribution center. Increasingly, importance is being attached to the study of freight loading. With the standardization and normalization of the logistics and distribution services, especially the wide use of great tonnage container transportation, in order to improve the loading efficiency, more and more standardized packaging has been used. Therefore, the research of the fixed-size loading is an urgent issue. However, in previous researches, loading problem with different sizes is explored more, such as in [1–10], the algorithms proposed are for different-size 3-dimension loading problem. The optimization algorithms are very complicated, and are not suitable for the fixed-size loading problem. Loading problem is a combination optimization problem with complex constraints. It is an NP hard problem which is difficult to solve. However, it is much easier for fixed-size loading problem, whose object is to load as much as possible. Currently, there is little research focus on fixed-size loading problem. George [11] proposed a Heuristic algorithm framework and gives an upper bound for optimal, Shulin Sui et al. [12] describe a Heuristic approach using nest loop in 2D matrix arrangement. For the cargo where side tumbling is allowed, Derong Yang [13] introduced a mirror copy method for 2D arrangement, optimizing each possible layout, then optimizing of trends in the three coordinate directions, choosing the best coordinate direction to maximize the space utilization. Lili Xu et al. [14] improved Yang's method, proposing a simpler method,and reducing 3-direction optimization to height direction optimization. In 2D arrangement optimization, they take the model proposed in Yuling Niu et al. [15]. However, practical loading needs consideration of the stability and convenience for loading and unloading. The algorithms above only consider maximization of the volume of cargo accommodated, which is equivalent to the high space usage; the arrangement is usually irregular or instable, and is not good for loading and unloading. From the intuition that objects should be kept even and stable, this paper proposes a Heuristic Algorithm of Stacking Layer for Three-Dimensional Packing of Fixed-Size Cargoes, which is convenient and stable for loading and unloading.

54.2 Problem Definition

There are two research directions for loading problem, one focuses on minimizing the number of containers, the other focuses on the arrangement of objects in a container, to maximize the total volume of cargo for each container. For objects with fixed-size, the main consideration is how to fill the car to maximize the objects in the car, such that the number of cars is minimal. So for fixed-size loading problem, the two directions have the same goal and we only need to find the best arrangement of objects.

The fixed-size of object is a box, so the problem can be described as: given a infinite set of boxes with fixed-size, try to find a best loading method to a container with known size, to maximize the efficiency of the loading space utilization (or the total volume of cargo), with the consideration of the convenience and stability of loading and unloading. There are several assumptions in the following discussion:

(1) the box size is less than the container size;
(2) boxes can be arbitrarily rotated and arranged in the container;
(3) the boxes can be stacked.

54.3 Stacking Layer Method Design

54.3.1 The Concept of Stacking Layer Method

The stacking layer method simulates the idea that each layer is ensured to be more "straight" during the real packing process. Each pile is packed along the same direction, and each layer cannot be filled until the previous layer is fully filled. Each layer must be smooth and tidy. There are two categories according to the direction when packing: the parallel layer stacking and vertical layer stacking. Parallel layer stacking keeps each layer of the cargo paralleled with the compartment door and vertical layer stacking keeps each layer vertical to the compartment door. The vertical layer stacking method in this paper specifically means that the stacking is along the direction of the compartment height.

The compartment door of the truck is usually set in three ways:

(1) in the rear compartment;
(2) in the side;
(3) two doors both in the rear and the side.

Vertical layer stacking is preferred for the stability of the load because the load-bearing surface holds the whole bottom of the compartment. For the parallel layer stacking, each layer bears a much smaller area where only a small strip of the bottom of the compartment is taken. But for (54.1) and (54.2), workers need to load and unload the cargos layer by layer since there is only one door and step on the cargos when working on the layers far away from the door. Loading and unloading cannot be done if there is not enough space in the compartment. Therefore, in order to facilitate loading and unloading easily, the parallel layer stacking method is more appropriate. The layer farthest from the door can be filled first, which means the filling is from inside out and the cargos can be unloaded from the outside layer by layer during the distribution. Because there are two doors in (54.3), it is much easier when loading and unloading. The layers far away from the back door can be loaded and unloaded through the side door, so vertical layer stacking method is preferred.

54.3.2 The Two-Step Solving Algorithm for Stacking Layer Method

Based on the earlier packing algorithms, a two-step solving algorithm for stacking layer method is introduced in this paper. First, a triangular combinatorial optimization along a certain direction of the compartment is operated. Then the layout

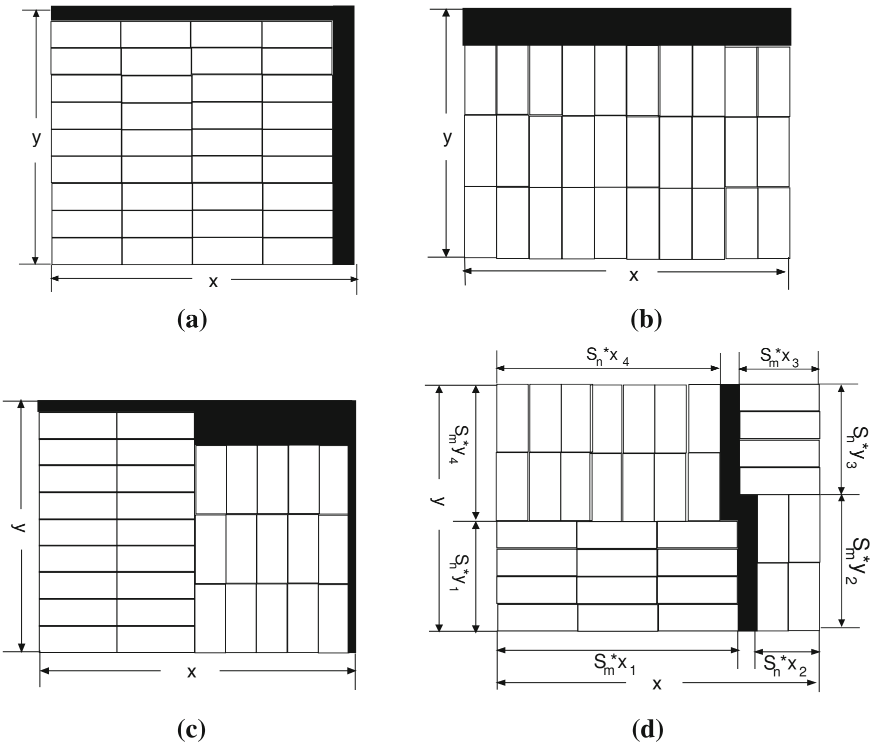


Fig. 54.1 General packing model **a** Stacking Model 1 **b** Stacking Model 2 **c** Stacking Model 3 **d** Stacking Model 4

is further optimized. Solving the first step is similar to the one-dimensional cutting stock problem. The length, width and height of the compartment is ordered from largest to smallest, denoted as s_1, s_2 and s_3 $(s_1 \geq s_2 \geq s_3)$. The loading direction of each layer is denoted as z (stands for the length or width or height of the compartment). We seek for the linear combination of s_1, s_2 and s_3 that minimizes the remaining space along the loading direction. The objective function is

$$\text{Min } FZ = Z - (a \times s_1 + b \times s_2 + c \times s_3) \tag{54.1}$$

which subjects to the constraints condition of $0 \leq a \leq \lfloor z/s_1 \rfloor$, $0 \leq b \leq \lfloor z/s_2 \rfloor$, $0 \leq c \leq \lfloor z/s_3 \rfloor$ (a, b, c are integers), where FZ is the remaining space and "$\lfloor\ \rfloor$" means to round down.

Step two is actually a layout problem that the layout space is located in (x, y) and the objective rectangular to be filled is (s_m, s_n). Suppose $\mathrm{s}_m \geq \mathrm{s}_n$, there are four types of stacking models illustrated in Fig. 54.1. The black filler represents the remained gap.

In fact, the former three stacking models are the special case of the fourth packing model. When $x_2 \sim x_4, y_2 \sim y_4$ or $x_1, x_2, x_4, y_1, y_2, y_4$ or x_2, x_4, y_2, y_4 are all zeros, it becomes stacking model 1; when $x_1, x_3, x_4, y_1, y_3, y_4$ or $x_1, x_2, x_3, y_1, y_2, y_3$ or x_1, x_3, y_1, y_3 are all zeros, it becomes stacking model 2; when x_3, x_4, y_3, y_4 or x_1, x_2, y_1, y_2 are all zeros, it becomes stacking model 3. So the stacking method can be summarized as: find the parameters in packing model 4 $x_1 \sim x_4$, $y_1 \sim y_4$ to minimize the remaining space and the maximum number of rectangles embedded whose edges are donated as s_m and s_n.

Stacking method four combines the length of side edge to fill the space, which is equivalent to variety stacking models (see Ref. [14]) and mutation models. The stacking model in [12] is a mutation model of stacking method 4, which is similar to this paper. But in [12], y_2 is not bound to $s_n \times y_1$ and x_3 is not bound to $s_n \times x_2$. But they are bound to each other in this paper, which is more consistent to the idea that people maintain the stability, tidily and smooth in the actual loading. In the stacking model of this paper, if the parameters x_1, y_1 are known, then

$$x_2 = \left\lfloor \frac{(x - s_m x_1)}{s_n} \right\rfloor \tag{54.2}$$

$$y_4 = \left\lfloor \frac{(y - s_n y_1)}{s_m} \right\rfloor \tag{54.3}$$

When using parallel stacking method, to maintain the stability of the loading, there is

$$y_2 = \min(\lceil s_n y_1 / s_m \rceil, \lfloor \mathrm{y} / s_m \rfloor) \tag{54.4}$$

"⌈ ⌉" means to round up. To use the smaller value is to ensure that y_2 is not greater than $\lfloor \mathrm{y}/s_m \rfloor$.

When $y_2 = \lfloor \mathrm{y}/s_m \rfloor$, there is $x_3 = y_3 = 0$ and

$$x_4 = \left\lfloor \frac{(x - s_n x_2)}{s_n} \right\rfloor \tag{54.5}$$

Otherwise

$$y_3 = \left\lfloor \frac{(y - s_m y_2)}{s_n} \right\rfloor \tag{54.6}$$

and for the pile of (x_3, y_3), when $s_n x_2 - s_m \lfloor s_n x_2 / s_m \rfloor \le s_m/2$, if we load cargos horizontally, the center of gravity will fall outside the cargo supporting surface (x_2, y_2), so we let

$$x_3 = \left\lfloor \frac{s_n x_2}{s_m} \right\rfloor \tag{54.7}$$

where x_4 is equal to (54.5).When the gap between pile(x_3, y_3) and pile(x_4, y_4) is satisfied with $x - s_m x_3 - s_n x_4 > s_n$, the cargos can be placed vertically (as illustrated in Fig. 54.2), we name this cargo heap as pile(x_5, y_5).

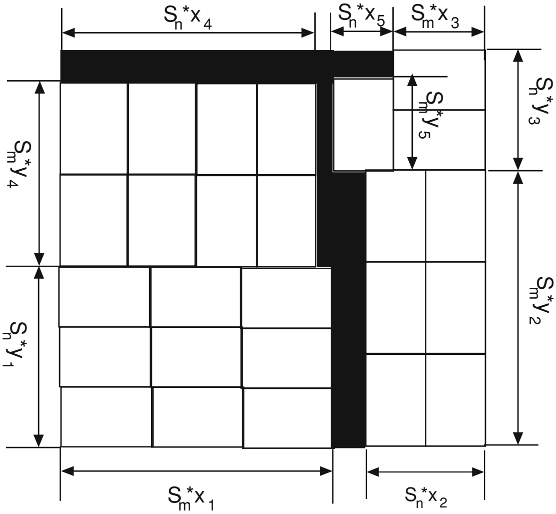


Fig. 54.2 Example of appearing (x_5, y_5) cargo heap

54.4 Space Utilization Test

In this section, the space utilization of packing with the layer stack heuristic algorithm is mainly tested. We realize the algorithm combining C$^{\#}$ with Matlab2009a and run the program on a PC with the Core Duo processor 2.27 GHz, 4G of memory. Table 54.1 compares the space utilization using the algorithm described in [12], general batch method and major domestic packing software. Using our layer stack method to solve the above packing problem, the result is shown in Table 54.2.

From Tables 54.1 and 54.2, we can find that the heuristic algorithm of [12] has the best solution with the highest utilize rate and loading number. The utilize rate shown in Table 54.2 is generally higher than that of general batch method, and maximum utilization of each row is about the same or even higher than that of [12]. For example, when the size of container is (1201, 233, 239) and the size of the small box is (60, 50, 30), the maximum number obtained by our algorithm is 720, which is more than the result of [12] by 20, and the utilization rate is 96.89%, which is 20 more than the result of [12] in number and higher by 2.97%. Especially when the size of container is (1201, 233, 239) and the small box is (61, 56, 39), our algorithm achieves the highest utilization rate of 99.20%.

Also, from Table 54.2, we can conclude that the utilization rate of our algorithm relates with the position of the compartment door. Different positions of the door determines different directions of the layer, and further produce different utilization rates. In the view of the experiment, when packing the layer in the length direction of the compartment, the utilization rate will generally be lower than the case along with the width and height direction. However, sometimes the layer packing along with the length direction may result in high utilization rate, for

Table 54.1 Results comparison of some algorithms

Size of container	Size of box	Trivial algorithm		Great fox software		[11]'s heuristic	
		Numbers	Spc util (%)	Numbers	Spc util (%)	Numbers	Spc util (%)
5800	390,320,310	756	89.49	804	95.17	806	95.41
2300	510,330,290	560	83.63	640	95.57	645	96.32
2450	530,310,470	390	92.15	401	94.75	402	94.98
	560,370,310	432	84.9	490	96.30	492	96.69
	600,530,310	288	86.87	312	94.1	316	95.31
1201	60,50,30	644	86.66	692	93.12	700	93.92
233	60,50,40	460	82.50	524	94.02	524	94.02
239	61,56,39	476	94.82	480	95.62	487	97.01
	63,87,32	316	82.87	361	94.67	361	94.67
	87,68,52	172	78.81	196	90.16	212	93.1

Table 54.2 Computing results of stacking layer method

Container	Size of box	Back door		Side door		Back and side door	
		Numbers	Spc util (%)	Numbers	Spc util (%)	Numbers	Spc util (%)
5800	390,320,310	782	92.57	801	94.82	806	95.41
2300	510,330,290	580	86.61	636	94.98	632	94.38
2450	530,310,470	380	89.78	408	96.40	400	94.51
	560,370,310	486	95.51	492	96.69	483	94.92
	600,530,310	301	90.79	310	93.50	316	95.31
1201	60,50,30	720	96.89	648	87.20	653	88.55
233	60,50,40	500	89.71	524	94.02	525	94.20
239	61,56,39	498	99.20	488	97.21	487	97.01
	63,87,32	329	86.28	350	91.79	358	93.88
	87,68,52	174	80.04	196	90.16	191	87.86

example in the experiment with the highest utilization rate 99.20% and the experiment in which the size of container is (1201, 233, 239), the small box is (60, 50, 30) and its highest utilization rate is 96.89%.

54.5 Conclusion

On packing the specifications of the goods, the previous algorithms only consider the maximization of space utilization. Although high utilization rate can be obtained, yet it is hard to actually achieve due to irregular arrangement. Our algorithm takes the convenience of loading before distribution, and the convenience of unloading in the delivery into account, and put forward the layer along with the direction of easy loading and unloading, according to the location of the compartment doors. If the car has only one door, make the layer parallel to the door, and if the car is equipped with two doors, make the layer perpendicular to the height direction of compartment door. The algorithm considers the actual packing factors better and is easier for loading and unloading. Compared with other packing algorithm, it is a more practical algorithm with high space utilization rate.

References

1. Bortfeldt A, Gehring H (2001) A hybrid genetic algorithm for container loading problem. Eur J Oper Res 131:143–161
2. Pisinger D (2002) Heuristic for the container loading problem. Eur J Oper Res 141:143–153
3. Bortfeldt A, Gehring H, Mack D (2003) A parallel tabu search algorithm for solving the container loading problem. Parallel Comput 29(5):641–662
4. Mack D, Bortfeldt A, Gehring H (2004) A parallel hybrid local search algorithm for the container loading problem. Int Trans Oper Res 11(5):511–533
5. Gendreau M (2006) A tabu search algorithm for a routing and container loading problem. Transp Sci 40(3):342–350
6. Crainic T, Perboli G, Tadei R (2009) TS2PACK: a two-level tabu search for the three-dimensional bin packing problem. Eur J Oper Res 195(3):744–760
7. Egeblad J, Pisinger D (2009) Heuristic approaches for the two-and three-dimensional knapsack packing problem. Comput Oper Res 36(4):1026–1049
8. Yang H, Shi J (2010) A hybrid CD/VND algorithm for three-dimensional bin packing. The 2nd International Conference on Computer Modeling and Simulation. IEEE Press, Sanya
9. de Almeida A, Figueiredo MB (2010) A particular approach for the three-dimensional packing problem with additional constraints. Comput Oper Res 37(11):1968–1976
10. Bu L, Yin C, Pu Y (2002) A genetic and simulated annealing algorithm for optimal sequential casing of less-than-carload freights. J Southwest Jiaotong Univ 37(5):531–535
11. George JA (1992) A method for solving container packing for a single size of box. J Oper Res Soc 43(4):307–312
12. Sui S, Shao W, Gao Z (2005) A heuristic algorithm for dimensional container packing problem of fixed-size cargoes. Inf Control 34(4):490–494
13. Yang D (2007) Optimum algorithm for container loading one type of objects. J Transp Eng Inf 5(2):17–23
14. XU L, Ji Z, Xia J (2008) The optimum algorithm for the container loading problem with homogeneous cargoes. J Shangdong Univ (Eng Sci) 38(3):1–4
15. Niu Y, Fan Y, Xu E (2004) Method of container loading in order model. Logist Technol 5:47–49

Chapter 55
Research on Advanced Web-Based Education via eLML-Structure-Based Mobile Learning

Jinguang Chen and Ding Wei

Abstract E-learning is emerging as a popular learning way all over the world. Mobile learning is one of the most promising solutions for e-learning. However, the practice use of mobile learning platforms is inefficient. This is because most mobile learning platforms serve specific types of mobile devices, which prevents the universal application of a certain platform. To enhance the generality ability of the mobile learning platform, a new solution for the mobile learning is proposed based on the e-lesson markup language (eLML) framework. ELML framework is independent of the platforms and abides by SCORM standard. It is compatible for the current mainstream e-learning platforms. Therefore, it can provide a convenient way to most types of mobile devices. The design of the mobile learning system based on eLML-structure is presented in this paper. The analysis result indicates that the proposed platform provides an effective solution to develop mobile learning system, and thus has application importance.

Keywords ELML · Network education · Mobile learning

J. Chen (✉)
School of Teacher Education, Huzhou Teachers College,
Huzhou 313000, China
e-mail: cjg2003@hutc.zj.cn

D. Wei
Henan Institute of Science and Technology,
Xinxiang 453003, China
e-mail: bingheku6@sina.com.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_55,

55.1 Introduction

Web-based education is the important part and great supplement of the national education system. The role it plays in the national education is greater and greater. Frankly speaking, almost every person has experienced the joy and convenience of the web-based education. It provides people more alternative ways to study at any time and any place. As one of the most promising solutions, mobile learning has been widely studied in the field of web-based education. By the use of mobile devices, mobile learning can offer learning activities at any time and any place [1]. By doing so, the learning cost is decreased while the learning efficiency is improved. In addition, the contradiction between lack of learning resources and increasing learners in developing countries can be relieved [2].

Currently, most mobile learning platforms are limited to serve the specific applications. The generality of different platforms is yet less to satisfying. Many countries and international organizations have already invested the application of integration of multi-mobile learning platforms and made a significant progress [3, 4]. However, the point is that, mobile learning application does not cope with the development of network technologies and computer breakthroughs. More embarrassed, as an advanced education solution, the mobile learning has been seldom adopted in practice. The reasons can be concluded as follows [5–7]: (1) Mobile devices cannot support SCORM standard, which leads to the impossibility of reuse of the e-learning materials; (2) The development of mobile learning platforms corresponds to the specific application and learning environment, and hence makes different mobile devices be different in sharing learning resources; (3) Computer had a significant advantage for recurrence-free serial to the mobile devices. Thus, how to present e-learning resources in a reasonable display to different mobile devices must be solvable in mobile learning research [8–12].

To develop more practical mobile learning, a new mobile learning platform based on eLML is presented in this work. ELML released by University of Zurich is a kind of e-lesson markup language based on XML framework [5]. It abides by SCORM standard and is available for multiple learning platforms. In addition, eLML can be integrated with mainstream e-learning platforms easily. Therefore, it is reasonable to use eLML for the mobile learning practice. The paper discusses the possibility of eLML applied to mobile education, and integrates e-learning resources with mobile learning environment for different devices. The configuration of the eLML-based mobile learning platform is demonstrated, and the analysis result shows that the proposed mobile learning system is feasible and available for practical use.

This paper is organized as follows. The recent progress on network education is described in Sect. 55.2. In Sect. 55.3, the suggestions for network education implement using mobile learning are discussed. The design of the proposed new mobile learning platform based on eLML is presented in Sect. 55.4. Some conclusions are drawn in Sect. 55.5.

55.2 Recent Progress on Network Education

55.2.1 The Connotation of Network Education

Network education is a polyphyletic object. The UNESCO report in 1998 that in both developed and developing countries, there exists a certain degree of gaps between education and social needs, which are extremely serious in the third world countries. So distance education, especially the network education popularization is not only the effective way to solve the problem, but also will become the innovation power on the traditional education mode. The survey report of Chinese social sciences academy points out, due to the rapid development of information network technology and the penetration and application in various professions, online education will become a new education mode with faster dissemination and larger space in our country, at the same time constitute diverse of system together with class education form and broadcasting television education means. It is a new education form developed by using network technology, multimedia technology, modern information technology and developing modern distance education will bring more benefits such as expand scale, improve education quality, strengthen the efficiency, establish lifelong education system and so on. Modern distance education is network education, i.e. e-learning.

55.2.2 Development of E-learning

For building an education website, the resource cost is far from imagining. The cost of education website is much higher than general website. The schools of network learning website face many problems, such as resource scarcity and technology weakness in the construction. There are more than 70 elementary education website currently in our country, but the quality is uneven. One of the main reasons is the lack of sufficient education budget, and another reason is probably because the elementary school education is compulsory, hence the sources is limited and cannot be popularized. We discuss some details as follows:

(1) The ways of construct websites are diverse. One effective way is give the authorization to school to establish website and register domain-name. Another way is the donated websites. The third way is to make websites by themselves. The fourth way is to attend the district portal education website and there have web pages for propaganda. At present universities all have their own websites, but the development of network institute is far behind other countries.

(2) The maintenance modes include schools' self-maintenance, company-hired maintenance and maintenance by education agency. The operation is including two kinds. One of the operation is schools' own technical support ability and

this becomes a platform for notice issue and resource management. The other is that it belongs to a static platform for introducing school basically.

(3) School websites are concentrated in many universities and key middle-school. Since universities usually have their professional maintenance, the maintenance is guaranteed. However, the maintenance ability in middle-school is relatively weak, leading to lag of the network construction of elaborate course. The incentive policy for teachers is not perfect and the society popularization cannot complete. The propaganda is mainly conducted by school and school management, and thus ignores the social requirements. In addition, the budget is discontinuous, which makes the website operation difficult. As a result, significant gaps exist between organizational interests and social needs, leading to poor technology transformation ability.

It can be seen that the education technology and website construction are both backward to national level in e-learning. This will cause the technology falling behind and the talent shortage. In the national mobile learning development trend, the mobile communications company should join with the network education and build mobile learning platform.

55.3 Network Education Implement Using Mobile Learning

55.3.1 The Definition of Mobile Learning

The authority of international remote education, Ireland education technology expert Desmond Keegan divided the remote learning into three stages: d-learning (distance learning), e-learning (electronic learning) and m-learning (mobile learning). He believes that mobile learning will be the main way of remote education in the future. For the remote education, successful education technology is not those suitable for the teaching characteristics, but has widespread popularization. Mobile communications technology is so far the most widely popularized technology.

Currently, in the world population of 6 billion, 15 million people have cell phones. Mobile growth speed of China is much faster, especially in the countryside, and the popularization rate of mobile phone is far higher than the popularization of computer. Therefore, mobile learning is the future of remote education. Mobile learning is not only for subscribing short messaging service (SMS), but is for comprehensive migration for traditional teaching and network teaching which faces wireless internet and mobile terminal environment. Hence, the teaching mode, teaching content, teaching management, teaching services and other aspects must be mobilized all-round.

New Oriental is the most successful education enterprise for face to face teaching in China. The total scale of students is 200 million. Yet its online registration subscriber is 2.5 times of face to face users, and is about 500 million.

New Oriental English online phone launched in 2009 has reached nearly 60 million users in a few months, which demonstrates that it is more easily to accept for students by using cell phones to study. It may become a habit of Chinese children with handheld devices in crucial moments to learn things in the future.

Many universities and research institutions also begin to introduce mobile learning to the traditional education system. Novel mobile learning networks based on mobile application to provide the campus network course have been developed. China mobile learning is subtly changing the perception of traditional education and will reach to the industry mature step by step.

55.3.2 Suggestions and Strategies

China mobile communications company has great passion on the development of mobile education. They suggest that we should rely on local education resources, introduce national excellent education resources and establish reliable and efficient mobile learning system. The system should contain several sections, including mobile assistant platform, mobile courseware release platform and mobile teaching interaction center.

Mobile assistant platform can realize many teaching organization function, including teaching resource distribution, teaching effect testing, online discussion and school assignment management, etc. Besides it can also complete educational management function, such as public announcement information, students register, organization management, registration management and examination/grade management, etc. It allows educational administrator, teachers and students to use different function.

Mobile assistant platform can maximize the use of cell phones and wireless network to realize the education process, and analyze the teaching process and the teaching achievement. It not only has improved the deficiency of traditional teaching methods, but also made education into reality that can proceed anytime and anywhere.

As the core of the solution, the making and release platform of multimedia courseware can conveniently transfer the content under the network environment into courseware under the mobile environment and release the network platform, which can provide the greatest support for the mobile environment teaching. On this platform, we can design and develop courseware according to the application characteristics of moving learners and hope to make fully advantage of moving learning. If all the mobile courseware can be controlled within 10 min, the basis of knowledge, on one hand it can just meet the gap learning needs when waiting for the car or people because of the dapper characteristics of mobile courseware; On the other hand is also full considering the efficient use of mobile terminal battery. As we know, the portable terminal screen is small, so the three split-screen courseware (outline, teacher video, notes) cannot be adopted which is often used in online learning stage, and must adopt by way of longitudinal thrust to apply resources to make materials courseware.

In addition, the group of XinTong message application platform, which is suitable for education agencies and mobile multimedia real-time interactive system, are also the important component solution of the mobile learning. However, being influenced by factors such as charges and the bad image of mobile value-added services, the watchers are many and mostly stay in prospective study, and network charge is the biggest threshold for the reason. Only network charges come down, mobile learning applications will develop in large-scale. We believe that under the promotion of professional education institutions, technology providers and mobile network, the day mobile learning taking the place of future learning overall is not far.

55.4 Design of eLML-Based Mobile Learning Platform

Since most of mobile devices can access to the internet, we adopt the Browser/Server mode as the overall structure of the mobile learning platform. It consists of the mobile terminals, wireless gateways, Web servers, database servers, etc. Figure 55.1 shows the proposed eLML-based mobile learning platform. The platform consists of four main components, including course storage management submodule, syntax analysis submodule, user device identification and connection submodule and e-learning serving support submodule. It also has the maintenance and management tools.

The course storage management submodule is to collect and record eLML courses in XML format. This procedure can be realized in the form of XML Spy, XML Editor or Firedocs eLML Editor. The teachers are asked to prepare eLML contents, and submit eLML lessons to the course storage management submodule. The module handles with uploaded eLML courses and manages eLML course database. Once students request the eLML lesson service, the related information (course name, teacher, eLML storage url, etc.) then can be registered to eLML course database. Thus the students can download those eLML courses in the uploading address.

The syntax analysis submodule is the core of the whole system. It is triggered to be active when the course storage management submodule finishes the uploading. Then it receives the eLML courses from the uploading address and decomposes the eLML courses into various small learning server script files. Through syntax analysis, the eLML courses will be divided down into small units of various content-related learning files at the server side.

The user device identification and connection submodule is to recognize the client. Both new registers and old students could be legally identified by this submodule. Hence, the legal users could consume all the learning contents in the education system.

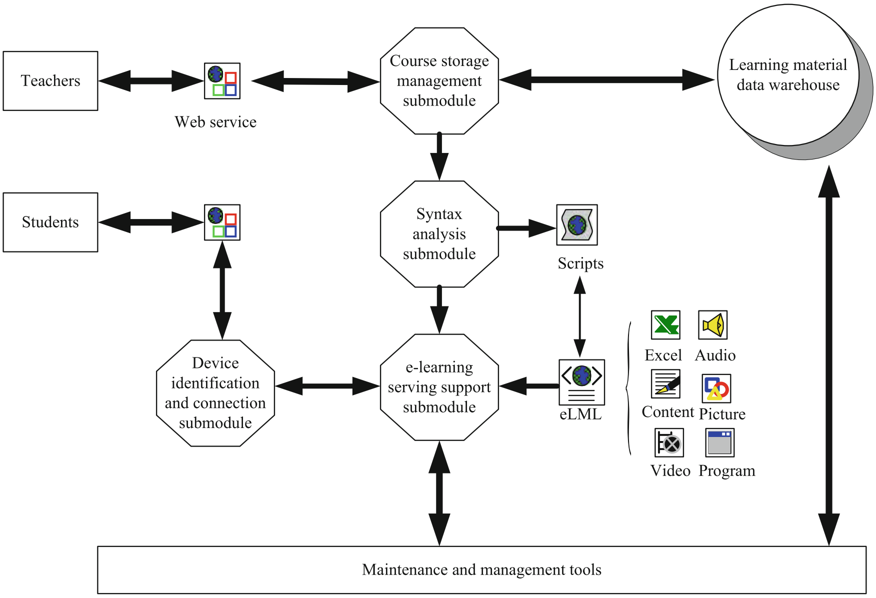

Teachers
Web service
Course storage management submodule
Learning material data warehouse
Students
Syntax analysis submodule
Scripts
Device identification and connection submodule
e-learning serving support submodule
eLML
Excel
Audio
Content
Picture
Video
Program
Maintenance and management tools

Fig. 55.1 The eLML-structure-based mobile learning platform

The e-learning serving support submodule provides learning contents to the students according to device codes and server-side files from users' requests. It also serves as the beeper to order learning materials for the learners. In addition, it processes the learning files and sends the readable and formatted learning courses to the students.

The maintenance and management tools is to manage the mobile interface, maintain the whole system, including database backup, course management, user management, mobile device management, log management, etc.

55.5 Conclusions

With the rapid development of wireless network, the participants of mobile learning increase more and more. The Internet has characteristics of equality, individuation, interactivity and interesting, which play important role in mobile learning areas. As the international mobile learning association (IAmLearn) prospect in the report in January 2009, the cell phone will become the power of learning in the upcoming decade.

However, there are two problems that lie in the practice of mobile learning. One is the generalization. A mobile learning platform only feeds special mobile devices. The other one is the optimized e-lesson contents in accordance with capabilities of mobile devices. To overcome these problems, a new mobile learning platform based on eLML-structure is proposed in this work. The advantages of the proposed mobile learning platform are that the eLML is available for any mobile devices, and the eLML framework is independent of the mobile learning platform. As a result, the process of the transformation of eLML lessons and mobile learning contents could be controlled. By doing so, the main problems for the practice of mobile learning application could be solved and the eLML framework-based mobile learning platform is feasible and available for network education.

References

1. Kukulska-Hulme A, Traxler J (2005) Mobile learning: a handbook for educators and trainers. Routledge, London
2. Traxler J (2004) Mobile learning—evaluating the effectiveness and the cost, Learning with mobile devices—research and development. LSDA 183–188
3. Kim I (2008) Seamless mobile learning: possibilities and challenges arising from the Singapore experience. Educ Technol Int 9:97–121
4. Keegan D (2005) The incorporation of mobile learning into mainstream education and training. In: Proceeding of MLearn
5. Chen Y (2010) Research and implementation of multiserving-oriented mobile learning application based on eLML. In: Proceeding of 2010 international conference on computer application and system modeling, pp 8619–8624
6. Fisler J, Bleisch S (2005) ELML, the e-lesson markup language: developing sustainable e-learning content using an open source XML framework. In: Proceeding of 2nd international conference on web information systems and technologies, pp 180–187
7. Arnold S, Fisler J (2010) OLAT: The Swiss open source learning management system. In: Proceedings of 2010 international conference on e-Education, e-Business, e-Management and e-Learning, pp 632–636
8. Lee M, Tsai K, Wang T (2008) A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. Comput and Educ 50:1240–1257
9. Loots C, Osborne M, Seagraves L (1998) Learning at work: work-based access to higher education. J Contin High Educ 46:16–30
10. Moon S, Birchall D, Williams S, Charalambos V (2005) Developing design principles for an e-learning programme for SME managers to support accelerated learning at the workplace. J Workplace Learn 17:370–384
11. Gangemi A, Catenacci C, Ciaramita M, Lemann J (2006) Modelling ontology evaluation and validation. Lect Notes Comput Sci 4011:140–154
12. Gladun A, Rogushina J, Martinez-Bejar R, Fernandez-Breis JT (2009) An application of intelligent techniques and semantic web technologies in elearning environments. Expert Syst Appl 36:1922–1931

Chapter 56
An Approach to Structure Simplifying for Large-Scale Workflows

Jie Cheng and Guangzhou Zeng

Abstract In a large-scale workflow, the workflow structure needs to be simplified before execution so as to improve the completion performance. This paper puts forward an approach to structure simplifying for structured workflows. First, we present a task planning method based on differential evolution algorithm to map the tasks into available resources; then, based on the mapping relationship, the workflow structure will be simplified by task clustering. To evaluate the performance of the proposed approach, the proposed algorithms are evaluated through a comparison study using simulated workflows executed on a prototype workflow platform. The simulation results prove the effectiveness of our approach.

Keywords Workflow simplifying · Task planning · Task clustering · Differential evolution algorithm

56.1 Introduction

With distributed resources, large-scale workflows have faced many execution challenges. For example, a workflow is usually composed of thousands of fine computational granularity tasks. Since the resource sites are shared and often

J. Cheng (✉) · G. Zeng
School of Computer Science and Technology, ShanDong University, Jinan, China
e-mail: chjie@sdu.edu.cn

G. Zeng
e-mail: gzzeng@sdu.edu.cn

J. Cheng
School of Mechanical, Electrical and Information Engineering,
Shandong University at Weihai, Weihai, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_56,

managed using queue-based management systems, tasks are usually processed with queuing in a resource site, which leads to a great deal of extra time on queue waiting. Moreover, the communication cost consumed for intermediate data I/O from one computing site to another is significant. To this end, the workflow structure needs to be simplified before execution so as to reduce the execution complexity and improve the performance of the whole workflow.

In the past years, a lot of work has been done in workflow structure transform. For example, [1–4] presented their approaches for flexibility and parallelism in decentralized execution environment. Andrea [5] and Neyem [6] gave their methods for the feasibility of execution in resource constrained execution environments. Choi [7] and Li [8] simplified the workflow structure to cater for the process understanding and analysis. However, these researches are not suitable for large-scale workflows where large numbers of tasks have strong impact on the execution performance. In this field, Pegasus [9, 10] first presented the concept of performance optimization, where tasks are grouped into clusters by level-based or label-based approaches so as to be executed as a single task. But these approaches cannot automatically determine the clustering granularity. The tasks to be clustered need to be statically specified in advance, which is not reasonable in most large-scale workflow applications.

This paper focuses on the workflow structure simplified for large-scale structured workflow. The objective is to improve the execution performance. In this paper, we first represent a DAG as an expression in terms of the character of a structured workflow, which can greatly reduce the storage cost and computing complexity. Based on the workflow expression, we put forward an approach to process structure simplifying. First, map the tasks into available resources based on differential evolution algorithm; and then, we simplify the workflow structure by task clustering.

The remainder of this paper is organized as follows. In Sect. 56.2, we introduce some related concepts. The process expression and the problem model are described here. In Sect. 56.3, the workflow structure simplifying approach is detailed. Section 56.4 contains the experiment procedure and the analysis of experimental results. Section 56.5 concludes the paper and gives some research perspectives.

56.2 Problem Formulations

56.2.1 Related Concepts

A task t in a workflow is defined as $t = (ca, \text{input}, \text{output})$, where, ca denotes the execution scale, which is also the estimated computation cost; input and output denote the input and output parameters respectively.

A workflow process is defined as $P = (T,E)$, where, $T = \{t_i | \, i = 1,2,\ldots, n\}$ is a set of tasks, where n is the task number. It is assumed that a workflow always has a unique starting node t_1, and a unique end node t_n; E is a set of form $<t_i, t_j>$, where t_i is called the prior task of t_j, and t_j the successor task of t_i.

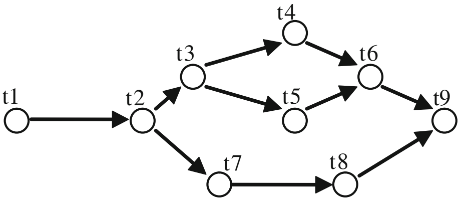


Fig. 56.1 An instance of structured workflow

Generally, a workflow process can be represented as a Directed Acyclic Graph (DAG), where vertexes and directed arcs express the tasks and the dependence relations respectively. In this paper, a workflow process is defined according to the following rules:

1. A task t is a process;
2. If P_1 and P_2 are processes, $P_1 \rightarrow P_2$ is a process, where, connector → denotes sequential relationship;
3. If P_1 and P_2 are processes, $P_1{*}P_2$ is a process, where, the connector * expresses parallel relationship;

According to this recursive definition, a process can be refined by replacing a task with sequential or parallel structure level by level. We call a workflow process generated in this way a structured workflow.

Process Expression (PE) of a process P, denoted as PE(P), is the abstract representation of the workflow structure in terms of the process definition. For example, in Fig. 56.1, the process expression of workflow process P can be expressed as: $\mathrm{PE}(P) = t_1 \rightarrow t_2 \rightarrow (t_3 \rightarrow (t_4 * t_5) \rightarrow t_6 * t_7 \rightarrow t_8) \rightarrow t_9$. For simplicity, process expression can be simplified by omitting the sequence connector →, then PE(P) can be simplified as:

$$\mathrm{PE}(P) = t_1 t_2 (t_3 (t_4 * t_5) t_6 * t_7 t_8) t_9.$$

As we can see, taking advantage of process expression, a workflow can be stored with a link list or a array with a linear complexity, which thus will be much more simpler than using adjacency matrix or an adjacent table. This is meaningful in a large-scale workflow.

A resource r is a service site defined as $r_{id} = (ab, S)$, where, id is the unique identification; ab_i is the execution capability of r_i; S is a set of tasks, in which each task can be executed by r. Let $R = \{r_i | \ i = 1,2,\ldots,m\}$ be the set of available resource, thus the precondition that a process $P = (T,E)$ can be executed by resource R is that $\bigcup_{i=1}^{m} r_i.S = T$

Let $D = (d_{ij})_{m \times m}$ express the communication bandwidth matrix among the execution resources, in which d_{ij} denotes the communication bandwidth between r_i and r_j.

A task block v is a sub-graph of the original workflow DAG, denoted as $v_{id} = \{t_1, t_2, \ldots, t_k\}$, where, id is the unique identification, k is the number of tasks contained in v. Given a workflow process P, the dependence relationship L between two task blocks v_i and v_j is defined as: $L = \{ <v_i, v_j> | (\exists t_p \in v_i) \wedge (\exists t_q \in v_j) \wedge (<t_p, t_q> \ \in P.E)\}$. Let V denote the set of blocks, then the process P can be simplified as $P = (V,L)$. It is assumed that $\forall (v_i, v_j \in V \wedge i \neq j) \rightarrow v_i \cap v_j = \varphi$ and all tasks contained in a cluster must be executed in the same resource.

56.2.2 Problem Model Description

Based on the concepts mentioned above, we describe the problem model of workflow simplifying as follows:

Input: The given process $P(T,E)$, resource set R and communication bandwidth matrix D.
Output: The simplified process $P(V,L)$.
Objective: The minimal completion time of the whole workflow, which comprises queue waiting time, execution time and communication time.

56.3 Workflow Structure Simplifying

In this section, we will describe the process structure simplifying in detail, which includes two steps: task planning and task clustering.

56.3.1 Task Planning

Task planning is to search an optimal mapping between tasks and the execution resources, the objective is to minimize the execution and communication time. In this paper, we employ discrete differential evolution (DE) to address this issue.

A. Solution Representation

We denote the target population, which is a potential solution of the problem, with an n-dimensional vector $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$, where, $x_{i,j}$, the jth dimension of the vector, represents a specific resource in which task t_j can be executed. Thus the vector represents a sequence of execution resources. The order of $x_{i,j}$ in X_i should be set as the order of t_j in the process expression PE(P) generated in Sect. 56.3.1. For each task t_i, its corresponding resource will be chosen in a resource set R^i in which each execution resource $r \in R^i$ satisfies $t_i \in r.S$.

B. Fitness Calculation

According to the problem model presented in Sect. 56.2.2, the fitness function should take into account the execution and communication cost, which is defined as formula (56.1):

$$\text{Fitness}(X_i) = \alpha \cdot \sum_{j=1}^{n} \frac{t_j.ca}{x_{i,j}.ab} + (1-\alpha) \cdot \sum_{j=1}^{n-1} \frac{e_{i,j}}{d_{x_{i,j},x_{i,j+1}}} \tag{56.1}$$

where, $\alpha(0 \leq \alpha \leq 1)$ is the weight of executing overhead.

C. Mutation and Crossover Operations

Assuming that $X_i^k = (x_{i,1}^k, x_{i,2}^k, \ldots, x_{i.n}^k)$ represents an individual of the *k*th generation, in which the best solution is denoted as $G^k = (g_1^k, g_2^k, \ldots, g_n^k)$. The main idea of DE algorithm is that each individual X_i^k will compete with a trial one $U_i^k = (u_{i,1}^k, u_{i,2}^k, \ldots, u_{i.n}^k)$ by comparing their fitness function value to determine who can survive for the next generation. According to the DE/rand/1/bin schemes of Storn [11], the trial vector U_i^k is generated as follows:

$$u_{i,j}^k = \begin{cases} x_{a,j}^{k-1} + F(x_{b,j}^{k-1} - x_{c,j}^{k-1}), \text{ if } r < \text{CR or } j = \text{D}(j) \\ x_{i,j}^{k-1}, \text{ otherwise} \end{cases} \tag{56.2}$$

where X_a^{k-1}, X_b^{k-1} and X_c^{k-1} are three different individuals which are randomly chosen from the (*k*-1)th generation population; *F* is the mutation factor which will affect the differential variation between the two individuals; *r* is a uniform random number between 0 to 1, CR is a user-specified crossover constant in the range [0,1), and *D*(*j*) is a randomly chosen integer in range [1,*n*] to ensure that the trial vector U_i^{k-1} differs from X_i^{k-1} by at least one parameter [12]. Finally, X_j^k will be determined as formula (56.3).

$$X_j^k = \begin{cases} U_j^{k-1}, \quad \text{If Fitness}(\ V_j^{k-1}) \ > \text{Fitness}(X_j^{k-1}) \\ X_j^{k-1}, \quad \text{otherwise} \end{cases} \tag{56.3}$$

Referencing [13], we propose a discrete DE algorithm for the resource allocation. The mutation and crossover operation of the algorithm are defined as follows:

$$U_i^k = F_1(c_1, X_i^{k-1}, F_2(c_2, G^{k-1}, F_3(t, X_a^{k-1}, X_b^{k-1}))) \tag{56.4}$$

Equation 56.4 consists of three components. The first one is $P^k = F_3(t, X_a^{k-1}, X_b^{k-1})$, in which, X_a^{k-1}, X_b^{k-1} are two different individuals ($a \neq b \neq i$) randomly chosen from the (*k*-1)th generation population; *t* is the mutation strength; and F_3 is the crossover operator which is conducted in such a way that it generates a uniform random start number *e* between 1 and *n*, and then determines the $P^k = (p_1^k, p_2^k, \ldots, p_n^k)$ as follows.

$$p_i^k = \begin{cases} x_{a,i}^k \text{ ,if } e \leq i < e + t \\ x_{b,i}^k \text{ ,otherwise} \end{cases} \tag{56.5}$$

The second component is $R^k = F_2(c_2, G^{k-1}, P^k)$, where G^{k-1} is the best individual of the (*k*-1)th generation population; c_2 is the mutation probability, and F_2 is the mutation operator which is employed to accept information from the global best to the temporary member P^k. That is, if a uniform random number r generated between (0,1) is less than c_2, a start number f ($1 \leq f \leq n$) and a length number e ($1 \leq e \leq$ differ (G^{k-1}, P^k), differ (G^{k-1}, P^k) represents the difference between $G^{k-1}$and P^k) will be randomly selected, and then $p_e^k, p_{e+1}^k, \ldots, p_{e+f-1}^k$ will be replaced with $g_e^{k-1}, g_{e+1}^{k-1}, \ldots, g_{e+f-1}^{k-1}$.

The third component is $U_i^k = F_1(c_1, X_i^{k-1}, R^k)$, where c_1 is the choice probability, and F_1 is the selection operator which is applied to determine the generation of the trial individual. If a uniform random number r generated between (0,1) is less than c_1, there will be $U_i^k = R^k$, else $U_i^k = X_i^{k-1}$.

Finally, the selection is based on the comparison of the fitness between U_i^k and X_i^{k-1} such that,

$$X_i^k = \begin{cases} U_i^k, & \text{if Fitness}(\ U_i^k) \ > \text{Fitness}(X_i^{k-1}) \\ X_i^{k-1}, & \text{otherwise} \end{cases} \tag{56.6}$$

D. Algorithm Description

To sum up, the task planning algorithm is described as follows:

```
Input: the given process $P$, resource set $R$, Matrix $D$ and the set of initial
        population $\{X_1^1, X_2^1, \ldots, X_M^1\}$
Output: the optimal planning solution.
Begin
{Initialize parameters for DE Algorithm;
 Initialize iteration number $k = 1$, the maximum iteration number is $K$;
 Calculate the fitness of each individual of the initial population and find the
 global best $G^1$;
 While ($k < K$) Do
    {For each individual $X_i^k$
        {Find two different members $X_a^k$, $X_b^k$, ($a \neq b$);
         $U_i^{k+1} = F_1(c_1, X_i^k, F_2(c_2, G^k, F_3(t, X_a^k, X_b^k)))$;
         If (Fitness($U_i^{k+1}$)>Fitness ($X_i^k$))
            {$X_i^{k+1} = U_i^k$;
            Update $G^{k+1}$;}
            Else $X_i^{k+1} = X_i^k$;
            } EndFor
      $k = k+1$;
    } EndDo
Output the global best $G^k$ as the optimal task allocation;
}End
```

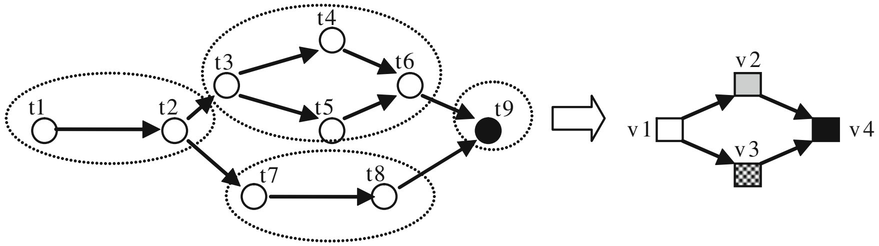


Fig. 56.2 Process structure simplifying (Different color denotes the blocks of different resources)

56.3.2 Task Clustering

After task planning, we get the optimal solution $G = (g_1, g_2,\ldots,g_n)$, where $\Psi(t_i) = g_i$. Then we simplify the workflow structure by the task clustering, which includes the following steps:

(1) Substitute t_i with $\Psi(t_i)$ in PE(P) and generate PE'(P).
(2) Scan PE'(P) from left to right, if there exists a string like "r_ir_i", "$(r_i{*}r_i)$" or "$(r_i + r_i)$", then cluster it into r_i and record the position of the clustering.
(3) Substitute each of the r_i with a block and simplify the workflow structure.

For example, in Fig. 56.2, tasks are clustered into 4 task blocks denoted as v_1 to v_4, the relationship between the structure of the task blocks and their mapping resources are as in Table 56.1. After task clustering, the structure of the workflow is simplified.

56.4 Experiments

To test the performance of the proposed approach, we developed a module to generate random structured DAGs which are simulated as workflow cases. According to the process definition mentioned in Sect. 56.2.1, we generate a simulate process with different structure and scale by such a way that, starting from a single-task-structure process, randomly select the nesting order and nesting depth of sequence and parallel structures. After generating test cases, we brought them into our workflow platform [14] where 10 execution resources attend the simulation. We test the performance of the proposed approach by comparing the workflow execution of two situations: with simplifying and without simplifying.

First, we test the process execution performance with different resources. The workflow case was generated with the process scale $n = 40$. The number of the resource was setting from 3 to 10. The queue waiting time of each executing unit (a task or a block) was simulated as 10 ms per interval. The weight α was set

Table 56.1 Task blocks and their corresponding resources

PE(v_i), $v_i \in V$	$\Psi(v_i)$
PE(v_1) = t_1t_2	$\Psi(v_1) = r_1$
PE($v2$) = $t_3(t_4$ *$t_5)t_6$	$\Psi(v_2) = r_2$
PE($v3$) = t_7 t_8	$\Psi(v_3) = r_3$
PE($v4$) = t_9	$\Psi(v_4) = r_4$

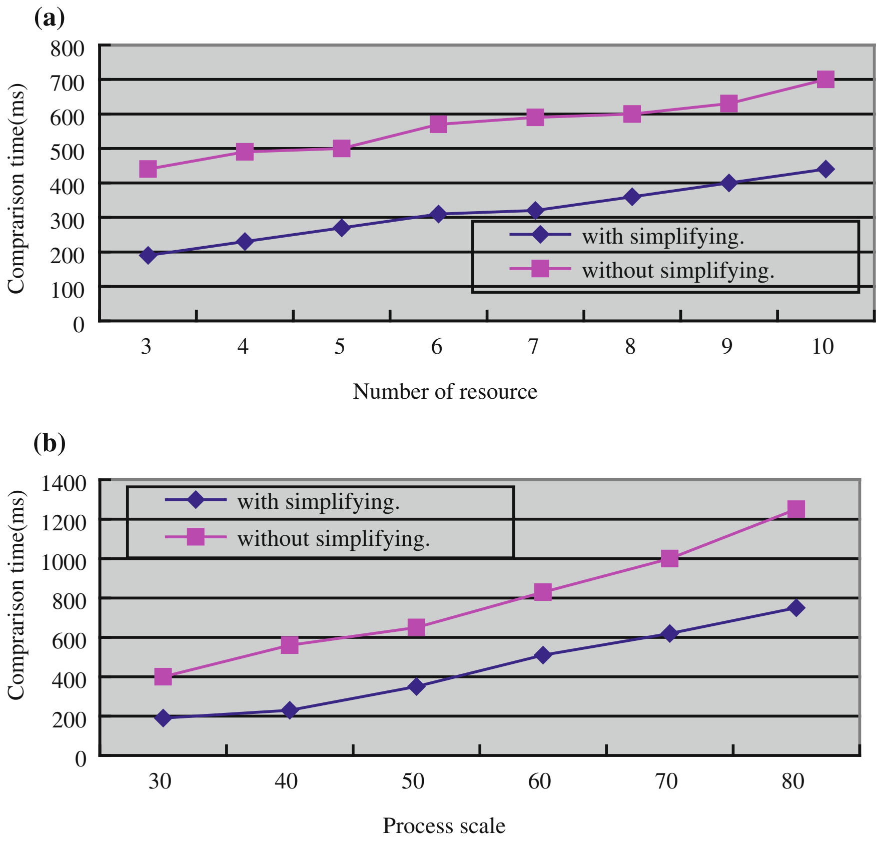


Fig. 56.3 Comparison of the execution performance

to 0.5, and other parameters are set randomly. Each experiment was run 10 times. The average result obtained is as shown in Fig. 56.3a, from which it is evident that there is an obvious advantage with the structure simplifying compared to without simplifying.

Second, we test the execution performance with different process scales. The process scale n was set from 30 to 80. The number of resources was set to 5. Similar to the first experiment, we simulate the queue waiting time of each executing unit as 10 ms per interval and the weight α was also set to 0.5. The test result is as shown in Fig. 56.3b, from which we also find that the performance with structure simplifying is superior to that without simplifying.

56.5 Conclusions and Future Works

Due to the increasing scale of workflow applications, structure optimization is deemed as an effective step before execution. The contribution of this paper is that we present a workflow structure simplifying approach for large-scale structured workflows. The proposed approach includes a task planning algorithm and a set of task clustering principles, with the objective of minimizing the overall completion time and improving the whole workflow execution performance.

In the near future, we will give a non-structured workflows optimization approach and study the convexity of structure simplifying. Furthermore, we will apply the workflows optimization into the specific scientific workflow applications.

References

1. Tan W, Fan YS (2007) Dynamic workflow model fragmentation for distributed execution. Comput Ind 58(5):381–391
2. Nanda MG, Chandra S, Sarkar V (2004) Decentralizing execution of composite web services. In: Proceedings of the 19th annual ACM SIGPLAN conference on object oriented programming, systems, languages, and applications, pp 170–187
3. Liu BX, Wang YF, Jia Y, Wu QY (2005) A role-based approach for decentralized dynamic service composition. J Softw 16(11):1859–1867
4. Bokhari SH (1988) Partitioning problems in parallel, pipelined, and distributed computing. IEEE Trans Comput (C-37):48–57
5. Andrea M, Stefano M (2005) Partitioning rules for orchestrating mobile information systems, personal and ubiquitous computing. Springer, London 9(5):291–300
6. Neyem A, Franco D, Ochoa SF, Pino JA (2007) Supporting mobile workflow with active entities. In: Proceedings of the 2007 11th international conference on computer supported cooperative work in design, pp 795–800
7. Choi Y, Zhao L (2005) Decomposition-based verification of cyclic workflow. Lect Notes Comput Sci. Springer, Berlin 3707:84–98
8. Li JQ, Fan YS (2002) Timing boundedness verification and analysis of workflow model. Comput Integr Manuf Syst 8(10):770–775
9. Deelman E (2010) Grids and clouds: making workflow applications work in heterogeneous distributed environments. Int J High Perform Comput Appl 24(3):284–298
10. Singh G, Kesselman C, Deelman E (2006) Optimizing grid-based workflow execution. J Grid Comput 3:201–219
11. Storn R, Price K (1995) Differential evolution—a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical report TR-95-012, ICSI
12. Beaumont O, Boudet V, Robert Y (2002) The iso-level scheduling Heuristic for heterogeneous processors. In: Proceedings of the 10th euromicro workshop on parallel, distributed and network-based processing, pp 335–342

13. Tasgetiren MF, Pan QK, Liang YC, Suganthan PN (2007) A discrete differential evolution algorithm for the total earliness and tardiness penalties with a common due date on a single-machine. In: Proceedings of the IEEE symposium on computational intelligence in scheduling, pp 271–278
14. Wu XG, Zeng GZ (2010) Goals description and application in migrating workflow system. Expert Syst Appl 37(12):8027–8035

Chapter 57
Generation of Pairwise Test Sets Using a Novel DPSO Algorithm

Sun Jia-Ze and Wang Shu-Yan

Abstract The pairwise test suite generation is one of key issues of combinatorial testing. This paper presents a novel discrete particle swarm optimization algorithm (DPSO) to generate pairwise test data of combinatorial testing. In the algorithm, a particle represents a test suite, fitness function is evaluated by the uncovered number of combination pair, and the position of the particle is produced by stochastic algorithm, which is randomly generated by the frequency of discrete values of all factors in test suite, then optimal test suite which covers all combination pairs is generated. Finally, the classic example is used to illustrate the performance of the proposed algorithm. Compared with the existing algorithms, this paper provides an effective pairwise test suite generation method which has nothing to do with the initial value and can generate the most effective test suit with fast convergence, less calculation and stability.

Keywords Combinatorial testing · Discrete particle swarm optimization algorithm · Test case generation

S. Jia-Ze (✉) · W. Shu-Yan
School of Computer Science and Technology,
Xi'an University of Post and Telecommunications,
Xi'an 710061, China
e-mail: sunjiaze@126.com

W. Shu-Yan (✉)
e-mail: wsylxj@126.com

S. Jia-Ze
Institute of Visualization Technology,
Northwest University, Xi'an 710127, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_57,

57.1 Introduction

Software testing is an important but expensive part in the software development. In many software testing situations, exhaustive testing using all possible combinations of input values for a system is not feasible. Combinatorial testing, as a practical software testing approach, aims to detect the faults that triggered by interactions among factors in SUT by designing and executing a small combinatorial test suite in situations where exhaustive testing with all possible inputs is not feasible to cover the required combinations of these factors. Pairwise testing is a combinatorial technique which selects a subset of all possible test case input combinations to reduce the number of test case inputs to a system.

The generation of pairwise test sets with a minimal size is a combinational optimization problem and can be described by the set covering problem which is well known to be NP-complete. And several deterministic algorithms have been published. While NP-complete problems do not admit efficient deterministic solutions in practice, generally speaking, the NP-complete problems can be solved approximately by heuristic approximate algorithm.

Recently, particle swarm optimization (PSO) [1] developed by Dr. Eberhart and Dr. Kennedy in 1995, with the strong global convergence ability and robustness, are not dependent on the characteristics of problem such as gradient information. It is a kind of random optimistic algorithm of swarm intelligence. Kennedy used a discrete binary version of particle swarm optimization (DPSO) to resolve combinatorial optimization problems in engineering practice [2]. A hybrid algorithm combining PSO with a cross-entropy method optimization method for solving the pairwise test case generation problem was given in [3]. However, search process is blind and slow.

In this paper, a modified discrete particle swarm optimization algorithm is proposed to solve the problem of pairwise test case generation. From testing results of the classic example of pairwise test case generation problem, the novel DPSO algorithm is obviously feasible.

In the next section, we outline basic conception and related work. In Sect. 3, the novel DPSO algorithm is described concretely. In Sect. 4, the typical examples are employed to evaluate the performance of the novel DPSO. Finally, the conclusion is given in the last section.

57.2 Background and Related Work

57.2.1 Pairwise Test Case Generation Problem

The pairwise test case generation problem can be stated as follows:

The fundamental notion behind pairwise testing is the premise that most software faults result from either single-value inputs or by an interaction between pairs of input values. Any pair relation of two factors in the software test system can use binary relation matrix to express as $T = (t_{i,j})_{P \times P}$. P is the number of the discrete values factors in the test system. i and j are the number of the discrete values, $t_{i,j} = 1$ show that the pair which is between ith discrete value and jth discrete value should be covered, $t_{i,j} = 0$ show that the pair which is between ith discrete value and jth discrete value do not need to be covered.

So far, many algorithms have been proposed for pairwise test case generation problem. Mandl [4] first used pair-wise coverage in the software industry for testing an Ada compiler by introducing the method of orthogonal Latin squares. AETG system, developed by Cohen et al. [5] Burrougbs et al. identify several cases of major imbalance in the results of AETG. This indicates that some testers are facing problems because of the results. IPO, based upon covering arrays are discussed in Lei and Tai [6]. The approach uses separate algorithms for horizontal growth and vertical growth. The research literature [7] extends that feasibility study and demonstrates the use of a genetic algorithm to generate pairwise test sets. The research literature [8] adopts ant colony arithmetic to solve the pair-wise test data generating question with fast calculating speed. The research literature [9] presents the results of generating pairwise test sets using a simulated bee colony algorithm. The research literature [3] uses cross-entropy method of statistics and PSO to generate pair-wise test data of combinatorial testing. In this paper, a novel discrete particle swarm optimization algorithm (CDPSO) is proposed, which can avoid slow search speed and premature convergence.

57.2.2 Discrete Particle Swarm Optimization

PSO is an algorithm inspired by the social behavior of bird flocking which is used for finding optimal regions of complex search spaces through the interaction of individuals in a population of particles. When integer solutions are needed, the optimal solution can be determined by rounding off the real optimum values to the nearest integer DPSO has been developed specifically for solving discrete problems. Kennedy and Eberhart put forward the DPSO in 1997 [2]. The velocities and positions of particles in DPSO are updated as follows:

$$v_{\mathrm{id}}^{t+1} = wv_{\mathrm{id}}^{t} + c_1 r_1^t (P_{\mathrm{bestid}}^t - x_{\mathrm{id}}^t) + c_2 r_2^t (G_{\mathrm{bestid}}^t - x_{\mathrm{id}}^t). \tag{57.1}$$

$$\begin{cases} x_{\mathrm{id}}^{t} = \begin{cases} 0, & \text{rand } \geq \mathrm{sig}(v_{\mathrm{id}}^{t+1}) \\ 1, & \text{others} \end{cases} \\ \mathrm{sig}(v_{\mathrm{id}}^{t+1}) = 1/(1+\exp(-v_{\mathrm{id}}^{t})) \end{cases} \tag{57.2}$$

respectively. v_{id}^t is the velocity of the particle in the tth iteration, w is the inertia weight which provides a balance between global and local exploration. r_1^t and r_2^t are the random values between [0, 1] in the tth iteration, and c1 and c2 are the acceleration coefficients. x_{id}^{t} is the position of the optimization domain in the space particle in the tth iteration, $\mathrm{Pbest}_{\mathrm{id}}$ is the best position that the particle ever had in the tth iteration, $\mathrm{Gbest}_{\mathrm{id}}$ is the best position that the group particles ever had in the tth iteration, and rand is a random number selected from a uniform distribution in [0, 1].

57.3 Method

57.3.1 Overview

Pair-wise is a combination way and an economical effective test method, which is produced by principle of combining two levels of all factors. It aims at covering all the possible combinations of all exterior input parameter of software one time at least with the final test suite. The research indicates that, the pair-wise testing is an effective and practical way in testing different software [10].

There are many variations of the basic algorithm structure which are possible. DPSO algorithms merely provide a basic framework for solving a problem and the implementation of a specific DPSO algorithm which solves a specific discrete problem requires several design decisions [11]. Some of the major design decisions include the following. First, a particle representation of a solution to the target problem must be designed. Second, a fitness function which measures how well a particle solves the target problem must be constructed. Third, stochastic algorithms to implement position movement of particle must be designed. Additional DPSO design parameters include selection of the population size, a method for determining how the movement of the particle is produced by stochastic algorithm.

Directly using traditional DPSO to solve pairwise test case generation problem, we may encounter some problems as follows: on one hand, the fitness function usually can not accurately evaluate the particle; on the other hand, as the update of particle position is limited in several discrete values from factors, traditional DPSO will have large amount of computation.

57.3.2 Fitness Function Suitable for the Pairwise Test Case Generation

Pairwise test case generation is general in the sense that the technique applies to any type of discrete input parameter values. When using a DPSO algorithm each parameter value corresponds to a position. The algorithm initially maps all possible input parameter values to consecutive integer values, and these integer values are used as individual position values. This approach results in a particle representation with an array of integer values. For example, suppose some system under test has three input parameters, p1, p2, and p3. Parameter p1 can take on one of two string values, “windows” or “Linux”. Parameter p2 can take on one of three numeric values, 61.5, 82.1, or 345.3. Parameter p3 can take on one of two Boolean values, false or true. These seven values are mapped to arbitrary integer IDs 1, 2, 3, 4, 5, 6, and 7 and used as position values for “ windows “, “ Linux “, 61.5,… true. A particle represents a test set. Suppose the test set size is set to 4 test vectors. Then the array [1–4, 6, 7] is a particle modeling a test set of size 4, where each test vector has size 3.

An individual in the DPSO implementation is defined as a particle and a fitness value. The fitness function is straightforward and is defined simply as the total number of distinct pairs non-captured by the particle representation of test vectors. Notice that this particle representation introduces implicit test vector boundary locations which can serve as target locations for position update operations. Because the total number of non-covered pairs of parameter values can be computed for any given set of parameters and their associated possible values, the fitness value can be used to identify situations where a given individual captures all pairs.

57.3.3 Position Updating for the Pairwise Test Case Generation

The key population design decision was the choice of the update method of position of particle population. For the characteristics of pairwise test case generation problem, the update of position of particle is to directly set several discrete values from factors to the location of each dimension for the particle, meanwhile the new algorithm does not randomly set to discrete values, but produces the position of the particle by stochastic algorithm, which is randomly generated by the frequency of discrete values of all factors in test suite. In this way the test cases with higher coverage will be selected with less probability in the next step, which contributes to improve the search success rate and accelerate the speed of evolution to the optimal solution.

57.3.4 A Novel DPSO Algorithm for the Pairwise Test Case Generation

In summary, the novel DPSO algorithm for the pairwise test case generation is as following:

Step 1. Initialize the number of particles in the particle swarm, the position of individual particle in d dimensions of the problem space. Then the non-covered pairs of parameter values of each particle can be calculated based on binary relation matrix T.

Step 2. Evaluate the fitness of each particle in the particle swarm according to fitness function.
Step 3. For each iteration, compare each particle's fitness with its previous best fitness (Pbest) obtained. If the current value is better than Pbest, then set Pbest equals to the current value and the Pbest location equals to the current location in the d-dimensional space.
Step 4. Compare Pbest of particles with each other and update the swarm global best location with the greatest fitness (Gbest)
Step 5. update the position of the particle in particle swarm.
Step 6. Repeat steps (3)–(5) until the termination t condition is satisfied reached, stop search process and put out he best particle as the best solution. The termination condition maybe a maximum number of iterations or a satisfactory fitness value.

57.4 Numerical Experiments

In this paper, in order to show the process of the novel algorithm, we take a typical and simple example of pairwise test case generation. Simulation experiment parameters are as follows: the particle swarm size is 10.

Suppose the software system has four parameters, "Database", "Client", "WebServer", and "OS". Further, suppose that parameter "Database" can accept one of two possible values, {DB/2, Oracle}. Suppose that parameter "Client"can accept one of two possible values, {Firefox, IE, Opera,Google}, suppose that parameter "WebServer"can accept one of two possible values, {WebSphere, Apache,NET}. Suppose that parameter "OS"can accept one of two possible values,{windows,linux} and suppose that the constraint is if([Client] =="IE")then ([OS]! = "linux").

In order to enhance the DPSO for pairwise test case generation we can map the factor values to arbitrary integer IDs 1, 2, 3, 4, 5, 6,7,8,9,10, and 11. We also note $H_1 = \{1,2\}$, $H_2 = \{3,4,5,6\}$, $H_3 = \{7,8,9\}$, and $H_4 = \{10,11\}$. For this situation there are a total of 2 * 4* 3 * 2 = 48 combinations of input values. For example, one arbitrary test vector is {Oracle, IE, Apache, windows}. Additionally, for this situation there are a total of 43 pairs of input values. So the binary relation matrix $T = (t_{i,j})_{11\times 11}$ of the software system to be tested can be expressed in Table 57.1:

The length of a particle structure is equal to the product of number of parameters and the test set size. Suppose that a particle structure is $l_1 = <(2,5,8,11)$, $(1,5,7,11),(1,3,9,10),(2,6,8,11),\quad(1,4,9,10),(1,5,7,11),(1,3,8,10),(1,5,8,10),(2,5,9,10),(2,5,7,10),(2,5,7,11)\,,(2,6,9,11)>$. The particle illustrated in the above particle represents the 12 test vectors $(2,5,8,11),(1,5,7,11),(1,3,9,10),(2,6,8,11),(1,4,9,10)(1,5,7,11),(1,3,8,10),(1,5,8,10),(2,5,9,10),(2,5,7,10),(2,5,7,11),$ and (2,6,9,11).

Table 57.1 Initial binary relation matrix T

0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	0
0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

Table 57.2 Updated binary relation matrix T_1

0	0	−1	0	−2	1	−1	−1	−1	−3	−1
0	0	1	1	−3	−1	−1	−1	−1	−1	−3
0	0	0	0	0	0	1	0	0	−1	1
0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	−3	−1	0	−2	−3
0	0	0	0	0	0	1	0	0	1	−1
0	0	0	0	0	0	0	0	0	0	−2
0	0	0	0	0	0	0	0	0	−1	−1
0	0	0	0	0	0	0	0	0	−2	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

So the particle l_1 has binary relation matrix $T_1 = (t_{i,j})_{11\times 11}$, showed in Table 57.2

The fitness of a particular is calculated by scanning through binary relation matrix T_1, and counting the number of distinct pairs which are not captured. For example, from T_1, we can add all the one values $\text{UnCount}(l_4) = 1 + (1 + 1) + (1 + 1) + (1 + 1) + 0 + (1 + 1) + 0 + 0 + 0 + 0 + 0 = 9$ so the fitness value of the fitness is 9, that is to say, there are nine pairs which are not covered.

Finally after seven iterations, the *LGBest* of the particle swarm reach zero, that is to say, the best position has been found, the best position is $l = lt; (2, 4, 7, 10), (1, 6, 7, 10), (1, 5, 7, 11), (2, 3, 7, 10), (1, 3, 9, 10), (1, 6\ 8, 10), (2, 6, 9, 11), (1, 4, 9, 10), (1, 5, 9, 10), (1, 4, 8, 10), (2, 5, 8, 11), (2, 3, 8, 11) >$.

The best particle has binary relation matrix $T = (t_{i,j})_{11\times 11}$, showed in Table 57.3.

Figure 57.1 is a test of the experiment time and the number of iterations corresponding map. As can be seen, the number of iterations carry out at least two iterations to a maximum of 20 iterations, average of 7.3336 times. Generally speaking the number of iterations is less and the convergence speed is faster.

Table 57.3 Final binary relation matrix T_1

0	0	0	−1	−1	−1	−1	−1	−2	−5	0
0	0	−1	0	0	−1	−1	−1	0	−1	−2
0	0	0	0	0	0	0	0	0	−1	0
0	0	0	0	0	0	0	0	0	−2	0
0	0	0	0	0	0	0	0	0	0	−1
0	0	0	0	0	0	0	0	0	−1	0
0	0	0	0	0	0	0	0	0	−2	0
0	0	0	0	0	0	0	0	0	−1	−1
0	0	0	0	0	0	0	0	0	−2	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0

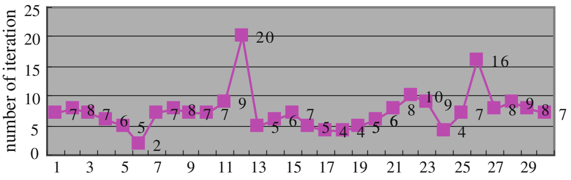


Fig. 57.1 Number of experiments and corresponding iteration map

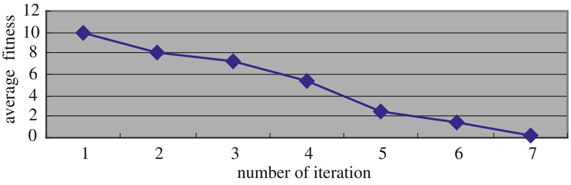


Fig. 57.2 Numbers of iterations and the global optimal fitness corresponding map

Figure 57.2 is the number of iterations and the global optimal location corresponding map during an iterative process. As can be seen, the uncovered pair gradually decreases from 10 to 0-the full coverage. Eventually the optimal position is obtained. So the convergence of this algorithm is effective.

57.5 Conclusions

This paper presents a novel discrete particle swarm optimization algorithm (DPSO) to generate pairwise test data of combinatorial testing. In the algorithm, a particle represents a test suite, fitness function is evaluated by the uncovered number of combination pair, and the position of the particle is produced by stochastic algorithm, which is randomly generated by the frequency of discrete values of all factors in test suite. The classic example is proved that this paper provides an effective pairwise test suite generation method can generate the most effective test suit with fast convergence, less calculation, and stability. Therefore, the DPSO algorithm provides pairwise test case generation problem with a novel and efficient solution.

Acknowledgments This work was supported in part by project "Research on Key Problem of Combinatorial Software Testing optimization Based on Swarm Intelligence" (61050003) from National Natural Science Foundation of China, by project "Smart Combinatorial Soft Testing method " (ZL2009-9) from Natural Science Foundation of XUPT, by project "Smart Combinatorial Embedded Soft Testing Platform" (2009K08-26) from Key Technologies R and D Programmed Foundation of Shan xi Province.

References

1. Kennedy J, Eberhart R (1945) Particle swarm optimization. Proc IEEE Int Conf Neural Netw 4:1942–1948
2. Kennedy J, Eberhart RC (1997) A discrete binary version of the particle swarm optimization. Proc Conf Syst Man Cybern 5:4104–4109
3. Zha R-J, Zhang D-P (2010) Test data generation algorithms of combinatorial testing and comparison based on cross-entropy and particle swarm optimization method. Chinese J Comput 10(33):1896–1908
4. Mandl R (1985) Orthogonal latin squares: an application of experiment design to compiler testing. Commun ACM 28(10):1054–1058 October
5. Cohen DM, Dalal SR, Fredman ML, Patton GC (1997) The AETG system: an approach to testing based on combinatorial design. IEEE Trans Softw Eng 23(7):437–444
6. Lei Y, Tai KC (1998) A test generation strategy for pairwise testing. In: Proceedings of the 3rd IEEE high-assurance systems engineering symposium, pp 254–261
7. McCaffrey JD (2009) Generation of pairwise test sets using a genetic algorithm. In: Proceedings of the 2009 33rd annual IEEE international computer software and applications conference, pp 625–628
8. Li K, Yang Z (2008) Generating method of pair-wise covering test data based on ACO 2008 international workshop on education technology and training, pp 776–780
9. McCaffrey JD (2009) Generation of pairwise test sets using a simulated Bee colony algorithm IEEE IRI 2009, 10–12 July, pp 115–120
10. Lei Y, Tai KC (2002) A test generation strategy for pair-wise testing. IEEE Trans Softw Eng 28(1):109–111
11. Sun J, Wang S (2010) A novel chaos discrete particle swarm optimization algorithm for test suite reduction. In: Proceedings of 2010 2nd international conference on information science and engineering (ICISE), pp 1–4

Chapter 58
Applied Research with Portable Cloud Computing

Mingxiang Sui, Zhihua Zhang and Juanli Hu

Abstract Cloud computing is a focus point in the field of research and application. With the continuous developing and in-depth of cloud computing applications, the handheld portable computing based on cloud computing will be the hot point of research and application. In this paper, we investigate the present primary demands and problems facing portable computing based on cloud computing, through analysis and compare with the related projects and the relevant papers at home and overseas, in combination with cheapness, high efficient and high stability of the computing power in the underlying server. We then propose new solutions based on cloud computing for portable computing. In the system platform, vendors and developers can store and deploy the information and the application into cloud servers unifiedly, and the complex data could be directly processed by the browser installed in handheld portable devices without any installation of other client softwares that is "The Cloud Processing". It can not only greatly reduce the burden of portable terminal, but also form the foundation of the direction and development of research and application for future handheld devices to complex computing.

M. Sui (✉) · Z. Zhang · J. Hu
Zhongshan Polytechnic, Zhongshan, China
e-mail: hjlfoxes@163.com

Z. Zhang
e-mail: hugodunne@yahoo.com.cn

J. Hu
e-mail: hugodunne24@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_58,

Keywords Cloud computing · Portable cloud computing · FaaS · IaaS · SaaS · PaaS

58.1 Introduction

In recent years, with the improvement of portable devices and hardware technology, network coverage, and bandwidth of wireless network, the users' requests for handheld portable devices have changed from simple conversations, SMS interaction and simplex type data storage to complex information query, heterogeneous data sharing and comprehensive application of multimedia. The requirement of diversification and complication makes the portable client resource to be consumed rapidly and the application more complex. More and more scholars and users pay attention to portable performance limitation and uncertainty. Reasonable use of the resources of portable client and the bandwidth of Internet has become the core problem in portable computing popularity and application.

In October 2007, IBM first announced the plan for cloud computing [1], and then many other companies and universities actively participated in relevant products, such as Microsoft [2], Google [3] and UC Berkeley [4]. There is no unanimous definition for cloud computing, and it is regarded as the result of mutual development with pervasive computing, distributed computing, parallel computing and grid computing, virtualization, on-demand calculation, and the theory of SOA WEB2.0. Its core concepts are two aspects: (1) It is mainly used to construct the application as the infrastructure with the equivalent of the status as independent PC (or x86 server) and Internet equipment. (2) The infrastructure above the software applications [5]. Obviously, the biggest characteristic of cloud computing is to integrate all resources seamlessly and adapt on-demand by using a typical B/S structure. The complicating information process can be processed by browser exchange through backend server cluster of cloud. Based on the core conception and character of cloud computing, the portable cloud computing can be described as follows: It is an IT resource or the deliver and use mode of information services which is the foundation established by portable network, platform and software.

58.2 The Present Situation of Portable Computing

Through the development in the past 20 years, the handheld portable devices have improved from cell phones to the PDA, the smart phone and various products with complex functions, which are widely used. At present, the handheld portable computing has become one of the mainstream computing models. Its software and hardware technology has realized the rapid development: ① such as the CPU clock speed has risen up to above 500 MHz and the capacity of RAM has reached between 256 Mb and 512 Mb. Their physical sizes of handheld portable devices have become larger, basically meeting the daily simple documents and small multimedia data processing. ② For the current different devices and OS, manufacturers and individuals provide tens of thousands of applications respectively. For example, Apple has introduced the iPhone, in which the applications have reached 1 million or more.

Handheld devices in the wireless portable environment bear more technical challenges in order to realize the Web access. In this environment for Wireless Web Access, it needs an efficient and reliable mechanism to give the support to overcome the difficulties that are brought by the limited capacity of the handheld device's own hardware and the unreliability of wireless networks. With the emergence and widespread application of the 3rd-generation (3G) [6], Wireless Fidelity (Wi-Fi) [7] and WiMax (IEEE 802.16) [8], the network speed and stability have improved very quickly. The handheld portable devices and the interactive Internet have increasingly become complex and diverse, giving us the premise and the methods to solve the bottleneck of portable devices in the client.

There is a huge contradiction. On the one hand, users require the diversity and high expectations of the handheld portable device functions. On the other hand, handheld portable devices only have self-limited storage, low computing and unstable software and hardware performance. For example, when a user's hardware and software go wrong, important data stored in the device is often lost, or when the user needs the more complex search in a large number of documents, general equipments not only consume too much time, but may even finish the tasks with difficultly. These are serious constraints to complete the portable computing development.

58.3 Framework of Portable Cloud Computing

58.3.1 Architecture of Portable Cloud Computing System

Portable computing framework based on the cloud computing have mainly three major components, as shown in Fig. 58.1.

(1) The user's display interface is the lightweight equipment of the front, including a variety of handheld portable devices: PDAs, smart phones and tablet computers, known collectively as the portable clients. In the portable client of the cloud computing, only the operating system, driver and browser will be installed, so the client is a lightweight terminal.

(2) The network in the middle integrates all the wired and wireless network transmission facilities. Cable network transmission equipment is mainly used to connect the underlying server-side; the wireless transmission equipment is

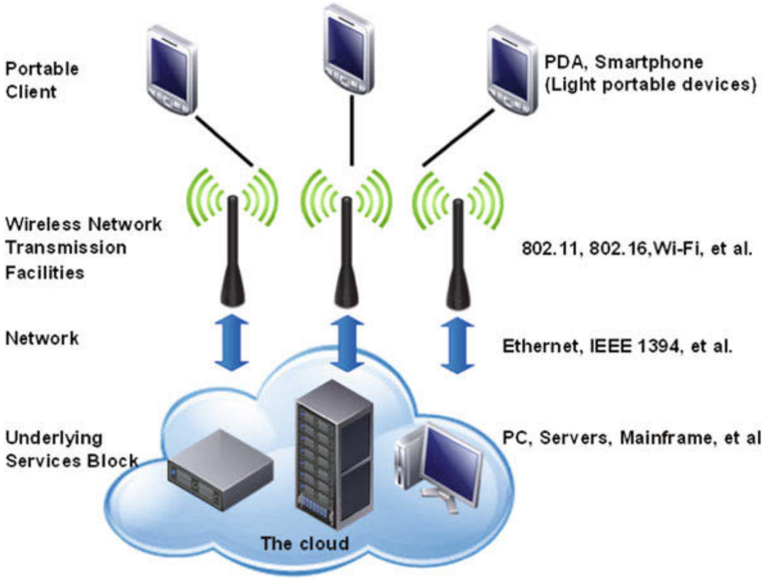

Portable Client
PDA, Smartphone (Light portable devices)
Wireless Network Transmission Facilities
802.11, 802.16,Wi-Fi, et al.
Network
Ethernet, IEEE 1394, et al.
Underlying Services Block
PC, Servers, Mainframe, et al
The cloud

Fig. 58.1 Architecture of portable cloud computing system

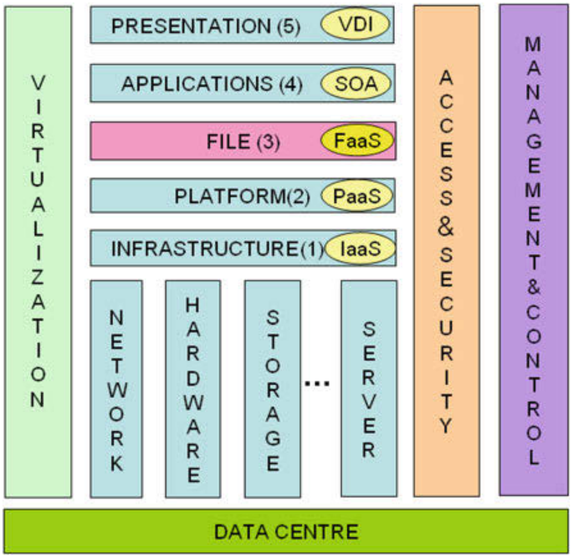
primarily used to connect portable terminals, and its main functions are to complete the data reception and transmission among every portable client and the background cloud.

(3) The underlying basic services can provide a variety of service delivery methods, either through the popular x86 server clusters, or through the large-scale servers etc., but generally tend to use a large number of low-cost servers. In the cloud computing framework, almost all computing tasks are executed through the underlying server clusters. The users can select the corresponding service through a variety of devices in the network according to their requirements; do not need to know the specific location and the technical details of the operating system, middleware and applications etc.

58.3.2 The Model and Layers of Portable Cloud Computing

The portable cloud computing framework model has a position similar to the desktop cloud computing, but it also has its own uniqueness. The model is mainly composed of five different layers. The key to resolve this confusion is the realization that the various offerings fall into different levels of abstraction, as shown in Fig. 58.2, aimed at different market segments.

Fig. 58.2 The Model and Layers of Portable Cloud Computing



(1) *Data Centre*. As the bottom layer of the Portable Cloud Computing model, Data Centre separates the server hardware and the server operating system through the virtualization and gets the hardware, then combines the data center environment and the hardware into a unified whole. As a whole physical environment, it is also known as "doing the physical resources and infrastructure management" (PRIM). In this concept, all the physical resources in the data center will be viewed as a whole, taking it as a container to carry the whole of the IT basic infrastructure and logical IT basic infrastructure.

(2) *Infrastructure*. Infrastructure layer realizes the application of package in the equipments supporting operations mainly through hardware virtualization technology. The equipments include the memory, hardware, the server and the high-speed Internet component, then they are integrated, and deployed, known as Infrastructure-as-a-Service (IaaS). After integrating, the underlying infrastructure can be looked upon as the whole large-scale server to provide the uniform support service for the upper layer, and it can also call the resource according to the needs of the users freely.

(3) *Platform*. The second layer is Platform-as-a-Service (PaaS). Vendors provide the development environment, the server platform and hardware resources to developers. Users can develop the programs based on the service providers to provide the operating system and related services without having to download or install. Services, such as the Force.com and Google App Engine, provide the programming environment which abstracts machine instances and other technical details from the developers. The programs are realized based on data centers, not concerning the developers with matters of allocation. For this, the developers have to handle some constraints that the environment imposes on

their application design, for example the use of key-value stores instead of relational databases.

(4) *File*. The third layer is a special framework of the portable computing with private cloud. The layer can be seen as a symbolic component of the portable computing with cloud. Because the ability of processing the information on the handheld portable devices is weak, the speed of the wireless network, when data is interacting, is slow and susceptible to cause instability. So in this layer, the processing speed about the documents in this layer should be improved, mainly including the file mapping and cutting. So the layers can be seen as a service (FaaS).

(5) *Application*. The fourth layer is the application layer. This layer is mainly through the SOA approach to develop the highly scalable software. By the software virtualization technology, it can provide customers with on-demand software services, which run on different operating systems and hardware platforms. It is also known as Software-as-a-Service (SaaS). SaaS is a channel to provide the software for users through the Internet. It applies the unified arrangements on the cloud server. Users can make custom applications on their own demands, according to software modules and the length of time, without the need to pay to buy the software as a whole. Users also do not have to manage, maintain and upgrade operations on the software. SaaS providers or managers can carry out such operations on the software located in the cloud server.

(6) *Presentation*. The Presentation Virtualization is the top layer of the whole framework. It is the hardware and software technology which makes any handheld portable device to access the application and the users do not to know too much about the other. The application can recognize a device it is used to working with. When the device recognizes an application it knows how to display. In some situations, the special purpose hardware is used on each side of the network connection in order to increase the performance; allow a lot of users to use a single client system or allow a single individual to see multiple displays.

The access, security and system management must also be considered as an important part for many systems. In a virtualized run-time environment, it is about the software technology that manages the access to systems, users and file level security arrangements, and management which makes it possible for some systems to be provisioned and managed as they were on a computing resource.

After the successful implementation of the five layers, the Cloud infrastructure is finished and it is for the execution of application streaming.

58.4 In the Portable Cloud Computing

When we have covered the fundamental model and architecture of the Portable Cloud Computing, its practical application may still be unclear. So, this section discusses the cases of LiveMesh, Android MobileMe and Blackberry Enterprise

Application Server Program, where the application of Portable Cloud Computing would yield big benefits.

Microsoft's "LiveMesh" and Google's mobile search demonstrate the model that cloud computing service providers provide services to the user by mobile phones (or other portable terminals). Apple's "MobileMe" and RIM's Blackberry e-mail service represent the model that the cell phone manufacturers provide services directly to users. The former model relies on the market dominance and the technological leadership of service providers in their superior areas. The latter pattern depends on the attractiveness of the public by the "star cell phone". The two models have achieved cross-cutting, cross-level integration of resources and services. The applications and services provided by the two models have the synchronization of the information storage and the application consistency, thereby ensuring the seamless interface about the user service experience. The open cloud platform will allow users to create and run their own applications.

58.5 Conclusion

With the increasing popularity of intelligent terminals, the rapid development of the wireless broadband and the mature application of virtualization technologies, capabilities of portable devices and user requirements are constantly improving. These changes make the Portable Cloud Computing promote to richer forms, broader applications and more powerful functions, which has brought a huge development space to the portable Internet. Although all sorts of obstacles exist, portable cloud computing has not still become the mainstream of portable Internet services. However, the above examples have demonstrated the fact that cloud computing and the portable Internet have produced broad application prospects by the combination of the two. Cloud computing will be more satisfied with the application development and customize business requirements based on the portable broadband, which enhances the intrinsic value of portable information services to help complete the transformation from portable carriers to integrated information service providers.

In the next step, we will continue to study the FaaS layer of Cloud Computing model deeply. For performance characteristics of handheld mobile devices, we can improve the data processing of the various files, which not only make the seamless interaction with customers and meet users' requirements, but also effectively reduce the resource consumption of terminal equipments to achieve the objective of low energy consumption and high efficiency.

Acknowledgment This paper is supported by the 2010 Education Scientific Research Plan of Zhongshan City (D10143).

References

1. Sims K (2007) IBM introduces ready-to-use cloud computing collaboration services get clients started with cloud computing on http://www-03.ibm.com/press/us/en/pressrelease/22613.wss
2. David C (2009) Introducing Windows Azure: http://download.microsoft.com/documents/uk/mediumbusiness/products/cloudonlinesoftware/IntroducingWindowsAzure.pdf
3. Barroso LA, Dean J, Hölzle U (2003) Web search for a planet: The Google cluster architecture. IEEE Micro Vol 2:22–28
4. Armbrust M,Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M (2009) Above the Clouds: A Berkeley View of Cloud Computing on http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.628&rep=rep1&type=pdf
5. Boss G, Malladi P, Quan D, Legregni L, Hall H (2007) Cloud computing. IBM White paper on http://download.boulder.ibm.com/ibmdl/pub/software/dw/wes/hipods/Cloud_computing_wp_final_8Oct.pdf
6. Arkko J, H. Haverinen (2004) Extensible authentication protocol method for 3rd generation authentication and key agreement (EAP-AKA). Internet Draft draft-arkko-pppext-eap-aka-15.txt
7. Lansford J, Stephens A, Nevo R (2002) Wi-Fi (802.11 b) and Bluetooth: enabling coexistence. IEEE Network 15:20
8. Ghosh A, Wolter DR, Jeffrey AG, Chen R (2005) IEEE Communications Magazine 43:129

Chapter 59
Sustainable Ecological Sanitation System

**Ma Wenlin, Liu Jianwei, Zhang Junzhi,
Dai Shuiwen and Zhang Qi**

Abstract For traditional drainage system used in city and town, there are high consumptions of water and energy, but nutrient and water resource were lost. For widely undeveloped countries and regions, there was lack of sanition. Ecological sanitation (Ecosan) system was built based on that domestic sewage and waste material contained in it were regarded as carriers of resources and energy. It was in accordance with the current economic strategy model of sustainable development. The common features of ecosan system are simply treatment ways, nutrients recycling and water recycling. Eco-toilet is main technological unit of the narrow sense of ecosan, which was fit for low population density of areas and most of them were built in rural villages. The general sense of ecosan system is a modification to traditional drainage system used in city and town, which was fit for developed economic level and high population density of areas. In the end, the development trend of ecosan system was forecasted.

Keywords Ecosan system · Theoretical basis · Concept · Classification · Development

M. Wenlin (✉) · L. Jianwei · Z. Junzhi ·
D. Shuiwen · Z. Qi
Key Laboratory of Urban Stormwater System
and Water Environment, Beijing University of Civil
Engineering and Architecture, Ministry of Education,
Beijing, China
e-mail: mawenlin1130@126.com

M. Wenlin · L. Jianwei ·
D. Shuiwen · Z. Qi
Department of Environmental Engineering,
Beijing University of Civil Engineering and Architecture, Beijing, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_59,

59.1 Introduction

With the development of society and economic, the water consumption of human beings is increasing, but the natural water resource which can be used directly continues to lose. At present, the world is facing severe water shortage problems. There was 1.2×10^{10} of people who were short of water. There was 3.0×10^{10} of people who were lack of sanitation. There was 3×10^{6}–4×10^{6} of people dying from water-related diseases each year. It was forecasted that water crisis would spread to forty-five countries and about 3.5×10^{10} of people would be short of water by 2025.

Problems of drainage systems. Traditional drainage system used in cities belong to end-pipe control. Human excreta and other domestic sewage were simply considered as waste material. They were diluted with clean water several times their volume and then discharged through traditional drainage system. There is a great deal of water wasted. However, it also increases the difficulty in wastewater treatment. There are high consumptions of water and energy. But nutrients and water resources were lost. In a word, the ways that people use nutrients and water are not recycling but extensive and straight line-like for a long time.

In undeveloped countries and regions, the water supply and drainage systems are not perfect. Cesspool is used widely. It is not safety to human health. The infectious diseases caused by fecal pollution often happen.

Utilizing ways to domestic wastewater. Domestic wastewater is mainly composed of washing water and human excreta. The pollution degree of washing with is light. It can be treated by absorption, filtration and biological purification from natural plant ecosystem. Treated water can reused instead of water supply for domestic and industrial production and agriculture production partly.

Human excreta, including feces and urine, contain great deal of nitrogen, phosphorous and organic matter. There aren't bacteria in urine of healthy people. A special pipeline is necessary to collect urine into a pool to storage. Urine can be only diluted before reused as farmland fertilizer. Human feces can be used to produce biogas, a clean energy. It is also used to make organic compost. Biogas residual is a kind of liquid organic fertilizer source. Both biogas fermentation and composting can change human excreta into agricultural fertilizer, which contribute to the recycling of nutrients.

Theoretical basis for ecosan system. Sustainable development is an organic unity of ecological sustainability, economic sustainability and social sustainability. Ecological sustainability is the foundation of sustainable development, economic sustainability is dominant, and social sustainability is goal. They are mutually dependent and reinforcing. All they combine to form a whole of the sustainable development system [1]. One of the conditions of sustainable development is that the resources should be used by humans continuously. Social development and evolution can only be carried out under an affordable range of resources and environment. Once the eco-cycle is interrupted, economic and social development would be significantly affected.

A prerequisite of ecological sanitation idea given was that domestic sewage and waste material contained in it were regarded as carriers of resources and energy. According to the philosophy, human life hardly produces "waste", which is a link of the matter and energy flows in a closed loop, all waste material can be cycled to useful material. So, the concept of ecosan is in accordance with the current economic strategy model of sustainable development [2].

Ecosan system is built on the foundation of regarding domestic pollutants as renewed resource. On the head of domestic pollution source, human excreta and gray water are collected separately, and appropriate disposal measures are taken. Water resource can achieve recycling by reusing of gray water treated. Nutrients from human excreta are reused in agriculture to carry out recycling [3].

59.2 Concept of ecosan system

Origin and development. In the early twentieth century, Leberrecht Migge, a Germany architect, was the first to bring up the concept of ecological sanitation (drainage) system and put it into practice in urban areas [4]. The ecological sanitation system or ecological drainage system was shortly called as "Ecosan" [5]. The development of ecosan systems were discussed frequently in the international seminars world-wide. It showed that there were more and more countries and regions to accept the concept of ecosan system all over the world [6].

Essential features. Since the occurrence of "ecological sanitation (drainage) system" concept, a discussion on how to define was being spread out deeply and extensively, but a generally acknowledged definition was not achieved.

However, the common features of ecosan system are [7]:

Simply treatment ways: Human waste was treated by using technologies near to nature and low energy consumption, such as composting and anaerobic biogas;

Nutrients recycling: Nutrients from human excreta can be safely recycled. A closed nutrient circular system is formed;

Water recycling: Domestic wastewater treated was reused as domestic water or environmental water. A high efficient, safe, reasonable water circular system is achieved.

Winbiad Konsulf, a Sweden researcher, thought that ecosan systems should meet three simple rules: preventing from environment pollution, killing pathogens in human excreta, nutrients contained in human excreta used as fertilizer.

The core ideology. Ecosan, a new concept of eco-economy, is a kind of wastewater management and sanitation systems with characteristics of ecological and economic sustainability. It is guided by the principles of ecology. Its criterions are harmlessness, reduction and resource. It can promote existing sewage collection and treatment system to develop toward a sustainable cycle economy direction. It makes human excreta treated by technologies with characteristics of saving water, energy and land and reused as fertilizer. It possesses performances of protecting health and maintaining ecological balance [8].

To sum up, ecosan systems is a material metabolism system. It links human and nature. It is led by the technology and social behavior, maintained by the natural life system, and activated by the ecological processes.

59.3 Classification of ecosan system

The narrow sense of ecosan system. It contains a series of decentralized, source separation domestic wastewater treatment systems (Fig. 59.1) They are fit for low population density of areas and most of them were built in rural villages. Eco-toilet is their main technological unit.

Bio-toilet refers to those ones they possess strong self-purify capability and resource recycling performances. They don't make environmental pollution and can take full advantage of a variety of resources, involving water, energy and nutrients. Development of eco-toilets was a useful exploration to balance the conflicts between human development and resource sustainable utilization. It was an effective way to solve the sanitation problems in undeveloped areas.

The general sense of ecosan system. Under the conditions of developed economic, high population density, artificial buildings, shortage land resources and degraded agriculture, the ecosystem in urban areas is lack of primary producers and decomposers. So, its development must depend on the interchanging of matter and energy with its surrounding area. For example, with traditional drainage system, its water was from the upper reaches, and its wastewater was released into the lower reaches. However, the system neglected to save water and recycle nutrients of human excreta. A large amount of useful material was diluted with clean water and transferred to centralized treatment plants. It wasn't ecological because it not only increased the difficulty of sewage treatment but also expanded the polluted area, as well as wasting water.

The general eco-san system is a modification to traditional drainage system in city and town (Fig 59.2). It was built according to the natural material circulation pattern, involving water supply based on different water quality standard, source separation system in the buildings, wastewater reused system in residential district, rainstorm storage and utilization device, resource utilization from organic solid waste, etc.

Ecosan prospects. Although the ideology of ecosan is unlikely to negate the existing urban sanitation systems fundamentally completely now, its occurrence confirmed that the traditional sanitation model in Chinese countries was with some ecological features, as well as giving an advice to resolved pollution problems of sewage wastewater in towns and small cities with decentralized treatment thought.

On the way to ecosan sustainable development, there are still a lot problems necessary to be researched deeply, especially its connection with traditional drainage system in densely populated urban areas. Today, a new Socialist countryside is being constructed in China. To the countryside, centralized wastewater collection and treatment system is not economically feasible. During the thousands years of agriculture production, Chinese farmers were used to use human and

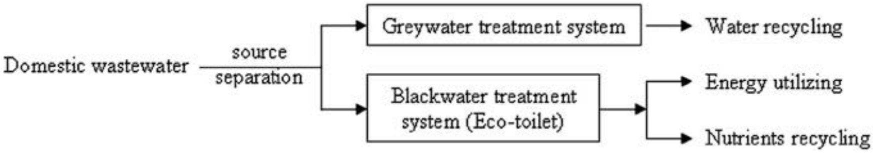


Fig. 59.1 The diagram of narrow sense of ecosan system

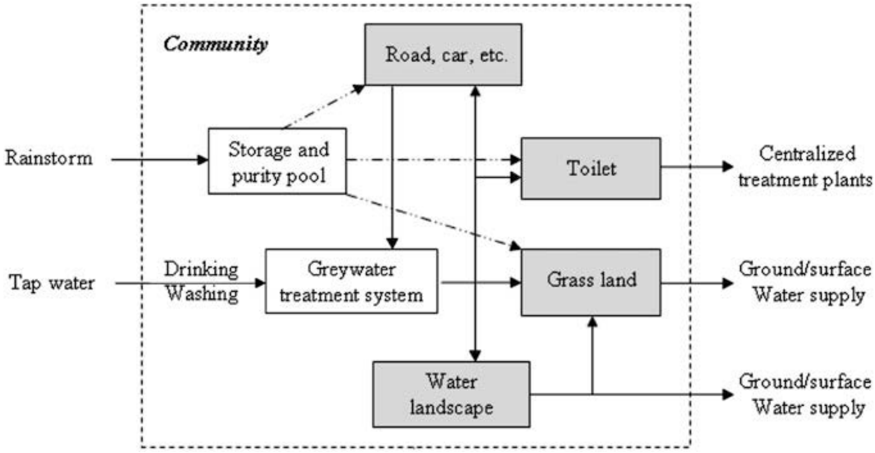


Fig. 59.2 The diagram of general sense of ecosan system

livestock manure to make organic fertilizers. But sometimes sanitation conditions did not meet the full criteria for ecological health. It will be one of important tasks to improve sanitation condition under maintaining healthy ecological circulation of matter and energy in the widely rural area.

Acknowledgements Supported by the Science and Technology Ministry (2009BAC57B0405), and Beijing Academic Innovation Group in Urban Stormwater System and Water Environmental Eco-Technologies (PHR201106124).

References

Xu D (2004) An inquiry into the relation between biodiversity and sustainable development. J Chongqing Three Gorges Univ (Sci) 5(20):105–107 (in Chinese)

Hao X, Song H (2005) Ecological sanitation: the new concept of sustainable and deconcentrated wastewater treatment. Water Wastewater 31(6):42–44 (in Chinese)

Wang X, Wang Q, Hu X (2006) Ecological sanitation system: face to sustainable water usage in China. China water wastewater 22(Suppl):198–201 (in Chinese)

Development of ecosan systems, Uno Winblad, Sweden, Ecosan-Closing the loop in wastewater management and sanitation. In: Proceedings of international symposium, pp 58–62

Esrey SA et al (1998) Ecological sanitation. Swedish International Development Cooperation Agency, Stockholm

Ecological sanitation and wastewater management systems in North America and the Pacific Islands, David del Porto, USA, Ecosan-Closing the loop in wastewater management and sanitation. In: Proceedings of international symposium, pp 204–208

Chen Li, Wei B (2005) Progress on research of ecological sanitation system in rural area. J Environ Health 22(4):306–308 (in Chinese)

Li W, Wang R (2001) System association of ecological sanitation in China. The abstracts of the first international symposium on the science of ecological sanitation. Guangxi Province, China, pp 1–5 (in Chinese)

Chapter 60
Plan and Design of Intelligent Citrus Orchard Irrigation System

Xuejun Yue, Tiansheng Hong, Jianian Li and Tongbiao Sun

Abstract The shortage of water resources is attracting more and more attention. The development of computer technology, controlling technology, electronic technology provides a new solution for efficient use of water. In this paper, an intelligent citrus orchard irrigation system is planned and designed, the wireless network is responsible for data transmission of the entire system, sensors which are controlled by SCM collect data of irrigation areas, then control information of irrigated areas is sent to control irrigation equipment to normally operate according to data analysis, which can effectively use water resources.

Keywords Intelligent · Irrigation system · Wireless network · Control technology

60.1 Introduction

Since the 1980s, the rapid development of computer technology, information technology, electronic technology, control technology, communication technology, has greatly promoted the improvement of the social productive forces, also rapidly changing people's lives [1]. In China, water resource is facing a serious

X. Yue · J. Li · T. Sun
Key Laboratory of Key Technology on Agricultural
Machine and Equipment, Ministry of Education College of Engineering,
South China Agricultural University, Guangzhou, China
e-mail: yuexuejun@scau.edu.cn

T. Hong (✉)
Tiansheng Hong, Professor, College of Engineering,
South China Agricultural University, 483 Wushan, Tianhe, Guangzhou, China
e-mail: tshong@scau.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_60,

shortage. Scarce water resources are not conducive to the development of agriculture. Soil for the growth of crops needs to maintain certain humidity [2]. In traditional agriculture, people often base on experience to irrigate, which if not timely or accurate, often results in inadequate irrigation or floods, and underutilization or waste of water resources.

How to use the limited water resources and taking the path of water-saving agriculture has become the only way for sustainable development of agriculture. With technology development, new technologies can be applied to irrigation areas [3]. Use of an intelligent irrigation system can effectively reduce the leakage of irrigation and the loss of evaporation process. In this paper, an intelligent citrus orchard irrigation system based on the wireless network is planned and designed, which could effectively control the irrigation time to realize water-saving irrigation purposes.

60.2 Principle and Overall Structure of the System

In this study of the intelligent citrus orchard irrigation system, the wireless sensor network as the cornerstone is responsible for data transmissions of the entire system. A large number of acquisition nodes and a few control nodes are included in the system. Acquisition nodes will collect humidity, temperature, water evaporation, light intensity and other information through the wireless sensor network to the computer database which is the control center [4, 5]. The computer produces the irrigation plan and calculates the amount of water according to the irrigation district information of the citrus orchard and citrus water requirement characters stored in the system itself, and sends control information to control nodes through the wireless sensor network. Control nodes control pumps to irrigate citrus orchard according to the received control information.

This intelligent citrus orchard irrigation system uses a three-layer control structure. Figure 60.1 is the framework of an intelligent citrus orchard irrigation system. The entire irrigation system according to the functions and structure can be divided into three layers from top to bottom: the data acquisition and control execution layer, the data aggregation layer, the system control layer [6]. Data acquisition and control execution layer are composed by a large number of acquisition nodes, control nodes and corresponding equipment. Acquisition node contains a variety of types of sensors which gather humidity, temperature, water evaporation, light intensity and other information; control node according to received information from the wireless network controls irrigation equipment connected with control node to run. The data aggregation layer includes wireless network, sink node, network management and database. Aggregation layer is responsible for data collection, sorting and storage. In the data aggregation layer, environment data is collected and control information is sent through the wireless network; network management is responsible for sorting and saving collected data to the database as a basis for generating control information [7]. The system

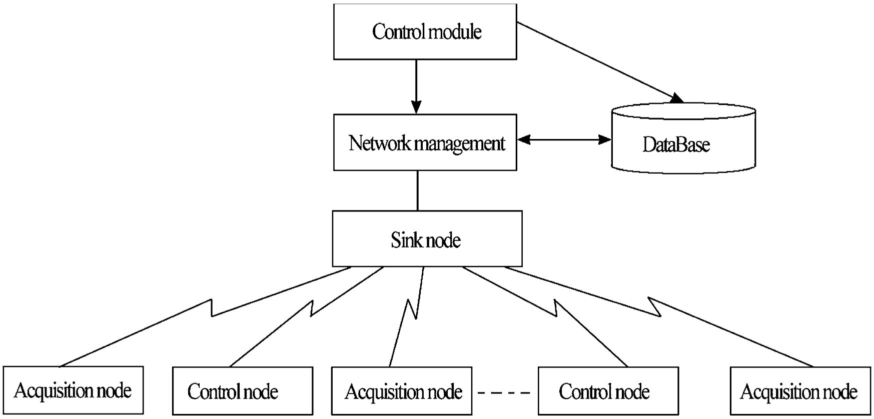


Fig. 60.1 The framework of intelligent citrus orchard irrigation system

control layer includes the control module in Fig. 60.1, the control module reads citrus orchard data from the database, sends corresponding control information to the irrigated areas according to different humidity, temperature, water evaporation, light intensity of irrigated area.

60.3 The Wireless Network Environment Built

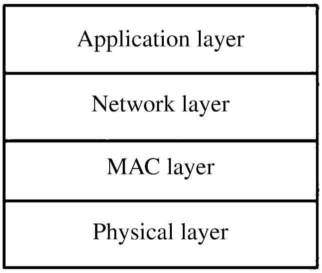
The wireless network is the cornerstone in the intelligent citrus orchard irrigation systems, it is responsible for data transmission of the entire system. In this paper, the wireless network is based on the ZigBee protocol stack. It has advantages of low power consumption, low cost, delay short, great network capacity, reliability and safety.

Network Architecture. ZigBee protocol stack includes physical layer, MAC layer, network layer and application layer. Figure 60.2 is the ZigBee architecture. Functions of protocol layers are as follows:

Physical Layer. Physical layer defines wireless channels and the interface between MAC layers, its function is to realize data transmission and management of physical channels in the hardware drivers. It includes the distinction and choice between channels, monitoring radio signals, modulation/demodulation, sending and receiving data, link quality indication, idle channel assessment.

MAC layer. MAC layer is defined by IEEE802.15.4, provides data transmission and data management services, achieves different users to share the available media resources, provides a unified service for network layer and shields under layers differences [8]. MAC layer management entity service access point

Fig. 60.2 The ZigBee architecture



(MLME-SAP) is that the MAC layer provides management services to the network layer interface to provide management services; public part of the sub-layer service access point is the data service interface to the network layer to provide data services. Yet a service interface exists between the MLME and the MCPS to make MLME use of MAC data service.

Network layer. The internal logical structure of the network layer is divided into two parts: the network layer data entity (NLDE) and network layer management entity (NLME) [9]. The main functions include neighbor discovery, route generation and routing selection. Network layer data entity service access point (NLED-SAP) provides data services for the application support sublayer, network layer through network layer management entity service access point (NLME-SAP) provides the network layer management services for the application support sublayer, in addition to being responsible for maintaining information base (NIB) in network layer. The network layer is responsible for the network topology establishment and network connection maintenance, it is necessary to complete the defined functions of MAC layer, but also provide the appropriate service interface for the application layer.

Application layer. The main function is to provide a variety of user-oriented applications. In this layer, application programming interfaces are provided to the developers, developers can develop corresponding applications according to functional needs.

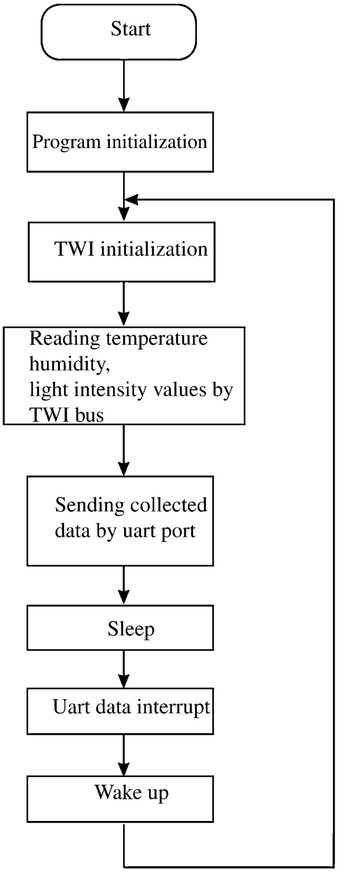
The task processing mechanism. In the ZigBee wireless network, it includes coordinator node, routing nodes and terminal nodes. In this system, the coordinator corresponds to cluster node which is only one in the network, a small amount of routing nodes correspond to acquisition nodes and control nodes, terminal nodes correspond to acquisition nodes and control nodes. Each node has the task handling mechanism to handle node events, including sending and receiving data processes. The main code of task processing mechanism is as follows:

```
UINT16 GenericApp_ProcessEvent( byte task_id, UINT16 events ){
……………………………………………………..
if ( events & SYS_EVENT_MSG) {
 MSGpkt = ( afIncomingMSGPacket_t*)osal_msg_receive( GenericApp_TaskID );
 while ( MSGpkt ) {
  switch ( MSGpkt- > hdr.event ) {
  case ZDO_CB_MSG:
   GenericApp_ProcessZDOMsgs((zdoIncomingMsg_t *)MSGpkt);
   break;
  case AF_DATA_CONFIRM_CMD:
   //This message is received as a confirmation of a data packet sent.
   afDataConfirm = (afDataConfirm_t *)MSGpkt;
   sentEP = afDataConfirm- > endpoint;
   sentStatus = afDataConfirm- > hdr.status;
   sentTransID = afDataConfirm- > transID;
   break;
  case AF_INCOMING_MSG_CMD:
   GenericApp_MessageMSGCB(MSGpkt);
   break;
  case ZDO_STATE_CHANGE:
   GenericApp_NwkState = (devStates_t)(MSGpkt- > hdr.status);
   if ((GenericApp_NwkState == DEV_ZB_COORD)|| (GenericApp_NwkState
 == DEV_ROUTER)|| (GenericApp_NwkState == DEV_END_DEVICE)) {
    //Start sending "the" message in a regular interval.
    osal_start_timerEx(
GenericApp_TaskID,GENERICAPP_SEND_MSG_EVT,
GENERICAPP_SEND_MSG_TIMEOUT);
   }
   break;
  default:
   break;
 }
 ……………………………………………………
}
//return unprocessed events
return (events ^ SYS_EVENT_MSG);
}
```

60.4 Research and Achievement of Intelligent Control

Intelligent control processes collect environmental data of each irrigation area, generate control information and send to the control node to achieve the control of irrigation equipment. It can be divided into environmental data acquisition and control information generated and sent.

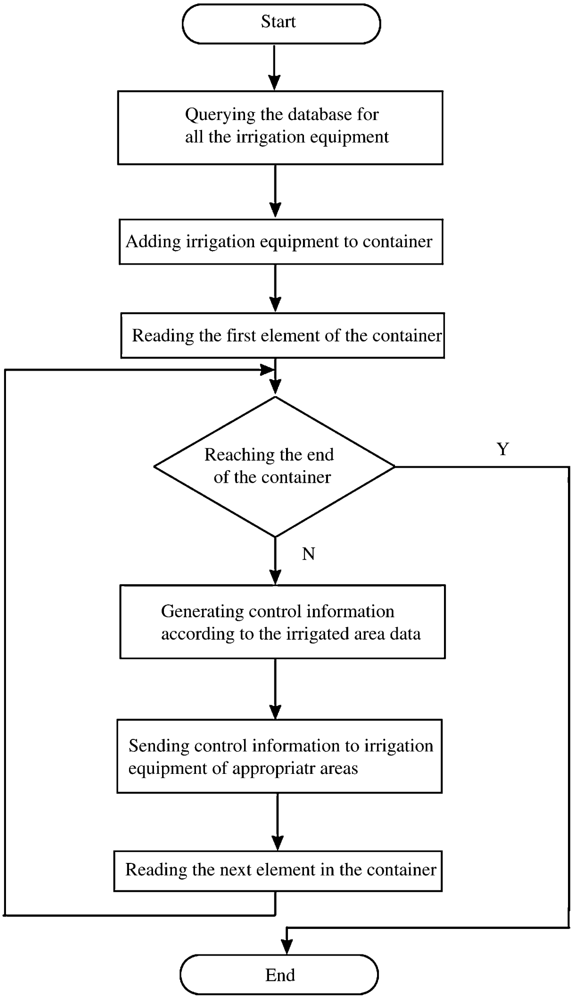
Fig. 60.3 The flowchart of data collection procedure


Start
Program initialization
TWI initialization
Reading temperature humidity, light intensity values by TWI bus
Sending collected data by uart port
Sleep
Uart data interrupt
Wake up

Environmental data collection. Environmental data collection is that where a microcontroller controls a variety of sensors by the TWI bus to collect humidity, temperature, water evaporation, light intensity and other information,and then the microcontroller sends collected data to terminal nodes in the wireless network through the serial uart port, terminal nodes sends out the data through wireless network. Figure 60.3 is a flowchart of the data collection procedure.

Control information generation and transmission. Control information generation and transmission is mainly achieved by the control module. First, the control software reads how many areas the citrus orchard is divided into from the database; then for each area, it reads environmental data of the area and deals with the data, such as the average value, etc.; according to the data processed result of each region, the regional control information is generated and by the wireless network sent to the control node of the corresponding region; control node executes control command to control irrigation equipment of the corresponding region operation in citrus orchard. Figure 60.4 is the flowchart of the control procedure.

Fig. 60.4 The flowchart of control procedure


Start
Querying the database for all the irrigation equipment
Adding irrigation equipment to container
Reading the first element of the container
Reaching the end of the container
Y
N
Generating control information according to the irrigated area data
Sending control information to irrigation equipment of appropriatr areas
Reading the next element in the container
End

60.5 Conclusion

Water is the source of life, but the shortage of freshwater resource is becoming increasingly a serious problem, and how to effectively use freshwater resource is a growing concern. In this paper, an intelligent citrus orchard irrigation system is planned and designed. It gathers the data of irrigation areas by the wireless network and realizes intelligent irrigation to irrigation areas, which both ensures adequate water to crops and saves water. This system provides good solution ideas and implementation methods for intelligent management and water conservation.

Acknowledgments The work was supported by grants supported by Special Fund for Agro-scientific Research in the Public Interest of China (No. 200903023), Earmarked Fund for Modern Agro-industry Technology Research System of China ([2007]431), China national spark program (2010GA780046), Science and technology projects in Guangdong Province (2010B020314007) and Guangdong province agricultural mechanization scientific research issue project scheduling([2010]605).

References

1. Robert PC (1999) Precislon Agriculture: an information revolution in agriculture[M]. Agri Outlookforum
2. Akkaya K, Younis M (2005) A survey on routing protocols for wireless sensor networks[J]. Elsevier Ad Hoc Netw J (03):325–349
3. Akyildizian F, Su W, Sankarasubramaniam Y et al (2002) A surey on sensor networks[J]. IEEE Commun Mag (8):1022114
4. Sinha A, Chandrakasan A (2001) An dynamic power management in wireless sensor networks [J]. IEEE Design TEX Comput 18(2):62–74
5. Warneke B, Last M, Liebowitz B et al (2001) Smart dust: communicating with a cubic-millimeter computer [J]. IEEE Comput Mag 34(1):44–51
6. QiangFeng J, Manivannan D (2004) Routing protocols for sensor networks[C]. USA:CCNC pp 93–98, 271–278
7. Hightower J, Borriello G (2001) Location systems for ubiquitous computing[J]. IEEE Comput 34:57–66
8. Guohua L, Shuqun S (2004) Self-organizing wireless sensor network research[J]. Data Commun Beijing(4):124
9. Fujun W, Songfeng P (2004) Application and design of single chip microcomputer principle system [M]. China Sci Technol Press, Hefei

Chapter 61
The Design and Realization of Underground Location System Based on ZigBee Technology

Yingxi Xu, Zeyu Sun and Chuanfeng Li

Abstract In order to solve problems that existed in the underground positioning systems, e.g., low precision of node positioning, and potential conflicts during data transfer, this study proposed an improved received signal strength indicator (RSSI) underground positioning algorithm which is based on wireless sensor networks. This algorithm analyzed and improved the distance measuring principles of the received signal strength indicator. Iteration computation was applied to calculate data, resulting in the improvement of positioning precision and a considerable decrease of the positioning error of the node position. Accuracy and stability of the data communication among different nodes were also guaranteed and real-time positioning monitoring of the underground worker was realized. Therefore, highly efficient rescue and the safety coefficient of the underground operation were guaranteed. The simulation experiment showed that this algorithm was able to finish complex computation process with fewer parametric variables, thus effectively reducing the communication overhead among network nodes, extending the life cycle of the network, and strengthening the robustness and stability of the entire network system.

Keywords Wireless sensor network (WSN) · Location algorithm · Monitoring system · Received signal strength indicator (RSSI)

Y. Xu (✉)
Electronics Information Engineering College, Henan University
of Science and Technology, Luoyang 471003, Henan, China
e-mail: yingxixu@163.com

Z. Sun · C. Li
Luoyang Institute of Science and Technology, Luoyang 471023, China
e-mail: lylg_sun1977@163.com

C. Li
e-mail: lichuanfeng@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_61,

61.1 Introduction

Because of the complexity of the underground environment, underground workers are often confronted with the threat from temperature, pressure, humidity, toxic gas, and even the life security. In addition, during the traditional mine operation, the dynamic distribution of the underground workers has great randomness, whereupon the management personnel can hardly know the dynamic operation conditions of the underground workers. Moreover, there are plenty of electric cables and wires in the mine. Once mine accidents occur, the consequences are disastrous.

The wireless sensor networks are widely applied to a variety of fields such as national defense, military affairs, environmental monitoring, target tracking, rescuing and relief work, and medical treatment, for its exclusive low cost, perception ability, computation ability, and wireless telecommunications ability [1]. One of the basic tasks of the wireless sensor networks is to monitor and track the information acquisition within the target area. When they are used in underground rescue operations, the number of victims and their specific locations can be tracked accurately through the monitoring of the host computer, making time for rescue operation.

61.2 The Design of the Positioning System

61.2.1 ZigBee Technology

ZigBee is a novel communication technology for short-range communications based on IEEE802.15.4 standards. Its characteristics include low power consumption, low cost, high capacity, short time-delay, and flexible network mode. [2]. Frequency-hopping technology with 2.4 GHz band is used by ZigBee which can form a wireless data transmission network platform consisting of as many as 6,500 wireless data transmission modules. Each transmission module is equal to a base station. Within the whole network range, intercommunication among different networks can be realized. The distance between nodes can be as long as hundreds of meters or even thousands of meters. As a result, they can connect with any existing networks. The network structure can be classified into three types: the star structure, the network structure, and the clustering structure.

61.2.2 The Principles and Design of the Location System

Due to the own limitations of the monitoring system, the complexity of the geologic structure of mines and the variability of environmental parameters, it is hard for the monitoring system to effectively monitor some data.The practical

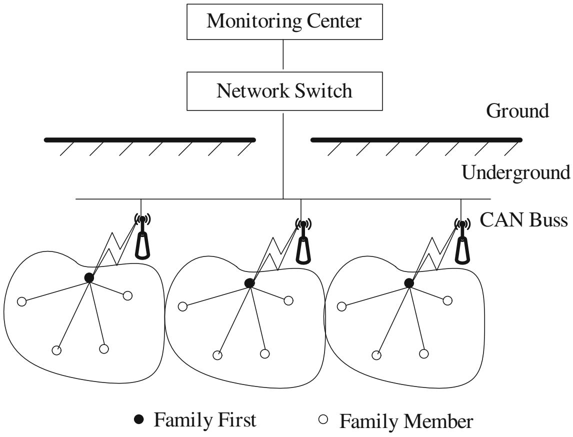


Fig. 61.1 Connection diagram of the location system

situations are taken into consideration, the aboveground part and the underground part, in order to make some improvements. The aboveground part comprises the controller area network (CAN) and the monitoring center while the underground part consists of the power supply, the bus controller, the bus converter, the reader-writer, and the marker. The function of the monitoring center is to transmit the underground data to the host computer and carry out dynamic monitoring and analysis on those data [3]. See Fig. 61.1 for details.

In the clustering structure, the cluster head selection mainly depends on the strength of the signal that sensor nodes can receive. The other nodes which can communicate with the selected cluster head are classified into the same clustering area [4]. In order to prevent the underground environment from interfering with the data signal, ZigBee network modules can be placed at the location needing to position and the location needing to connect with the network. Each network module has the directive function to indicate the strength of the received signal (RSSI). Meanwhile, network module can automatically form a ZigBee communications network. Each cluster node has a unique network identification code. When the cluster head node finishes data collection from the cluster nodes, the data will be packed again in the base station and transmitted to the CAN bus.

61.2.3 The Framework of the System

Two physical equipments used together are defined by ZigBee protocol: the full function device (FFD) and the reduced function device (RFD) [5]. Supporting any types of topological structure, the full function device can communicate with any types of communication equipment. The reduced function device, however, can only communicate with the full function device as it only supports star topology. The underground positioning base station is a FFD. The underground positioning

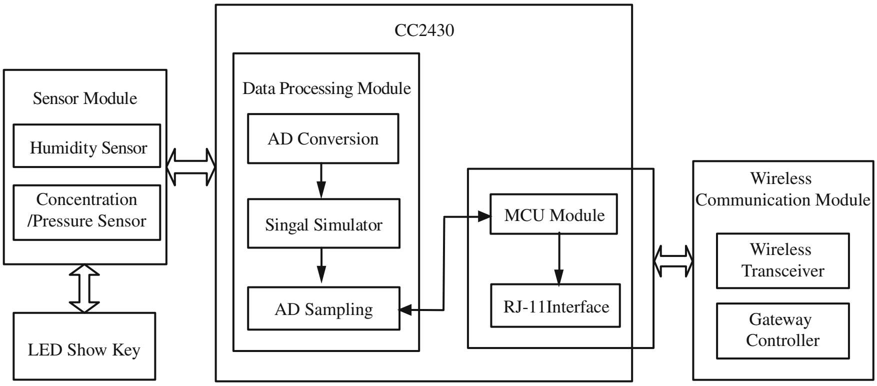


Fig. 61.2 Block diagram of system

station is responsible for building the underground wireless communications network. Broadcast messages will be sent every 3 min to establish wireless communications link with neighboring personnel positioning end devices, in order to realize personnel positioning [6]. Personnel position end device is a reduced function device with two main functions, i.e., sending current position information of the miner, and sound and light alarm. Its core module is CC2430 module. Each terminal is powered by batteries. A unique ID is assigned to each terminal for identification. Location of the personnel is determined through communications between terminal and beacon nodes, i.e., underground positioning of base station. The RFD terminal has a button which can send SOS message, a LED alarming light, and a buzzer. It can realize sound and light early warning when aboveground monitoring center sends early warning signals. The framework diagram is shown in Fig. 61.2.

61.3 The Positioning Algorithm

61.3.1 RSSI Positioning Method by Measuring Distance

The attenuation during sign transmission is used by received signal strength indicator (RSSI) to estimate the distance. The signal strength keeps decreasing as the transmission distance increases when the signal is transmitting in the space. The distance between the signal sending node and the signal receiving node can be estimated by measuring the strength of the received signal [7]. The mathematical model for signal transmission is:

$$P_r(d) = P_r(d_0) + 10n\lg\left(\frac{d}{d_0}\right) + X_\sigma \tag{61.1}$$

where $P_r(d)$ is the received signal strength at the place with a distance of d from the signal sending node, d_0 is the reference distance, $P_r(d_0)$ is the path attenuation coefficient at the reference distance d_0, n is the path attenuation index, and X_σ is the Gaussian distribution random index with the mean value of 0. In the practical application, the positioning accuracy is greatly affected by the path attenuation of wireless relay broadcasting when RSSI signal strength is used to measure distance.

61.3.2 The Location Algorithm Improvement

The measurements are repeated for several times and every measurement is assumed to be independent. Namely, the RSSI measuring values obtained are independently distributed. According to Bayesian probability algorithm, the signal strength is considered as a probability model with normal distribution. As the received signal strength will reach balance finally in RSSI, the Bayesian probability model statistic algorithm has to preserve the distribution information of signal strength as complete as possible, in order to improve the positioning accuracy. The model is:

$$p(l|o) = \frac{p(o|l)p(l)}{p(o)} = \frac{p(o|l)p(l)}{\sum\limits_{l' \in L} p(o|l')p(l')} \tag{61.2}$$

where $p(l|o)$ is the distribution probability of the observing signal strength model o of an unknown node at the place l, $p(l)$ is the prior probability of the unknown node at the place l, which can be assumed as a normal distribution model. L is the set of all beacon nodes and $p(o)$ can be considered as a standard constant. When the signal strength is computed, it is assumed that they all had Gaussian distribution with the mean value of x_i and the standard deviation of σ, where σ is the adjustable parameter in the model. Compared to x_i, the measurement result x has a density function:

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-x_i)^2}{2\sigma^2}} \tag{61.3}$$

When the random disturbance occurs, it can be considered to be reflected in the signal in the form of an extraneous variable with Bayesian probability distribution. When $p_r(d)$ is subject to Gaussian distribution, $p_r(d) \sim N(m, \sigma)$, where $m = \sum_{i=1} x_i \big/ n$, $\sigma^2 = \sum_{i=1} (x_i - m)^2 \Big/ n - 1$, x_i is the ith signal strength value, and n is the signal number. The RSSI value to be transmitted to the unknown node will be obtained when the RSSI received Bayesian filtering processing calculates the mean value.

61.3.3 The Refinement Process

The refinement process is essentially a recursive iterative algorithm. Node positions are updated through a series of procedures and rewritten in the database. At the beginning at each step, the node networks its own position and makes computation. Meanwhile, the position information and corresponding distance evaluation of the neighboring nodes are also received by it. The distance error will be reduced gradually. Finally the refinement algorithm will terminate and the final position will be reported. When the neighboring nodes of an anchor node in a three-dimensional space are over four, the inferred linear equation set are overdetermined. Errors can be balanced using the least square method [8].

Errors are inevitable. Roughly they can be classified into three types. The first type of errors is errors of the position coordinates of the reference nodes selected for node positioning. The second type of errors is measurement errors of RSSI distance measurement technology. The third type of errors is caused during the computation. Accordingly, further correction and evaluation are needed for the estimated position coordinates. The correction function is:

$$\Delta f(x) = \sum e_{ij}\left(l_{ij} - d_{ij}\right)p(i|j) \tag{61.4}$$

where, l_{ij} is the distance between the node i and the node j, d_{ij} is the RSSI measuring distance between the node i and the node j, e_{ij} is the direction unit vector, $p(i|j)$ is the conditional probability of the node i and the node j. Order $E(X)$ and $D(X)$ are the mean value and variance of Gaussian distribution function, respectively. The covariance is:

$$\mathrm{Cov}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x - x_0)(x - x_0)^T \tag{61.5}$$

Then the jth probability density of the posterior probability density is:

$$E(Z_{ij}) = p\left(C_j|X_i\right) = \frac{p\left(c_j\right)p\left(x_i|c_j\right)}{p(x_i)} = \frac{p\left(c_j\right)p\left(x_i|c_j\right)}{\sum_{j=1}^{k} p\left(c_j\right)p\left(x_i|c_j\right)} \tag{61.6}$$

The position coordinate of the unknown node can be considered as $p_i = p_i + \Delta f(x)$, where p_i is the coordinate figure of the node i. Repeat the adjustment process. The adjustment of position coordinate is considered to finish when the correct value of the node position tends to the actual value.

$$E_{rr} = \frac{\sum\limits_{i,j}\left(d_{ij} - \overset{\Lambda}{d_{ij}}\right)^2 \Big/ d_{ij}}{N} \tag{61.7}$$

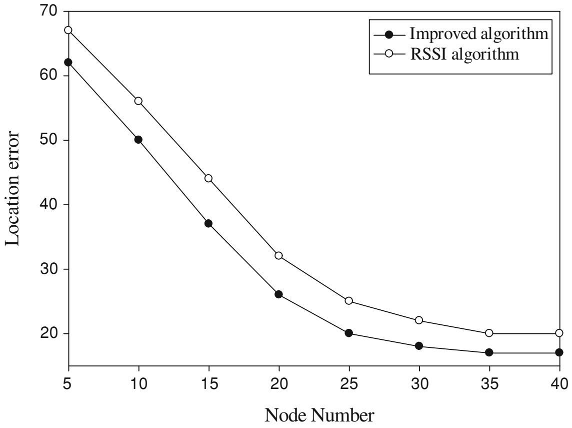


Fig. 61.3 Node location error when the total number of 40

where N is the total number of the neighboring nodes. Since the absolute errors are directly affected by the length of the measuring distance, the E_{rr} at different communication radius varies too greatly to be used as a general positioning error standard. Generally speaking, the measuring error of the distance between nodes increases as the distance increases and the corresponding node positioning error also increases. Thereupon, using iterative method can make the distance between the known node and the unknown node closer to the value calculated theoretically.

61.4 The Evaluation System

In order to verify the correctness of the corrected algorithm and the accuracy of the Bayesian wave filtering probability model and the refinement iteration process, a simulation test was carried out in a two-dimensional space with five sets of coordinate figures for the convenience of the test. MATLAB6.5 was used as the simulation platform. The simulation area was 100 × 100 , where the anchor points were evenly distributed. It was assumed that the network environment was 100 × 100, the beacon nodes and unknown nodes were distributed randomly in the simulation area, the number of nodes was 40, and the communication radius of each node was 10. The experimental data were obtained from the average values of 50 simulation experiments. When determining the number of the nodes, the positioning error would decrease as the beacon nodes increased dynamically, as shown in Fig. 61.3.

The network environment in Fig. 61.3 was that 40 nodes were randomly distributed and the total number of the nodes remained constant. If the number of the nodes changed, the positioning accuracy of both the two algorithms was improved with the increase of the nodes. When the number of the nodes was between 25 and 30, the positioning accuracy of the improved algorithm tended to be balanced. The positioning accuracy of the improved algorithm was significantly higher than that of the RSSI algorithm. For the improved algorithm, the positioning accuracy was calculated through several iterative refinement processes while for the RSSI algorithm, the positioning accuracy was obtained by consuming considerable time and network energy.

61.5 Conclusions

Underground operation safety is directly related to underground mine personnel safety. Highly efficient rescue mainly depends on personnel positioning. Accurate positioning can reduce casualties. Firstly, the design of the positioning system was analyzed and studied carefully and a system framework model was proposed. Secondly, some improvements were made to the RSSI distance measuring algorithm. Stochastic waves were filtered out using the Bayesian probability model whereupon the RSSI algorithm was improved. Then data were corrected through the recursive refinement process. All these measures were expected to effectively determine the specific locations of the underground workers and reduce positioning errors. Experimental results showed that compared with the traditional positioning, the accuracy was improved and underground measurement with high precision positioning was achieved.

Acknowledgments This work was supported in part by the Nature Science Foundation Research Project of Henan Province (2011A590001, 2010B510016), the Funding Scheme for youth teacher of Henan Province (2009GGJS-122), and the Science and Technology Key Project of Luoyang City (1001053A).

References

1. Xu P, Liu B (2009) Topological structure with multiple levels for P2P VoD systems. Comput Eng Appl 45:111–114
2. Li Z, Miao SG (2010) Application of WSNs in mine laneway monitoring system based on ZigBee. Instrum Tech Sens 23:57–61
3. Bulusu N, Heidemann J, Estrin D (2008) GPS-less low cost outdoor localization for very small devices. IEEE Pers Commun Mag 5:28–34
4. Zhang XH, Deng ZD (2008) Localization method used for underground coal mine based on wireless sensor network. Comput Meas Control 16:2003–2005
5. Zhou Y, Jing B (2008) Embedded remote monitoring system based on ZigBee wireless sensor networks. Instrum Tech Sens 21:47–50
6. Li M, Li GH, Diao WG (2010) Wireless sensor network for coal mine personnel positioning system design. Saf Coal Mines 11:77–81
7. Jiang B, Wu YZ, Xie DM (2007) Electronic nose for formaldehyde detection in air. Chin J Sens Actuators 20:1381–1384
8. Lin JZ, Liu HB, Li GJ, Liu ZJ (2009) IPv6 based wireless sensor network design. Appl Res Comput 26:1272-1277

Part VI
Forensics, Recognition Technologies and Applications

Chapter 62
Assessment of Urban Ecosystem Health Based on Attribute Recognition Theory

Mingfu Li

Abstract The deterioration of urban ecological conditions has aroused broad attention during recent years. Assessment of urban ecosystem health is the basis of forecasting and warning of ecosystem conditions. It is also an effective tool to manage urban ecosystem. This chapter constructed an assessment indicator system composed of vigor, organization, resilience, maintenance of ecosystem service and human health. Then it established an urban ecosystem health assessment model based on attribute recognition theory. Finally, it assessed urban ecosystem health conditions of Chongqing city from 1998 to 2007. The results show that there is a tendency of gradual improvement in urban ecosystem health conditions of Chongqing. Health conditions of urban ecosystem belong to relatively unhealthy grade from 1998 to 2004 and critically healthy grade from 2005 to 2007. The assessment model can assess urban ecosystem health conditions objectively and provide a new way for urban ecosystem health assessment.

Keywords Attribute recognition · Urban ecosystem · Health assessment · Analytical hierarchy process · Entropy method

62.1 Introduction

Ecosystem health is a comprehensive science which focuses on human activities, social economic organization and human health and has become one of hotspots in ecology. The concept of ecosystem health was first proposed by

M. Li (✉)
Henan Institute of Engineering, Zhengzhou 451191, People's Republic of China
e-mail: f1937@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_62,

Leopold in 1941 [1]. After that, many scholars advance different definitions of ecosystem health. Haskell states that an ecosystem is healthy if it is stable and sustainable. Costanza summarizes ecosystem health as homeostasis, as absence of disease, as diversity or complexity, as stability or resilience, as vigor or scope for growth and as balance between system components. International Society for Ecosystem Health defines ecosystem health as a science to study the precautionary, diagnostic and prognostic characters of ecosystem management and the relations between ecosystem health and human health. As for the concept of urban ecosystem health, Colin thinks that a healthy urban ecosystem includes the health and integrity of urban ecosystem composed of natural and man-made environments, the health of urban dwellers and social health. Hancock summarizes the six dimensions of urban ecosystem which include health condition of urban dwellers; social welfare state, the effectiveness of government administration and social justice in the city; the built environment quality including housing, traffic, water supply, public park and recreational facilities conditions; natural environment quality including pollution conditions of air, water, soil and noise; the influence of urban ecosystem on generalized natural ecosystem, namely the ecological footprint brought by human beings on natural environment.

During recent years, the rapid acceleration of urbanization has caused many ecological and environmental problems. These problems weaken the service functions of natural ecosystem and damage urban ecosystem health. Therefore, it has become an urgent task to discuss the mechanism of urban ecosystem sickness so as to promote the virtuous cycle of urban ecosystem. Currently study on urban ecosystem health focuses on the construction of assessment indicator system and selection of assessment method. International achievements about assessment indicator system of urban ecosystem health mainly include: sustainable development indicators proposed by International Institute for Sustainable Development, indicators of ecosystem health advanced by International Joint Communion, indicators of urban sustainable development and living quality conducted by the Canada Housing and Mortgage Corporation. Chinese scholars Ma, Wang, Li, Guo, Hu, Ceng and Zhou establish different assessment indicator systems of urban ecosystem health. The methods to assess urban ecosystem health include indicator summarization, fuzzy evaluation, composite index, fuzzy matter-element, attribute reorganization, energy analysis and so on. As a mathematic method to quantify the qualitative descriptions on things and natural phenomena, attribute recognition theory advanced by Professor Cheng [2] can recognize and compare different things effectively and overcome the shortcomings of other assessment methods. Currently attribute recognition theory has been applied to areas such as air quality assessment and urban environmental quality assessment . We introduce attribute recognition theory into urban ecosystem health assessment and develop a case study on Chongqing city.

62.2 Construction of Assessment Indicator System of Urban Ecosystem Health

62.2.1 Selection of Assessment Indicators

Urban ecosystem has the characteristics of multi-variable, so assessment indicator system of urban ecosystem health should have the characteristics of multi-scale and dynamics. According to Rapport [3], health condition of natural ecosystem is measured in terms of vigor (productivity), organization, resilience, maintenance of ecosystem service, management choice, decrease of external input, harm to neighboring system and human health. The above eight items belong to biophysics, social economy, human health, certain time and space categories. We adopt vigor, organization, resilience, maintenance of ecosystem service and human health as main elements of urban ecosystem health in this chapter. Vigor mainly refers to the input and output of materials and energies and the capacity of recycling nutrients in an urban ecosystem. Organization refers to the complexity of ecosystem structure. Generally speaking, with the increase of species number and the complexity of the interactions between the species, the organization of urban ecosystem tends to be healthier. Resilience (which is also called resistance ability) mainly refers to the ability of ecosystem to maintain its structure, namely the ability to gradually rebound after the threat on urban ecosystem diminishes. Maintenance of ecosystem service reflects the services provided by urban ecosystem to human beings, such as water conservation, water purification, providing recreation, reducing soil erosion and so on. Human health reflects the existence and health condition of human beings. Following the principles of availability, comprehensiveness and comparability, we choose representative indicators in the five elements, as shown in Table 62.1.

62.2.2 Grading Standards of Indicators

Referring to [4–6], we divide the levels of urban ecosystem health into five grades: healthy, relatively healthy, critically healthy, relatively unhealthy and unhealthy. We consult the suggested value of ecological city and environment protection model city commonly recognized as the standard value of healthy grade and the international or national minimum value as the limitation value of unhealthy grade. Then we acquire the boundary value between relatively healthy and critically healthy grade by downwardly fluctuation 20% on the former basis and the boundary value between relatively unhealthy and critically healthy grade by upwardly fluctuation 20% on the latter basis. Subsequently, the value of critically healthy grade determined by former and latter means is mutually regulated to yield the termination value. Grading standards of the indicators corresponding to the five grades are shown in Table 62.2.

Table 62.1 Assessment indicator system of urban ecosystem health

Destination layer	Element layer	Indicator layer
Exponent of urban ecosystem health	Vigor	Per capita GDP, GDP growth, Per capita cultivated land, Per capita daily water consumption for living
	Organization	Ratio of tertiary industry to GDP, Ratio of imports and exports values to GDP, Rate of urbanization, Registered urban unemployment rate, Per capita public green land, Rate of afforestation covered area to developed area
	Resilience	Innocent treatment rate of urban garbage, Attainment rate of industrial waste water discharge, Ratio of fixed assets investment to GDP, Ratio of education investment to GDP, Rate of industrial solid wastes comprehensively utilized
	Maintenance of ecosystem service	Proportion of high air quality days in downtown, Rate of quality of drinking water sources up to standards, Per capita residential space of urban households, Engle coefficient of urban households
	Human health	Average life expectancy, Natural growth rate of population, Average education length of urban population, Number of undergraduates per 10,000 persons

62.2.3 Calculation of Indicators Weights

We adopt AHP and entropy methods to determine weights of indicators. The steps of AHP and entropy methods are illustrated in [7, 8].

62.3 Assessment Model of Urban Ecosystem Health Based on Attribute Recognition Theory

Suppose X is the set of x_1, x_2,…, x_n (x_i is the ith city) and m is number of assessment indicators. G is an attribute space of X and is divided into K attribute sets (namely health grades): G_1, G_2,…, G_k. If $G_1 < G_2 < \cdots < G_k$ or $G_1 > G_2 > \cdots > G_k$, where $\cup_{i=1}^{K} G_i = G$ and $G_i \cap G_j = \phi(i \neq j)$, $\{G_1 G_2, \ldots, G_k\}$ is called an ordered division of G [9]. As for each city, ecosystem health condition can be reflected in terms of the m indexes.

62.3.1 Attribute Measure of a Single Indicator

Suppose x_{ij} is measured value of the ith city to the jth indicator, then there is an original data matrix $X = (x_{ij})_{n \times m}$. To measure the attribute of a single indicator, attribute rank division must be confirmed first (as shown in Table 62.3).

Table 62.2 Grading standards of indicators

Indicator	Grading standards of indicators				
	Healthy	Relatively healthy	Critically healthy	Relatively unhealthy	Unhealthy
Per capita GDP [RMB]	>10	[10,5)	[5,3)	[3,0.7)	≤0.7
GDP growth [%]	>10	[10,8)	[8,5)	[5,3)	≤3
Per capita cultivated land [hc]	>0.08	[0.08,0.05)	[0.05,0.03)	[0.03,0.02)	≤0.02
Per capita daily water consumption [liter]	>300	[300,240)	[240,180)	[180,120)	≤120
Ratio of tertiary industry to GDP [%]	>75	[75,60)	[60,40)	[40,20)	≤20
Ratio of imports and exports values to GDP [%]	>12	[12,10)	[10,9)	[9,5)	≤5
Rate of urbanization [%]	>50	[50,40)	[40,30)	[30,20)	≤20
Registered urban unemployment rate [%]	<2.5	[2.5,3.5)	[3.5,4.5)	[4.5,5.5)	≥5.5
Per capita public green land [m^2]	>16	[16,10)	[10,7)	[7,4)	≤4
Rate of afforestation covered area [%]	>50	[50,40)	[40,30)	[30,20)	≤20
Innocent treatment rate of urban garbage [%]	>95	[95,85)	[85,75)	[75,70)	≤70
Attainment rate of industrial waste water [%]	>95	[95,70)	[70,50)	[50,30)	≤30
Ratio of fixed assets investment to GDP [%]	>40	[40,35)	[35,25)	[25,20)	≤20
Ratio of education investment to GDP [%]	>5.4	[5.4,4.5)	[4.5,3.5)	[3.5,2)	≤2
Rate of industrial solid wastes utilized [%]	>95	[90,70)	[70,50)	[50,30)	≤30
Proportion of high air quality days [%]	>90	[90,85)	[85,74)	[74,60)	≤60
Rate of drinking water up to standards [%]	>97	[97,92)	[92,85)	[85,80)	≤80
Per capita urban residential space [%]	>20	[20,16)	[16,11)	[11,7)	≤7
Engle coefficient of urban households [%]	<30	[30,40)	[40,50)	[50,59)	≥59
Average life expectancy [year]	>80	[80,73)	[73,68)	[68,65)	≤65
Natural growth rate of population [%]	<5	[5,8)	[8,9)	[9,10)	≥10
Average education length of population [year]	>16	[16,14)	[14,9)	[9,7)	≤7
Number of undergraduates per 10^4 persons [person]	>1200	[1200,1000)	[1000,600)	[600,300)	≤300

Suppose μ_{ijk} $(x_{ij} \in G_k)$ is attribute measure value of x_{ij} with attribute G_k, where $1 \leq i \leq n$, $1 \leq j \leq m$, $1 \leq k \leq K$, $\mu_{ijk} \geq 1$. To calculate attribute measure value of a single indicator, we choose normal distribution function as attribute measure function. Formula of μ_{ijk} $(x_{ij} \in G_k)$ is:

Table 62.3 Attribute rank dvision of indicators

Indicator	G_1	G_2	…	G_K
I_1	a_{11}–a_{12}	a_{12}–a_{13}	…	a_{1K}–a_{1K+1}
I_2	a_{21}–a_{22}	a_{22}–a_{23}	…	a_{2K}–a_{2K+1}
…	…	…	…	…
I_m	a_{m1}–a_{m2}	a_{m2}–a_{m3}	…	a_{mK}–a_{mK+1}

$$\mu_{ijk}\left(x_{ij}\middle|x_{ij} \in C_k\right) = e^{-\left[\left(x_{ij}-b_{ijk}\right)/c_{ijk}\right]^2} \tag{62.1}$$

where $b_{ijk} = \left(a_{ijk} + a_{ijk+1}\right)/2$ and $c_{ijk} = \left|a_{ijk} - a_{ijk+1}\right|/\left(2\sqrt{\ln 2}\right)$.

62.3.2 Comprehensive Attribute Measure of Multiple Indicators

Suppose w ($w = (w_1, w_2,\ldots, w_m)$) is weight vector of indicators against destination layer, where $w_j \geq 0$ and $w_1 + w_2 + \cdots + w_m$. After weights and attribute values of indicators are calculated, comprehensive attribute measure of the ith city with attribute G_k is calculated by:

$$\mu_{ik} = \mu(x_i \in G_k) = \sum_{j=1}^{m} w_j \times \mu_{ijk}. \tag{62.2}$$

Vector $\mu_i(\mu_i = (\mu_{i1}, \mu_{i2},\ldots, \mu_{iK}))$ is comprehensive attribute measure of multiple indicators of the ith city with each grade. If attribute measure is aimed at element or destination layer, the result is comprehensive assessment of element layer or destination layer, and the latter final result.

62.3.3 Judgment of Ecosystem Health Condition of a Single City

The aim of attribute recognition is to judge health grade of the ith urban ecosystem according to comprehensive attribute measure vector μ_i. In terms of confidence criterion, for confidence degree λ ($0.6 < \lambda < 0.7$ under most circumstances), k_i can be calculated by the following formulas:

For $G_1 < G_2 < \cdots < G_k$, the formula is:

$$k_i = \max\left\{k : \sum_{t=k}^{K} \mu_{ik} \geq \lambda,\ 1 \leq k \leq K\right\}. \tag{62.3}$$

For $G_1 > G_2 > \cdots > G_k$, the formula is:

$$k_i = \min\left\{k : \sum_{t=1}^{k} \mu_{ik} \geq \lambda,\ 1 \leq k \leq K\right\}. \tag{62.4}$$

If the value of k satisfies the above formulas, ecosystem health condition of the ith city belongs to G_{ki} level [10].

62.3.4 Sorting of Ecosystem Health Conditions of Different Cities

To sort ecosystem health grades of different cities, vector μ_i must be sorted firstly. Suppose n_i is the score of attribute set G_i (as for $G_1 < G_2 < \cdots <G_k$, $n_i = i$; as for $G_1 > G_2 > \cdots >G_k$, $n_i = K+1-i$), then ecosystem health score of the ith city is calculated by:

$$q_i = \sum_{k=1}^{K} n_k \mu_{ik}. \tag{62.5}$$

In terms of (62.5), ecosystem health scores of different cities are calculated. If $q_i > q_j$, we consider the ith city is healthier than the jth city.

62.4 Case Study

Chongqing is one of the four municipalities directly under the central government in China. During recent years, owing to ignoring the protection of resource and environment, urban ecosystem health is endangered. Therefore, it is of great significance to assess ecosystem health conditions of Chongqing. The original data is derived from Chongqing Statistical Yearbook (1998–2007) and Environment Situations Bulletin of Chonqing (1998–2007).

According to AHP and entropy methods, the weight vector of indicators is: $w =$ (0.0338,0.0344,0.0424,0.0355,0.0406,0.0346,0.0348,0.0357,0.0356,0.0366, 0.03440,0.0335,0.0327,0.0384,0.0339,0.0371,0.0360,0.0353,0.0339, 0.0389,0.0350,0.053,0.0366,0.0354,0.0344,0.0341,0.0352,0.0360). In this chapter, attribute space is $G = \{\text{health grade}\}$ and is divided into five attribute sets: $G_1 = \{\text{healthy}\}$, $G_2 = \{\text{relatively healthy}\}$, $G_3 = \{\text{critically healthy}\}$, $G_4 = \{\text{relatively unhealthy}\}$, $G_5 = \{\text{unhealthy}\}$. Obviously, $\{G_1, G_2, G_3, G_4, G_5\}$ is an ordered division of G. Let $\lambda = 0.6$ and $n_i = i$, ecosystem health conditions of Chongqing are calculated, as shown in Table 62.4.

According to Table 62.4, the sequence of ecosystem health conditions of Chongqing from poor to good is: 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007. Health conditions from 1998 to 2004 and from 2005 to 2007 belong to relatively unhealthy and critically healthy grade, respectively. Obviously, there is a tendency of improvement year by year in urban ecosystem health conditions of Chongqing. However, urban ecosystem health conditions of Chongqing are unsatisfactory and need to be improved.

Table 62.4 Ecosystem health conditions of Chongqing city from 1998 to 2007

Year	Comprehensive attribute measure					Health grade	q_i
	Healthy	Relatively healthy	Critically healthy	Relatively unhealthy	Unhealthy		
1998	0.0658	0.1182	0.2673	0.2807	0.2681	Relatively unhealthy	2.4329
1999	0.0685	0.1330	0.2600	0.2819	0.2566	Relatively unhealthy	2.4748
2000	0.0798	0.1479	0.2603	0.2814	0.2307	Relatively unhealthy	2.5649
2001	0.0971	0.1598	0.2562	0.2766	0.2103	Relatively unhealthy	2.6569
2002	0.1102	0.2235	0.1809	0.2692	0.2162	Relatively unhealthy	2.7422
2003	0.1332	0.2240	0.1923	0.2512	0.1993	Relatively unhealthy	2.8406
2004	0.1769	0.2216	0.1705	0.2651	0.1658	Relatively unhealthy	2.9786
2005	0.1610	0.2500	0.1968	0.2432	0.1490	Critically healthy	3.0307
2006	0.1900	0.2397	0.2304	0.2405	0.0994	Critically healthy	3.1804
2007	0.1988	0.2180	0.2777	0.2339	0.0717	Critically healthy	3.2383

62.5 Conclusions

This chapter uses attribute recognition theory to assess ecosystem health conditions of Chongqing city from 1998 to 2007. There is a tendency of improvement year by year in ecosystem health conditions of Chongqing during the study period, which increases from relatively unhealthy to critically healthy grade. It indicates that service function of urban ecosystem has deteriorated, ecological environment is endangered to a certain extent, ecosystem structure has changed and is apt to degrade after perturbation, problems of ecological environment appear and ecological disasters occur from time to time. On the whole, ecosystem health conditions of Chongqing city are unsatisfactory and need to be improved. The results conform to actual situations and can provide references for regional development policies.

References

1. Jerry MS, Mariano B, Annalee Y (2001) Developing ecosystem health indicators in Centro-Habana: a community-based approach. Ecosyst Health 7:15–26
2. Cheng QS (1997) Attribute sets and attribute synthetic assessment system (in Chinese). Syst Eng Theory Pract 17:1–8
3. Rapport DJ (1989) What constitutes ecosystem health. Perspect Biol Med 33:120–132
4. Guo XR, Yang JR, Mao XQ (2002) Primary studies on urban ecosystem health assessment (in Chinese). China Environ Sci 22:525–529
5. Zhou WH, Wang RS (2005) An entropy weight approach on the fuzzy synthetic assessment of Beijing urban ecosystem health (in Chinese). Acta Ecol Sinica 25:3244–3251
6. Sang YH, Chen XG, Wu RH (2006) Comprehensive assessment of urban ecosystem health (in Chinese). Chin J Appl Ecol 17:1280–1285

7. Li B (1998) Weighting and the accuracy of weight estimation in delphi and AHP (in Chinese). Syst Eng Theory Pract 16:78–82
8. Jin LJ, Wu KY, Li RZ (2007) Region water security evaluation method based on information entropy and improved fuzzy analytic hierarchy process (in Chinese). J Hydroelectr Eng 26:56–59
9. Yan WT (2007) Research on urban ecosystem health attribute synthetic assessment model and application (in Chinese). Syst Eng Theory Pract 27:137–145
10. Yan WT, Yuan XZ, Xing Z (2007) Urban ecosystem health assessment based on attribute theory: a case study in the new district of northern Chongqing city (in Chinese). Chin J Ecol 26:1679–1684

Chapter 63
Mobile Single Sign-On Systems Against Guessing Attack

Yung-Cheng Lee

Abstract Single sign-on system allows users to login networks for services with only a set of password. Many single sign-on solutions available today are lack of mobility or even insecure. In this article, we propose two new single sign-on solutions with the merits of mobility and resisting password guessing attack. The proposed first scheme is based on smart card, and the second one is based on password. Both the schemes are simple and practical for non-PKI users.

Keywords Single sign-on · Smart card · Password authentication · Guessing attack

63.1 Introduction

Networks are very important platforms for people to obtain various kinds of information or services. In these days, for obtaining various services, people usually have many different network accounts such as webshops, work accounts, ISP, e-mail and online-banking. Each account needs at least one set of password to login, as a result people have to manage a lot of passwords. For a network without single sign-on mechanism, if a user chooses a set password for all accounts, it is very dangerous to security since all accounts will be opened in case of the password is disclosed. On the other hand, having different password for each account will lead to trouble on password management. People are probable to write down

Y.-C. Lee (✉)
Department of Security Technology and Management,
WuFeng University 117,
Sec. 2, Chiankuo Rd., Minhsiung,
Chiayi County 62153, Taiwan
e-mail: yclee@wfu.edu.tw

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_63,

their passwords or to use the same password for different accounts, unfortunately, systems with these approaches always have security flaws since the passwords can be easily revealed. Therefore, it needs to find better solutions for users to login their accounts [1].

Single sign-on system allows users to login networks for services with only a set of password. Several researchers have proposed solutions for single sign-on systems [2, 3, 4]. In 2008, Mauro et al. proposed a single sign-on solution for the management of healthcare in Germany hospitals with smart cards [5]. Many of the single sign-on solutions available on the market today are either too expensive or lack of mobility. Pashalidis et al. proposed a single sign-on system for GSM [2]. Their scheme changes the GSM architecture such that it is impractical even though the scheme is secure. Moreover, their scheme cannot prevent smart card stolen-attack. That is the attacker can impersonate the legal user to login system if he/she obtains the smart card.

Short and simple passwords are vulnerable to guessing attacks [6–8]. Password guessing attacks include online and offline attacks. An online password guessing attack happened when an attacker attempts to guess the password during an online transaction, while an offline password guessing attack occurred when an attacker guesses the password and verifies his/her guess offline. Single sign-on systems need passwords for users to login, thus how to develop a secure single sign-on mechanisms that can resist various attacks including guessing attack is an important issue.

Smart cards, because of their portability and tamper-free features, are convenient and secure devices for remote authentication. Though smart cards have deficiency in computation and memory, they are widely used in modern networks [9, 10]. Using smart cards to login networks is a solution for single sign-on problem. In order to protect smart card free from thieves to login, it requires PIN or passwords to resist smart card stolen-attack.

In this article, we propose two new single sign-on solutions that can resist guessing attack. The preliminaries and notations are presented in next section. The proposed smart-card-based mobile single sign-on system is described in Sect. 63.3. The password-based single sign-on is presented in Sect. 63.4. Finally, we make conclusions.

63.2 Preliminaries and Notations

A mobile single sign-on system denotes a user login network for service through a terminal at any place if it connects to the networks. In the mobile single sign-on systems, suppose that there are users, service providers and a trusted authentication center in the system. The service providers and authentication center have joined in the public-key infrastructure (PKI). The service providers offer users services through networks, and the authentication center is used to verify users. A user, even though he/she is not a PKI participant, can perform symmetric and asymmetric cryptographic computations and send confidential messages to the PKI users. The notations of the paper are as follows:

PW	a password
ID	the identity of user
U_i	a user
SP	service provider
Auc	authentication center
pub_A	public key of A
pri_A	private key of A
K	a session key with at least 128-bit-length
x	the secret of the authentication center
$(M)_K$	the message M is encrypted or decrypted with session key K by using symmetric cryptosystem
$\{M\}_{\text{pub\_}A}$	the message M is encrypted with A's public key by using asymmetric cryptosystem
$\{M\}_{\text{pri\_}A}$	the message M is signed with A's private key by using asymmetric cryptosystem
$A \rightarrow B{:}M$	A sends message M to B

63.3 The Smart-Card-Based Mobile Single Sign-On System

In this Section, we propose a smart-card-based mobile single sign-on system. The smart-card-based system includes registration and login phases. All users should register to the authentication center before they request for services. The center sends each user a smart card after registration. The registration phase and login phase are as follows.

63.3.1 Registration Phase

In the registration phase, the user forwards his identity and password to the authentication center (Auc), the Auc sends a smart card to the user if he/she is authenticated.

Step R.1. Ui → Auc: ID, PW

The user forwards his/her identity ID and password PW to the authentication center personally or through a secure channel.

Step R.2. Auc → Ui: smart card

The Auc verifies (ID, PW) to authenticate user. After the user is authenticated, Auc computes $S = h(\text{ID} \oplus x) \oplus \text{PW}$ and stores it along with a secure hash function $h(\cdot)$ into smart card. Then the center sends the smart card to the user.

63.3.2 Login phase

The login phase provides user to ask for services. When user logins the first service provider (SP), the service provider authenticates the user with the help of authentication center. Hereafter, the user can request services for different service providers.

63.3.2.1 User Login to the First Service Provider

When a user wants to login a service provider for services, he/she inserts smart cart to the cardholder. The steps of the login phase are as followings.

Step L.1. Ui → SP: ID, C1, C_2, T_1

The user inputs password to the smart card. The smart card generates a random number R with at least 128 bit-length, and computes $C_1 = S \oplus PW \oplus R$ and $C_2 = h(R, T_1)$, where T_1 is the timestamp. Note that $C_1 = S \oplus PW \oplus R$ $= h(\mathrm{ID} \oplus x) \oplus R$. Next, the user forwards (ID, C_1, C_2, T_1) to service provider.

Step L.2. SP → Auc: ID, C_1, C_2, T_1

Because secret x is unknown by the SP, the SP cannot authenticate the user with (ID, C_1, C_2, T_1). The SP forwards the message to the Auc.

Step L.3. Auc → SP: C_3

After receiving (ID, C_1, C_2, T_1), Auc checks the freshness of T1 and computes $R' = C_1 \oplus h(ID \oplus x)$. Auc can authenticate the user if $h(R', T_1) = C_2$. Next, Auc computes Token $= \{\mathrm{ID}, h(R), \mathrm{Assertion}, \mathrm{Auc}, T_2\}_{\mathrm{pri\_Auc}}$ and sends C_3 to the user, where $C_3 = \mathrm{Token}, h(R), ID, SP_{\mathrm{pubSP}}$ and T_2 is the new timestamp.

Note that other parameters such as serial number, date and version, can also be included in the Token. Since Token is signed by Auc, everyone even not a PKI user can verify Token by using the public-key of Auc.

Step L.4. SP → Ui: C_4

After receiving C_3, the service provider obtains $(\mathrm{Token}, h(R), \mathrm{ID}, \mathrm{SP})$ by using private key. After Token and $h(R)$ is verified, the service provider generates a partial session key K_1. Next, the service provider computes and forwards $C_4 =$ $(\mathrm{Token}, K_1, \mathrm{ID}, \mathrm{SP})_{h(R)}$ to the user.

Step L.5. Ui → SP: C_5

After receiving C_4, the user recovers $(\mathrm{Token}, K_1, \mathrm{ID}, \mathrm{SP})$ with $h(R)$. The user generates another partial session key K_2 after Token is verified. The user obtains the session key K by$K = K_1 \oplus K_2$. Then the user sends C_5 to the service provider, where $C_5 = (K_2, h(K))_{\mathrm{pub\_SP}}$.

After receiving C_5, the service provider decrypts it to obtain $(K_2, h(K))$. Next, the service provider obtains the session key K' by $K' = K_1 \oplus K_2$. Finally, the user will be authenticated if $h(K) = h(K')$. Hereafter, the service provider can provide service to the user securely through the session key K.

63.3.2.2 User Logins to Another Service Provider

If the user wants to login another service provider (SP') for service, the procedures are as follows:

Step L.6. Ui → SP': C_6

The user generates a new partial session key K_1' and sends C_6 to the service provider, where $C_6 = (\mathrm{Token}, K_1', h(R), \mathrm{ID}, \mathrm{SP}')_{\mathrm{pub\_SP}'}$.

Step L.7. SP' → Ui: C_7

After receiving C_6, the service provider recovers $(\mathrm{Token}, K_1', h(R), \mathrm{ID}, \mathrm{SP}')$ and verifies Token and $h(R)$. The service provider generates another partial session key K_2' and obtains session key by $K' = K_1' \oplus K_2'$. Next, SP computes and sends C_7 to the user, where $C_7 = (ID, SP', K_2', h(K'))_{h(R)}$.

After receiving C_7, the user decrypts it to recover $(\mathrm{ID}, \mathrm{SP'}, K_2', h(K'))$ with $h(R)$. The service provider is authenticated if $h(K') = h(K_1' \oplus K_2')$. Thereafter, the user and the service provider can communicate securely with session key K$'$.

63.3.3 Security Discussions

The proposed smart-card-based single sign-on system provides mutual authentication. It is very simple and has merits as follows:

Adversary cannot know secret x of the authentication center. Since smart card is a tamper-free device and $S = h(\mathrm{ID} \oplus x) \oplus \mathrm{PW}$, x cannot be revealed even if password is known. Moreover because C_1 and C_2 come from R, x, and T_1 with hash function, R and x have at least 128 bit-length, the adversary is infeasible to obtain x from C_1 and C_2. If adversary wants to obtain R from C_2 and then using R to find x, it is also infeasible due to the long bit-length of R and irreversible hash function.

Adversary cannot masquerade as a legal user to login the system. Since $C_1 = h(\mathrm{ID} \oplus x) \oplus R$ and $C_2 = h(R, T_1)$, without knowing x, the adversary cannot forge C_1 and C_2 consistently for a successful login. Similarly, due to $K = K_1 \oplus K_2$ and K_1 cannot be obtained from C_4 without knowing $h(R)$, the adversary cannot forge C_5 for authentication. Thus the adversary cannot login the system.

It can resist online password guessing attacks. If the adversary uses a stolen smart card to login, the probability for a successful login is $p = 2^{-|\mathrm{PW}|}$ which is quite low, where |PW| denotes bit-length of the password. If the attacker continues to guess the password, the server will reject the illegal attempt by limiting the number of fail trials. If the adversary has no smart card, as described in (2), he/she cannot forge C_1 and C_2 to login the system.

It can resist offline password guessing attacks and protect the smart card. In general, the offline guessing attack can be avoided if there are no enough messages for the attacker to verify his/her guessing. The attack falls into either of the following cases:

(a) The attacker has no smart card. As mentioned above, since R and x are unknown, an attacker cannot generate C_1 and C_2 consistently to masquerade a legal user. That is, the scheme can withstand off-line guessing attacks when attackers have no smart card.

(b) The attacker has stolen or somehow obtained a smart card. If an adversary obtains smart card, he/she tries to guess password offline. Because R is a random number changed on each login session, C_1 and C_2 are renewed on each session even if timestamp remains the same. If an attacker obtains the smart card and knows C_1 and C_2 at previous successful login session, he/she still cannot verify his/her guess because C_1 and C_2 are changed on each trial. That is the adversary cannot know whether his guessing is right or wrong. Thus our new scheme can resist offline password guessing attacks and protect the cardholder even if the smart card is lost.

It can resist replay attacks. Since both C_2 and Token include timestamps, the server can detect replay attacks if an attacker resubmits an intercepted message. The nonce can also be used to resist the replay attack.

It does not require password verification table. For an authentication system with verification table, a hacker usually attacks the system by obtaining or modifying the verification table. The proposed scheme can resist such attacks since there is no verification table at the server end.

It can resist service provider impersonate attack. In the system, Token is signed by Auc, the service provider, the user and the adversary cannot forge C_6 for a successful login. Moreover, since ID is included in the Token, the service provider cannot impersonate a legal user to login other service providers even though he knows user's Token.

63.4 The Passwords-Based Mobile Single Sign-On System

The proposed second single sign-on system is based on password. In the system, users need no smart cards and it provides users to login networks only by passwords. The password verification table is required in authentication center. The registration phase and login phase are described as follows.

63.4.1 Registration Phase

In this system, the user should also register to Auc beforehand. The user sends his/her ID and password to Auc for registration. After registration, the user and the authentication center share the common password PW.

63.4.2 Login Phase

Similar to the steps of the smart-card-based single sign-on system, the user logins the service provider when he/she asks for services, and thereby the service provider releases services after the authentication center verifies the user. The password- based single sign-on system is simple and no smart cards are required.

63.4.2.1 User Login to Some Service Provider

When user asks for services from a service provider, the login phase is as follows.

Step L.1. Ui → SP: C_1

The user inputs password PW, and generates a random number R. Next, he/she computes C_1 by $C_1 = \{\mathrm{ID}, \mathrm{SP}, \mathrm{PW},\ R,\ T_1\}_{\mathrm{pub\_Auc}}$.

Note that the random number R is updated on each login session and T_1 is the timestamp. Next, the user sends C_1 to the service provider.

Step L.2. SP → Auc: C_1

The service provider forwards the message C_1 to Auc.

Step L.3. Auc → SP: C_2

After receiving C_1, Auc decrypts it to obtain $(\mathrm{ID}, \mathrm{SP}, \mathrm{PW},\ R,\ T_1)$. Then Auc generates a Token and sends C_2 to the service provider if password PW is verified and $T_1$is fresh, where $C_2 = \{\ \mathrm{Token}, h(R), \mathrm{ID}, \mathrm{SP}\}_{\mathrm{pub\_SP}}$ and $\mathrm{Token} = \{ID, h(R), \mathrm{Assertion}, \mathrm{Auc}, T_2\}_{\mathrm{pri\_Auc}}$.

Step L.4. SP → U: C_3

After receiving C_2, the service provider will obtain $(\mathrm{Token}, h(R), \mathrm{ID}, \mathrm{SP})$. Then the service provider generates a partial session key K_1 after Token and $h(R)$ are verified. Next, the service provider computes and forwards $C_3 = (\mathrm{Token}, K_1, \mathrm{ID}, \mathrm{SP})_{h(R)}$ to the user.

Step L.5. Ui → SP: C_4

After receiving C_3, the user recovers $(\mathrm{Token}, K_1, \mathrm{ID}, \mathrm{SP})$ with $h(R)$. Then the user generates another partial session key K_2 after Token is verified. The user obtains the session key K by $K = K_1 \oplus K_2$. Then the user sends C_4 to the service provider, where $C_4 = (K_2, h(K))_{\mathrm{pub\_SP}}$.

After receiving C_4, the service provider obtains K_2. With the partial session key K_1 generated previously and the recovered partial keyK_2, the service provider will obtain session key K' by $K' = K_1 \oplus K_2$.

Finally, the user will be authenticated if $h(K) = h(K')$. Thereafter, the user and the service provider can communicate and provide services securely with session key K.

63.4.2.2 User Logins to another Service Provider

If the user asks for another service provider SP' for service, the procedures are the same as the smart-card-based system.

63.5 Security Discussions

Like the smart-card-based single sign-on system, the proposed password based mobile single sign-on system also has the advantages as follows:

It can resist online password guessing attacks. Since the message in Step L.1 and L.2 is C_1 which is $\{\,\mathrm{ID}, \mathrm{SP}, \mathrm{PW},\ R,\ T_1\}_{\mathrm{pub\_Auc}}$, the attacker should guess password PW to obtain C_1 for a successful login. The probability to guess password is only $p = 2^{-|\mathrm{PW}|}$. If an attacker continues to guess the password, the Auc will reject the illegal attempt by limiting the number of error trials.

It can resist offline password guessing attacks. Since R is renewed and thereby C_1 is changed in each login session, the attacker cannot verify his/her guess even he/she intercepts message at the previous login session. Thus our new scheme can resist offline password guessing attacks.

It can resist replay attacks. Since both C_1 and Token include timestamps, the server can detect replay attacks if an attacker resubmits an intercepted message.

It can resist service provider impersonate attack. Since ID and $h(R)$ are included in the Token, the service provider cannot impersonate a legal user to login other SP though he knows user‘s Token after a successful login.

No smart card is needed in the system. The proposed system only uses password for authentication, it is simple and low cost for single sign-on systems. Note that, for security reason, the stored password in the database is h(PW) instead of PW.

63.6 Conclusions

We have proposed two new single sign-on systems with the merits of mobility and resisting password guessing attack. One of the proposed systems is based on smart cards, and the second one only uses passwords to login. The smart-card-based system does not need password verification table and can resist smart stolen-attack, the password-based system is simple and cost is quite low. Both the systems are simple and can stand online and offline guessing attacks.

References

1. Mauro C, Sunyaev A, Leimeister JM, Schweiger A, Krcmar H (2008) A proposed solution for managing doctor’s smart cards in Hospitals using a single sign-on central architecture. In: Proceedings of the 41st Hawaii International Conference on Systems Science
2. Pashalidis A, Mitchell C. (2003) Using GSM/UMTS for Single Sign-On, Joint First Workshop on Mobile Future and Symposium on Trends in Communications 2003, SympoTIC ‘03, pp. 138–145
3. Peyravian M, Zunic N (2000) Methods for Protecting Password Transmission. Comput Secur 19(5):466–469

4.. Zhao G, Zheng D, Chen K. (2004) Design of Single Sign-On, IEEE International Conference on E- Commerce Technology for Dynamic E- Business (CEC-East'04), pp 253–256
5. Mitchell CJ, Pashalidis A (2003) A Taxonomy of Single Sign-On Systems, ACISP 2003. LNCS 2727:249–264
6. Chien HY, Jan JK, Tseng YM (2002) An Efficient Solution to Remote Authentication: Smart Card. Comput Secur 21(4):372–375
7. Hwang MS, Lee CC, Tang YL (2002) Simple Remote User Authentication Scheme. Math Comput Model 36:103–107
8. Kwon T, Song J (1998) Efficient and Secure Password-Based Authentication Protocols against Guessing Attack. Comput Commun 21:853–861
9. Sun HM (2000) An Efficient Remote User Authentication Scheme Using Smart Cards. IEEE Trans on Consumer Electron 46(4):958–961
10. Yang WH, Shieh SP (1999) Password Authentication Schemes with Smart Card. Comp Secur 18(8):727–733

Chapter 64
The Implementation of Face Detection Algorithm AdaBoost Based in the Embedded System

Wenxian Xiao, Zhen Liu, Ning Li and Wenlong Wan

Abstract To solve the real-time problem of face detection, for the realization bottleneck of AdaBoost pure software algorithm, FPGA-based hardware acceleration platform strategy is proposed, using pipeline processing technology to achieve rapid calculation of integral image. Using PowerPC 405 processor VIrtex TM II Pro platform FPGA in the experiment, under the conditions of the size of the input image 352 × 288 pixels, detection speed reaches 50 frames per second; detection rate was 98%, false detection rate of about 1%, and thus achieving the requirement of real-time face detection.

Keywords Detection · Algorithm · Hardware acceleration

64.1 Introduction

Face detection is defined as the process of the identification of all of the face region in an image, regardless of the location of the face, direction and background conditions. Face detection technology can be used in face recognition, video conferencing, image and video retrieval, intelligent human–computer interaction and other fields. With the development of embedded technology and smart device, in some special applications fields [1], the requirements of mobilization and outdoor work make the needs of embedded face detection technology more and more urgent [2].

W. Xiao (✉) · Z. Liu · N. Li · W. Wan
Henan Institute of Science and Technology,
Xinxiang 453003, Henan, China
e-mail: Xwenx@yeah.net

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_64,

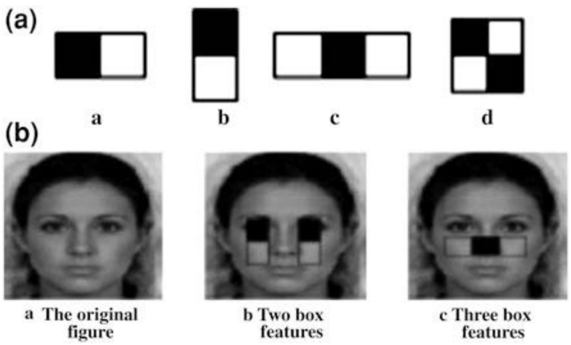
Measures the performance of face detection are two key indicators, namely, accuracy and speed. By the Freund and Scrapire AdaBoost proposed target detection algorithm in the precision and speed has reached a higher level. P. Viola, etc. AdaBoost face detection algorithm is applied and to achieve a true real-time detection, which makes real-time embedded platform, face detection possible. There is already research and practice in this area, but the research work, mostly in software, however, AdaBoost algorithm is a computational load and data throughput of the algorithm are large, pure software implementation of the access points chart the bottleneck of the algorithm, therefore, has to implement the detection algorithm embedded platform, the ability of the processor alone can not achieve real-time requirements, which need to find hardware acceleration AdaBoost algorithm approach [3]. Theo arid proposes a framework for array-based algorithm, but the structure of hardware resources consuming too much, there are certain difficulties to implement. In this chapter, based on its design improved algorithms for software implementation bottlenecks, hardware-accelerated method to achieve the square and the integral image and integration of real-time calculation and classification features of value to achieve a deep pipeline. Power architecture using processors in IBM PowerPC 405 VirtexTM II Pro implemented on FPGA platform [4].

64.1.1 AdaBoost Face Detection Algorithm Principle

AdaBoost basic idea of the algorithm is to use the classification capability of a large number of simple and general methods of classification by some together [5], and form a very strong classification ability classifier, then a number of strong classifier hierarchical classification series completes image search detection. According to AdaBoost algorithm, an image is classified as a sub-regional human face region if and only if the sub-region through the algorithm, all of the cascade classifier; and a sub-region through a classification only when the promoter region of the classifier on all the characteristics of the Eigen values is greater than the threshold of the classifier. In this chapter, based on Haar-like AdaBoost features face detection algorithm.

64.1.2 Harr-Like Characteristics and Points of Calculation

Characteristics of the so-called Harr-like structure is composed of a series of rectangles. This feature representation in the form of simple, fast calculation in favor, while in various forms. Now many researches have been established based on the positive characteristics of Harr-like face detection system, in Fig. 64.1a shows the four basic Haar-like rectangular features, the white areas and black areas, and the difference between the pixel gray characteristics for the Harr-like the



a The original figure

b Two box features

c Three box features

Fig. 64.1 Harr-like characteristics. **a** Basic Harr-like features, **b** the typical features of face in Harr-like

characteristics of value [6]. Figure 64.1b shows typical facial features of Haar, application of these typical characteristics of the face can be effectively separated from the non-human face region.

Integral map is a transformation of the original image, which is defined as follows:

$$\mathrm{ii(x,y)} = \sum_{\mathrm{x'<x,y'<y}} \mathrm{i(x',y')}. \tag{64.1}$$

Where is the integral figure in the point of the integral value, the image gray values at the point? Visualization of the calculation process is shown in Fig. 64.2. Figure 64.2 points the black dot in the figure a point [7], its value is equal to the original image from the upper left corner to the point (not including the black spots where the rows and columns) of all pixels and, in the gray area of figure gray and. Figure and calculations indicate points in Fig. 64.2.

Using the points chart, any rectangle feature regional and use of the four reference points can be calculated. Figure 64.2 in the rectangular area ABCD pixel gray value and can be A, B, C, D four reference points

$$\mathrm{sum} = \mathrm{ii(C)} + \mathrm{ii(A)ii(B)}\ \mathrm{ii(D)} \tag{64.2}$$

which, are the gray pixels of the area of ABCD.

64.2 Adaboost Algorithm for Bottleneck Analysis

According to the algorithm theory, integral graphs and data classification algorithm for the implementation of critical data. The face detection involves four layers from the inside to the outside loop: traverse the classification characteristics of each Harr and calculate the characteristic values of the cascade classifier ∀ traverse each level through each window classifier ∀ detection window progressive amplification. One of the central part of the innermost of the two cycles (execution

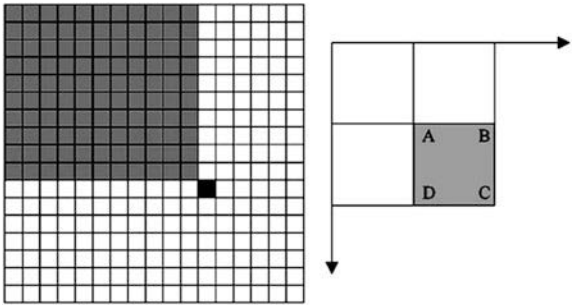


Fig. 64.2 Diagrams and calculation of plot points indicate

time mainly by the classifier Harr number of features in the detection window and progression through the classification decision), computing capacity, memory access frequency, it is difficult to achieve real-time requirements, therefore, Harr characteristics and the integral method of calculation is the bottleneck in a software implementation, through testing, the calculation of integral figure accounted for about 40% of the detection time [8], and that through algorithm analysis, the calculation of the relative integral map more independent, can be used as a separate module to design the hardware acceleration.

64.3 Hardware Implementation of Face Detection

64.3.1 System Structure

System structure shown in Fig. 64.3. First, FPGA grayscale image under test into the SRAM 1, at the same time, integration plans and square and have been calculated in real time and stored in SRAM 2, the classifier ROM (CROM) device data storage cascade. Then, FPGA using the integral image and the square and through the classifier to detect face detection results and write SRAM 3, then the interface chip through the PCI have agreed to use the test results output. SRAM 1, SRAM 2 have the same structure, under the control of the frame sync signal interchangeable functions, equivalent to a PINGPONG cache structure, thus forming a top two lines, which greatly accelerated the speed of memory access.

64.3.2 Hardware-Accelerated Computing Integral Images

For compute-intensive and high parallel application acceleration hardware acceleration is usually a better solution for complex algorithms can obtain better performance [9]. This design uses an array of cell-based hardware architecture. To make the implementation of the algorithm on Haar process parameters without

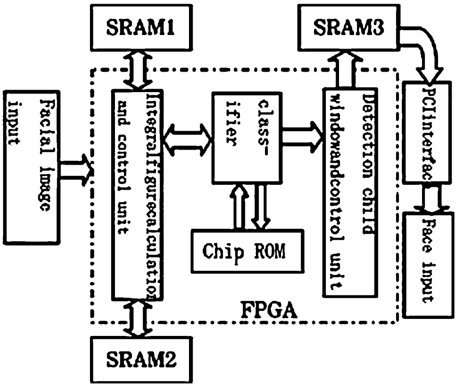

SRAM1
SRAM3
Facial image input
Integralfigurecalculation and control unit
class-ifier
Detection child windowandcontrol unit
PCIInterfac
Chip ROM
Face input
FPGA
SRAM2

Fig. 64.3 System structure

the need to make any adjustments (you can use the ROM stored data for hardware used classification), using the same training samples when the classification images, the fixed size of 24 × 24 array unit, no longer used to expand the window of detection methods, but by changing (reducing) the size of the image is detected to detect, disguise window to expand testing to achieve the image at any fixed size face detection window, while avoiding large calculations image points and, so greatly reduced the scale of integration plans, thereby reducing the hardware scale and simplifying the hardware functions.

Array structure, 24 × 24 units composed of two-dimensional array of interconnected and form to the right lateral movement of the vertical lines and lines of upward mobility.

Array structure, 24 × 24 units composed of two-dimensional array of interconnected and form to the right lateral movement of the vertical lines and lines of upward mobility. Array structure, 24 × 24 units composed of two-dimensional array of interconnected and form to the right lateral movement of the vertical lines and lines of upward mobility.

Haar horizontal lines used to calculate the characteristic value; the line integral logic input data for horizontal integration, streamlined and gray rectangles to complete the calculation is complete

$$\mathrm{IP}_i = \sum_{j=1}^{i} \mathrm{p}_j (1 \le i \le 24).$$

Vertical lines in the input data flow process to complete the vertical integration, integration of data will be sent after the vertical lines accumulated and were passed, used to achieve rapid calculation of integral graphs. Principle of vertical integration calculation shown in Fig. 64.4.

Each circle in Fig. 64.4 represents a pixel, one pixel vertical integration is the top of the column with the same gray value of all the pixels and, for example, point C3 (i.e. a point) of the vertical integration is C3 + C2 + C1; point F6

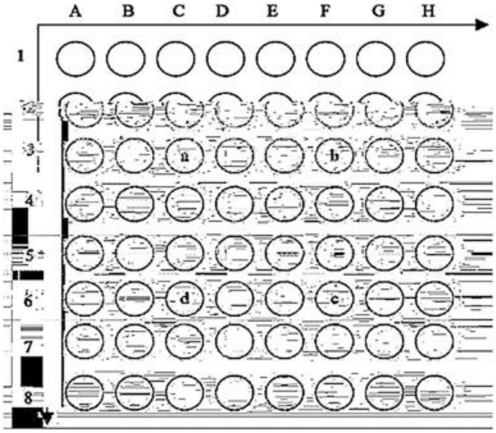


Fig. 64.4 Calculation of vertical integration

(that c points) of vertical integration is F6 + F5 + F4 + F3 + F2 + F1, so the rectangle ABCD, the integral and the (C6–C2) + (D6–D2) + (E6–E2) + (F6–F2).

In theory, the calculation of an integral figure to be seized of the window that is a gray rectangle, and at least 24 cycles, but two-way lines, each cycle can have a rectangular array of parameters from the left to center, but from the right by Integral map information. Therefore, once the pipeline fills the average gray level of a rectangle and only one cycle, a pure software approach to solve the bottleneck problem. This is the hardware accelerator that can greatly accelerate the process of face detection being the key reason.

64.3.3 Loading and Testing Process Optimization

When loading the basic unit of behavior, each window into 24 × 24 pixel array units, and complete the integration plans and square, and the calculations. The order of a vertical scan window scan, the y coordinate of priority, the advantages of doing the same column with two different consecutive one-line window, and the remaining 23 rows of data are identical, can be shared, just read a line of new data that may be, that is, read the review window to be compared with the previous window, the line increment, and placed FIFO, thus greatly reducing the frequency of the image SRAM access times. Because the data SRAM, and ROM width is 64 bits, while the line formed by the 24 bytes, so a row of data read at least 3–4 times SRAM access. In order to avoid acceleration module delay to load images as pause, in the SRAM or ROM, and array elements as a buffer between the both FIFO. This ensures speed Module in 24 cycles without pause to read the entire window of 24 lines. When the module starts to accelerate post-test, test sub-window control module to read the next few lines into the FIFO in order to reduce the data bandwidth is limited and result in a short time delay to read large

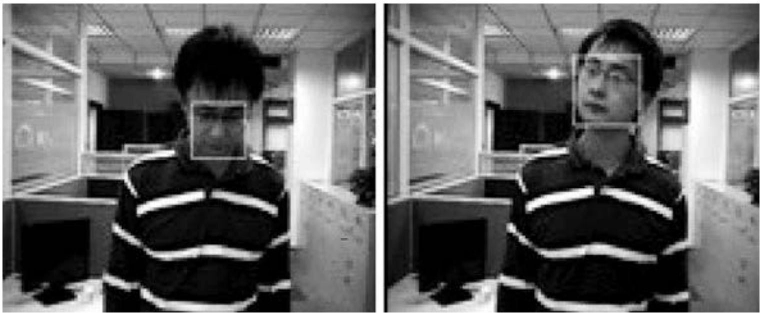


Fig. 64.5 Face detection results

amounts of data. In the detection phase, the accelerator is read from the CROM cascade classifier data, and placed in the horizontal line array unit constituted in a rectangular gray and fast calculation speed up the detection process to achieve the purpose.

64.4 Analysis of Experimental Platform and Results

The system uses PowerPC405 processor VIrtex TMII Pro platform FPGA implementation of face detection, PowerPC405 300 MHz processor embedded Harvard (Harvard) RISC nuclear structure, with five data channels pipeline, hardware multiply and divide unit, 32 32-bit general registers, memory management unit (MMU) and a dedicated on-chip memory interface [10].

In this study, Open CV the AdaBoost fixed-point processing algorithms ported to embedded platforms, the classifier using the Open CV library provides the training data, a total of 25, containing 2913 Haar features. Tested under the 300 MHz clock detect a human face 352 × 288 image requires only 20 ms (Paul VIolaetal PIII700 MHz of the PC on the need to 67 ms), because the system algorithm and PC platforms using exactly the same algorithm, so the test results are correct and exactly the same PC platform, but the detection rate significantly improved, reaching 50 fps performance; another face with a video containing 360 individuals tested, were detected effective face 353, of which non-human face 3, detection rate was 0.98, false detection rate of about 1% to the real-time face detection. Test results shown in Fig. 64.5.

In order to verify the image points and computing hardware-accelerated effects, the AdaBoost face detection algorithm is a pure software implementation and collaboration with hardware acceleration hardware and software programs to compare the detection results, the two programs included 14 test positive with a human face 352 × 288 pictures of the performance test data as shown in Table 64.1.

Table 64.1 Data comparing the performance test program Case detection time side

Scheme	Testing time (cycle)	Detection accelerating rate	Other time (cycle)	Total time (cycle)	General acceleration ratio
Pure software	823,950,935	1	21,942,756	845,893,691	1
Hardware acceleration	60,265,450	13.67	19,325, 351	79,590,801	10.63

Test results can be seen, compared to pure software, hardware accelerators can reach more than 13 times speedup and greatly improve the detection performance of the system, and mainly in the detection of the acceleration part to part, or other parts of the software, so there is no too much time on the increase.

64.5 Conclusion

In this chapter, aiming to the implementation bottlenecks of AdaBoost face detection algorithm software, in the embedded platform by leveraging the advantages of parallel processing hardware, reducing hardware size; to grayscale and calculate in the form of flowing water to improve the detection rate; full account access memory access bandwidth and storage optimization order to enhance the system's processing performance, etc., and ultimately realize a high precision real-time face detection.

References

1. Podlubny I (1999) Fractional differential equations [M]. Academic, New York
2. Hartley TT, Lorenzo CF, Qammer HK (1995) Chaos in a fractional order Chua's system. IEEE Trans [Z] CAS-I 42:485–490
3. Petrá I (2006) Method for simulation of the fractional order chaoticsystems [J]. ActaMontanisticaSlovaca 11(4):273–277
4. Ahmad WM, Sprott JC (2003) Chaos in fractional-order autonomous nonlinear systems [Z]. Chaos, Solitons Fractals 16:339–351
5. Lvjin H, Jun-An L, Chen S (2002) Chaotic time series analysis and its applications [M]. Wuhan University Press, Wuhan
6. Chen G, Lvjin H (2003) Lorenz system family dynamics analysis, control and synchronization of [M]. Science Press, Beijing
7. Chen G, Yu X (2003) Chaos control: theory and applications [M]. Springer, Berlin
8. Tavazoei MS, Haeri M (2007) Unreliability of frequency-domain approximation in recognising chaos in fractional-order systems [J]. IET Sig Process 1(4):171–181
9. Tavazoei MS, Haeri M (2007) A necessary condition for doublescroll attractor existence in fractional-order systems [J]. Phys Lett A 367 (1–2):102–113
10. Tavazoei MS, Haeri M (2008) Chaos control via a simple fractional order controller [Z]. Phys Lett A 372:798–807

Chapter 65
An Enhanced Authenticated 3-round Identity-Based Group Key Agreement Protocol

Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

Abstract In 2008, Gang Yao et al. proposed an authenticated 3-round identity-based group key agreement protocol, which is based on Burmester and Desmedt's protocol proposed at Eurocrypt 94. However, their protocol can only prevent passive attack. If active attack is allowed, the protocol is vulnerable and an internal attacker can forge her neighbor's keying material. It is obvious that the protocol does not achieve the aim of authentication. In this chapter, we propose an enhanced provably secure protocol based on their protocol. Finally, we make a detailed security analysis of our enhanced authenticated identity-based group key agreement protocol.

Keywords Authentication · Identity-based · Key agreement · Bilinear pairing · Cryptanalysis · Attack

65.1 Introduction

Secure and reliable communications [1] have become critical in modern society. Centralized services such as file sharing, can be changed into distributed or collaborated systems based on multiple systems and networks. Basic cryptographic

W. Yuan · L. Hu · H. Li · J. Chu (✉)
Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_65,

functions such as data confidentiality, data integrity, and identity authentication are required to construct these secure systems.

Key agreement protocol [2–4] allows two or more participants, each of whom has a long-term key, respectively, to exchange information over a public communication channel with each other. However, the participants cannot ensure others' identity. Though Alice wants to consult a session key with Bob, Alice cannot distinguish it if Eve pretends that she is Bob. The authenticated key agreement protocol overcomes this flaw and makes unfamiliar participants to ensure others' identities and consult a common session key in the public channel.

Shamir [5] introduced an identity-based public key cryptosystem in 1984, in which a user's public key can be calculated from his identity and defined as hash function, while the user's private key can be calculated by a trusted party called Private Key Generator (PKG). The identity-based public key cryptosystem simplifies the program of key management and increases the efficiency. In 2001, Boneh and Franklin [6] found bilinear pairing positive applications in cryptography and proposed the first practical identity-based encryption protocol with bilinear pairings. Soon, the bilinear pairings became important tools in constructing identity-based protocols and a number of identity-based encryption or signature schemes [7–12] and authenticated key agreement protocols [13–17] were proposed.

In 2008, Gang et al. [18] proposed an authenticated three-round identity-based group key agreement protocol. The first round is for identity authentication, the second round is for key agreement, and the third round is for key confirmation. Their protocol is based on the protocol of Burmester and Desmedt [19] which was proposed at Eurocrypt 94. They declared the proposed protocol provably secure in the random oracle model.

In this chapter, we show that an authenticated 3-round identity-based group key agreement protocol proposed by Gang Yao et al. is vulnerable: an internal attacker can forge her neighbors' keying material. Then we propose an improved provably secure protocol based on Burmester and Desmedt's as well. Finally, we summarize several security attributes of our improved authenticated identity-based group key agreement protocol.

65.2 Paper Preparation

65.2.1 Bilinear Pairing

Let P denote a generator of G_1, where G_1 is an additive group of large order q and let G_2 be a multiplicative group with $|G_1| = |G_2|$. A bilinear pairing is a map $e : G_1 \times G_1 \rightarrow G_2$ which has the following properties:

1. *Bilinearity*. Given $Q, W, Z \in G_1$, $e(Q, W + Z) = e(Q, W) \cdot e(Q, Z)$ and $e(Q + W, Z) = e(Q, Z) \cdot e(W, Z)$. There, for any $q, b \in Z_q : e(aQ, bW) = e(Q, W)^{ab}$ $= e(abQ, W) = e(Q, abW) = e(bQ, W)^{a}$.
2. *Non-degenerative*. $e(P, P) \neq 1$, where 1 is the identity element of G_2.
3. *Computable*, If $Q, W \in G_1$, one can compute $e(Q, W) \in G_2$ in polynomial time efficiently.

65.2.2 Computational Problems

Let G_1 and G_2 be two groups of prime order q, let $e : G_1 \times G_1 \to G_2$ be a bilinear pairing and let P be a generator of G_1.

- *Discrete logarithm problem* (*DLP*). Given $P, Q \in G_1$, find $n \in Z_q$ such that $P = nQ$ whenever such n exists.
- *Computational Diffie–Hellman problem* (*CDHP*). Given$(P, aP, bP) \in G_1$ for $a, b \in Z_q^*$, find the element abP.
- *Bilinear Diffie–Hellman problem* (*BDHP*). Given$(P, xP, yP, zP) \in G_1$ for $x, y, z \in Z_q^*$, compute $e(P, P)^{xyz} \in G_2$

65.2.3 Introduction of BR Security Model

To describe the security model for entity authentication and key agreement aims, Bellare and Rogaway proposed the BR93 model [13] for a two-party authenticated key agreement protocol in 1993 and the BR95 model [14] for three-party authenticated key agreement protocol in 1995. In BR model, the adversary can control the communication channel and interact with a set of $\Pi^{i}_{U_x,U_y}$ oracles, which specify the behavior between the honest players U_x and U_y in their it instantiation. The predefined oracle queries are described informally as follows:

- *Send* (U_x, U_y, *i, m*). The adversary sends message m to the oracle $\Pi^{i}_{U_x,U_y}$. The oracle $\Pi^{i}_{U_x,U_y}$ will return the session key if the conversation has been accepted by U_x and U_y or terminate and tell the adversary.
- *Reveal* (U_x, U_y, *i*). It allows the adversary to expose an old session key that has been previously accepted. After receiving this query, $\Pi^{i}_{U_x,U_y}$will send this session key to the adversary, if it has accepted and holds some session key.
- *Corrupt* (U_x, K). The adversary corrupts U_x and learns all the internal state of U_x. The corrupt query also allows the adversary to overwrite the long-term key of corrupted principal with any other value K.
- *Test* (U_x, U_y,*i*). It is the only oracle query that does not correspond to any of the adversary's abilities. If $\Pi^{i}_{U_x,U_y}$ has accepted with some session key and is being

asked a Test (U_x, U_y, i) query, then depending on a randomly chosen bit b, the adversary is given either the actual session key or a session key drawn randomly from the session key distribution.

Freshness. The notion is used to identify the session keys about which adversary should not know anything because she has not revealed any oracles that have accepted the key and has not corrupted any principals knowing the key. Oracle $\Pi^i_{A,B}$ is fresh at the end of execution, if, and only if, oracle $\Pi^i_{A,B}$ has accepted with or without a partner oracle $\Pi^i_{B,A}$, both oracle $\Pi^i_{A,B}$ and its partner oracle $\Pi^i_{B,A}$have not been sent a Reveal query, and the principals A and B of oracles $\Pi^i_{A,B}$ and $\Pi^i_{B,A}$ (if such a partner exists) have not been sent a Corrupt query.

Security is defined using the game *G*, played between a malicious adversary and a collection lection of $\Pi^i_{U_x,U_y}$ oracles and instances. The adversary runs the game simulation *G*, whose setting is as follows.

Phase 1. Adversary is able to send any Send, Reveal, and Corrupt oracle queries at will in the game simulation *G*.

Phase 2. At some point during *G*, adversary will choose a fresh session on which sxit is to be tested and send a Test query to the fresh oracle associated with the test session. Note that the test session chosen must be fresh. Depending on a randomly chosen bit *b*, adversary is given either the actual session key or a session key drawn randomly from the session key distribution.

Phase 3. Adversary continues making any Send, Reveal, and Corrupt oracle queries of its choice.

Finally, adversary terminates the game simulation and outputs a bit *b*', which is its guess of the value of *b*. Success of adversary in *G* is measured in terms of adversary's advantage in distinguishing whether adversary receives the real key or a random value. A wins if, after asking a Test (U_x, U_y, *i*) query, where $\Pi^i_{U_x,U_y}$is fresh and has accepted, adversary's guess bit *b*' equals the bit *b* selected during the Test (U_x, U_y, *i*) query.

A protocol is secure in the BR model if both the validity and indistinguishability requirements are satisfied:

- *Validity*. When the protocol is run between two oracles in the absence of a malicious adversary, the two oracles accept the same key.
- *Indistinguishability*. For all probabilistic, polynomial-time (PPT) adversaries A, AdvA (k) is negligible.

65.3 Improvement of Gang Yao et al.'s Protocol

In this section, we first review Bermester and Desmedt's group key exchange protocol. Then we propose a non-authentication protocol based on their protocol with bilinear pairing. Finally, we improve the non-authentication group key agreement protocol to an authentication group key agreement protocol.

65.3.1 Bermester and Desmedt's Group Key Exchange Protocol

Let$n$ be the size of the group, the Bermester and Desmedt's group key exchange protocol works as follows:

- Each participant U_i chooses a random number x_i and broadcasts $z_i = g^{x_i}$;
- Each participant computes $Z_i = z_{i-1}^{x_i}$ and $Z_{i+1} = z_i^{x_{i+1}} = z_{i+1}^{x_i}$, and broadcasts $X_i = {}^{Z_{i+1}}/_{Z_i}$;
- Each participant computes his session key as $K_i = Z_i^n X_i^{n-1} X_{i+1}^{n-2} \ldots X_{i+n-2}$. It is easy to see that each $U_i$can compute the same session key $K_i = \sum_{j=1}^{n} Z_j = g^{x_1x_2+x_2x_3+\cdots+x_nx_1}$

65.3.2 Non-Authentication Protocol Transformed from Bermester and Desmedt's Protocol

G_1 and G_2 are two cyclic groups of order q for some large prime q. $G_1$is a cyclic additive group and G_2 is a cyclic multiplicative group. Let P be an arbitrary generator of G_1, $e : G_1 \times G_1 \rightarrow G_2$ be a bilinear pairing and n be the size of the group, the non-authentication protocol works as follows:

- Each user U_i chooses a random number $r_i \in Z_q^*$ and broadcasts $z_i = r_iP$
- Each user U_i computes $Z_i = r_iz_{i-1}, Z_{i+1} = r_iz_{i+1}$, and broadcasts $X_i = Z_{i+1} - Z_i$
- Each player $U_i$can computes his session key as:

$$K_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$$

It is easy to see that for each U_i, $K_i = \sum_{j=1}^{n} Z_j = (r_1r_2 + r_2r_3 + \cdots + r_nr_1)P$

65.3.3 Our Authenticated Identity-Based Group Key Agreement Protocol

Let $U_1, \ldots, U_n$ be n participants, and PKG be the private key generator. Let ID$_i$ be the identity of U_i. Suppose that G_1 and G_2 are two cyclic groups of order q for some large prime q. $G_1$is a cyclic additive group and G_2 is a cyclic multiplicative group. Let P be an arbitrary generator of G_1, and $e : G_1 \times G_1 \rightarrow G_2$ be a bilinear pairing.

Our protocol is described as follows:

Setup. The PKG chooses a random numbers $s \in Z_q^*$, sets $R = sP$, chooses two hash functions, H_0 and H, where $H_0 : \{0,1\}^* \rightarrow G_1^*$. Then the PKG publishes system parameters$\{q, G_1, G_2, e, P, R, H_0, H\}$, and keeps the master key s as a secret.

Extract. Given a public identity ID $\in \{0,1\}^*$, the PKG computes the public key $Q = H_0(\mathrm{ID}) \in G_1$and generates the associated private key $S = sQ$. The PKG outputs S as the private key to the user via some secure channel. Let n users $U_1, \ldots, U_n$ with respective public key $Q_i = H_0(\mathrm{ID}_i)(1 \le i \le n)$ decide to agree upon a common secret key. $S_i = sQ_i$ is the long term secret key of U_i sent by the PKG on submitting U_i's public identity $(1 \le i \le n)$. Let U denote $U_1 || \cdots || U_n$.

We assume that U_1 is the protocol initiator. The protocol may be performed in three rounds as follows:

Round 1. Each participant U_i chooses a random number $r_i \in Z_q^*$, computes $z_i = r_i P, B_i = H(\mathrm{ID}_i, z_i), v_i = B_i S_i$ and broadcasts$(\mathrm{ID}_i, z_i, v_i)$.

Round 2. After receiving each $(\mathrm{ID}_i, z_i, v_i)(1 \le i \le n)$, each user can compute

$$B_i = H(\mathrm{ID}_i, z_i), Q_i = H_0(\mathrm{ID}_i)(1 \le i \le n)$$

and verify whether the equation

$$e(v_i, P) \overset{?}{=} e(B_i Q_i, R)$$

sets or not. If the equation sets, U_i can ensure that $(\mathrm{ID}_i, z_i, v_i)$is not modified or forged by attackers. Then he computes $Z_i = r_i z_{i-1}, Z_{i+1} = r_i z_{i+1}, X_i = Z_{i+1} - Z_i$, $C_i = H(\mathrm{ID}_i, X_i), w_i = C_i S_i$ and broadcasts$(\mathrm{ID}_i, X_i, w_i)$.

Round 3. After receiving each$(\mathrm{ID}_i, X_i, w_i)(1 \le i \le n)$, each user can compute

$$C_i = H(\mathrm{ID}_i, X_i), Q_i = H_0(\mathrm{ID}_i)(1 \le i \le n)$$

and verify whether the equation

$$e(w_i, P) \overset{?}{=} e(C_i Q_i, R)$$

sets or not. If the equation sets, U_i can ensure that $(\mathrm{ID}_i, X_i, w_i)$ is not modified or forged by attackers. Then he computes the keying material

$$D_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$$

Actually,

$$\begin{aligned} D_i &= nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2} \\ &= \sum_{j=1}^{n} Z_j = (r_1 r_2 + r_2 r_3 + \cdots + r_n r_1)P \end{aligned}$$

Then each user computes the session key as

$$K_i = H(U || z_1 || \cdots || z_n || X_1 || \cdots || X_n || D_i)$$

65.4 Security Analysis of our Protocol

Theorem 4.1 *Any modification can be found by the short signature if the hash function H is collision resistance.*

Proof In the function of $e(v_i, P) \overset{?}{=} e(B_iQ_i, R)$, the parameters P and R are public, which cannot be forged or modified, and $B_i = H(\mathrm{ID}_i, z_i)$,$Q_i = H_0(\mathrm{ID}_i)$are computed by the receiver. Though v_i and z_i may be modified by the attacker, the collision resistance hash function H will make it impossible to gain suitable pairs of v_i and z_i to pass the verification function. So if the attacker modifies any elements of $(\mathrm{ID}_i, z_i, v_i)$, other users can find it. The function $e(w_i, P) \overset{?}{=} e(C_iQ_i, R)$ has a similar situation with $e(v_i, P) \overset{?}{=} e(B_iQ_i, R)$. That is why any modification can be found by the short signature.

Theorem 4.2 *The attacker cannot obtain the session key from the intermediate messages if CDH problem is hard.*

Proof Suppose the challenger C wants to solve the CDH problem. That is, given(aP, bP), C should compute abP. In our protocol, the intermediate messages transmitted in the public channel are$(\mathrm{ID}_i, z_i, v_i)$ in the first round and $(\mathrm{ID}_i, X_i, w_i)$ in the second round. The efficient elements are z_i and X_i, and other elements are used to protect them. Supposed that the attacker can obtain the session key $K_i = H(U||z_1||\ldots||z_n||X_1||\ldots||X_n||D_i)$. That is, she can obtain the keying material D_i. For $D_i = nZ_i + (n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}$, she can obtain Z_i according to the equation $Z_i = {D_i - [(n-1)X_i + (n-2)X_{i+1} + \cdots + X_{i+n-2}]}/_{n}$, where X_i and n had been obtained by the attacker. As it is known to us, $z_i = r_iP$ and $Z_i = r_iz_{i-1} = r_ir_{i-1}P$. Define $z_i = aP$ and $z_{i-1} = bP$, which is given to the attacker. If she can obtain the session key, she can compute $Z_i = abP$ and she solves the CDH problem. That is to say, if CDH problem is hard, the attacker cannot obtain the session key.

65.5 Conclusions

In this chapter, we propose an improved provably secure protocol based on Burmester and Desmedt's protocol as well. Finally, we summarize some security attributes of our improved authenticated identity-based group key agreement protocol.

Acknowledgments The authors thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Kulkarni SS, Bruhadeshwar B (2010) Key-update distribution in secure group communication. Comput Commun 33:689–705
2. Diffie W, Hellman ME (1976) New directions in cryptography. IEEE Trans Inf Theory 22:644–654
3. Diffie W (1988) The first ten years of public-key cryptograph. Proc IEEE 76(5):560–577
4. Zhao J, Gu D, Li Y (2010) An efficient fault-tolerant group key agreement protocol. Comput Commun 33:890–895
5. Shamir A (1984) Identity-based cryptosystems and signature schemes. Advances in cryptology, CRYPTO'84, LNCS 196. Springer, Berlin, pp 47–53
6. Boneh D, Franklin M (2001) Identity-based encryption from the Weil pairing, advances in cryptology, CRYPTO'2001, LNCS 2139. Springer, Berlin, pp 213–229
7. Chin J-J, Heng S-H, Goi B-M (2008) An efficient and provable secure identity-based identification scheme in the standard model, LNCS 5057, Springer, Berlin, pp 60–73
8. Liu Z, Hu Y, Zhang X, Ma H (2010) Certificateless signcryption scheme in the standard model. Inf Sci 180:452–464
9. Zhang J, Yang Y, Niu X, Gao S, Chen H, Geng Q (2009) An improved secure identity-based on-line/off-line signature scheme, ISA 2009, LNCS 5576, Springer, Berlin, pp 588–597
10. Chang T-Y (2009) An ID-based group-oriented decryption scheme secure against adaptive chosen-ciphertext attacks. Comput Commun 32:1829–1836
11. Kiayias A, Zhou H-S (2007) Hidden identity-based signatures, LNCS 4886, Springer, Berlin, pp 134–147
12. Li C-T (2010) On the security enhancement of an efficient and secure event signature protocol for P2P MMOGs, ICCSA, LNCS 6016, pp 599–609
13. Lu R, Cao Z (2005) A new deniable authentication protocol from bilinear pairings. Appl Math Comput 168:954–961
14. Lu R, Cao Z, Wang S, Bao H (2007) A new ID-based deniable authentication protocol. Informatics 18(1):67–78
15. Cao T, Lin D, Xue R (2005) An efficient ID-based deniable authentication protocol from pairings, AINA'05, pp 388–391
16. Chou JS, Chen YL, Huang JC (2006) An ID-based deniable authentication protocol on pairings, cryptology ePrint archive: report (335)
17. Hwang JY, Choi KY, Lee DH (2008) Security weakness in an authenticated group key agreement protocol in two rounds. Comput Commun 31:3719–3724
18. Yao G, Wang H, Jiang Q (2008) An authenticated 3-round identity-based group key agreement protocol, the third international conference on availability, reliability, and security. ACM 2008, pp 538–543
19. Burmester M, Desmedt Y (1994) A Secure and Efficient Conference Key Distribution System, EUROCRYPT'94, LNCS 950. Springer, Berlin, pp 275–286

Chapter 66
Modal Identification of Crane Structure Based on Output-Only Data

Weiguo Zhang, Huiqing Qiu, Kailiang Lu, Zhiyong Hao and Yuan Liu

Abstract In this chapter, the Balance realization (BR) method is employed to identify the modal parameters of a crane structure from the output-only data obtained from the FEM simulation with Newmark method. Comparison and validation studies show that the BR method can extract the modal parameters of the crane structure accurately from in-operation output-only data.

Keywords Stationary random · Modal identification · Output-only data · Crane structure

66.1 Introduction

Most of the industrial structures such as bridge and crane are very large in size. It is difficult to carry out artificial excitation, and in most cases, only the response of the structures can be measured while the actual loading conditions are unknown. Therefore, it is necessary for the system identification process to be conducted based on the output-only data from the system itself. Over the past decades, several modal parameter identification techniques have been proposed to study modal parameter extraction from output-only data. They include natural excitation technique (NEXT) [1–3], autoregressive moving averaging models (AMAM) [4] and stochastic subspace technique (SST) [5–8].

W. Zhang (✉) · H. Qiu
School of Mechanical Engineering, Tongji University, Shanghai, China
e-mail: wgzhang@shmtu.edu.cn

W. Zhang · K. Lu · Z. Hao · Y. Liu
Logistics Engineering College, Shanghai Maritime University, Shanghai, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_66,

Among all of the modal estimation techniques, NEXT and SST are very important and effective. Under the assumption that the system is excited by stationary white noise, James [1] deduced that the cross-correlation functions between the responses signals are a sum of decaying sinusoids of the same form as the impulse response function of the original system. Each decaying sinusoid has a damped natural frequency and damping ratio which is identical to that of a corresponding structural mode.

The balance realization (BR) method is a kind of stochastic subspace identification method [9]. In this chapter, the BR method is employed to estimate the modal parameters of a crane structure by the simulated 'output-only data'. This chapter is organized as follows: First, the stationary random load spectra are simulated by the Shinozuka unary multidimensional stationary random process simulation method. Second, the output data of a crane structure is acquired by applying the simulated spectra on the FEM model. Third, the BR method is used to identify the modal parameters from in-operation output-only data which is obtained from FEM by applying the Newmark method. Finally, comparisons between BR and FEM are conducted to verify the modal identification of crane structure based on output-only data.

66.2 Balance Realization Method

The BR method is based on "subspace" techniques to identify frequency, damping and mode shapes of the structure.

For the "subspace" techniques, the discrete-time output vector $\{y_k\}$ is defined by a discrete-time stochastic state-space model:

$$\begin{aligned} \{x_{k+1}\} &= [A]\{x_k\} + \{w_k\} \\ \{y_k\} &= [C]\{x_k\} + \{v_k\} \end{aligned} \tag{66.1}$$

where $\{x_k\}$ represents the state vector of dimension n, $\{w_k\}$ and $\{v_k\}$ are zero-mean white-noise vector sequences, representing the process noise and measurement noise, respectively. The matrices $[A]$ and $[C]$ are the state space matrix and the output matrix, respectively.The dynamics of the system are completely characterized by the eigenvalues and the observed parts of the eigenvectors of the [A] matrix.

Let $[H_{p,q}]$ be the following block-Hankel matrix filled up with p block rows and q block columns of the correlation matrix $[R_k]$:

$$[H_{p,q}] = \begin{bmatrix} [R_1] & [R_2] & \dots & [R_q] \\ [R_2] & [R_3] & \dots & [R_{q+1}] \\ \dots & \dots & \ddots & \vdots \\ [R_p] & [R_{p+1}] & \dots & [R_{p+q-1}] \end{bmatrix} \tag{66.2}$$

where

$$[R_k] = E(\{y_{k+m}\}\{y_m\}_{\mathrm{ref}}^T)$$

$E(.)$ denotes the expected value operator and $\{y_m\}_{\mathrm{ref}}$ is a subset of the output vector $\{y_m\}$ containing N_{ref} outputs which are serving as references.

And the $[H_{p,q}]$ can be factored by

$$[H_{p,q}] = [O_P][C_q] \tag{66.3}$$

For large enough$p$ and$q$, along with this model, the observability matrix $[O_p]$ of order p and the controllability matrix $[C_q]$ of order q are defined as:

$$[O_p] = \begin{bmatrix} [C] \\ [C][A] \\ \vdots \\ [C][A]^{p-1} \end{bmatrix} : [C_q] = [G][A][G] \cdots [A]^{q-1}[G] \tag{66.4}$$

where

$$[G] = E(\{x_{k+1}\}\{y_k\}^T) \tag{66.5}$$

$E(.)$ denotes the expectation operator. The matrices $[O_p]$ and $[C_q]$ are assumed to be of rank $2N_m$, and N_m is the number of system modes.

Let $[W_1]$ and $[W_2]$ be two user-defined invertible weighting matrices of size pN_{resp} and qN_{resp} ,respectively. Pre- and post-multiplying the Hankel matrix with the matrices $[W_1]$ and $[W_2]$ and performing a SVD decomposition on the weighted Hankel matrix gives the following:

$$[W_1][H_{p,q}][W_2] = [[U_1][U_2]] \begin{bmatrix} [S_1] & [0] \\ [0] & [0] \end{bmatrix} \begin{bmatrix} [V_1]^T \\ [V_2]^T \end{bmatrix} = [U_1][S_1][V_1]^T \tag{66.6}$$

where $[S_1]$ contains n non-zero singular values in decreasing order,the n columns of $[U_1]$are the corresponding left singular vectors and the n columns of $[V_1]$ are the corresponding right singular vectors.

On the other hand, the factorization property of the weighted Hankel matrix results in

$$[W_1][H_{p,q}][W_2]^T = [W_1][O_p][C_q][W_2]^T \tag{66.7}$$

From (66.6) and (66.7), it can be easily observed that the observability matrix can be recovered, up to a similarity transformation, as

$$[O_p] = [W_1]^{-1}[U_1][S_1]^{1/2} \tag{66.8}$$

The system matrices are then estimated, up to similarity transformation, using the shift structure of $[O_p]$. So,

$$[C] = \{\mathit{first\ block\ row\ of}\ [O_p]\} \tag{66.9}$$

And $[A]$ is computed as the solution of

$$[O^{\uparrow}_{p-1}] = [O_{p-1}][A] \tag{66.10}$$

where $[O_{p-1}]$ is the matrix obtained by deleting the last block row of $[O_p]$ and$[O^{\uparrow}_{p-1}]$ is the upper shifted matrix by one block row.

In case of BR, the weighting is as follows:

$$[W_1] = [I] \text{ and } [W_2] = [I] \tag{66.11}$$

So, no weighting is involved.

66.3 Stationary Random Road Spectra Simulation

When running on the rail, the vibration of the crane structure occurs due to the track irregularities. From the testing results of high speed railway from ZHENGZHOU to GUANGZHOU by China Academy of Railway Science, the power spectra of track irregularities are fitted by (66.12),

$$s(f) = \frac{A(f^2 + Bf + C)}{f^4 + Df^3 + Ef^2 + Ff + G} \tag{66.12}$$

where, $s(f)$ is power spectrum density, mm^2/(1/m); f is special frequency of rail irregularity, (1/m); *A, B, C, D, E, F, G* are characteristic parameters which can refer to [10];

Taking the track irregularity as a stationary random process, and considering the track irregularities in different directions having the weak correlation, vertical and lateral track irregularity curves (Fig. 66.1) of the left and right track are simulated by the Shinozuka unary multidimensional stationary random process simulation method [11]. Making the mathematical statistics for the track irregularity, the simulated PSD agrees well with the theory of PSD, see Fig. 66.2.

66.4 Output-Only Data Simulation

The output data of crane structure is simulated by Newmark method, for the linearity structure, the dynamic equation is:

$$[M]\{\ddot{u}\} + [C]\{\dot{u}\} + [K]\{u\} = \{F^a\} \tag{66.13}$$

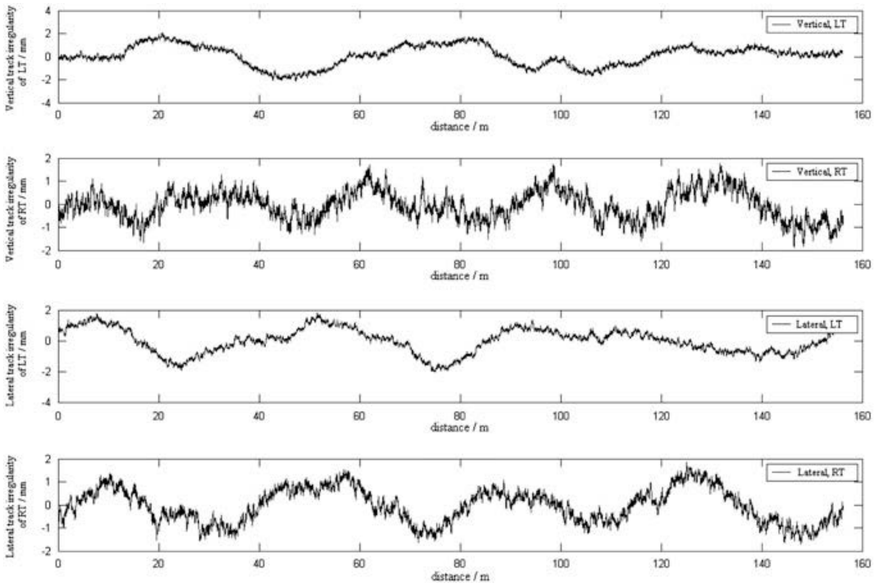
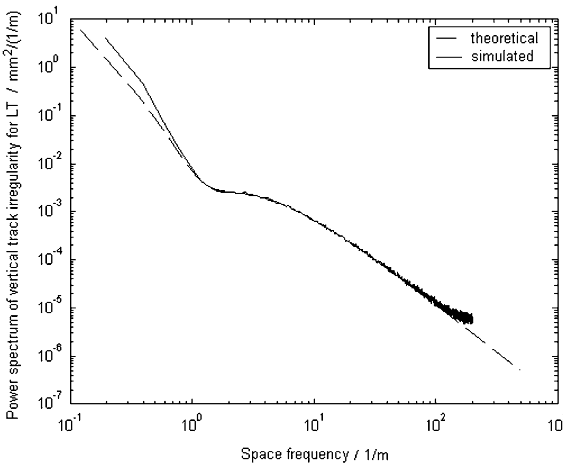


Fig. 66.1 Vertical and lateral track irregularity curve of the left and right track

Fig. 66.2 Comparison of the theoretical spectrum and the simulated spectrum of the left track's vertical irregularity



Where, $[M]$, $[C]$, $[K]$, $\{\ddot{u}\}$, $\{\dot{u}\}$, $\{u\}$ are the mass matrix, damping matrix, stiff matrix, nodal acceleration vector, nodal velocity vector and nodal displacement vector, respectively. $\{F^a\}$ is the applied load vector.

The Newmark method uses finite difference expansions in the time interval Δt, in which it is assumed that:

$$\{\dot{u}_{n+1}\} = \{\dot{u}_n\} + [(1-\delta)\{\ddot{u}_n\} + \delta\{\ddot{u}_{n+1}\}]\Delta t \qquad (66.14)$$

$$\{u_{n+1}\} = \{u_n\} + \{\dot{u}_n\}\Delta t + \left[\left(\frac{1}{2} - \alpha\right)\{\ddot{u}_n\} + \alpha\{\ddot{u}_{n+1}\}\right]\Delta t^2 \tag{66.15}$$

where α, δ are Newmark integration parameters; $\Delta t = t_{n+1} - t_n$; u_n, $\dot{u}_n, \ddot{u}_n$ are respectively the nodal displacement vector, velocity vector and acceleration vector at time t_n; $u_{n+1}, \dot{u}_{n+1}, \ddot{u}_{n+1}$ are respectively the nodal displacement vector, velocity vector and acceleration vector at time t_{n+1}.

The primary aim is to calculate the displacements u_{n+1}, so at time t_{n+1}:

$$[M]\{\ddot{u}_{n+1}\} + [C]\{\dot{u}_{n+1}\} + [K]\{u_{n+1}\} = \{F^a\} \tag{66.16}$$

The solution for the displacement at time t_{n+1} is obtained by rearranging (66.14), (66.15) and (66.16), such that:

$$\{\ddot{u}_{n+1}\} = a_0(\{u_{n+1}\} - \{u_n\} - a_2\{\dot{u}_n\} - a_3\{\ddot{u}_n\} \tag{66.17}$$

$$\{\dot{u}_{n+1}\} = \{\dot{u}_n\} + a_6\{\ddot{u}_n\} + a_7\{\ddot{u}_{n+1}\} \tag{66.18}$$

$$\begin{aligned}(a_0[M] + a_1[C] + [K])\{u_{n+1}\} = \{F^a &+ [M](a_0\{u_n\} + a_2\{\dot{u}_n\} + a_3\{\ddot{u}_n\})\\ &+ [C](a_1\{u_n\} + a_4\{\dot{u}_n\} + a_5\{\ddot{u}_n\})\end{aligned} \tag{66.19}$$

where,

$$a_0 = \frac{1}{\alpha\Delta t^2},\ a_1 = \frac{\delta}{\alpha\Delta t},\ a_2 = \frac{1}{\alpha\Delta t},\ a_3 = \frac{1}{2\alpha} - 1,\ a_4 = \frac{\delta}{\alpha} - 1,$$
$$a_5 = \frac{\Delta t}{2}\left(\frac{\delta}{\alpha} - 2\right),\ a_6 = \Delta t(1 - \delta),\ a_7 = \delta\Delta t$$

From Ref. [12] the unconditional stability can be achieved when:

$$\alpha > \frac{1}{4}\left(\frac{1}{2} + \delta\right)^2,\ \delta \geq \frac{1}{2},\ \frac{1}{2} + \delta + \alpha > 0 \tag{66.20}$$

$\delta = 0.5$ and $\alpha = 0.5$ in this chapter.

The FEM model (Fig. 66.3) of the crane has 3250 nodes and 1277 elements, the main element type being beam 44, beam 189, Mass 21 and link 10. Apply the stationary random road spectrum to the four legs of this crane structure, and then use the Newmark method to simulate the output data.

66.5 Modal Parameters Identification

It is well known that the reference should be selected in such a way that the outputs contain as much relevant modal information as possible. In the present study, the beam 19 (at the middle of the beam) in X-direction and lcmtz-1 (at the land side

Fig. 66.3 Crane FEM model

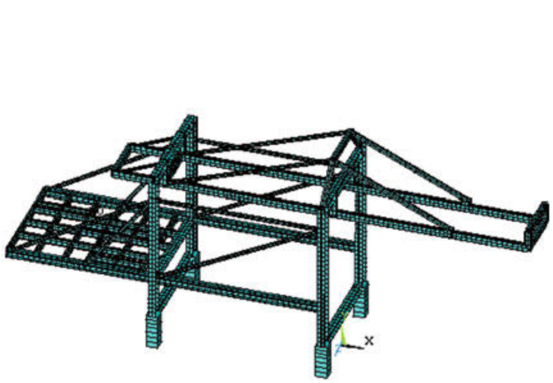
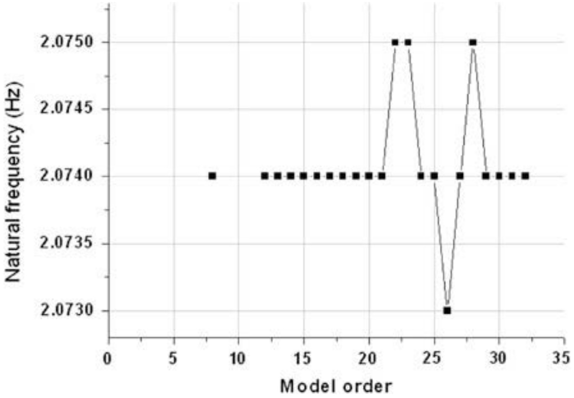


Fig. 66.4 Frequency of mode 1 identified by BR, plot as a function of the model order



leg) in Y-direction and beam 33 (at the end of the beam) in Z-direction are selected as the references.

With respect to the selection of the model order, the stability of modal parameters has been investigated for increasing model order in stabilization diagrams. Figures 66.4 and 66.5 show the frequency and damping ratio of mode 1 which is identified by BR as the function of the model order. From the figures, approximately constant values are found for model orders between 12 and 17. So we take 2.074 Hz and 0.371% as mode 1 frequency and damping ratio at model order 16, respectively. The modal parameters between mode 2 and mode 5 are determined by the same way. Comparisons between the results of FEM and BR have been listed in Table 66.1.

Further validations were conducted, for example:MAC (modal assurance criterion), MPC (modal phase collinearity), MPD (mean phase deviation) and MP (modal participation) values have been calculated. The MAC values in Table 66.2 show that the five identified modes are linear and normal which can also be

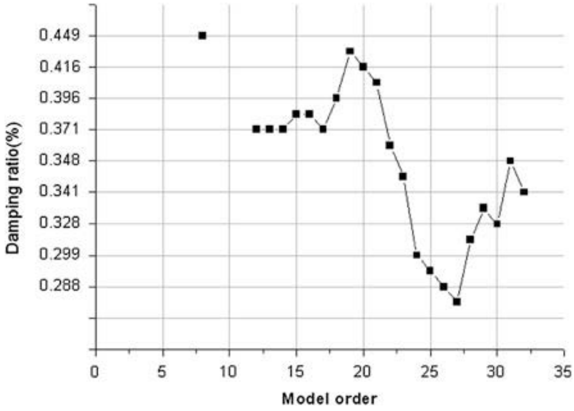


Fig. 66.5 Damping ratio of mode 1 identified by BR, plot as a function of the model order

Table 66.1 A comparison between FEM and BR

	Freq. (Hz)		Damping ratio (%)	Model order
	FEM	BR		
Mode 1	2.089	2.074	0.371	16
Mode 2	3.271	3.267	0.221	16
Mode 3	4.276	4.265	0.332	15
Mode 4	5.530	5.528	0.159	16
Mode 5	6.940	6.944	0.305	16

Table 66.2 Auto Modal Assurance Criterion (%)

	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
Mode 1	100	1.981	0.965	0.918	0.002
Mode 2	1.981	100	0.722	0.812	0.004
Mode 3	0.965	0.722	100	0.980	0.762
Mode 4	0.918	0.812	0.980	100	1.736
Mode 5	0.002	0.004	0.762	1.736	100

Table 66.3 Validations for MPC, MPD and MP (%)

	Freq. (Hz)	MPC (%)	MPD (0)	MP (%)
Mode 1	2.074	99.960	1.314	100
Mode 2	3.267	99.981	0.868	100
Mode 3	4.265	98.084	7.800	100
Mode 4	5.528	99.980	0.790	100
Mode 5	6.944	99.988	0.695	100

concluded from the MPC and MPD values in Table 66.3. Moreover, the MP factor values (see Table 66.3) are all 100% indicating that the corresponding mode is well excited by the stationary random road spectrum.

66.6 Conclusions

The stationary random road spectra are well simulated by the Shinozuka unary multidimensional stationary random process simulation method, and the crane structure of FEM is well excited by the stationary random input. The comparisons between the frequencies of FEM and BR show that the frequencies identified by BR agree well with those by FEM. Further validations (e.g. MAC, MPC, MPD) show that the identified five modes are exactly all normal modes, linearly independent and orthogonal. The present work indicates that the BR method can identify the modal parameters of the crane structure accurately by the output-only data.

Acknowledgment This work sponsored by National 863 plans projects (2009AA043001), supported by Shanghai Education Committee Projects (J50604&09ZZ163), supported by Shanghai Science & Technology Committee Research Project (08210510550, 09DZ2250400 & 10PJ1404700), and supported by Ministry of Communications Research Project (2009-329-810-020&2009-353-312-190).

References

1. James GH III, Carne TG, Lauffer JP (1993) The natural excitation technique (Next) for modal parameter extraction from operating wind turbines [R]. Sandia national laboratories
2. Shen F, Zheng M, Shi DF, Xu F (2003) Using the cross-correlation technique to extract modal parameters on response-only data [J]. J Sound Vib 259:1163–1179
3. Carne TG, James III GH (2010) The inception of OMA in the development of modal testing technology for wind turbines [J]. Mech Syst Signal Process 24:1213–1226
4. Smail M, Thomas M, Lakis A (1999) Arma models for modal analysis: Effect of model orders and sampling frequency [J]. Mech Syst Signal Process 13:925–941
5. Yu DJ, Ren WX (2005) EMD-based stochastic subspace identification of structures from operational vibration measurements [J]. Eng Struct 27:1741–1751
6. Bodeux JB, Golinval JC (2003) Modal identification and damage detection using the data-driven stochastic subspace and armav methods [J]. Mech Syst Signal Process 17:83–89
7. Brownjohn JMW, Magalhaes F, Caetano E, Cunha A (2010) Ambient vibration re-testing and operational modal analysis of the humber bridge [J]. Eng Struct 32:2003–2018
8. Liu W, Gao WC, Sun Y, Xu MJ (2008) Optimal sensor placement for spatial lattice structure based on genetic algorithms [J]. J Sound Vib 317:175–189
9. Hermans L, Van der auweraer H (1999) Modal testing and analysis of structures under operational conditions: industrial applications [J]. Mech Syst Signal Process 13:193–216
10. Xia H, Zhang N (2005) Vehicle And Structure Dynamic Interaction [M] 2nd edn. Science Press, Beijing
11. Shinozuka M, Jan CM (1972) Digital Simulation of Random Process And Its Application [J]. J Sound Vib 25:111–128
12. Zienkiewicz C (1977) The Finite Element Method, London

Chapter 67
CVS-MH: An Efficient Certificate Validation Scheme Based on Multi-Hashing

Mengbo Hou and Qiuliang Xu

Abstract In the recent years, the Public Key Infrastructure with the X.509 authentication is becoming a prominent security model for a variety of e-commerce applications and large-scale distributed computing, while it has not been sufficiently investigated in the certificate revocation and verification mechanism. We discuss its need and importance and analyze the limitations of several certificate validation schemes that are widely used. Then we propose an alternative scheme. The underlying idea is that the certificate holder provides certificate validation proof to the verifiers in manner of initiative. According to this scheme, The certificate validation proof is a proof issued by a trusted third party for the certificate stating whether it was revoked or not. For both parties in any transaction, the certificate holder provides the certificate validation proof to the verifier, the verifier knows about the validity status of the certificate by validation efficiently without any extra information except the certificate. The certificate validation proof is created by multi-operations with a HASH function and operations are associated with the current time. The suggested scheme is principally simple with characteristics of distributed processing, high security, low communication costs and good practicability.

Keywords Public Key Infrastructure · X.509 certificate · Certificate validation · Hash function

M. Hou (✉) · Q. Xu
School of Computer Science and Technology, Shandong University,
Jinan, People's Republic of China
e-mail: houmb@sdu.edu.cn

Q. Xu
e-mail: xuqiuliang@sdu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_67,

67.1 Introduction

As more and more security infrastructures developed, Public Key Infrastructure (PKI) [1, 2] gained considerable attention as it seems to hold a promising foundation for secure electronic commerce, Grid computation, cloud computing and Ad Hoc network. With cryptographic primitives, such as asymmetric encryption, symmetric encryption, hash function and message authentication code (MAC), it provides data confidentiality, data integrity, authentication and non-repudiation for applications. The wide use of public key cryptography requires the ability to verify the authenticity of public keys. This is achieved through the use of digital certificates to serve as a means for transferring trust, such as X.509 certificate [1] or PKIX certificate [2]. A digital certificate is a message signed by a publicly trusted Certification Authority (CA) which includes a subject entity, subject public key and additional data, such as expiration date, and information regarding the key and the subject entity.

When a digital certificate is issued, its validity is limited by a starting date and an expiration date. However, there are circumstances where a certificate must be revoked prior to its expiration date, such as when a private key is revealed or a subject's affiliation or position is changed. Thus, the verification process of a digital certificate is a necessary but not sufficient evidence for its validity, and a mechanism is needed for determining whether a certificate was revoked. Certificate revocation [2, 3, 4, 5] is the act of invalidating the association between the public key and attributes embodied in a certificate. However, certificate revocation is inherently difficult [6, 7]. No solution has been found that meets the timeliness and performance requirements of all applications and environments.

In a typical PKI environment, a certificate revocation and verification scheme needs to be fast, efficient, timely and particularly appropriated for large infrastructures. It is necessary to reduce the number of time-consuming calculations like verification processes of a digital signature and to apply other mechanisms, or to minimize the amount of data transmitted. In this paper, we suggest an alternative scheme which is principally simple with characteristics of distributed processing, high security, low communication costs and good practicability.

The remainder of this paper is organized as follows: In Sect. 67.2, we briefly review several schemes (CRL [2], OCSP [8], CRT [9], and CRS [10]). Our scheme called CVS-MH and its analysis are described in detail in Sects. 67.3 and 67.4. Finally, in Sect. 67.5, we give the conclusion.

67.2 Discussion of Several Certificate Validation Schemes

Several different schemes were proposed in the literature in the recent years. We discuss them below, including their advantages and disadvantages.

67.2.1 Certificate Revocation List (CRL)

A CRL [2] is a signed list issued by the CA identifying all revoked certificates by their serial numbers. It is sent to the directory on a periodic basis. The main advantages of the scheme are simplicity, easy to implement and deploy. However, it has several disadvantages. (1) The cost of CRL management and distribution is high. Because of the periodical distribution of CRL and potential size of CRL, scaling to large communities can be difficult. (2) It is inapt for transactions that require real-time revocation status. (3) The CRL distribution period is very hard to make certain. (4) It does not provide a positive response. Because it only identifies revoked certificates, the existence of a certificate cannot be determined solely from validity information.

67.2.2 Online Certificate Status Protocol (OCSP)

In order to overcome the limitation inherent to the CRL schemes, several approaches of online certificate validation have been proposed [8, 11, 12]. The most widely used of these is the Online Certificate Status Protocol (OCSP) [8]. It allows CA to set up responders that can, when given a certificate identifier, responds with either “good”, “revoked” or “unknown”. It overcomes the chief limitation of the CRL and makes verifying certificates happen in a rapid and online fashion. There are still some problems existing in the OCSP scheme: (1) The requester must know the proper OCSP responder to query in advance. (2) The responder needs to know about the certificate in question as well as the signing authority. (3) Wide-band network and high performance OCSP server are required to ensure the speed of requests and responses. (4) If the responder is centralized, it is vulnerable to DOS attack. (5) Compromise of responder’s private key affects the entire system.

67.2.3 Certificate Revocation Tree (CRT)

Kocher [9] suggested the use of Certificate Revocation Trees (CRT) in order to enable the verifier of a certificate to get a short proof that the certificate was not revoked. A CRT is a hash tree with leaves corresponding to a set of statements about certificate serial number n issued by a CA. The set of statements is produced from the set of revoked certificates of every CA. It provides the information whether a certificate n is revoked or not. To produce the CRT, the CRT issuer builds a binary hash tree with leaves corresponding to the above statements. A proof for a certificate status is a path in the hash tree. The main advantages of CRT over CRL are that the entire CRL is not needed for verifying a specific certificate and that a user may hold a succinct proof of the validity of his certificate. The main disadvantage of CRT is in the computational work needed to update the CRT. Any change in the set of revoked certificates may result in re-computation of the entire CRT.

67.2.4 Certificate Revocation System (CRS)

Micali [10] suggested the Certificate Revocation System (CRS) in order to improve the CRL communication costs. The underlying idea is to sign a message for every certificate stating whether it was revoked or not, and to use an offline/online signature scheme to reduce the cost of periodically updating these signatures. The directory is updated daily by the CA sending this signature for each certificate. The advantage of CRS over CRL is in its query communication costs. Although the daily update of the CRS is more expensive than a CRL update, the cost of CRS querying is much lower. The main disadvantage of this system is the increase in the CA-to-directory communication. Moreover, since the CA's communication costs are proportional to the directory update rate, CA-to-directory communication costs limit the directory update rate.

It has been argued at length by Rivest [13] that CRL is both semantically and technically inferior to other approaches, and then he asked whether CRL could, and should, be eliminated in favor of other mechanisms. According to his underlying idea, we proposed a new scheme whereby the certificate verifier can easily make sure of the certificate validity status by the certificate holder showing proper proof directly.

67.3 CVS-MH: Our Alternative Scheme

67.3.1 System Architecture of CVS-MH

The CA provides a trusted third party that can vouch for the validity of the credentials of both parties in any transaction. It is based upon open standards of which the most important is X.509 [1, 2] and PKIX [14, 15] thus allowing it to work with other CA systems that use X.509 certificates. Its architecture is designed to be modular with components defined in key areas of functionality. At the top of the tree-type structure, the CA is central to the viability of the system, responsible for generating, publishing and revoking digital certificates. The CA is managed by the CA Operator (CAO) and beneath the CA are Registration Authorities (RAs) which act as the interfaces between the end users and the CA, carrying the burden of enrolments and acting as intermediary for authentication. In turn, the RAs are managed by RA Operators (RAOs). Each of the CA, CAO, RA and RAO has its own certificate so that each component of the PKI is able to identify itself with other components and communicate securely. Figure 67.1 shows an example of a

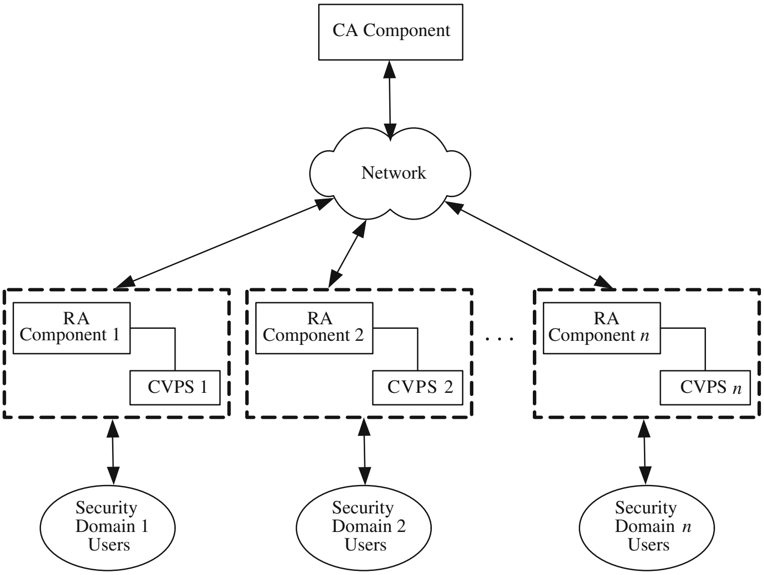


Fig. 67.1 Architecture of CVS-MH

classical CA system except for the Certificate Validation Proof Server (CVPS) components.

CVPS components are deployed in our scheme in order to issue user's Certificate Validation Proof (CVP) according to the demand of the sub-security domain users. The demand, here, is a query message, indicated to acquire a proof for the current validity status of user certificate.

The CVPS is deployed as an important component of the RA system, and is administrated by the RAO. For each of the RAs, there is a CVPS deployed. With the protected channel (such as SSL Channel), the CVPS communicates with the CA component securely. For the common certificate holder and its relying party in the application, the CVP is acquired from the CVPS by anyone at any time. There is no serious security requirement in the communication between the CVPS and the requesters, so the CVP of a certificate can be acquired by either the certificate holder or the relying party. In our scheme, the certificate holder is preferred, because the relying party can hardly know which CVPS to query in advance, while the certificate holder knows about that. Due to the distributed design of the CVPS system, high performance and convenience are enabled.

67.3.2 Generation of User Certificate

(1) Modification of X.509 certificate. The CA component defines a time interval n, according to the certificate validity period (e.g., with respect to one year for the certificate validity period, define $n = 365$ and increment i represents a

day). Using X.509 certificates, the number of extension fields needs to be extended by two fields: Y indicates the certificate validity and N indicates the certificate invalidity. Because of the CA's signature, the authenticity of both values is guaranteed.

(2) Generating secrets. The CA chooses two pseudo-random numbers Y0 and N0 (Y0 $\neq$ N0) for each certificate requester distinctly, keeps them secret and sends them to the responding CVPS at the RA end in a secure manner.

(3) Choosing one hash function. The CA constitutes a proper hash function H, then the CA calculates as below, and generates a certificate for the certificate requester:

$$Y \leftarrow Y_n = H^n(Y_0),\ N \leftarrow H(N_0)$$
$$(\text{define } H^1(x) = H(x),\ \text{ and } H^n(x) = H(H^{n-1}(x)), \quad n = 2, 3, \ldots)$$

67.3.3 Revocation and Reuse of User Certificate

When the certificate holder or the CA wants to revoke certificates for some reasons, proper revocation request message should be submitted to the responding CVPS, the CVPS will revoke the certificate after careful auditing. The revocation operation is merely a symbol marking to indicate the revoked status of the certificate; if the user wants to reuse the formerly revoked certificate, proper reuse request messages should be submitted to the responding CVPS, and the CVPS will recover the certificate after careful auditing. The reuse operation is merely a symbol marking to indicate the validity status of the certificate too. Although the revocation and reuse of a certificate are very simple, the authentication of operations should be considered seriously.

67.3.4 Generation of Users' Certificate Validation Proof (CVP)

Whenever the user wants to communicate with other relying parties in a secure manner, besides the certificate, a CVP should be provided to indicate the validity status of its certificate. The CVP is generated and acquired from the CVPS in the user's security domain according to the current validity status of the user's certificate. (1) If the certificate is currently valid and i days have passed from the very beginning of the certificate validity date, the CVPS calculates: $CVP_i = H^{n-i}(Y_0)$. Apparently, we have $H^i(CVP_i) = Y$. (2) If the certificate is currently revoked, then the CVPS calculates: $CVP_i = N_0$. Apparently, we have $H(CVP_i) = N$.

Theorem 1 *Forging of the Certificate validation proof (CVP) is computationally infeasible if the hash function H(x) is strongly collision-free.*

67.3.5 Operations of the Certificate Holders

If certificate validity verification is required in any application, the certificate holder can, at any time, acquire its CVP from the CVPS and submit it to the relying party, along with the corresponding certificate in manner of initiative. For example, within a Client/Server application system with authentication based on certificate that allows the client operators to interact with the server in a secure manner, the operator merely acquires the CVP for the first time login, even though he logs in many times in a day, hence only one interaction with the CVPS is required. The CVP can be cached locally for all the day until the next day to be refreshed.

67.3.6 Operations of the Relying Party

When the relying party gets the certificate and the CVP of the certificate holder, it verifies the validity of certificate status as the following steps (suppose the certificate is not out-of-date): The relying party computes the number of days between the issuing day and current day, noted as *i*, then computes $Y^{'} = H^{i}(CVP)$. If Y′ = Y, then the certificate is valid (not revoked); If H (CVP) = N, then the certificate is revoked before now.

67.4 Analysis of Security and Performance

We show that the proposed scheme is secure from attacks. The adversary can neither modify Y and N in the certificate, nor reveal Y0 and N0 from Y and N, because Y0 and N0 are privately generated and occupied by CA and CVPS. If the certificate has been revoked, the certificate holder can hardly acquire a valid CVP, even a valid CVP acquired before the certificate revoked, he still could not construct the valid CVP of current time by the old CVPs (by Theorem 1). Although the adversary could acquire the valid CVP of other users from the CVPS easily, he still could not gain any advantage from that. The security of the scheme mainly focuses on the confidentiality and randomicity of Y0 and N0. Accuracy of the system time is also a serious concern, especially at the end of the relying parties, as the validation of the verification is time-related. The proposed scheme has many advantages compared to other schemes:

The certificate holder submits the CVP to the relying party in manner of initiative. The relying party could verify the validity of certificate without extra information. It is quite suitable for the Client/Server or Browser/Web applications.

The CVP proof is a small value of hashing operations, so the communication costs are saved.

The verification processes are totally operations of hash function computations, so high speed can be guaranteed.

As for the CVPS, the CVP requesters are all from the local security domains, and each CVPS is only responsible for the users of its domain, the operations of CVPS are totally hash function computations. Additionally, for each CVPS, the CVP requests are periodical, so the burden of each CVPS is very limited. Due to the distributed architecture of CVPS, it is much more efficient than other schemes.

The CA component defines time interval n according to the validity time. Parameter n is variable(If n represents the number of days, then the certificate holder need only request CVP proof one time for one day. If n represents the number of weeks, then the CVP proof need not to be refreshed once a week). To some extent, the value of n is measurement of certificate verification security level.

The certificate holder merely interacts with one of the distributed CVPSs, but not the centralized CA, so the certificate revocation processes are distributed and without the CA's awareness.

67.5 Conclusion

In this chapter, we proposed a new certificate validity scheme according to the idea that the certificate holder providing the certificate validity proof in manner of initiative. For both parties in any transaction, the certificate holder provides the certificate validity proof to the verifier, the verifier knows about the validity status of the certificate by verifying certificate validity proof efficiently without any extra information except the certificate. The certificate validity proof is created by multi-operations with hash function computations; the operations are associated with the current time. The new scheme is principally simple with characteristics of distributed processing, high security, low communication costs and good practicability.

Acknowledgments This work was supported by National Natural Science Foundation of China (No. 60873232), also by Natural Science Foundation of Shandong Province, China (No. Y2007G37).

References

1. ITU-T Rec X.509 | ISO/IEC 9594-8: Information technology-open systems interconnection-the directory: public-key and attribute certificate frameworks (2001)
2. Housley R, Polk W, Solo D (2002) Internet X.509 Public key infrastructure certificate and CRL profile [RFC 3280]. IETF PKIX work group
3. Mo J, Wang XM (2010) Distributed certificate revocation scheme for Ad Hoc network. Comput Eng 36(10):149–151
4. Zhong H, Xu CX, Qin ZG (2007) A distributed certificate revocation scheme for Ad Hoc networks. J Univ of Electron Sci Technol China 36(3):496–499

5. Fox B, LaMacchia B (1998) Certificate revocation: mechanics and meaning. In: Proceedings of financial cryptology-FC'98, LNCS vol 1465, pp 158–164
6. Arnes A, Meijer H, Lloyd S, Just M (2000) Selecting revocation solutions for PKI. In: Proceedings of the fifth nordic workshop on secure IT systems
7. Myers, M (1998) Revocation: Options and challenges. In: Proceedings of the 2rd international conference on financial cryptography, LNCS vol 1465, pp 165–171
8. Myers M, Ankney R, Malpani A, Galperin S, Adams C (1999) X.509 Internet public key infrastructure online certificate status protocol–OCSP [RFC2560]
9. www.valicert.com/company/crt.html
10. Micali S (1996) Efficient certificate revocation. MIT Laboratory for Computer Science, Technical Memo 542b
11. Ambarish M, Paul H (2000) Simple certificate validation protocol (SCVP). Internet Draft IETF PKIX work group
12. Phillip HB (1999) OCSP Extensions. Internet Draft IETF PKIX work group
13. Rivest R (1998) Can we eliminate certificate revocation lists? In: Proceedings of financial cryptography'98, LNCS vol 1465, pp 178–183
14. Noar M, Nassim K (1998) Certificate revocation and certificate update. In: Proceedings of 7th USENIX Security Symposium, pp 217–228
15. Arboit G (2008) A localized certificate revocation scheme for mobile ad hoc networks. Ad Hoc Netw 6(1):17–31

Chapter 68
Improvement of Selvi et al.'s Identity-Based Threshold Signcryption Scheme

Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

Abstract Signcryption can realize the function of encryption and signature in a reasonable logic step, which can lower computational costs and communication overheads. In 2008, Selvi et al. proposed an identity-based threshold signcryption scheme. Our previous analysis has been showed that the threshold signcryption scheme of Selvi et al. is vulnerable if the attacker can replace the group public key and has pointed out that the receiver uses the senders' public key without any verification in the unsigncrypt stage cause this attack. Further, we propose a probably secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

Keywords Identity-based · Signcryption · Bilinear pairing · Cryptanalysis · Attack

68.1 Introduction

Encryption and signature are the two basic cryptographic tools offered by public key cryptography for achieving confidentiality and authentication. Signcryption can realize the function of encryption and signature in a reasonable logic step

W. Yuan · L. Hu · H. Li · J. Chu (✉)
Department of Computer Science and Technology,
Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_68,

which is proposed by Zheng [1] in 1997. Comparing to the traditional way of signature then encryption or encryption then signature, signcryption can lower the computational costs and communication overheads. As a result, a number of signcryption schemes [2–8] were proposed following Zheng's work. The security notion for signcryption was first formally defined in 2002 by Baek et al. [9] against adaptive chosen ciphertext attack and adaptive chosen message attack. The same as signature and encryption, signcryption meets the attributes of confidentiality and unforgeability as well.

In 1984, Shamir [10] introduced identity-based public key cryptosystem, in which a user's public key can be calculated from his identity and defined hash function, while the user's private key can be calculated by a trusted party called Private Key Generator (PKG). The identity can be any binary string, such as an email address and need not to be authenticated by the certification authentication. As a result, the identity-based public key cryptosystem simplifies the program of key management to the conventional public key infrastructure. In 2001, Boneh and Franklin [11] found bilinear pairings positive in cryptography and proposed the first practical identity-based encryption protocol using bilinear pairings. Soon, many identity-based [12, 14–16] and other relational [13, 17, 18] schemes were proposed and the bilinear pairings became important tools in constructing identity-based protocols.

Group-oriented cryptography [19] was introduced by Desmedt in 1987. Elaborating on this concept, Desmedt and Frankel [20] proposed a (t, n) threshold signature scheme-based RSA system [21]. In such a (t, n) threshold signature scheme, any out of n signers in the group can collaboratively sign messages on behalf of the group for sharing the signing capability.

Identity-based signcryption schemes combine the advantages of identity-based public key cryptosystem and signcryption. The first identity-based threshold signature scheme was proposed by Baek and Zheng [22] in 2004. Then Duan et al. proposed an identity-based threshold signcryption scheme [23] in the same year by combining the concepts of identity-based threshold signature and encryption together. However, in Duan et al.'s scheme, the master-key of the PKG is distributed to a number of other PKGs, which creates a bottleneck on the PKGs. In 2005, Peng and Li proposed an identity-based threshold signcryption scheme [24] based on Libert and Quisquater's identity-based signcryption scheme [25]. However, Peng and Li's scheme dose not provide the forward security. In 2008, another scheme [26] was proposed by Fagen Li et al., which is more efficient compared to previous scheme. However, Selvi et al. pointed out that Fagen Li et al.'s scheme is not equilibrium between the usual members and a dealer called clerk in Fagen Li et al.'s scheme and proposed an improved scheme [27].

In this chapter, we propose a probably secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

68.2 The Improvement of Selvi et al.'s Scheme

The scheme involves four roles: the PKG [28], a trust dealer, a sender group $U_A = \{M_1, M_2, \ldots, M_n\}$ with identity ID_A and a receiver Bob with identity ID_B.

Setup. Given a security parameter k, the PKG chooses groups G_1 and G_2 of prime order q (with G_1 additive and G_2 multiplicative), a generator P of G_1, a bilinear map $e : G_1 \times G_1 \to G_2$, a secure symmetric cipher (E, D) and hash functions $H_1 : \{0,1\}^* \to G_1,\ H_2 : G_2 \to \{0,1\}^{n_1},\ H_3 : \{0,1\}^* \times G_1 \times \{0,1\}^* \times G_1 \to Z_q^*$. The PKG chooses a master-key $s \in {}_R Z_q^*$ and computes $P_{\text{pub}} = sP$. The PKG publishes system parameters $\{G_1, G_2, n_1, e, P, P_{\text{pub}}, E, D, H_1, H_2, H_3\}$and keeps the master-key s secret.

Extract. Given an identity ID, the PKG computes $Q_{\text{ID}} = H_1(\text{ID})$ and the private key $S_{\text{ID}} = sQ_{\text{ID}}$. Then PKG sends the private key to its owner in a secure way.

Keydis. Suppose that a threshold t and n satisfy $1 \le t \le n < q$. To share the private key S_{ID_A} among the group U_A, the trusted dealer performs the steps below.

(1) Choose $F_1, \ldots, F_{t-1}$ uniformly at random from G_1^*, construct a polynomial $F(x) = S_{\text{ID}_A} + xF_1 + \cdots + x^{t-1}F_{t-1}$.
(2) Compute $S_i = F(i)$ for $i = 0, \ldots, n$. ($S_0 = S_{\text{ID}_A}$). Send S_i to member M_i for $i = 1, \ldots, n$ secretly.
(3) Broadcast $y_0 = e(S_{\text{ID}_A}, P)$ and $y_j = e(F_j, P)$ for $j = 1, \ldots, t-1$.
(4) Each M_i then checks whether his share S_i is valid by computing $e(S_i, P) = \prod_{j=0}^{t-1} y_j^{i^j}$. If S_i is not valid, M_i broadcasts an error and requests a valid one.

Signcrypt. Let $M_1, \ldots, M_t$ be the t members who want to cooperate to signcrypt a message m on behalf of the group U_A.

(1) Each M_i chooses $x_i \in {}_R Z_q^*$, computes $R_{1i} = x_iP, R_{2i} = x_iP_{\text{pub}}, \tau_i = e(R_{2i}, Q_{\text{ID}_B})$ sends (R_{1i}, τ) to the clerk C.
(2) The clerk C (one among the t cooperating players) computes $R_1 = \prod_{i=1}^{t} R_{1i}, \tau = \prod_{i=1}^{t} \tau_i, k = H_2(\tau), c = E_k(m)$ and $h = H_3(m, R_1, k, Q_{\text{ID}_A})$.
(3) Then the clerk C sends h to M_i for $i = 0, \ldots, t$.
(4) Each M_i computes the partial signature $W_i = x_iP_{\text{pub}} + h\eta_iS_i$ and sends it to the clerk C, where $\eta = \prod_{j=1, j\neq i}^{t} -j(i-j)^{-1} \text{mod } q$.
(5) Clerk C verifies the correctness of partial signatures by checking if the following equation holds:

$$e(P, W_i) = e\big(R_{1i}, P_{\text{pub}}\big)\left(\prod_{j=0}^{t-1} y_j^{i^j}\right)^{h\eta_i}$$

If all partial signatures are verified to be legal, the clerk C computes $W = \sum_{i=1}^{t} W_i$; otherwise rejects it and requests a valid one.

(6) The final threshold signcryption is $\sigma = (c, R_1, W)$.

Unsigncrypt. When receiving σ, Bob follows the steps below.

(1) Compute $\tau = e(R_1, S_{\mathrm{ID}_B})$ and $k = H_2(\tau)$.
(2) Recover $m = D_k(c)$.
(3) Compute $h = H_3(m, R_1, k, Q_{\mathrm{ID}_A})$ and accept σ if and only if the following equation holds:

$$e(P, W) = e(\left(P_{\mathrm{pub}}, R_1 + hQ_{\mathrm{ID}_A}\right).$$

68.3 Security Analysis of Our Improved Scheme

In this section, we will give a formal proof on Unforgeability and Confidentiality of our scheme under CDH problem and DBDH problem [29, 30].

Theorem 1 (Unforgeability) *Our improved scheme is secure against chosen message attack under the random oracle model if CDH problem is hard.*

Proof Suppose the challenger C wants to solve the CDH problem. That is, given (aP, bP), C should computes abP.

C chooses system parameters $\left\{G_1, G_2, n_1, e, P, P_{\mathrm{pub}}, E, D, H_1, H_2, H_3\right\}$, sets $P_{\mathrm{pub}} = aP$ and sends parameters to the adversary E (the hash functions H_1, H_2, H_3 are random oracles).

H_1 query: C maintains a list L_1 to record H_1 queries. L_1 has the form of $(ID, \alpha, Q_{\mathrm{ID}}, S_{\mathrm{ID}})$. Suppose the adversary Eve can make H_1 queries less than q_{H_1} times. C selects a random number $j \in [1, q_{H_1}]$. If C receives the jth query, he will return $Q_{\mathrm{ID}_j} = bP$ to Eve and sets $\left(\mathrm{ID}_j, \perp, Q_{\mathrm{ID}_j} = bP, \perp\right)$ on L_1. Else C selects $\alpha_i \in Z_q^*$, computes $Q_{\mathrm{ID}_i} = \alpha_i P, S_{\mathrm{ID}_i} = \alpha_i P_{\mathrm{pub}}$, returns Q_{ID_i} to E and sets $(\mathrm{ID}_i, \alpha_i, Q_i, S_i)$ on L_1.

H_2 query: C maintains a list L_2 to record H_2 queries. L_2 has the form of (τ, k). If C receives a query about τ_i,, selects $k_i \in Z_q^*$, returns k_i to E and sets (τ_i, k_i) on L_2.

H_3 query: C maintains a list L_3 to record H_3 queries. L_3 has the form of (m, R, k, Q, h). If C receives a query about $(m_i, R_{1i}, k_i, Q_{\mathrm{ID}_i})$, selects $h_i \in Z_q^*$, returns h_i to Eve and sets $(m_i, R_{1i}, k_i, Q_{\mathrm{ID}_i}, h_i)$ on L_3.

Signcrypt query: if C receives a query about signcrypt with message m_i, identity ID_i

1. Select $x_i \in Z_q^*, W_i \in G_1$.
2. Look-up L_1, L_2, set $Q_{\mathrm{ID}_i} = \alpha_i P$ in $L_1, k_i = k_i$ in L_2, and compute $R_i = x_i Q_{\mathrm{ID}_i}$.
3. Set $h_i = H_3(m_i, R_i, k_i, Q_{\mathrm{ID}_i})$.
4. Return (h_i, W_i) to Eve.

Finally, Eve output a forged signcryption $(m, h_i, W_i, Q_{\mathrm{ID}_i})$. If $Q_{\mathrm{ID}_i} \neq Q_{\mathrm{ID}_j}$, Eve fails. Else, if $Q_{\mathrm{ID}_i} = Q_{\mathrm{ID}_j}$, Eve succeeds in forging a signcryption.

As a result, C gains two signcryption ciphertexts which meet:

$$e(P, W_i) = e\left(P_{\text{pub}}, R_i + h_i Q_{\text{ID}_i}\right)$$

$$e(P, W_j) = e\left(P_{\text{pub}}, R_j + h_j Q_{\text{ID}_j}\right)$$

Thus,

$$e\left(P, \left(W_i - W_j\right)\right) = e\left(P_{\text{pub}}, \left(R_i + h_i Q_{\text{ID}_i}\right) - \left(R_j + h_j Q_{\text{ID}_j}\right)\right) \tag{68.1}$$

Note $Q = Q_{\text{ID}_i} = Q_{\text{ID}_j}$,

(1) can be expressed as

$$e\left(P, \left(W_i - W_j\right)\right) = e\left(P_{\text{pub}}, \left(R_i - R_j\right) + \left(h_i - h_j\right)Q\right) \tag{68.2}$$

$$\because P_{\text{pub}} = aP, Q_{ID_j} = bP$$

(2) can be expressed as $e\left(P, \left(W_i - W_j\right)\right) = e\left(aP, \left(\left(\alpha_i - \alpha_j\right) + \left(h_i - h_j\right)\right)bP\right)$

$$\therefore W_i - W_j = \left(\left(\alpha_i - \alpha_j\right) + \left(h_i - h_j\right)\right)abP$$

Hence, the CDH problem $abP = \frac{W_i - W_j}{(\alpha_i - \alpha_j) + (h_i - h_j)}$ can be computed by C with aP and bP.

Theorem 2 (Confidentiality) *Our improved scheme is secure against adaptive chosen ciphertext and identity attack under the random oracle model if DBDH problem is hard.*

Proof Suppose the challenger C wants to solve the DBDH problem. That is, given (P, aP, bP, cP, τ), C should decide whether $\tau = e(P, P)^{abc}$ or not. If there exists an adaptive chosen ciphertext and identity attacker for our improved scheme, C can solve the DBDHP.

C chooses system parameters $\left\{G_1, G_2, n_1, e, P, P_{\text{pub}}, E, D, H_1, H_2, H_3\right\}$, sets $P_{\text{pub}} = aP$ and sends parameters to the adversary E (the hash functions H_1, H_2, H_3 are random oracles).

H_1 query: C maintains a list L_1 to record H_1 queries. L_1 has the form of $(\text{ID}, \alpha, Q_{\text{ID}}, S_{\text{ID}})$. Suppose the adversary Eve can make H_1 queries less than q_{H_1} times. C selects a random number $j \in [1, q_{H_1}]$. If C receives the jth query, he will return $Q_{\text{ID}_j} = bP$ to Eve and sets $\left(\text{ID}_j, \perp, Q_{\text{ID}_j} = bP, \perp\right)$ on L_1. Else C selects $\alpha_i \in Z_q^*$, computes $Q_{\text{ID}_i} = \alpha_i P$, $S_{\text{ID}_i} = \alpha_i P_{\text{pub}}$, returns Q_{ID_i} to E and sets $(\text{ID}_i, \alpha_i, Q_i, S_i)$ on L_1.

H_2 query: C maintains a list L_2 to record H_2 queries. L_2 has the form of (τ, k). If C receives a query about τ_i, selects $k_i \in Z_q^*$, returns k_i to E, and sets (τ_i, k_i) on L_2.

H_3 query: C maintains a list L_3 to record H_3 queries. L_3 has the form of (m, R, k, Q, h). If C receives a query about $(m_i, R_{1i}, k_i, Q_{\mathrm{ID}_i})$, selects $h_i \in Z_q^*$, returns h_i to Eve and sets $(m_i, R_{1i}, k_i, Q_{\mathrm{ID}_i}, h_i)$ on L_3.

Signcrypt query: if C receives a query about Signcrypt with message m_i, identity ID_i

1. Select$c_i \in Z_q^*, W_i \in G_1$.
2. Look-up L_1, L_2, set $Q_{\mathrm{ID}_i} = \alpha_i P$ in $L_1, k_i = k_i$ in L_2. Compute $R_i = c_i P$, if ID_i / $= \mathrm{ID}_j$. Else, if $\mathrm{ID}_i = \mathrm{ID}_j$, compute $R_i = cP$.
3. Set $h_i = H_3(m_i, R_i, k_i, Q_{\mathrm{ID}_i})$.
4. Return $(h_i, W_i)$to Eve.

After the first stage, Eve chooses a pair of identities on which he wishes to be challenged on $\left(\mathrm{ID}_i, \mathrm{ID}_j\right)$. Note that Eve cannot query the identity of ID_A. Then Eve outputs two plaintexts m_0 and m_1. C chooses a bit $b \in \{0, 1\}$ and signcrypts m_b. To do so, he sets $R_1^* = cP$, obtains $k^* = H_2(\tau)$ from the hash function H_2, and computes $c_b = E_{k_1^*}(m_b)$. Then C chooses $W^* \in G_1$ and sends the ciphertext $\sigma^* = \left(c_b, R_1^*, W^*\right)$to Eve. Eve can perform a second series of queries like at the first one. At the end of the simulation, he produces a bit b'. for which he believes the relation $\sigma^* = \text{Signcrypt}\left(m_{b'}, \{S_i\}_{i=1,\ldots,t}, \mathrm{ID}_j\right)$ holds. If $b = b'$, C outputs $\tau = e\left(R_1^*, S_{\mathrm{ID}_j}\right) = e(cP, abP) = e(P, P)^{abc}$. Else, C outputs $\tau \neq e(P, P)^{abc}$. So C can solve the BDDH problem.

68.4 Conclusion

In this chapter, we show that the threshold signcryption scheme of Selvi et al. is vulnerable if the attacker can replace the group public key. Then we point out that the receiver uses the senders' public key without any verification in the unsigncrypt stage cause this attack. Further, we propose a probably secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

Acknowledgments The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318) and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Zheng Y (1997) Digital signcryption or how to achieve cost (signature and encryption) << cost (signature) + cost (encryption). In: Proceedings of advances in CRYPTO'97. LNCS, vol 1294. Springer, Berlin, pp 165–179
2. Bao F, Deng RH (1997) A signcryption scheme with signature directly verifiable by public key. PKC'98. LNCS, vol 1431. Springer, Berlin, pp 55–59
3. Chow SSM, Yiu SM, Hui LCK, Chow KP (2004) Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. ICISC'03. LNCS, vol 2971. Springer, Berlin, pp 269–352
4. Boyen X (2003) Multipurpose identity based signcryption: a swiss army knife for identity based cryptography. CRYPT'03. LNCS, vol 2729. Springer, Berlin, pp 383–399
5. Mu Y, Varadharajan V (2000) Distributed signcryption. INDOCRYPT'00. LNCS, vol 1977. Springer, Berlin, pp 155–164
6. Yang G, Wong DS, Deng X (2005) Analysis and improvement of a signcryption scheme with key privacy. ISC'05. LNCS, vol 3650. Springer, Berlin, pp 218–232
7. SteinFeld R, Zheng Y (2000) A signcryption scheme based on integer factorization. ISW'00. LNCS, vol 1975. Springer, Berlin, pp 308–322
8. Libert B, Quisquater J (2004) Efficient signcryption with key privacy from gap Diffie–Hellman groups. PKC'04. LNCS, vol 2947. Springer, Berlin, pp 187–200
9. Baek J, Steinfeld R, Zheng Y (2002) Formal proofs for the security of signcryption. PKC'02. LNCS, vol 2274. Springer, Berlin, pp 80–98
10. Shamir A (1984) Identity-based cryptosystems and signature schemes. CRYPTO'84. LNCS, vol 196. Springer, Berlin, pp 47–53
11. Boneh D, Franklin M (2001) Identity-based encryption from well pairing. CRYPTO'01. LNCS, vol 2139. Springer, Berlin, pp 213–229
12. Barreto PSLM, Libert B, Mccullagh N, Quisquater JJ (2005) Efficient and provably-secure identity-based signatures and signcryption from bilinear maps. ASIACRYPT'05. LNCS, vol 3788. Springer, Berlin, pp 515–532
13. Huang X, Susilo W, Mu Y, Zhang E (2005) Identity-based ring signcryption schemes: cryptographic primitives for preserving privacy and authenticity in the ubiquitous world. 19th international conference on advanced information networking and applications, Taiwan, pp 649–654
14. Li F, Xiong H, Nie X (2009) A new multi-receiver ID-based signcryption scheme for group communications. ICCCAS'2009. IEEE Press, San Jose, pp 296–300
15. Han Y, Gui X (2009) Multi-recipient signcryption for secure group communication. ICIEA 2009, pp 161–165
16. Jin Z, Wen Q, Du H (2010) An improved semantically-secure identity-based signcryption scheme in the standard model. Comput Electr Eng 36:545–552
17. Liu Z, Hu Y, Zhang X, Ma H (2010) Certificateless signcryption scheme in the standard model. Inf Sci 180:452–464
18. Yu Y, Yang B, Sun Y, Zhu S (2009) Identity based signcryption scheme without random oracles, Computer Standards Interfaces 31:56–62
19. Desmedt Y (1987) Society and group oriented cryptography: a now concept. CRYPTO'87. LNCS, vol 293. Springer, Berlin, pp 120–127
20. Desmedt Y, Frankel, Y (1991) Shared generation of authenticators and signatures. CRYPTO'91. LNCS, vol 576. Springer, Berlin, pp 457–469
21. Rivest RL, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. Commun ACM 21(2):120–126
22. Baek J, Zheng Y (2004) Identity-based threshold signature scheme from the bilinear pairings. International conference on information technology 2004, Las Vegas, Nevada, USA, pp 124–128

23. Duan S, Cao Z, Lu R (2004) Robust ID-based threshold signcryption scheme from pairings. International conference on information security, Shanghai, China, pp 33–37
24. Peng C, Li X (2005) An identity-based threshold signcryption scheme with semantic security. Computational intelligence and security 2005. LNAI, vol 3902. Springer, Berlin, pp 173–179
25. Libert B, Quisquater JJ (2003) Anew identity based signcryption schemes from pairings, 2003 IEEE information theory workshop, Paris, France, pp 155–158
26. Li F, Yu Y (2008) An efficient and provably secure ID-based threshold signcryption scheme. ICCCAS'2008, 4657820. IEEE Press, Chengdu, China pp 488–492
27. Selvi SSD, Vivek SS, Rangan CP (2008) Cryptanalysis of Li et al.'s identity-based threshold signcryption scheme. Proc EUC 2008, vol 2, Shanghai, China, pp 127–132
28. Malone Lee J (2002) Identity based signcryption. In: Cryptology ePrint archive. Report 2002/098
29. Chow SSM, Yiu SM, Hui LCK, Chow KP (2004) Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. In: Lin J-I, Lee D-H (eds) ICISC 2003. LNCS, vol 2971. Springer, Heidelberg, pp 352–369
30. Boyen X (2003) Multipurpose identity based signcryption: a Swiss army knife for identity based cryptography. In: Boneh D (ed) CRYPTO 2003. LNCS, vol 2729. Springer, Heidelberg, pp 383–399

Chapter 69
Analysis of an ID-Based Threshold Signcryption Scheme

Wei Yuan, Liang Hu, Xiaochun Cheng, Hongtu Li, Jianfeng Chu and Yuyu Sun

Abstract Signcryption can realize the function of encryption and signature in a reasonable logic step, which can lower computational costs and communication overheads. In 2008, Li et al. proposed an efficient secure id-based threshold signcryption scheme. The authors declared that their scheme had the attributes of confidentiality and unforgeability in the random oracle model. In this paper, we show that scheme is insecure against malicious attackers and give our attacker method to forge the ciphertext.

Keywords Identity-based · Signcryption · Bilinear pairing · Cryptanalysis

W. Yuan · L. Hu · H. Li · Jianfeng Chu (✉) · Y. Sun
Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

Y. Sun
e-mail: sunyy@ccu.edu.cn

X. Cheng
The School of Computing Science, Middlesex University, London, UK
e-mail: xiaochun.cheng@gmail.com

Y. Sun
Software Institute, Changchun University, Changchun, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_69,

69.1 Introduction

Encryption and signature are the two basic cryptographic tools offered by public key cryptography for achieving confidentiality and authentication. Signcryption can realize the function of encryption and signature in a reasonable logic step which is proposed by zheng [1] in 1997. Comparing to the traditional way of signature then encryption or encryption then signature, signcryption can lower the computational costs and communication overheads. As a result, a number of signcryption schemes [2–8] were proposed following Zheng's work. The security notion for signcryption was first formally defined in 2002 by Baek et al. [9] against adaptive chosen ciphertext attack and adaptive chosen message attack. The same as signature and encryption, signcryption meets the attributes of confidentiality and unforgeability as well. In 1984, Shamir [10] introduced identity-based public key cryptosystem, in which a user's public key can be calculated from his identity and defined hash function, while the user's private key can be calculated by a trusted party called private key generator (PKG). The identity can be any binary string, such as an email address and that need not be authenticated by the certification authentication. As a result, the identity-based public key cryptosystem simplifies the program of key management to the conventional public key infrastructure. In 2001, Boneh and Franklin [11] found bilinear pairings positive in cryptography and proposed the first practical identity-based encryption protocol using bilinear pairings. Soon, many identity-based [12–15] and other relational [16–18] schemes were proposed and the bilinear pairings became important tools in constructing identity-based protocols. Group-oriented cryptography was introduced by Desmedt [19] in 1987. Elaborating on this concept, Desmedt and Frankel [20] proposed a (t, n) threshold signature scheme based RSA system [21]. In such a (t, n) threshold signature scheme, any to out of n signers in the group can collaboratively sign messages on behalf of the group for sharing the signing capability. Identity-based signcryption schemes combine the advantages of identity-based public key cryptosystem and Signcryption. The first identity-based threshold signature scheme was proposed by Baek and Zheng [22] in 2004. Then Duan et al. [23] proposed an identity-based threshold signcryption scheme in the same year by combining the concepts of identity-based threshold signature and encryption together. However, in Duan et al.'s scheme, the master-key of the PKG is distributed to a number of other PKGs, which creates a bottleneck on the PKGs. In 2005, Peng and Li [24] proposed an identity-based threshold signcryption scheme based on Libert and Quisquater's [25] identity-based signcryption scheme. However, Peng and Li's scheme dose not provide the forward security. In 2008, another scheme was proposed by Li et al. [26] which is more efficient comparing to previous scheme.

In this paper, we show that the threshold signcryption scheme of Li et al. [26] is vulnerable if the attacker can replace the group public key or even the attacker can intercept the intermediate messages. Further, we propose a probably-secure improved scheme to correct the vulnerable and give the unforgeability and confidentiality of our improved scheme under the existing security assumption.

69.2 Review of Fagen Li's ID-Based Threshold Signcryption Scheme

In this section, we review the identity-based threshold signcryption scheme as proposed by Li and Yu [26]. The scheme involves four roles: the PKG, a trust dealer, a sender group $U_A = \{M_1, M_2, \ldots, M_n\}$ with identity ID_A and a receiver Bob with identity ID_B.

Setup. Given a security parameter k, the PKG chooses groups G_1 and G_2 of prime order q (with G_1 additive and G_2 multiplicative), a generator P of G_1, a bilinear map $e : G_1 \times G_1 \to G_2$, a secure symmetric cipher (E, D) and hash functions $H_1 : \{0,1\}^* \to G_1,\ H_2 : G_2 \to \{0,1\}^{n_1},\ H_3 : \{0,1\}^* \to Z_q^*$. The PKG chooses a master-key $s \in {}_R Z_q^*$ and computes $P_{\mathrm{pub}} = \mathrm{sP}$. The PKG publishes system parameters $\{G_1, G_2, n_1, e, P, P_{\mathrm{pub}}, E, D, H_1, H_2, H_3\}$ and keeps the master-key secret.

Extract. Given an identity ID, the PKG computes $Q_{\mathrm{ID}} = H_1(\mathrm{ID})$ and the private key $S_{\mathrm{ID}} = sQ_{\mathrm{ID}}$. Then PKG sends the private key to its owner in a secure way.

Keydis. Suppose that a threshold t and n satisfy $1 \le t \le n < q$. To share the private key S_{ID_A} among the group U_A, the trusted dealer performs the steps below.

(1) Choose $F_1, \ldots, F_{t-1}$ uniformly at random from G_1^*, construct a polynomial $F(x) = S_{\mathrm{ID}_A} + xF_1 + \cdots + x^{t-1}F_{t-1}$ and compute $S_i = F(i)$ for $i = 0, \ldots, n$. Note that $S_0 = S_{\mathrm{ID}_A}$.

(2) Send S_i to member M_i for $i = 1, \ldots, n$ secretly. Broadcast $y_0 = e(S_{\mathrm{ID}_A}, P)$ and $y_j = e(F_j, P)$ for $j = 1, \ldots, t-1$.

(3) Each M_i then checks whether his share S_i is valid by computing $e(S_i, P) = \prod_{j=0}^{t-1} y_j^{i^j}$. If S_i is not valid, M_i broadcasts an error and requests a valid one.

Signcrypt. Without loss of generality, we assume that $M_1, \ldots, M_t$ are the t members who want to cooperate to signcrypt a message m on behalf of the group U_A.

(1) Each M_i chooses $x_i \in {}_R Z_q^*$, computes $R_{1i} = x_iP$ and $R_{2i} = x_iP_{\mathrm{pub}}$, and sends (R_{1i}, R_{2i}) to the clerk C.

(2) The clerk C computes $R_1 = \prod_{i=1}^{t} R_{1i},\ R_2 = \prod_{i=1}^{t} R_{2i},\ \tau = e(R_2, Q_{\mathrm{ID}_B})$, $k = H_2(\tau)$, $c = E_k(m)$, and $h = H_3(m, R_1, k)$. Then the clerk C sends h to M_i for $i = 0, \ldots, t$.

(3) Each M_i computes the partial signature $W_i = x_iP_{\mathrm{pub}} + h\eta_iS_i$ and sends it to the clerk C, where $\eta = \prod_{j=1, j\ne i}^{t} -j(i-j)^{-1} \bmod q$.

(4) When receiving M_i's partial signature W_i, the clerk C verifies its correctness by checking if the following equation holds:

$$e(P, W_i) = e(R_{1i}, P_{\mathrm{pub}})\left(\prod_{j=0}^{t-1} y_j^{i^j}\right)^{h\eta_i}$$

If all partial signatures are verified to be legal, the clerk C computes $W = \sum_{i=1}^{t} W_i$; otherwise rejects it and requests a valid one. The final threshold signcryption is $\sigma = (c,\ R_1,\ W)$.

Unsigncrypt. When receiving σ, Bob follows the steps below.

(1) Compute $\tau = e(R_1,\ S_{\mathrm{ID}_B})$ and $k = H_2(\tau)$.
(2) Recover $m = D_k(c)$.
(3) Compute $h = H_3(m,\ R_1,\ k)$ and accept σ if and only if the following equation holds:

$$e(P,\ W) = e\left(P_{\mathrm{pub}},\ R_1 + hQ_{\mathrm{ID}_A}\right)$$

69.3 Analysis of Fagen Li's ID-Based Threshold Signcryption Scheme

69.3.1 Forgery Attack

Suppose that an attacker can control the communication channel, which means that s/he can gain each user's corresponding ciphertext in the channel and modify or forge it to replace the original one. Then s/he will try to disrupt the scheme as follows:

All the attack process will be finished in the Signcrypt stage. We describe it as follows:

(1) The attacker records $(R_{1i},\ R_{2i})$ sent from M_i for $i = 0,\ \ldots,\ t$.
(2) The attacker intercepts h sent from clerk C. Then s/he computes $R_1 = \prod_{i=1}^{t} R_{1i}$ and $R_2 = \prod_{i=1}^{t} R_{2i}$ using $(R_{1i},\ R_{2i})$, computes $\tau = e(R_2,\ Q_{\mathrm{ID}_B})$, and $k = H_2(\tau)$. Further, s/he selects a message $m^{'}$ which s/he wants to forge, computes $c^{'} = E_k\left(m^{'}\right)$, and $h^{'} = H_3\left(m^{'},\ R_1,\ k\right)$. Finally, s/he sends $h^{'}$ to M_i for $i = 0,\ \ldots,\ t$.
(3) The attacker intercepts W_i sent from M_i, for $i = 0,\ \ldots,\ t$. Note that the message $W_i = x_i P_{\mathrm{pub}} + h^{'} \eta_i S_i$ here. Then s/he computes

$$\begin{aligned} W_i^{'} &= R_{2i} + (W_i - R_{2i}) \cdot h \Big/ h^{'} \\ &= x_i P_{\mathrm{pub}} + \left(x_i P_{\mathrm{pub}} + h^{'} \eta_i S_i - x_i P_{\mathrm{pub}}\right) \cdot h \Big/ h^{'} \\ &= x_i P_{\mathrm{pub}} + h \eta_i S_i \end{aligned}$$

and send $W_i^{'}$ to clerk C.

(4) Because $W_i' = x_i P_{\text{pub}} + h\eta_i S_i$. The verification function $e(P, W_i') = e(R_{1i}, P_{\text{pub}})\left(\prod_{j=0}^{t-1} y_j^{i^j}\right)^{h\eta_i}$ will hold. Then the clerk C will send $\sigma = (c, R_1, W')$, where $W' = \sum_{i=1}^{t} W_i'$, to the receiver.

The attacker intercepts $\sigma = (c, R_1, W')$, computes $W = \sum_{i=1}^{t} W_i$, and sends $\sigma' = (c', R_1, W)$ to the receiver.

In the Unsigncrypt stage:

After receiving σ', the receiver Bob executes the following steps

(1) He will compute $\tau = e(R_1, S_{\text{ID}_B})$ and $k = H_2(\tau)$.
(2) He will recover $m' = D_k(c')$.
(3) He will compute $h' = H_3(m', R_1, k)$ here. Then the equation $e(P, W) = e(P_{\text{pub}}, R_1 + h' Q_{\text{ID}_A})$ will hold. Because

$$\begin{aligned} e(P, W) &= e\left(P, \sum_{i=1}^{t} W_i\right) \\ &= e\left(P, \sum_{i=1}^{t}\left(x_i P_{\text{pub}} + h'\eta_i S_i\right)\right) \\ &= e\left(P, R_2 + h' S_{\text{ID}_A}\right) \\ &= e\left(P_{\text{pub}}, R_1 + h' Q_{\text{ID}_A}\right) \end{aligned}$$

So the receiver accepts the forged message m'.

69.3.2 Key Replacement Attack

Li et al. scheme is insecure from the view of a malicious attacker who can control the communication channel.

The attacker intercepts the ciphertext $\sigma = (c, R_1, W)$ from sender.

(1) Randomly choose $\alpha, x \in Z_q^*$ and prepare a forged message m'.
(2) Compute $R_1' = xP$, $R_2' = xP_{\text{pub}}$, $\tau' = e(R_2', Q_{\text{ID}_B})$, $k' = H_2(\tau)$, $c' = E_{k'}(m')$, $h' = H_3(m', R_1', k')$.
(3) Compute $W' = \alpha P_{\text{pub}}$, set $Q_A' = (\alpha - x)P/h'$ as a public key of U_A.
(4) The final ciphertext is $\sigma' = (c', R_1', W')$.

(5) Attacker sends the forged ciphertext and the replaced public key to the receiver.

After receiving the ciphertext $\sigma^{'} = \left(c^{'},\ R_1^{'},\ W^{'}\right)$, the receiver.

(1) Compute $\tau = e\left(R_1^{'},\ S_{\mathrm{ID}_B}\right) = e\left(R_2^{'},\ Q_{\mathrm{ID}_B}\right) = \tau^{'}$, $k = H_2(\tau) = H_2\left(\tau^{'}\right) = k^{'}$.
(2) Recover $m = D_k\left(c^{'}\right) = D_{k^{'}}\left(c^{'}\right) = m^{'}$, $h = H_3\left(m^{'},\ R_1^{'},\ k^{'}\right) = h^{'}$.
(3) Verify $e\left(P,\ W^{'}\right) \overset{?}{=} e\left(P_{\mathrm{pub}},\ R_1^{'} + hQ_{\mathrm{ID}_A}^{'}\right)$

$$\because e\left(P_{\mathrm{pub}},\ R_1^{'} + hQ_{\mathrm{ID}_A}^{'}\right) = e\left(P_{\mathrm{pub}},\ xP + h\cdot(\alpha - x)P/h^{'}\right) = e\left(P_{\mathrm{pub}},\ \alpha P\right) = e\left(P,\ W^{'}\right)$$

$\therefore$ The equation $e\left(P,\ W^{'}\right) = e\left(P_{\mathrm{pub}},\ R_1^{'} + hQ_{\mathrm{ID}_A}^{'}\right)$ set.

In the view of the attacker, [26] can be simulated as following basic Signcryption scheme:

A sender “Alice” with key pairs $\{Q_{\mathrm{Alice}} = H_1(\mathrm{Alice}),\ S_{\mathrm{Alice}} = sH_1(\mathrm{Alice})\}$.

A receiver “Bob” with key pairs $\{Q_{\mathrm{Bob}} = H_1(\mathrm{Bob}),\ S_{\mathrm{Bob}} = sH_1(\mathrm{Bob})\}$.

Alice chooses $x \in Z_q^*$, $R_1 = xP$, $R_2 = xP_{\mathrm{pub}}$, $\tau = e(R_2,\ Q_{\mathrm{Bob}})$, $k = H_2(\tau)$, $c = E_k(m)$, $h = H_3(m,\ R_1,\ k)$, $W = xP_{\mathrm{pub}} + hS_{\mathrm{Alice}}$ and sends $\sigma = (c,\ R_1,\ W)$ to Bob as the ciphertext of his message.

There is a small mistake of the definition $H_3 : \{0,\ 1\}^* \rightarrow Z_q^*$. We think the authors’ real intention is $H_3 : \{0,\ 1\}^* \times G_1 \times \{0,\ 1\}^* \rightarrow Z_q^*$ to meet $h = H_3(m,\ R_1,\ k)$. In this hash function, any message about the sender is not contained. If an attacker Eve says “I am Alice” to Bob, Bob can not distinguish with only the hash value h. Our attacker just utilizes this attribute of Li’s scheme.

Suppose that H_3 is defined as $H_3 : \{0,1\}^* \times G_1 \times \{0,1\}^* \times G_1 \rightarrow Z_q^*$, and $h = H_3(m,\ R_1,\ k,\ Q_{\mathrm{Alice}})$. The attacker Eve intercepts the ciphertext $\sigma = (c,\ R_1,\ W)$ from sender Alice and s/he runs the algorithm of forging ciphertext like:

(1) Randomly choose $\alpha, x \in Z_q^*$ and prepare a forged message $m^{'}$.
(2) Compute $R_1^{'} = xP$, $R_2^{'} = xP_{\mathrm{pub}}$, $\tau^{'} = e\left(R_2^{'},\ Q_{\mathrm{Bob}}\right)$, $k^{'} = H_2(\tau)$, $c^{'} = E_{k^{'}}\left(m^{'}\right)$, $h^{'} = H_3\left(m^{'}, R_1^{'},\ k^{'},\ Q_{\mathrm{Alice}}^{'}\right)$.
(3) Compute $W^{'} = \alpha P_{\mathrm{pub}}$, set $Q_{\mathrm{Alice}}^{'} = (\alpha - x)P/h^{'}$ as a public key of U_A.
(4) The final ciphertext is $\sigma^{'} = \left(c^{'},\ R_1^{'},\ W^{'}\right)$.
(5) Send the forged ciphertext and the replaced public key to the receiver.

S/he will meet a hard problem that if s/he wants to compute $h^{'}$, $Q_{\mathrm{Alice}}^{'}$ is necessary or if s/he wants to computes $Q_{\mathrm{Alice}}^{'}$, $h^{'}$ must be known. As a result, if s/he can succeed in forging the ciphertext, s/he must own the ability to solve the DL problem.

69.4 Conclusion

In this paper, we show that the threshold signcryption scheme of Li et al. [26] is vulnerable if the attacker can replace the group public key. Then we point out that the receiver uses the senders' public key without any verification in the unsigncrypt stage to cause this attack.

Acknowledgments The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318), and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Zheng Y (1997) Digital signcryption or How to achieve cost (signature & Encryption) $\ll$ cost (signature) + cost (encryption). In: Proceedings of the Advances in CRYPTO'97, LNCS 1294:165–179
2. Bao F, Deng RH (1997) A signcryption scheme with signature directly verifiable by public key. PKC'98. LNCS 1431:55–59
3. Chow SSM, Yiu SM, Hui LCK, Chow KP (2004) Efficient forward and provably secure ID-based signcryption scheme with public verifiability and public ciphertext authenticity. ICISC'03. LNCS 2971:352–269
4. Boyen X (2003) Multipurpose identity based signcryption: a swiss army knife for identity based cryptography. CRYPT'03. LNCS, 2729:383–399
5. Mu Y, Varadharajan V (2000) Distributed signcryption, INDOCRYPT'00. LNCS 1977:155–164
6. Yang G, Wong DS, Deng X (2005) Analysis and improvement of a signcryption scheme with key privacy. ISC'05. LNCS 3650:218–232
7. SteinFeld R, Zheng Y (2000) A signcryption scheme based on integer factorization. ISW'00. LNCS 1975:308–322
8. Libert B, Quisquater J (2004) Efficient signcryption with key prevacy from gap Diffie–Hellman groups. PKC'04 LNCS 2947:187–200
9. Baek J, Steinfeld R, Zheng Y (2002) Formal proofs for the security of signcryption PKC'02. LNCS 2274:80–98
10. Shamir A (1984) Identity-based cryptosystems and signature schemes. CRYPTO'84 LNCS 196:47–53
11. Boneh D, Franklin M (2001) Identity-based encryption from well pairing CRYPTO'01. LNCS 2139:213–229
12. Barreto PSLM, Libert B, Mccullagh N, Quisquater JJ (2005) Efficient and provably-secure identity-based signatures and signcryption from bilinear maps ASIACRYPT'05 LNCS 3788:515–532
13. Li F, Xiong H, Nie X (2009) A new multi-receiver ID-based signcryption scheme for group communications, ICCCAS'2009, pp 296–300
14. Han Y, Gui X (2009) Multi-recipient signcryption for secure group communication, ICIEA pp 161–165
15. Jin Z, Wen Q, Du H (2010) An improved semantically-secure identity-based signcryption scheme in the standard model. Comput Electr Eng 36(2010):545–552

16. Huang X, Susilo W, Mu Y, Zhang E (2005) Identity-based ring signcryption schemes: cryptographic primitives for preserving privacy and authenticity in the ubiquitous world. 19th International Conference on Advanced Information Networking and Applications, Taiwan, pp 649–654
17. Liu Z, Hu Y, Zhang X, Ma H (2010) Certificateless signcryption scheme in the standard model. Inform Sci 180 (2010):452–464
18. Yu Y, Yang B, Sun Y, Zhu S-l (2009) Identity based signcryption scheme without random oracles. Comput Stand Interf 31(2009):56–62
19. Desmedt Y (1987) Society and group oriented cryptography: a now concept, CRYPTO'87. LNCS 293:120–127
20. Des Frankel Y (1991) Shared generation of authenticators and signatures, CRYPTO'91. LNCS 576:457–469
21. Rivest RL, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. Commun ACM 21(2):120–126
22. Baek J, Zheng Y (2004) Identity-based threshold signature scheme from the bilinear pairings. International conference on information technology 2004, Las Vegas, USA, pp 124–128
23. Duan S, Cao Z, Lu R (2004) Robust ID-based threshold signcryption scheme from pairings. International conference on information security. Shanghai, China, pp 33–37
24. Peng C, Li X (2005) An identity-based threshold signcryption scheme with semantic security. Comput Int Secur 2005, LNAI 3902:173–179
25. Libert B, Quisquater JJ (2003) Anew identity based signcryption schemes from pairings, 2003 IEEE information theory workshop. Paris, France, pp 155–158
26. Li F, Yu Y (2008) An efficient and Provably Secure ID-Based Threshold Signcryption Scheme, ICCCAS 2008, pp 488–492

Chapter 70
Provably Secure Cross-Realm Client-To-Client Password-Authenticated Key Exchange Protocol

Wenmin Li, Qiaoyan Wen and Qi Su

Abstract The cross-realm client-to-client password-authenticated key exchange protocol, which was proposed by Byun et al. in 2002, allows two clients in different realms with different passwords to agree on a common session key through their corresponding servers. From then on, many research works have been done. In this paper, we define a new formal security model of cross-realm C2C-PAKE which captures more desirable security requirements than the previous models. Then, we construct a new cross-realm C2C-PAKE key exchange protocol based on quadratic residues and prove its security in the model. In addition, both the communication complexity and the computational complexity are reduced in our protocol.

Keywords Password · Key exchange · Cross-realm · Provable security

70.1 Introduction

With rapid changes in the modern communication environment such as ad hoc networks, mobile network and ubiquitous computing, it is necessary to construct a secure end-to-end channel between clients in cross-realm setting. The first protocol in the setting is proposed in Ref. [1] by Byun et al. in 2002. Unfortunately, dictionary attack is found against this protocol by [2]. Subsequent works [3–6] include other attacks and a few variants either to resist existing attacks or to improve the efficiency. However, all these variants were designed with heuristic

W. Li (✉) · Q. Wen · Q. Su
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China
e-mail: liwenmin02@gmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_70,

security analysis. In Ref. [7], Byun et al. provided formal treatments for cross-realm C2C-PAKE protocol and introduced a protocol with security proof. Subsequently, Wang et al. [8] pointed out that protocol in Ref. [7] is vulnerable to password compromise impersonate attack and man-in-the-middle attack if the key between servers is compromised. In Ref. [9], Phan et al. showed undetectable online dictionary attack by any adversary on the same protocol. In addition, Ref. [10] pointed out the flaws in Byun et al.'s security model. Besides, [10] modified formal security model and proposed a new cross-realm C2C-PAKE protocol with security proof. However, the public-key encryption used in the protocol is time-consuming, the number of steps is not optimal and several security properties are not captured in the improved security model.

In this paper, our work is divided into two parts. First, we define a new stronger security model for cross-realm C2C-PAKE than previous models. In addition to this, we show how the model captures the security properties. Our model is based on the recent formal models [10–12]. Compared with previous models, our model can not only capture forward security but also resistance to password compromise impersonation, unknown-key share and so on. Second, we construct a new cross-realm client-to-client password-authenticated key exchange protocol using quadratic residues. In our protocol, the interaction between client and the corresponding server is based on the protocol in [13]. Number–theoretic techniques provide additional security assumption; therefore, password is no longer the only way to verify the identity of the client. So we can make counter measures against adversary's attacks without using public-key system. At the same time, both the communication complexity and the computational complexity are reduced.

The remainder of this paper is organized as follows. In Sect. 70.2, we introduce the security requirement of cross-realm C2C-PAKE and define a security model for the cross-realm C2C-PAKE in Sect. 70.2.2. In Sect. 70.3, we present the computational assumptions upon which the security of our protocol is based. Then, in Sect. 70.3.2, we describe our cross-realm C2C-PAKE protocol, along with security proof and efficiency analysis. Finally, we conclude in Sect. 70.4.

70.2 Security Model

In this section, we describe how an adversary will be allowed to interfere in the protocol. Our work integrate previous works [10–12] on password-authenticated key exchange protocols by considering the following properties in a single model:

Indistinguishability. This notion provides security properties with respect to session keys, i.e., key secrecy (KS), forward secrecy (FS), key privacy (KP) and resistance to password compromise impersonation (PCI) attack and unknown key share (UKS) attack.

Password Protection. This notion provides security properties with respect to passwords, i.e., resistance to undetectable on-line dictionary attack (UDonDA) and to off-line dictionary attack (offDA).

70.2.1 Adversary Capabilities

The adversary interacts with the protocol via the following oracle queries:

Execute $\left(\Pi^i_U, \Pi^j_{U'}\right)$. This query models passive attacks. It outputs the whole transcript of an honest execution of the protocol.

Send $\left(\Pi^i_U, m\right)$. This query models active attacks. It outputs the message that instance Π^i_U would generate upon receiving such a messagem.

Re*veal* $\left(\Pi^i_U\right)$. This query models misuses of session keys. The output of the query is the session key for client Π^i_U, if it has ever been defined. Otherwise, return $\perp$.

Corrupt $\left(\Pi^i_U\right)$. This query models exposure of the password held by the client U or the server U. The output of this query is the password of instance Π^i_U. Note that we consider weak corruptions only, where the long term secrets are revealed, but not the internal states.

Establish Party (U, S, pw_U). This query models the adversary to register a static secret pw_U on behalf of a client. In this way the adversary totally controls that client. Clients against whom the adversary did not issue this query are called honest.

TestPassword (U, pw'). This query does not model the adversarial ability, but no leakage of the password. If the guess password pw' is just the same as the client U's password pw, then return 1. Otherwise, returns 0. Note that, the adversary can TestPassword query only once at any time during the experiment.

Test $\left(\Pi^i_U\right)$. This query is only used to measure an adversary's knowledge about a session key. If no session key is defined for instance Π^i_U or if instance Π^i_U is not fresh (see notion of freshness below), then this query is answered by $\perp$. If this query has already been asked, then it outputs the same answer. Otherwise, a coin is flipped to generate a random bit b. If $b = 1$, it outputs the real session key $\left(\text{from Reveal } \left(\Pi^i_U\right)\right)$, and if $b = 0$, it outputs a random one of the same size.

70.2.2 Security Definitions

Following are some definitions needed in the security proof [10, 12]:

SID. The value of sid is taken to be the partial transcript of the communication between the clients before the session key has been accepted.

Partner. Two instances Π^i_U and $\Pi^j_{\bar{U}}$ are said to be partners if and only if the following four conditions hold: (1) Both Π^i_U and $\Pi^j_{\bar{U}}$ are accepted; (2) Both Π^i_U and $\Pi^j_{\bar{U}}$ share the same side; (3) The partner for Π^i_U is $\Pi^j_{\bar{U}}$ and vice versa; (4) No instance other than Π^i_U accepts partner $\Pi^j_{\bar{U}}$ and vice versa.

Freshness. We say an oracle instance Π_U^i is fresh if the following conditions hold: (1) It has accepted and generated a valid session key; (2) No Reveal queries have been made to Π_U^i or its partner; (3) U's partner (and the server of U's partner) is not corrupted (not being issued the Corrupt query); and (4) If U (or the server of U) is corrupted, then Send$\left(\Pi_{\bar{U}}^j, m\right)$ query and Send $\left(\Pi_{\bar{S}}^t, m\right)$ should not be made, where $\bar{U}$ is U's partner, $\bar{S}$ is the server of $\bar{U}$, and m is a message chosen by an adversary.

Oracle-Generated Messages. We say that m is an oracle-generated-message if there exists an instance Π_U^i, for a participant $\Pi_{\bar{U}}^j \in$ U such that $m =$ Send$\big(\Pi_U^i$, $\Pi_{\bar{U}}^j$; m') for some message m'.

Indistinguishability. An outside adversary A, against protocol P, is allowed to make Execute, Send, Corrupt, Reveal, Establish Party-queries, as well as Test queries to fresh instances only and outputs a guess bit b_0. Let Succ$^{\text{ind}}$ denote the event that $b_0 = b$, where b is the random bit chosen in the Test-query. Then, we define the advantage of A by:

$\text{Adv}_{\text{P}}^{\text{ind}}(\text{A}) = 2Pr\left[\text{Succ}^{\text{ind}}\right] - 1$ and $\text{Adv}_{\text{P}}^{\text{ind}}(t, R) = \max_{\text{A}}\left\{\text{Adv}_{\text{P}}^{\text{ind}}(\text{A})\right\}$, where the maximum is over all A with time-complexity at most t and using the number of queries to oracles at most R.

We say that a cross-realm client-to-client password-based key exchange protocol P satisfies indistinguishability of the session key, if $\text{Adv}_{\text{P}}^{\text{ind}}(t, R) < \frac{c \cdot q}{N} + \text{negl}()$, where N, is the size of the dictionary, where the passwords are uniformly drawn, and c is a small constant (ideally 1).

Password protection. The notion of indistinguishability cannot capture insider attacks; therefore, we consider the notion of password protection against malicious clients or malicious server from other realm. This notion provides security properties with respect to passwords, i.e., resistance to UDonDA and to offDA.

The adversary is allowed to make Execute, Send, Corrupt, Reveal, and Test-Password queries. Let Succ$^{\text{pw}}$ denote the event that TestPassword outputs 1. Note that, we restrict the adversary such that U and the corresponding server are honest, and neither of them is corrupt. If the adversary guesses a password of a client correctly, it is considered successful. The advantage of A is defined as following: $\text{Adv}^{\text{pw}}(\text{A}) = \text{Pr}[\text{Succ}^{\text{pw}}]$ and $\text{Adv}^{\text{pw}}(t, R) = \max_{\text{A}}\{\text{Adv}^{\text{pw}}(\text{A})\}$ where the maximum is over all A with time-complexity at most t and using the number of queries to its oracle at most R.

We say that a cross-realm client-to-client password-based key exchange protocol P satisfies password protection against malicious clients or malicious server from other realm, if $\text{Adv}^{\text{pw}}(t,\ R) < \frac{c \cdot q}{N} + \text{neg}l()$, where N is the size of the dictionary, the passwords are uniformly drawn, c is a small constant, and q is the number of sent queries in which messages are found as "invalid" by the target client.

Table 70.1 The comparison between previous security models and ours

	FS	PCI	UKI	UDonDA
Ref. [7]	Yes	No	No	No
Ref. [10]	No	Yes	No	Yes
Ours	Yes	Yes	Yes	Yes

70.2.3 Comparison

The security model in [7] cannot grasp the notion of UDonDA, and resistance to PCI and UKS attack. Besides, FS and resistance to UKS are also out of scope in the security model of [10]. The reason is that adversary capabilities do not include any query for corruption of parties in the test session. Therefore, conditions of UKS and FS cannot be represented. The comparison between previous security models and ours is shown in Table 70.1.

70.3 Our Protocol

70.3.1 Protocol Description

In this section, we propose an efficient C2C-PAKE protocol. In an instance of the protocol, there are four entities, denoted as A, S_A, B and S_B, respectively, where A (B) is a client in the realm of server S_A (S_B).

Initialization phase. Each client shares his (her) password pw with corresponding server S and learn the selected Blum number N by using algorithms $\mathrm{G_{pw}}$ and R. The initialization process also specifies a set of cryptographic function (e.g., hash function) and sets a number of cryptographic parameters.

Protocol description. The concrete protocol is described as follows:

1. A chooses random values $\alpha_A \in Q_{N_A}$, $r_A \in \{0,1\}^k$, $x_A \in Z_N^*$ and computes $\gamma_A = H(\mathrm{pw}_A,\ r_A,\ \mathrm{ID}_{S_A},\ \mathrm{ID}_A,\ N_A)$. If $\gcd(\gamma_A, N_A) = 1$, A assigns γ_A to λ_A; otherwise, A assigns a random number of $Z_{N_A}^*$to λ_A. Next, A computes $Z_A = \left(\lambda_A \alpha_A^2\right)^{2^{tA-1}} \bmod N_A$ and $k_A = H_1(\mathrm{pw}_A,\ \alpha_A,\ r_A,\ \mathrm{ID}_{S_A},\ \mathrm{ID}_A,\ N_A)$. Subsequently, A generates $C_A = [g^{x_A},\ \mathrm{pw}_A]_{k_A}$ and sends $\{\mathrm{ID}_A,\ r_A,\ z_A,\ C_A\}$ to S_A.
2. Upon receiving the message, S_A computes $\gamma_A = H(\mathrm{pw}_A,\ r_A,\ \mathrm{ID}_{S_A},\ \mathrm{ID}_A,\ N_A)$ and checks if $\gcd(\gamma_A,\ N_A) = 1$. If $\gcd(\gamma_A,\ N_A) \neq 1$, S_A chooses $\beta_A \in Z_{N_A}^*$ randomly; otherwise, S_A computes $\beta_A \in Q_{N_A}$ satisfying $\left(\gamma_A \beta_A^2\right)^{2^{tA-1}} = z_A \bmod N_A$. Next, S_A computes $k'_A = H_1(pw_A,\ \beta_A,\ r_A,\ \mathrm{ID}_{S_A},\ \mathrm{ID}_A,\ N_A)$ and obtain g^{x_A}, $pw'_A$by decrypting C_A with k'_A. If $pw'_A \neq pw_A$, S_A aborts the session,

otherwise, S_A chooses $s_A \in Z_N^*$, generates $\mathrm{Ticket}_B = [\mathrm{ID}_A, g^{s_A x_A}, g^{s_A}, L]_K$ and $[g^{s_A}]_{k'_A}$, and sends $\left\{Ticket_B,\ [g^{s_A}]_{k'_A}\right\}$ to A.

3. Upon receiving the message from S_A, A decrypts $[g^{s_A}]_{k'_A}$ by k_A. Then A computes $g^{s_A x_A}$ and forwards $\{\mathrm{ID_A}, \mathrm{Ticket_B}\}$ to B.
4. B performs similar operation as A did in step 1, generates k_B and sends $\{\mathrm{ID}_A,\ \mathrm{ID}_B,\ r_B,\ z_B,\ \mathrm{Ticket}_B\}$ to S_B.
5. Upon receiving the message from B, S_B computes γ_B and $\mathrm{k'_B}$ as S_A did in step 2. Next, S_B decrypts Ticket_B using K to obtain $ID_A,\ g^{s_A x_A},\ g^{s_A},\ L$. Then S_B verifies the validity of Ticket_B by checking the lifetime L and $\mathrm{ID_A}$. Finally, S_B chooses s_B randomly, generates $C_B = [\mathrm{ID}_A,\ \mathrm{ID}_B,\ g^{s_A x_A s_B},\ g^{s_A s_B},\ g^{s_A x_A},\ \mathrm{pw}_B]_{k'_B}$ and sends C_B to B.
6. B decrypts C_B to obtain $g^{s_A x_A s_B}, g^{s_A s_B}, g^{s_A x_A}, \mathrm{pw}'_B$. If $\mathrm{pw}'_B \neq \mathrm{pw}_B$, B aborts the session, otherwise, B selects $x_B \in Z_N^*$ randomly, computes $g^{s_A s_B x_A x_B}$, $g^{s_A s_B x_B}$ and $M_B = H_2(\mathrm{ID}_B, \mathrm{ID}_A, g^{s_A x_A}, \mathrm{cs})$, and sends $g^{s_A s_B x_B}$, M_B to A.
7. Upon receiving the message $g^{s_A s_B x_B}, M_B$, A computes $g^{s_A s_B x_A x_B}$ and verify the validity of M_B. If it is invalid, A aborts the session. Otherwise, A computes $\mathrm{sk} = H_3(\mathrm{ID}_A, \mathrm{ID}_B, \mathrm{ID}_{S_A}, \mathrm{ID}_{S_B}, g^{s_A s_B x_A x_B})$, generates $M_A = H_2(\mathrm{ID}_A, \mathrm{ID}_B, g^{s_A x_A}, g^{s_A s_B x_A x_B})$ and sends $M_A$to A for key conformation.
8. B authenticates A by the validity of the received message M_A. If it is invalid, B aborts the session, otherwise, generate a common session key sk.

70.3.2 Security Result

Theorem 3.1 *The proposed cross-realm C2C-PAKE protocol satisfies indistinguishability, provided that the CDH, DDH* [14] *and factoring assumption hold, the underlying symmetric encryption scheme is secure.*

Proof In this proof, we are interested in the event S_n which occurs if the adversary correctly guesses the bit b involved in the Test-queries. Our proof uses a sequence of games, starting with the real attack and ending in a game in which the adversary's advantage is 0. Each game addresses a different security aspect. Game G_0 is the real protocol in the random oracle and ideal cipher models. In Game G_1, we simulate hash oracles (H, H_1, H_2 and H_3) and the ideal encryption and decryption oracles (E, D) by maintaining corresponding lists Λ_H, Λ_{H_1}, Λ_{H_2}, Λ_{H_3}, Λ_E and Λ_D. The Execute, Reveal, Send, Test and Corrupt Oracles are also simulated as in the real attack in the same game. In Game G_2, we cancel games in which collisions on the transcripts, the output of H_1, H_2, H_3, H, E and D appear. The adversary's advantage in this game can be computed according to the birthday paradox. In Game G_3, we replace the ephemeral key k_A, k_B with random keys r_{k_A}, r_{k_B}. The passwords are no longer used in computing the ephemeral keys by client A (B) and the corresponding server $S_A(S_B)$. The success probability of the adversary in distinguishing the difference between k_A and r_{k_A} is smaller than $\frac{q_{\text{send}}}{|D|} + \mathbf{negl}()$ as shown in [13]. In Game G_4, we replace the pre-shared key K by r_K. The difference between Games G_4 and G_3 can not be detected as long as the symmetric encryption is security. In the last game G_5, we injected a random DDH triple into the protocol, then the triple is used to compute the session key. We show that the computed session key is indistinguishable from random values in this game.

Theorem 3.2 *Assuming the underlying symmetric encryption scheme is semantically secure, and then our scheme satisfies password protection.*

Proof In this proof, we show that the malicious client insiders, who can trivially know the session key, can not learn the password of other clients. The password is used only in two steps. First, it is used as the input of the Hash function, together with a private random number. Secondly, the password is been encrypted by the symmetric encryption. It is impossible for the malicious client insiders to obtain the private random number or the encryption key. Consequently, the proposed protocol has password protection against malicious insiders.

70.3.3 Efficiency analysis

In this section, we analyze the computational and communicational complexities of the proposed C2C-PAKE protocol, which need for deployment in a practical circumstance. First, the proposed protocol has at least one step less than the other protocols [7, 10]. Second, in order to improve computational efficiency, the proposed protocol does not require server's public key and use symmetric encryption scheme. Third, the interactive procedure during the registration process increases the communication overhead on participates and some power-consuming operations could also be pre-computed in the process.

70.4 Conclusion

In this paper, we first investigate the design of formal security model for C2C-PAKE and propose a new model. Then we present an efficient C2C-PAKE protocol based on number-theoretic techniques and prove its security in the model. In particular, the proposed protocol is implemented without the server's public key. Therefore, it is more suitable to an imbalanced computing environment where a low-end client device communicates with a powerful server over a broadband network.

Acknowledgments This work is supported by National Natural Science Foundation of China (Grant Nos. 60873191, 60903152, 60821001), Beijing Natural Science Foundation (Grant No. 4072020).

References

1. Byun J, Jeong I, Lee D, Park C (2002) Password-authenticated key exchange between clients with different passwords. In: Proceedings of ICICS02, LNCS, vol 2513, Springer, London, pp 134–146
2. Chen L. A weakness of the password-authenticated key agreement between clients with different passwords scheme, ISO/IEC JTC 1/SC27 N3716. Inf Sci 177(19):211–216
3. Byun J, Jeong I, Lee D, Park C (2003) Password-authenticated key agreement between clients with different passwords, working draft of ISO/IEC 11770-4 in document 27 N 3576
4. Kim J, Kim S, Kwak J, Won D (2004) Cryptoanalysis and improvements of password authenticated key exchange scheme between clients with different passwords. In: Proceedings of ICCSA 2004. LNCS 3044:895–902
5. Phan RC-W, Goi B-M (2005) Cryptanalysis of an improved client-to-client password-authenticated key exchange (C2C-PAKE) scheme. In: Proceedings of ACNS 2005. LNCS 3531:33–39
6. Wang S, Wang J, Xu M (2004) Weakness of a password-authenticated key exchange protocol between clients with different passwords. In: Proceedings of ACNS 2004. LNCS 3089: 414–425
7. Byun J, Lee D, Lim J (2007) EC2C-PAKA: an efficient client-to-client password-authenticated key agreement. Inf Sci 177(19):3995–4013
8. Wang J, Zhang Y. Cryptanalysis of a client-to-client password-authenticated key agreement protocol, eprint.iacr.org/2008/248.pdf
9. Phan RC-W, Goi B-M (2006) Cryptanalysis of two provably secure cross-realm C2C-PAKE protocols. In: Proceedings of INDOCRYPT 2006. LNCS 4329:104–117
10. Feng D, Xu J (2009) A new client-to-client password-authenticated key agreement protocol. In: Proceedings of IWCC 2009. LNCS 5557:63–76
11. Yoneyama K (2008) Efficient and strongly secure password-based server aided key exchange. In: Proceedings of INDOCRYPT 2008. LNCS 5365:172–184
12. Abdalla M, Izabachéne M, Pointcheval D (2008) Anonymous and transparent gateway-based password-authenticated key exchange. In: Proceedings of CANS'08. LNCS 5339:133–148
13. Zhang M (2007) Computationally-efficient password authenticated key exchange based on quadratic residues. In: Proceedings of INDOCRYPT 2007. LNCS 859:312–321
14. Abdalla M, Chevassut O, Fouque P-A, Pointcheval D (2005) A simple threshold authenticated key exchange from short secrets. In: Proceedings of Asiacrypt'05. LNCS 3788:566–584

Chapter 71
A Research on SRC Plus Homotopy in Disguised Face Recognition

Junying Gan and Peng Wang

Abstract Disguised face recognition is a great challenge to general face recognition systems, since a variety of disguises, to some degree, corrupt the information needed in identification. This chapter presents sparse representation-based classification (SRC) combined with homotopy algorithm to cope with disguised face recognition. To represent the face image sparsely, SRC constructs the over-complete dictionary using training samples as atoms, and we employ homotopy to compute the expansion coefficients effectively. Experimental results based on Aleix Martinez and Robert Benavente (AR) face database show the validity of SRC combined with homotopy algorithm in face recognition.

Keywords Disguised face recognition · SRC · Homotopy

71.1 Introduction

A practical face recognition system should be capable of identifying individuals who use disguises or occlusion to deliberately alter their appearances. Researchers have proposed several approaches to solve the problems of variation in pose, illumination, and expression [1, 2]. However, very few researchers have well handled the challenge of face recognition when an individual hides one's identity

J. Gan (✉) · P. Wang
School of Information Engineering, Wuyi University,
Jiangmen 529020, Guangdong, China
e-mail: junyinggan@163.com

P. Wang
e-mail: 2009wp2012@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_71,

using disguises [3]. In the fields of statistical signal processing, sparse linear representations with respect to an overcomplete dictionary of base elements or signal atoms [4, 5], as a preprocessing method before signal analysis, has been a hot topic in the recent five years [6].

In this chapter, sparse representation-based classification (SRC) algorithm is used to perform disguised face recognition. Experimental results show that whenever the optimal representation is sufficiently sparse, it can be efficiently solved by convex optimization [4], which can be formulated as an l_1-minimization problem [7, 8]. Moreover, the l_1-minimization problem can be well solved by homotopy algorithm [9]. In view of this, we combine SRC and homotopy to perform the disguised face recognition.

71.2 Approaches Presented

SRC. To convert every $w \times h$ face image into the corresponding vector $\mathbf{x} \in \mathbf{R}^m$, where $m = wh$, one can concatenate its columns orderly. Suppose there are n_i training samples belonging to the ith class, represented by $\mathbf{X}_i = \left[\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,j}, \ldots, \mathbf{x}_{i,n_i}\right] \in \mathbf{R}^{m \times n_i}$, where $\mathbf{x}_{i,j} \in \mathbf{R}^m, j = 1, 2, \ldots, n_i$ is the vector associated with the jth face image from the ith class. Taking sufficient training samples belonging to the ith class $\mathbf{X}_i = \left[\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,j}, \ldots, \mathbf{x}_{i,n_i}\right] \in \mathbf{R}^{m \times n_i}$ into account, any test sample $\mathbf{y} \in \mathbf{R}^m$ from the same class will approximately reside in the linear subspace spanned by the training samples associated with the ith class. Thus one can get

$$\mathbf{y} = a_{i,1}\mathbf{x}_{i,1} + a_{i,2}\mathbf{x}_{i,2} + \cdots + a_{i,n_i}\mathbf{x}_{i,n_i} = \mathbf{X}_i\mathbf{a}_i, \tag{71.1}$$

where $\mathbf{a}_i = [a_{i,1}, a_{i,2}, \cdots, a_{i,n_i}]^{\mathrm{T}} \in \mathbf{R}^{n_i}$ is a coefficient vector. However, the membership of the test sample $\mathbf{y}$ is initially unknown, and we construct an augmented matrix $\mathbf{X}$ that collects the n training samples of all k classes involved:

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k] \in \mathbf{R}^{m \times n}, \tag{71.2}$$

where $n = \sum_{i=1}^{k} n_i$. Then, the linear expansion of $\mathbf{y}$ can be rewritten with respect to all the training samples by

$$\mathbf{y} = \mathbf{X}_1\mathbf{a}_1 + \mathbf{X}_2\mathbf{a}_2 + \cdots + \mathbf{X}_k\mathbf{a}_k = \mathbf{XA} \in \mathbf{R}^m, \tag{71.3}$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k]^{\mathrm{T}} = \left[0, \ldots, 0, a_{i,1}, a_{i,2}, \ldots, a_{i,n_i}, 0, \ldots, 0\right]^{\mathrm{T}} \in \mathbf{R}^n$, whose elements should be zero in terms of theory except those associated with the ith class.

In practical face recognition, since $m < n$, the equation $\mathbf{y} = \mathbf{XA}$ is typically underdetermined, then its solution is not unique. Conventionally, this difficulty is resolved by choosing the solution in terms of the minimum of l_1-norm:

$$\mathbf{A}_{\text{opt}} = \arg\min\|\mathbf{A}\|_1 \text{ subject to } \mathbf{y} = \mathbf{X}\mathbf{A} \tag{71.4}$$

Homotopy. Obviously, formula (71.4) can be reconstructed as an unconstrained optimization problem in formula (71.5):

$$\begin{aligned}\mathbf{A}_{\text{opt}} &= \arg\min_a J(\mathbf{A}, \lambda)\\ &= \arg\min_a \left\{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{A}\|_2^2 + \lambda\|\mathbf{A}\|_1\right\}\\ &= \arg\min_a \{f(\mathbf{A}) + \lambda g(\mathbf{A})\},\end{aligned} \tag{71.5}$$

where $\lambda \neq 0$ is a scalar regularization parameter. For a fixed λ, the optimal solution is achieved by assigning $\frac{\partial J(\mathbf{A},\lambda)}{\partial \mathbf{A}} = \mathbf{0}$, then One can easily get $\frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\mathbf{A} - \mathbf{y}) = -\mathbf{c}(\mathbf{A})$. Hence, one maintains a sparse support set: $S = \left\{i \middle| \left|\mathbf{c}_i^l\right| = \lambda \neq 0\right\}$, where i labels the nonzero elements of $\mathbf{c}(\mathbf{A})$ associated with $\mathbf{A}$, and l is the loop index. The algorithm computes the solution $\mathbf{A}$ along the updated direction $\mathbf{d}^l$, which is the solution to the following system:

$$\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\mathbf{d}^l(S) = \operatorname{sgn}(\mathbf{c}^l(S)), \tag{71.6}$$

where $\mathbf{X}_S$ is a submatrix of $\mathbf{X}$ that collects the column vectors of $\mathbf{X}$ with respect to S, and $\mathbf{c}^l(S)$ is a vector that contains the coefficients of $\mathbf{c}^l$ in terms of S. Along the direction indicated by $\mathbf{d}^l$, an update on $\mathbf{A}$ may lead to a breakpoint $\gamma^l = \min\left\{\gamma_+^l, \gamma_-^l\right\}$, which occurs in two conditions:

The first condition is

$$\gamma_+^l = \min_{i \notin S}\left\{\frac{\lambda - c_i}{1 - \mathbf{x}_i^{\mathrm{T}}\mathbf{X}_S\mathbf{d}^l(S)}, \frac{\lambda + c_i}{1 + \mathbf{x}_i^{\mathrm{T}}\mathbf{X}_S\mathbf{d}^l(S)}\right\}. \tag{71.7}$$

The second condition is

$$\gamma_-^l = \min_{i \in S}\left\{-\frac{a_i}{d_i}\right\}. \tag{71.8}$$

One can update $\mathbf{A}$, with an initial value $\mathbf{A}^0 = \mathbf{0}$, by $\mathbf{A}^{l+1} = \mathbf{A}^l + \gamma^l\mathbf{d}^l$, then the algorithm shall terminate when the update approaches zero.

Classification procedure. For each class i, we define a mapping $\varphi_i : \mathbf{R}^n \to \mathbf{R}^n$ to select the coefficients associated with the i th class, namely, for $\mathbf{A}_{\text{opt}} \in \mathbf{R}^n$, $\varphi_i(\mathbf{A}_{\text{opt}}) \in \mathbf{R}^n$ is a new vector whose elements are all zero except those, only related to the ith class, maintaining the original value. We rebuild the test sample $\mathbf{y}$ by

$$\tilde{\mathbf{y}}_i = \mathbf{X}\varphi_i(\mathbf{A}_{\text{opt}}). \tag{71.9}$$

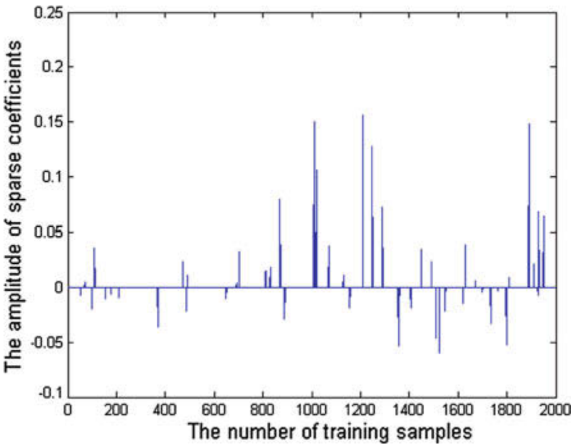


Fig. 71.1 Sparse coefficient extracted

Then **y** can be classified as the i^*th class according to the minimum residual between **y** and $\tilde{\mathbf{y}}_i$, expressed by

$$i^* = \arg\min_i \varepsilon_i(\mathbf{y}) = \|\mathbf{y} - \tilde{\mathbf{y}}_i\|_2. \tag{71.10}$$

71.3 Experimental Results and Analysis

Based on Aleix Martinez and Robert Benavente (AR) database, we chose 100 individuals consisting of 50 male subjects and 50 female subjects to conduct this experiment. For each individual, there exists 26 face images with the variations of illumination, expressions and disguises. We select the former 20 face images as training samples, and the others as test samples.

As mentioned before, precisely extracting sparse coefficients is a significant step. As portrayed in Fig. 71.1, for a random test sample, the distribution of sparse coefficients is disorderly and unsystematic. Thus, in order to identify the test sample, we should obtain the reconstructed residual error. As shown in Fig. 71.2, we can see that the energy of the residual error associated with the object class is obviously lower which can be used to determine the object class.

Moreover, the reconstructed results about two random test samples are demonstrated in Fig. 71.3. We can see that the corruption due to disguises is reduced greatly in our reconstructed images, which contribute to the improvement of recognition rate compared with other works [10].

Experimental results listed in Table 71.1 show that, as the iterations of homotopy increase, the time consumed is growing but the recognition rate are not always improved, and the highest recognition rate is 96.6667%, then we obtain the optimal iterations, 100 times.

Fig. 71.2 Residual error

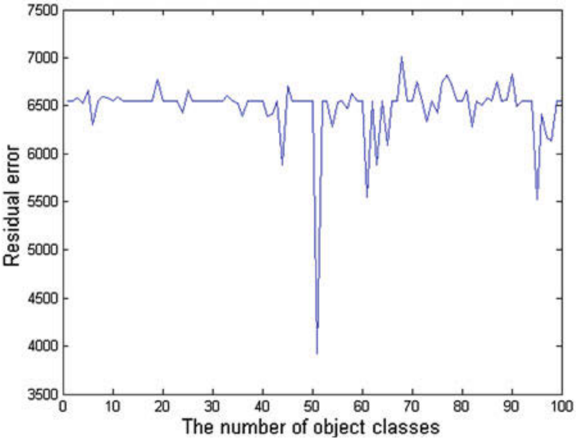


Fig. 71.3 Comparision of original images and reconstructed images

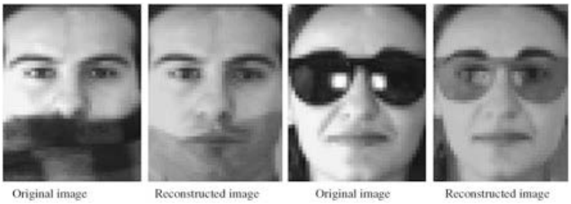


Table 71.1 Recognition rate and time consumed vary with the iterations

Iterations	Time consumed (s)	Recognition rate (%)
50	15.1720	88.3333
100	19.9840	96.6667
150	28.0160	96.6667
300	57.5310	95.8333
400	66.3750	95.8333

71.4 Conclusions

This chapter presents SRC combined with homotopy algorithm to perform disguised face recognition, which obtains favorable effect. SRC using residual represention based on an overcomplete dictionary ensures the stability of numerical calculation, which is robust to the corruption of face images attributed to disguise. Above all, the method of homotopy solves the l_1-minimization problem exactly needed in SRC. However, for the images in a more complex condition, how to improve the recognition rates need to be further studied.

Acknowledgments This work is supported by NNSF (No. 61072127, No. 61070167), NSF of Guangdong Province, P.R.C. (No. 10152902001000002, No. 07010869) and High Level Personnel Project of Guangdong Colleges (No. [2010] 79).

References

1. Yang J, Zhang Di, Xu Y, Yang JY (2005) Two-dimensional discriminant transform for face recognition. Patt Recog 38(7):1125–1129
2. Ramanathan N, Chowdhury AR, Chellappa R (2004) Facial similarity across age, disguise, illumination and pose. Proc Int Conf Image Process 3:1999–2002
3. Kanade T (1973) Picture processing system by computer complex and recognition of human faces. Doctoral dissertation, Kyoto University, November 1973
4. Donoho D (2006) For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. Comm Pure and Appl Math 59(6):797–829
5. Cande E, Tao T (2006) Near-optimal signal recovery from random projections universal encoding strategies. IEEE Trans Inf Theo 52(12):5406–5425
6. Chen S, Donoho D, Saunders M (1999) Atomic decomposition by basis pursuit. SIAM J Sci Comput 20(1):33–61
7. Yang J, Zhang Y (2009) Alternating direction algorithms for l_1-problems in compressive sensing. (preprint) arXiv:0912.1185–1224
8. Wright J, Ma Y (2010) Dense error correction via l_1-minimization. IEEE Trans Inf Theory
9. Osborne M, Presnell B, Turlach B (2000) A new approach to variable selection in least squares problems. IMA J Numer Anal 20:389–404
10. Singh R, Vatsa M, Noore A (2007) Face recognition with disguise and single gallery images. Lane Department of Computer Science and Electrical Engineering West Virginia University, USA

Part VII
Fuzzy System and Control

Chapter 72
Adaptive Neural Network Tracking Control for a Class of Nonlinear Pure-Feedback Systems

Hui Hu, Peng Guo and Cheng Liu

Abstract A new output feedback tracking control algorithm using neural network for a class of SISO pure-feedback nonlinear systems is presented, under the constraints that only the system output variable can be measured. The previous output-feedback control algorithms are all based on the backstepping scheme and the state observer, which makes the control law and stability analysis of the closed-loop system and real implementation very complicated. In this paper, the algorithm is not based on backstepping scheme and no state observer is employed. Only the output error is used in control laws and weights update laws and no robustifying control term is employed. The stability of closed-loop system and signals boundedness are demonstrated by Lyapunov stability theorem.

Keywords Output feedback · Pure-feedback · Nonlinear · Neural network

72.1 Introduction

In recent years, adaptive control of uncertain nonlinear systems has received increasing attention and many significant developments have been achieved. The existing methods for dealing with uncertainties include adaptive control and robust control. However, the real systems always contain some uncertain elements which cannot be modeled. So, most of the adaptive controllers involve certain types of function approximators such as neural networks (NNs) and fuzzy logic systems in

H. Hu (✉) · P. Guo · C. Liu
Department of Electrical and Information Engineering,
Hunan Institute of Engineering,
Xiangtan 411101, Hunan, China
e-mail: onlymyhui@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_72,

their mechanisms [1–4]. By combining the concepts of backstepping approach and neural network approximation, several adaptive neural backstepping approaches for strict- and pure-feedback nonlinear systems have been proposed in [1–8]. The suggested controllers achieved the good tracking performance, and also guaranteed uniform ultimate boundedness of all the signals in the closed-loop system. A major problem of the adaptive backstepping approach is that the assumption of linearity in the unknown parameters is required and tedious and complex analysis is needed to determine regression matrices. The complexity is inherited to the approximator-based adaptive backstepping controller. Moreover, the complexity exhibits an exponential increase as the order of the controlled system grows, so that the learning time tends to be unacceptably large for the higher order systems and the time-consuming process is unavoidable when the controllers above are implemented. Recently in [9], an adaptive predictive neural controller that is not based on backstepping has been proposed for a class of discrete pure-feedback systems. In [10], an adaptive neuron control algorithm for a SISO strict-feedback nonlinear system is proposed without backstepping. But the algorithm is based on the higher order observer which makes the stability analysis of the closed-loop system and real implementation very complicated.

In this paper, we propose an adaptive neural network controller for a class of pure-feedback nonlinear system. The proposed control scheme is not based on backstepping scheme and no state observer is employed in the algorithm. Only the output error is used in control laws and weights update laws. Based on this fact, it is shown that controller design and stability analysis are considerably simpler than the previous backstepping-based algorithms. Lyapunov theory is applied to guarantee the boundedness of all signals in the closed-loop system. Therefore, it is convenient to realize the algorithm in engineering.

72.2 Problem Formulation

Consider the following pure-feedback nonlinear system:

$$\begin{cases} \dot{x}_i = f_i(\bar{x}_i) \quad i = 1, \ldots, n-1 \\ \dot{x}_n = f_n(\bar{x}_n, u) \\ y = x_1 \end{cases} \tag{72.1}$$

where $\bar{x}_i = [x_1, \ldots, x_i] \in R^i$ and $u, y \in R$ are system state vector, system input and output, respectively. It is assumed that only y is measurable. $f_i(\cdot)$, $i = 1, 2, \ldots, n$ are unknown smooth functions. For the controllability issue, the following assumption must be made.

Assumption 1 The values of the $\partial f_i / \partial x_{i+1}$, $i = 1, 2, \ldots, n$ and $\partial f_n / \partial u$ are all nonzero. Without loss of generality, we assume that

$$\frac{\partial f_i}{\partial x_{i+1}} > 0, \quad i = 1, 2, \ldots, n \tag{72.2}$$

$$\frac{\partial f_n}{\partial u} > 0 \tag{72.3}$$

The control objective is to design an adaptive neural network tracking controller for system (72.1) such that the system output y follows a desired trajectory y_d without observer and all signals in the closed-loop system remain bounded.

The following process shows that the original system (72.1) can be viewed as the standard normal form with respect to the newly defined state variables. Let $z_1 \triangleq y$, $z_2 \triangleq \dot{z}_1 = f_1 + g_1 x_2$. The time derivative of z_2 is derived as

$$\begin{aligned}\dot{z}_2 &= \frac{\partial f_1}{\partial x_1}\dot{x}_1 + \dot{x}_2 \\ &= \frac{\partial f_1}{\partial x_1} f_1 + \frac{\partial f_1}{\partial x_2} f_2 \\ &\triangleq a_2(\bar{x}_2) + b_2(\bar{x}_3)\end{aligned} \tag{72.4}$$

where $a_2(\bar{x}_2) = \frac{\partial f_1}{\partial x_1} f_1$, $b_2(\bar{x}_3) = \frac{\partial f_1}{\partial x_2} f_2$. Let $z_3 \triangleq a_2(\bar{x}_2) + b_2(\bar{x}_3)$, whose time derivative is derived as

$$\begin{aligned}\dot{z}_3 &= \sum_{j=1}^{2} \frac{\partial}{\partial x_j}(a_2(\bar{x}_2) + b_2(\bar{x}_3))\dot{x}_j + \frac{\partial b_2(\bar{x}_3)}{\partial x_3}\dot{x}_3 \\ &= \sum_{j=1}^{2} \frac{\partial}{\partial x_j}(a_2(\bar{x}_2) + b_2(\bar{x}_3)) f_j + \frac{\partial f_1}{\partial x_2}\frac{\partial f_2}{\partial x_3} f_3 \\ &\triangleq a_3(\bar{x}_3) + b_3(\bar{x}_4)\end{aligned} \tag{72.5}$$

where $a_3(\bar{x}_3) = \sum_{j=1}^{2} \frac{\partial}{\partial x_j}(a_2(\bar{x}_2) + b_2(\bar{x}_3)) f_j$, $b_3(\bar{x}_4) = \frac{\partial f_1}{\partial x_2}\frac{\partial f_2}{\partial x_3} f_3$. In general, by induction, the following general formulas are derived for $i = 2, \ldots, n$

$$\begin{aligned} z_i &\triangleq a_{i-1}(\bar{x}_{i-1}) + b_{i-1}(\bar{x}_i) \\ \dot{z}_i &= a_i(\bar{x}_i) + b_i(\bar{x}_{i+1}) \end{aligned} \tag{72.6}$$

where $a_1 = 0, b_1 = f_1(\bar{x}_2)$ and for $i = 2, \ldots, n$

$$a_i(\bar{x}_i) = \sum_{j=1}^{i-1} \frac{\partial}{\partial x_j}(a_{i-1}(\bar{x}_{i-1}) + b_{i-1}(\bar{x}_i)) f_j(\bar{x}_{j+1}) \tag{72.7}$$

$$b_i(\bar{x}_{i+1}) = \left(\prod_{j=1}^{i-1} \frac{\partial f_j(\bar{x}_{j+1})}{\partial x_{j+1}}\right) f_i(\bar{x}_{i+1}) \tag{72.8}$$

while $\bar{x}_{n+1} = \left[\bar{x}_n^T, u\right]^T$. As a result the pure-feedback system (72.1) can be redescribed as the following normal form with respect to the newly defined state variables z_i is:

$$\begin{aligned} \dot{z}_i &= z_{i+1} \quad i = 1, 2, \ldots, n-1 \\ \dot{z}_n &= a_n(\bar{x}) + b_n(\bar{x}, u) \\ y &= z_1 \end{aligned} \tag{72.9}$$

A low-pass filter is employed to transform (72.9) into affine in the pseudo-input dynamics. The transfer function of the low-pass filter is

$$L(s) = \frac{u}{u_p} = \frac{\sigma}{s + \sigma} \tag{72.10}$$

where σ is a positive design constant. Define the augmented state variable as $\bar{z} = [z_1, z_2, \ldots, z_n, \dot{z}_n]^T = \left[y, \dot{y}, \ldots, y^{(n-1)}, y^{(n)}\right]^T$, and let $\eta = \left[\bar{x}_n^T, u\right]^T$, then

$$\begin{aligned} \dot{z}_i &= z_{i+1} \quad i = 1, 2, \ldots, n-1 \\ \dot{z}_n &= a_n(\bar{x}) + b_n(\eta) =: z_m \\ \dot{z}_m &= \frac{\partial}{\partial \eta}(a_n(\bar{x}) + b_n(\eta))\dot{\eta} \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(a_n(\bar{x}) + b_n(\eta))\dot{x}_i + \frac{\partial b_n}{\partial u}\dot{u} \\ &= \left(\sum_{i=1}^{n} \frac{\partial}{\partial x_i}(a_n(\bar{x}) + b_n(\eta)) f_i - a\frac{\partial b_n}{\partial u}u\right) + a\frac{\partial b_n}{\partial u}u_p \end{aligned} \tag{72.11}$$

Define the functions $a(\eta)$ and $b(\eta)$ as

$$\begin{aligned} a(\eta) &= \sum_{i=1}^{n} \frac{\partial}{\partial x_i}(a_n(\bar{x}) + b_n(\eta)) f_i - a\frac{\partial b_n}{\partial u}u \\ b(\eta) &= a\frac{\partial b_n}{\partial u}u_p \end{aligned} \tag{72.12}$$

Then the nonaffine nonlinear system becomes the mth-order affine in the pseudo-input nonlinear system:

$$\begin{aligned} \dot{z}_i &= z_{i+1} \quad i = 1, 2, \ldots, n \\ \dot{z}_m &= a(\eta) + b(\eta)u_p \end{aligned} \tag{72.13}$$

where the functions $a(\cdot)$ and $b(\cdot)$ are totally unknown. From the definitions of $b_n(\bar{x}, u)$, the function $b(\eta)$ is nonzero and positive according to Assumption 1. Thus, there exist positive constant $\bar{b} > 0$ such that $b(\eta) \geq \bar{b}$ for all $\eta \in R^m$.

72.3 Controller Design and Stability Analysis

Define vector x_d, $\bar{e}$ and tracking error e and a filtered tracking error s as

$$x_d = \begin{bmatrix} y_d & \dot{y}_d & \ldots & y_d^{(n)} \end{bmatrix}^T \tag{72.14}$$

$$e = y_d - y \tag{72.15}$$

$$\bar{e} = x_d - \bar{z} = \begin{bmatrix} e & \dot{e} & \ldots & e^{(n)} \end{bmatrix}^T \tag{72.16}$$

$$s = \left(\frac{\mathrm{d}}{\mathrm{d}t} + \lambda\right)^n e = \left[\tau^T 1\right]\bar{e} = \Lambda\bar{e} \tag{72.17}$$

where $\lambda > 0$ is a design constant and $\tau = \left[\lambda^n, (n)\lambda^{n-1}, \ldots, n\lambda\right]^T$, $\Lambda = [\tau^T 1]$.

A filtered tracking error is defined as

$$\begin{aligned} \dot{s} &= \mathit{\Lambda}_1^T\bar{e} + y_d^{(n+1)} - y^{(n+1)} \\ &= \mathit{\Lambda}_1^T x_d - \mathit{\Lambda}_1^T\bar{z} + y_d^{(n+1)} - a(\eta) - b(\eta)u_p \\ &= -a(\eta) - \mathit{\Lambda}_1^T\bar{z} - b(\eta)u_p + v_1 \end{aligned} \tag{72.18}$$

where $\Lambda_1 = [0\tau^T]^T$, $v_1 = y_d^{(n+1)} + \Lambda_1^T x_d$.

If $a(\eta)$ and $b(\eta)$ are known, and the ideal control input is determined as

$$u^* = k(t)\lambda^{n-1}e + \frac{\bar{a}(\bar{x}) + v}{b(\bar{x})} \tag{72.19}$$

where $k(t) > 1/2$ is a design parameter. $\bar{a}(\eta) = -a(\eta) - \Lambda_1^T\bar{z} - b(\eta)k\tau_2^T\bar{z}$, $v(\eta, v_1, v_2) = v_1 + b(\eta)v_2$, $v_2 = k\tau_2^T\bar{x}_d$, $\tau_2 = \left[0, n\lambda^{n-1}, \ldots, n\lambda, 1\right]^T$. Then, s converges to zero.

Proof Consider the Lyapunov function $V_s = \frac{1}{2}s^2$. Taking the time derivative of V_s along (72.18) yields

$$\begin{aligned} \dot{V}_s &= s\dot{s} = s\left(-a(\eta) - \mathit{\Lambda}_1^T\bar{z} - b(\eta)u_p + v_1\right) \\ &= s\left[-b(\eta)k(t)\lambda^n e - b(\eta)k\tau_2^T\bar{e}\right] \\ &= -b(\eta)k(t)s^2 \end{aligned} \tag{72.20}$$

According to the Lyapunov theorem, the result implies that $\lim_{t \to \infty} s = 0$.

However, in our case, since, $a(\eta)$, $b(\eta)$ are unknown and only output is measurable then the ideal controller u^*cannot be realized. Rewriting (72.19), we obtain

$$u^* = k(t)\lambda^{n-1}e + u^*_{\mathrm{ad}} \tag{72.21}$$

where $u^*_{\mathrm{ad}} = \bar{a}(\eta) + v/b(\eta)$ is an unknown function. Because NNs have the universal function approximation property, we employ an RBFNN to estimate the unknown function u^*_{ad} using universal function approximation property of the NNs as follows:

$$u^*_{\mathrm{ad}} = \frac{\bar{a}(\eta) + v}{b(\eta)} = W^{*^T}\phi(\xi) + \varepsilon \tag{72.22}$$

where W is the adjustable parameter vector, $\phi(\xi)$ is Gaussian function and ε is the approximation error which satisfies $|\varepsilon| \leq \varepsilon_0$. $\xi = [y(t), y(t-d_1), \ldots, y(t-nd_1), v_1(t), v_2(t)]^T$, $d_1 > 0$ is a positive time delay. Then the optimal parameter is defined as

$$W^* = \arg\min_{W \in \Omega_\omega}\left\{\sup\left|W^T\phi(\xi) - u^*_{\mathrm{ad}}\right|\right\} \tag{72.23}$$

where $\Omega_\omega = \{W| \|W\| \leq \varepsilon_\omega\}$, $\varepsilon_\omega > 0$ is the design constant. Then NN output feedback controller is

$$u = k\lambda^{n-1}e + \hat{u}_{\mathrm{ad}}(\xi) \tag{72.24}$$

where $\hat{u}_{\mathrm{ad}}(\xi) = \hat{W}^T\phi(\xi)$ is the output of the NNs.

The update law for $\hat{W}$ is determined as

$$\dot{\hat{W}} = \gamma(e\phi - \sigma|e|\hat{W}) \tag{72.25}$$

where the adaptive gains $\gamma, \sigma > 0$. Then, for adaptive algorithm (72.25), there exists a compact set

$$\Theta_\omega = \left\{\hat{W}| \|\hat{W}\| \leq \frac{\phi_m}{\sigma}\right\} \tag{72.26}$$

where $\|\phi(\xi)\| \leq \phi_m$, with ϕ_m constant, such that, if $\hat{W}(0) \in \Theta_\omega$, then $\hat{W}(t) \in \Theta_\omega$, $\forall t \geq 0$.

Proof Let the Lyapunov function $V_\omega = \frac{1}{2\gamma}\hat{W}^T\hat{W}$, whose time derivative is

$$\begin{aligned}\dot{V}_\omega &= \frac{1}{\gamma}\hat{W}^T\dot{\hat{W}} = \hat{W}^T\left(e\phi - \sigma|e|\hat{W}\right) \\ &= \hat{W}^Te\phi - \sigma|e|\left\|\hat{W}\right\|^2 \leq -|e|\left\|\hat{W}\right\|\left(\sigma\left\|\hat{W}\right\| - \phi_m\right)\end{aligned} \tag{72.27}$$

It follows that $\dot{V}_\omega \leq 0$ as long as $\|\hat{W}\| > \phi_m/\sigma$. Therefore, $\hat{W}(t) \in \Theta_\omega, \forall t \geq 0$.

According to (72.18), the time derivative of the filter tracking error can be derived as

$$\begin{aligned}\dot{s} &= -a(\eta) - \Lambda_1^T \bar{z} - b(\eta)u + v_1 \\ &= -a(\eta) - \Lambda_1^T \bar{z} - b(\eta)u - b(\eta)u^*_{\text{ad}} + b(\eta)u^*_{\text{ad}} + v_1 \\ &= b(\eta)\left(-ks - \tilde{W}^T\phi + \varepsilon\right)\end{aligned} \tag{72.28}$$

where $\tilde{W} = \hat{W} - W^*$.

Theorem 1 *Consider the pure-feedback system* (72.1) *with the control input* (72.24) *and adaptive law* (72.25). *Then all the signals in the closed-loop system are bounded and the state vector* $\bar{z}$ *remains in*

$$\Omega_z = \left\{\bar{z}(t)||e_i(t)| \leq 2^i \lambda^{i-n-1} \frac{b_\omega \phi_m + \varepsilon_0}{\sqrt{k - 0.5}}, i = 1, 2, \ldots, n+1\right\}, \quad \forall t \geq T$$

where $b_\omega = \phi_m/\sigma + \|W^*\|$.

Proof Let the Lyapunov function $V_s = \frac{1}{2}s^2$, whose time derivative is

$$\begin{aligned}\dot{V}_s &= s\dot{s} = b(\eta)\left(-ks - \tilde{W}^T\phi + \varepsilon\right)s \\ &\leq -b(\eta)ks^2 + b(\eta)|s|\left(\|\tilde{W}\|\|\phi\| + \varepsilon_0\right)\end{aligned} \tag{72.29}$$

since $\hat{W}$ is bounded, it follows that $\|\tilde{W}\| \leq b_\omega$, with $b_\omega = \frac{\phi_m}{\sigma} + \|W^*\|$. Then

$$\dot{V}_s \leq -b(\eta)ks^2 + b(\eta)|s|(b_\omega \phi_m + \varepsilon_0) \tag{72.30}$$

From the inequality $|\alpha||\beta| \leq (\alpha^2 + \beta^2)/2$, it follows that

$$\begin{aligned}\dot{V}_s &\leq -b(\eta)ks^2 + 0.5\, b(\eta)\left((b_\omega\phi_m + \varepsilon_0)^2 + s^2\right) \\ &= -2b(\eta)(k - 0.5)\left[V_s - \frac{(b_\omega\phi_m + \varepsilon_0)^2}{4(k - 0.5)}\right]\end{aligned} \tag{72.31}$$

Let $\bar{V}_s = V_s - \dfrac{(b_\omega\phi_m + \varepsilon_0)^2}{4(k - 0.5)}$. Using the comparison principle in [11], it follows that

$$\bar{V}_s \leq \bar{V}_s(0)\mathrm{e}^{-2(k-0.5)\int_0^t b(\eta(\tau))\mathrm{d}\tau} \tag{72.32}$$

This implies that

$$V_s - \frac{(b_\omega\phi_m + \varepsilon_0)^2}{4(k - 0.5)} \leq \left[V_s(0) - \frac{(b_\omega\phi_m + \varepsilon_0)^2}{4(k - 0.5)}\right]\mathrm{e}^{-2(k-0.5)\int_0^t b(\eta(\tau))\mathrm{d}\tau} \tag{72.33}$$

Since $b(\eta) \geq \bar{b} > 0$ and $-\frac{(b_\omega \phi_m + \varepsilon_0)^2}{4(k-0.5)} \mathrm{e}^{-2(k-0.5)\bar{b}t} \leq 0$, it follows that

$$V_s \leq \bar{V}_s(0)\mathrm{e}^{-2(k-0.5)\bar{b}t} + \frac{(b_\omega \phi_m + \varepsilon_0)^2}{4(k-0.5)} \tag{72.34}$$

Therefore

$$s^2 \leq s^2(0)\mathrm{e}^{-2(k-0.5)\bar{b}t} + \frac{(b_\omega \phi_m + \varepsilon_0)^2}{2(k-0.5)} \tag{72.35}$$

The boundedness of s implies that $\bar{z}$ is bounded. According to the properties of the filter tracking error, the state errors $\bar{z}$ will remain in Ω_z.

72.4 Conclusions

A new and simple output-feedback adaptive neural network controller for pure-feedback nonlinear system is proposed. The control problem of the pure-feedback system is treated as the same problem as the affine system in the normal form, which results in avoiding backstepping in the controller design. At the same time, the previous output-feedback control algorithms are all based on the state observer,which makes the stability analysis of the closed-loop system and real implementation very complicated. The proposed algorithm in this paper is that no state observer and robustfying term are employed. Only the output error is used to generate control input and update laws for neural network parameters. Only one RBFNN is employed to approximate unknown lumped nonlinear function. It is shown that the filtered tracking error and RBFNN weight vector are uniformly ultimately bounded in theory.

Acknowledgments This project is supported by The Research Foundation of Education Bureau of Hunan Province, China (Grant No. 09B022) and Provincial Natural Science Foundation of Hunan, China (Grant No. 09JJ3094) and the Science-Technology Project of Hunan Province, China (Grant No. 2008FJ3029). The Key Construction Discipline of Hunan Provinces. Science and Technology Innovation Team of Hunan Province: Complex Network Control.

References

1. Zhang C, Qi X (2006) Adaptive control of general nonlinear systems in strict feedback form. Cont Theo Appl 23:621–626
2. Zhang J, Ge SS, Lee TH (2005) Output feedback control of a class of discrete MIMO nonlinear systems with triangular form inputs. IEEE Trans Neural Netw 16:1491–1503
3. Ge SS, Wang C (2002) Direct adaptive NN control of a class of nonlinear systems. IEEE Trans Neural Netw 13:214–221

4. Wang C, Hill DJ, Ge SS, Chen G (2006) An ISS-modular approach for adaptive neural control of pure-feedback systems. Automatica 42:723–731
5. Yang Y, Zhou C (2005) Robust adaptive fuzzy tracking control for a class of perturbed strict-feedback nonlinear systems via small-gain approach. Inf Sci 170:211–234
6. Li Y, Qiang S, Zhuang X, Kaynak O (2004) Robust and adaptive backstepping control for nonlinear systems using RBF neural networks. IEEE Trans Neural Netw 15:693–701
7. Zhou L, Jiang CS, Qian CS (2008) A fast adaptive backstepping method based on neural networks. J Astronaut 29:1888–1894
8. Wang D, Huang J (2005) Neural network-based adaptive dynamics surface control for a class of uncertain nonlinear systems in strict-feedback form. IEEE Trans Neural Netw 16:195–202
9. Ge SS, Yang C, Lee TH (2008) Adaptive predictive control using neural network for a class of pure-feedback systems in discrete-time. IEEE Trans Neural Netw 19:1599–1614
10. Park JH, Kim SH, Moon CJ (2009) Adaptive neural control for strict-feedback nonlinear systems without backstepping. IEEE Trans Neural Netw 20:1204–1209
11. Lakshmikanthan VL, Leela S (1969) Differential and integral inequalities. Academic Press, New York

Chapter 73
Application of Double-Loop PID Controller in the Inversed Pendulum Real-Time Control System

Xiaokan Wang, Zhongliang Sun and Shouxiang Zai

Abstract The proposed double-loop PID control scheme may solve the problems of the unstable inverted pendulum system with single-input and double-outputs and strong nonlinear coupling. The feasibility of double-loop PID control's multiple combinations scheme is studied by simulating with MATLAB. The simulation results showed that double-loop PD–PD control scheme was the best options when comparing all kinds of double-loop PID control methods. The real-time control experiment of the actual single inverted pendulum device confirmed that the proposed control scheme could realize the double closed-loop control of the car position and the swing link angle.

Keywords Double-loop PID control · Real-time control · Inverted pendulum · Simulation

73.1 Introduction

The inverted pendulum system is a natural instable nonlinear system, and is the ideal verification platform for the control theory teaching and the research of various control strategies. The inverted pendulum system with high times, unstable, multivariable, nonlinear and strong coupling is always regarded as the best test object to verify the control theoretical method for many modern control theory researchers. The new control method was constantly discovered by researching the inverted pendulum control which applied in the many high-tech fields such as space technology and robotics [1].

X. Wang (✉) · Z. Sun · S. Zai
Henan Mechanical and Electrical Vocational Education Group,
Zhengzhou 450002, China
e-mail: wxkbbg@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_73,

At present, there are many successful application of inverted pendulum system reports with advanced control algorithms such as optimal control [2], adaptive control [3], intelligent control [3–6] and disturbance rejection control [7]; but it is difficult to see success with conventional PID control to control the inverted pendulum system. And it was also heard that the conventional PID control could not successfully control the inverted pendulum system. As we all know, PID controller is used most widely and the most common by far. In the actual control field, many researchers believe that PID controller is often better than the advanced controller. Therefore, whether you can confirm the conclusion that inverted pendulum cannot be successfully controlled by the conventional PID control is valid? This is the study subject of this chapter.

73.2 Mathematical Model of Linear First-Level Inverted Pendulum

The mathematical model of the first-level linear inverted pendulum is given for simulation and study [8]. The motion equations of linear inverted pendulum can be deduced from kinetic theory, and carries on approximate processing to its motion equation, then it may obtain the system state equation which is shown in the formula (73.1) [2]

$$\begin{bmatrix} \dot{x} \\ \ddot{x} \\ \dot{\phi} \\ \ddot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & \frac{-\left(I+ml^2\right)b}{I(M+m)+Mml^2} & \frac{m^2gl^2}{I(M+m)+Mml^2} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{-mlb}{I(M+m)+Mml^2} & \frac{mgl(M+m)}{I(M+m)+Mml^2} & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ \dot{x} \\ \phi \\ \dot{\phi} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{I+ml^2}{I(M+m)+Mml^2} \\ 0 \\ \frac{ml}{I(M+m)+Mml^2} \end{bmatrix} \cdot u \tag{73.1}$$

In the formula: x is the car displacement; ϕ is the angle between the swing link and the vertical upward direction; u is the input value which represents the input force F of the controlled object; M is the car quality; m is the swing link quality; b is the car friction coefficient; l is the length between the swing link rotation axle center and the pole centroid and I is the swing link inertia. These parameter values are: $M = 1.125$ kg, $m = 0.13$ kg, $l = 0.5$ m, $b = 0.14$ N/m/s and $I = 0.0036$ N*m.

73.3 Simulation and Research of Control Strategy

The controlled object is a single input (The force F) and the double outputs (the car displacement, the swing link's angle) which is known by the mathematical model equation (73.1) of linear inverted pendulum. If using the conventional PID control, we would consider a single-loop PID control, double-loop PID control and double-loop PID control with decoupling control scheme.

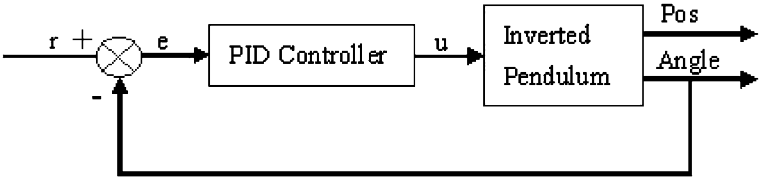


Fig. 73.1 The structure of single-loop PID control system

73.3.1 Single-Loop PID Control Scheme

In Fig. 73.1, Pos expresses the car displacement and Angle expresses swing link's angle. Seen from Fig. 73.1, we could know that the single-Loop PID control system does only control the swing link's angle and is not useful to car's position.

The reference input signal is $r = 0.1(\varepsilon(t) - \varepsilon(t-1))$ in the simulation response curve of Fig. 73.2. It is seen from the Fig. 73.2 that the pendulum's angle value tends to 0 after a period of time; all this explained that the pendulum may be upright. At the same time, the absolute value of the car's displacement increased over time becomes larger and larger (in the negative direction), indicating that the car has been offset in one direction. All of above showed that the single-loop PID control of inverted pendulum is successful in the vertical control, but the car displacement control is failure. In fact, pendulum device manufacturers provide PID control scheme which is the single-loop control scheme. The actual testing result of the pendulum is upright and drifts to one side, and finally the car will stop running when it hits the end limit.

73.3.2 Double-Loop PID Control Scheme

The double-loop PID control system is shown in Fig. 73.3 which not only has carried on the closed-loop control to the swing link's angle, but has also added on a closed-loop control to car's position.

In Fig. 73.4, the pendulum's angle tends to 0 after a period of time, while the car's displacement tends to a constant −0.6. This shows that the pendulum is in the upright position even when stable car is deviated from the rail in the middle of the 0.6. So we thought the double-loop PID control scheme of inverted pendulum is successful.

73.3.3 Research on Double-Loop PID Control Strategy

Double-loop PID control has many kinds of schemes according to two loops of PID control law. Suppose the car position is the fronted and the swing link angle is

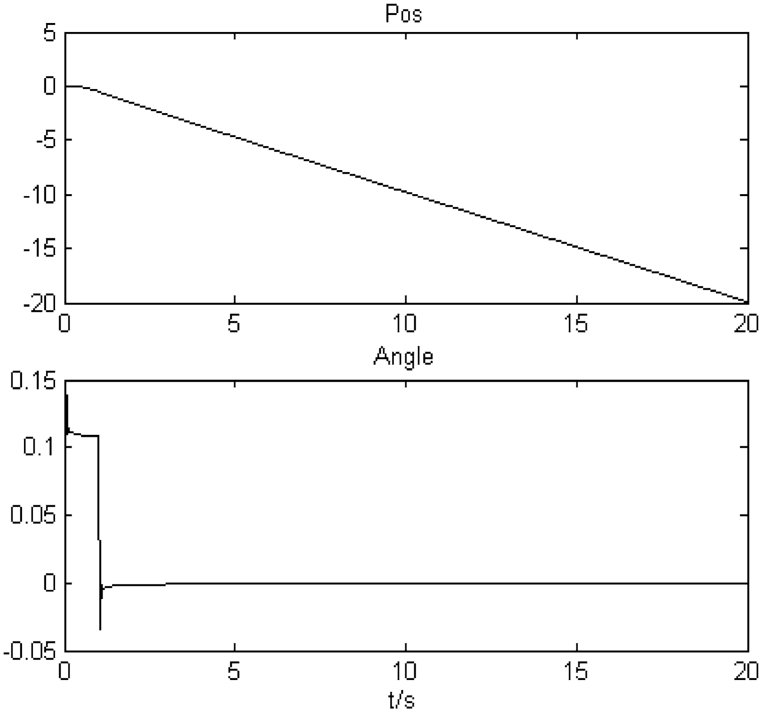


Fig. 73.2 The single-loop PID control response curve

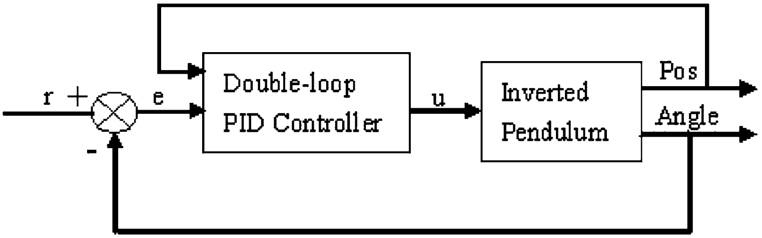


Fig. 73.3 Double-loop PID control system structure

the last, the combination types of the two loops control strategy has 16 kinds which are P–P, P–PI, P–PD, P–PID, PI–P, PI–PI, PI–PD, PI–PID, PD–P, PD–PI, PD–PD, PD–PID, PID–P, PID–PI, PID–PD and PID–PID. Which is the best combination? Seven specially selected combinations (P–PD, PI–PD, PD–P, PD–PI, PD–PD, PD–PID, PID–PD) of the simulation results show that: swing loop with P or PI control is not stability; position loop with P or PI control is unstable; PD–PID and PID–PD combination of control response are similar, but the combination of PD–PID

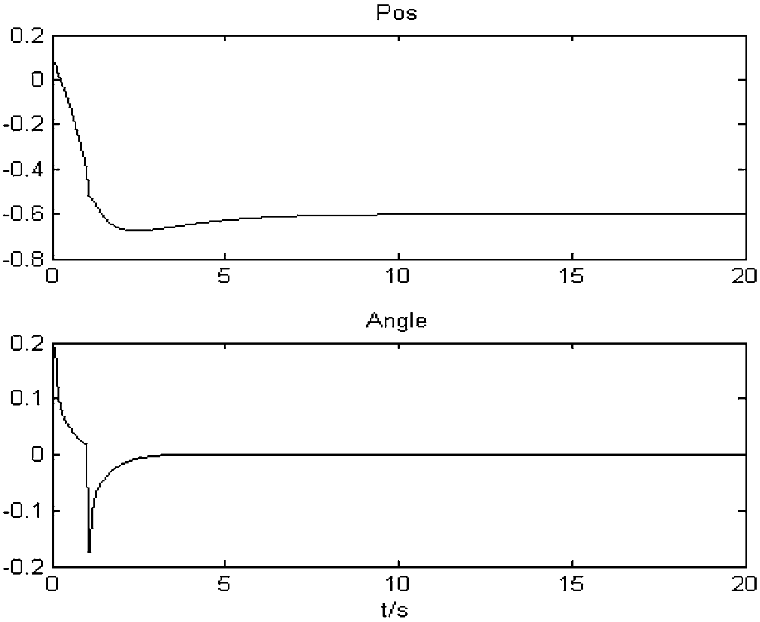


Fig. 73.4 The response curve of double-loop PID control

control of position has steady-state deviation (Fig. 73.5). So the best combination is PD–PD (Fig. 73.6).

From all study and analysis of above, the integral control is not suitable for the inverted pendulum and the function of pure proportional control is insufficient, so the double-loop control scheme of the proportion that adds the differential is the best.

73.3.4 Decoupling Control Scheme

From the above analysis, it is not necessary to add decoupling control for the inverted pendulum control system because double-loop PD–PD control has been able to perfectly control the inverted pendulum.

73.4 Actual Control Test

According to the above simulation results, we do the actual control test with double-loop PD–PD control in the linear first-level inverted pendulum device of some high-tech Co., Ltd.

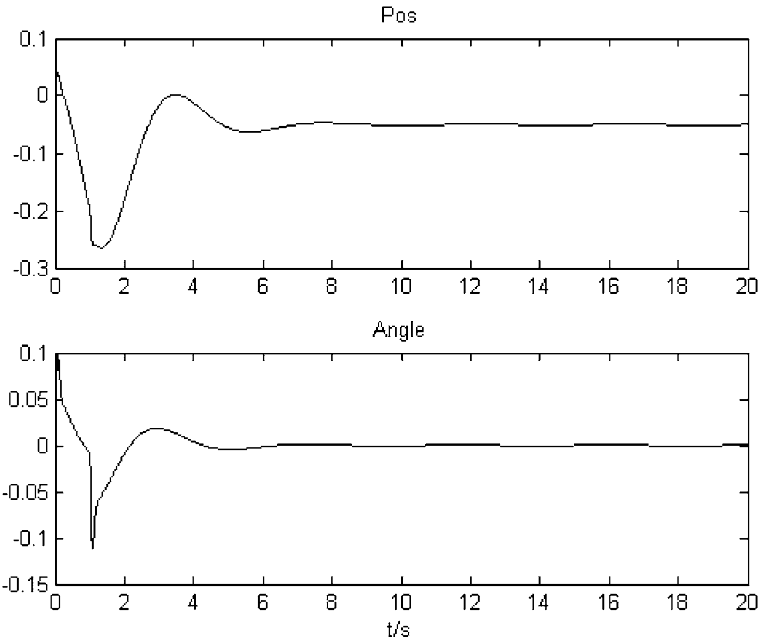


Fig. 73.5 The response curve of double-loop PD–PID control

The actual control test has obtained success by controlling loop configuration and debugging process parameters. The inverted pendulum quickly maintained stability upright and continued for 15 h.

Seen from the real-time response curve for the inverted pendulum in Fig. 73.7, the moves car back and forth in the near center, the largest deviation distance from the center rail is less than 0.06 m; the angle change of swing link is also very small, the angle between the maximum deflection direction and the vertical upward direction approximately is 0.02841 radian (1.63°).

73.5 Conclusion

In summary, we could achieve the following conclusions.

(1) The conventional PID control can be successfully applied in the inverted pendulum control system. The actual experiment described in this article has confirmed this conclusion for the linear first-level inverted pendulum set up.
(2) The double-loop PID control scheme is feasible for the first-level linear inverted pendulum, but the single-loop PID control scheme is not acceptable.
(3) The integral control is not suitable and the function of pure proportional control is insufficient for the inverted pendulum, the double-loop control

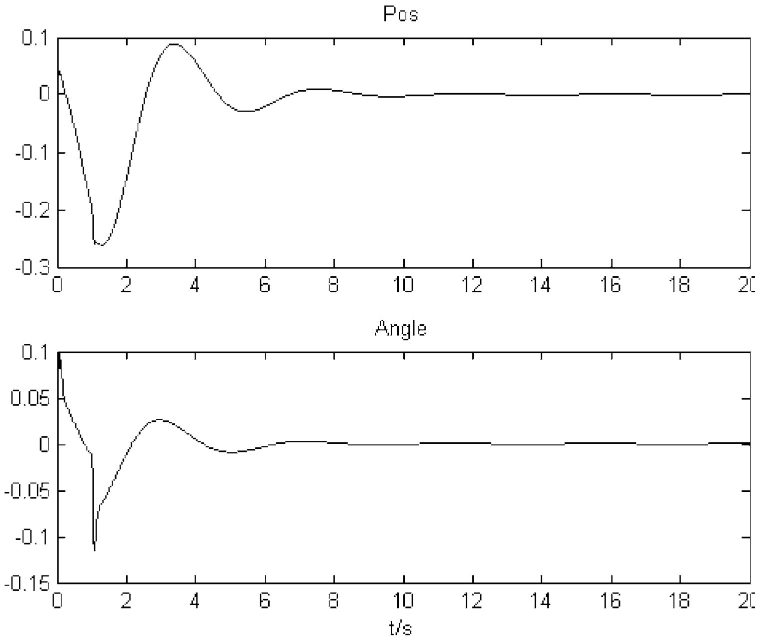


Fig. 73.6 The response curve of double-loop PD–PD control

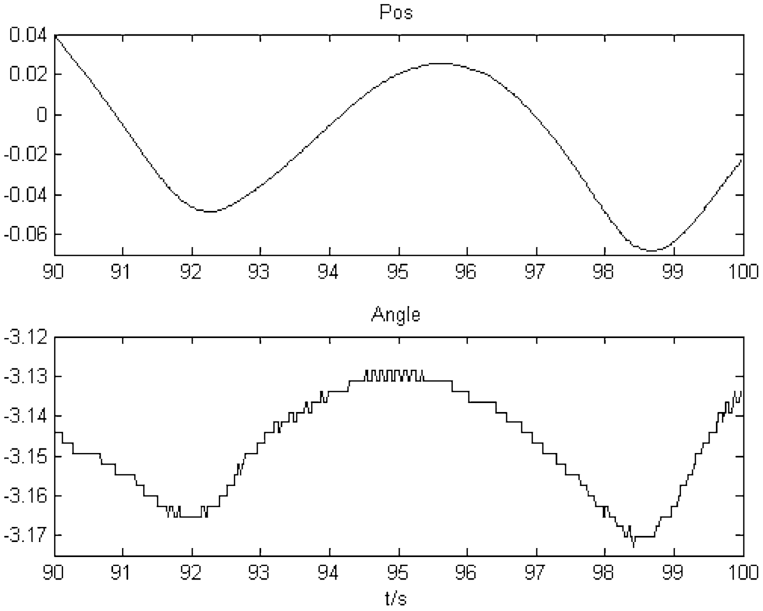


Fig. 73.7 The real-time response curve of the inverted pendulum

strategy of the proportion adds the differential is most suitable. That is, the control scheme of double-loop PD–PD combination is the best.

(4) It is not necessary to add decoupling control for the first-level linear inverted pendulum control system.

Acknowledgment The work was supported by the Innovation Fund for Small Technology-based Firms of China (No. 09C26214101775, No. 2009GJD00027) and Innovation Scientists and Technicians Troop Construction Projects of Zhengzhou City (10CXTD160).

References

1. Googol Technology (HK) Limited (2002) Googol inverted pendulum system and automatic control experiment
2. Xu L, Gao W, Qiu L (2009) A three-dimensional simulation of inverted pendulum based virtual reality modeling language. Comput Tech Autom 03:26–28
3. Feng D, Ma J, Zhou Y (2008) Design and simulation of digital optimal controller to single inverted pendulum system. Microcomput Inf 28 (10):85–86,174
4. Cui P, Weng Z, Patton R (2008) Novel active fault-tolerant control scheme and its application to a double inverted pendulum system. J Syst Eng Electr 01:138–144
5. Wang X, Feng D (2008) Research of LQR controller design method based on MATLAB. Microcomput Inf 10 (04):43–45
6. Wang X, Sun Z, Wang L, Feng D (2008) Design and research based on fuzzy PID-parameters self-tuning controller with MATLAB. In: International conference on advanced computer theory and engineering (ICACTE 2008), Phuket. IEEE CPS, Thailand, pp 996–999
7. Wang X, Wang L, Sun Z, Feng D (2009) Short-term load forecasting based on RBF adaptive neural fuzzy inference. In: Proceedings of the 14th youth conference on communication. Scientific Research Publishing, USA, pp 220–224
8. Wang X, Sun Z, Wang L, Huang S (2009) Simulation and optimization of parameters on DC motor double closed-loop control system based on simulink. International conference on intelligent human–machine systems and cybernetics (IHMSC09). IEEE CPS, Hangzhou, pp 253–256

Chapter 74
Using Ontology to Construct a Restricted Administrable Mechanism to Agents

Xin Cui

Abstract In many web application fields, multi-agent systems are required to be developed. One key issue in multi-agent systems is to develop a mechanism of being restricted administrable how to interact and exchange information autonomously across applications. Ontology, as a kind of knowledge base, has become the key technique to annotate semantics and provide a common foundation for various complex resources on the semantic web. Efficient administrable mechanism to agents is the fundamental basis based on ontology, but extant related mechanisms are insufficient to manage semantic information and enable interoperation based on them. It is well known that the current standard Web Ontology Language has no well-defined solution for administration among agents. So, this chapter proposes a restrictive administrable approach as a common framework to manage agents and its evolution information for semantic interoperation. Under the guidance of this framework, it is feasible to perform semantic restricted management to agents on the semantic web.

Keywords Ontology · OWL · Agent · Restricted Management

74.1 Introduction

The multi-agent systems field has been rapidly growing during the last decade. They are now used in the industrial context, and the standardization process initiated by the FIPA organization1 is gaining momentum. The knowledge representation field is

X. Cui (✉)
Business School of Hohai University, Nanjing 210098, China
e-mail: cuixin550042@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_74,

also gaining momentum with the semantic web supported by the W3C. More precisely, works that have been done on languages like RDF, DAML + OIL and OWL, show a strong trend towards a broadening of knowledge representation use in everyday Internet technologies. On the semantic web, a large-scale domainal application of ontology usually requests to communicate with many relative agents. Every formal ontology and agent which itself possesses belief, has both goal and functional behavior as unattached formalism. Therefore, there exists a crucial issue of how agents collaborate when an ontology acts as a knowledge-based formalism. While the second generation Web Ontology Language standard (OWL) [1] has no satisfying solution, it only provides an interface structure ‘owl:import’ that allows an ontology to import other entities. Because there is no restriction and filter by the means using a reference URI, the OWL for such a purpose is clearly insufficient and it is difficult to resolve the difficult problem.

In order to achieve this kind of communication in an effective way, we propose a way of ontological administration with agents by extending the Web Ontology Language OWL (i.e. IOWL) which is very necessary for the extensive application of the semantic web. Considering OWL-Full goes beyond the scope of the abstract description logic (ADL) which is the basics of OWL, our work is restricted into the area of OWL-Lite or OWL-DL. First, an ontological restricted administrable model with agents is proposed, where a clustering agent structure is constructed according the relation of subdomains. Then we add some axioms into OWL that includes the Agent-Link between the ontology and an agent, its property, data structure and trigger sign of agents’ modality and so on. Finally, the author also does some work about reasoning characteristic supported by the IOWL-DL.

74.2 Related Works

For the ontological administration with agents, many works have been done recently. They include mainly three classes.

The first, interoperation standard of Intelligent Physical Agents recommended by the FIPA organization [2], use a so-called OKBC (Open Knowledge Base Connectivity) as a complete open, no restricted knowledge base to support the heterogeneous agents communicating with each other, and use ACL(agent communication language) initiated by FIPA to achieve the ontological cooperation with agents in the sight of its semantic or functional sense. The literature [3] presents a method that holds the two characters above to construct ontology using FIPA ACL.

The second is usually implemented with DAML + OIL [4] technology. DAML (Darpa Agent Markup Language) [5] is based on the RDF, OIL (Ontology Interface Layer) is called ontology interface language, they unify the frame system based on RDF(s) and many methods based on description logics to support the administration. There is also a combination of the two methods above to transform the ontology to carry out the cooperation [6].

The third, under the open environment, uses lots of agents with distinct functions to bridge the gap of the administration between a formal ontology and some heterogeneous agents [7]. This solution defines several distinct agents such as mapping agent (MA), interface agent(IA), similarity agent(SA), ontology agent(OA), query agent (QA). OA, driven by various resources, takes charge of all relative task about ontology, operates on the structure of an ontology and its mapping file; IA can define its application domain, and MA runs after it to get all the relative information provided by OA and SA, the understandable knowledge can be obtained by SA.

All the above methods need to construct a shared intelligent interface between the two entities. This interface definition eliminates the semantic conceptual difference of the independent heterogeneous formalism, and bridges the gap to construct a mapping relation. But there are obvious shortcomings. On the one hand, ontology is so completely shared from the point of view of the agent that all operations on it are not restricted. On the other hand, ontology only utilizes the interface owl: import to import foreign information from agents without filtering and restriction, so the expression leads to indecision easily. In addition, there also exist larger costs of communication. Therefore it is difficult for all the above methods to be applied broadly.

Certainly, there are also many closely resembling works as before. The literature [8] supposes an architecture specially designed to model agent-based virtual organizations, offers mechanisms to take into account their structure, behavior, dynamic, norms and environment. The literature [9] presents a system to collect information through the cooperation of intelligent agent software, in addition to providing warnings after analysis to monitor and predict some possible error indications among controlled objects in the network. Using four main components: an Interface Agent, a Proxy Agent, a Monitoring Agent, and a Search Agent, it offers a kind of graphic network monitoring system to provide fast, convenient, and profound network solutions to the users. The literature [10] demonstrates how taking business, information and organizational perspectives are applied to help managers evaluate the impact of their decision-making by providing them with a set of "what-if" scenarios, which are enacted by intelligent agents, and by discussing a case example, demonstrates a kind of framework HABIO that can be applied effectively using a multi-agent system to analyze the outsourcing decision for a call centre sector in the retail industry. This work proposes a prototype of the multi-agent system which presents that intelligent agents are endowed with specific domain ontologies, appropriate to solve a specific problem, and with a reasoning mechanism to achieve specific goals and to collaborate together to produce a set of "what-if scenarios" to the outsourcing managers. The literature [11] proposes that an effective and efficient knowledge support system is crucial for a universal design process, as it has become a major design issue in the last decade with the growth of the elderly population and disabled people. The literature [12] proposes a multi-agent system for building indirect alignment between multilingual ontologies, and also presents a novel architecture to reuse and compose alignments between ontologies, also it has implemented an information retrieval system for searching for tourism information.

Our work will be defined inside the OWL-DL domain to resolve this difficult problem by adding administrable interface between an ontology and agents which include the link itself, its property, data structure, trigger sign, and so on. This OWL-based expansion (*IOWL*) constructs a restricted administrable model in itself, in which a clustered structure is constructed according to the relation of various subdomains. Finally, the author has also done some work about some reasoning characteristics supported by the IOWL-DL. Recently, OWL-DL has become the underlying of the ontological development. In order to implement this kind of administration in an effective way, extensive application of the semantic web to extend the standard is necessary.

This is a theoretical chapter to explain the basic idea of the mechanism of being restricted administrably on how to interact and exchange information autonomously across applications.

74.3 A Restricted Ontological Administration to Agents

An ontology serves as the service specification of an agent operating in the domain, and will be used in making ontological commitments among other software agents [13]. An ontological commitment is an agreement to use a vocabulary in a way that is consistent with respect to the theory specified by the ontology, i.e., an agreement on what local models are about to achieve user goals [7]. We build agents that commit to our ontology. Conversely, we design ontologies in order to share knowledge with and among these agents [13]. On the semantic web, a large-scale domainal application of ontology usually requests to communicate with many relative agents. In the course of ontological administration with agents, to reach the sharing and reusing of knowledge when they express their domainal knowledge, the ontology and agent are both required to keep their semantics identical. The ontology concentrates on the desired behaviors and management service descriptions. It serves both as a specification and the reference model to which the agents operating in the domain should comply to.

This kind of administration should follow the three basic semantic principles:

(1) *mutual semantic relationship*. If there is a mutual semantic relationship between an ontology and an agent, then this kind of mutual semantic relationship must be the two sides domainal intersection. Otherwise, when a clash appears, then its primary issue is to eliminate the clash.
(2) *semantic completeness*. If the semantic element from an agent accords with an applicational demand for ontology, it must be completeness.
(3) *semantic structure consistency*. Semantic structure of ontology and agent should set up an abstract data structure matched automatically.

On the basis of the principle above, we will construct an ontological administration model with an agent. it can be divided into five main components.

(1) *Registration.* All the agents cooperating with this ontology should be registered in the ontology utilizing *EnumeratedAgentClass* defined in *IOWL.*
(2) *Domain partition*. Ontology should offer unified data structure to all kinds of agencies, so it is easy to realize division of field, and every class has a unified *clusterID* as a chain beginning of agent-link for all registered agents based on all the principles above and unified semantic among *IOWL.*
(3) *Classification.* Ontology should classify agents by utilizing *AgentClass*. All agents carrying on mutual operation over the same semantic data structure belong to the same clustering (or class).
(4) *Modality*. Agent's active state should be written down, and the modality sign expresses that the ontology can arouse the cooperation under ready state, offering a kind of mechanism of triggering agent.
(5) *Restriction*. One integrated chain, not only defines subdomain to offer a manipulatable field to the other side, but also defines its own range. In addition, any foreign agent entity should also be identified. This kind of restriction has followed the context of the newly increased knowledge of the ontology.

In the following section, we will extend the standard of *OWL* on the basis of the mode above, this extension is called *IOWL*, and the whole cluster of an agent chain is called a cluster.

74.4 A Restricted Administrable Interface

The Ontology Web Language (OWL) is a W3C recommendation for an ontology description language that has gained widespread adoption and for which a considerable number of tools have been developed. Many health care processes, such as computer aided decision-making or disease diagnosis and treatment, are often best modeled using a declarative approach, leading to a very active interest in rule-based systems [14]. However, our work wants to use more basic changes by means of lower level of language rather than rules over upper level.

W3C gives an available method for agent in *RDF* (*s*), and it should be available for ontology too. At the same time, *OWL* reserves one interface *owl: import*, an ontology can be channeled into another ontology or other entities, this import is a kind of "ways of affirming totally" by using a *URI* reference. Both these ways are not to add any restraint, filter. Considered the factor of security etc., an ontology as a knowledge base, should not offer the overall and open operation mode for all agents. For this reason it is necessary to add an interface definition inside *OWL* frame, to extend the standard *OWL* and call the extension *IOWL*. When we define the interface, we not only think that some meta data structures should be reserved, but also some condition of being restrained attached to the interface. By means of the expanded expression method *EBNF* of *BNF* (brace { } shows that this item can appear arbitrary times, square bracket [] say this item appears less than or equal to one time, the ended sign agrees with quotation marks of adding, and no ended sign need not to add quotation marks). At first, we add axioms into *OWL* as follows:

Axiom:: = *'Agent-Link('agent-linkID'['deprecate']*
{'annotation'}{'domain('Description')'}{'datastruct('Description')'}
{'range ('Description')'}{'ForeignAgent ('{agentClassID}{agentID}')}')'

Here, domain is the definition field of this ontological interface, it limits the field where this interface can be employed, what range described is the field that foreign agent can operate on, the purpose to define *datastruct* provides a semantic data structure for foreign agent, it should be consistent with range, and every entity agent should have only one identification *agentID*. Certainly, entity agent should also observe the corresponding development standard. At the same time *agentClassID* expresses every chain, which includes corresponding *agentID*s inside, affiliated type. An *agent-linkID* regarded as a bridge between the ontology and one or more agents here, have dispelled the wide gap.

Axiom:: = *'AgentClass('agentClassID'['deprecated']* ***modality****{'annotation'}*
{'domain('Description')}datastruct('Description'))'

modality:: = *'active'|'inactive'*

agentPropertyID *:: = URIreference*

AgentClass can help to register for all agents that want to operate with an ontology, it will make the same semantic *datastruct* agents to be a cluster. At the same time, an agent's active state should be written down, and the modality sign expresses that the ontology can arouse the cooperation under ready state, offering a kind of mechanism of triggering agent.

Next, in the *IOWL*, an enumeratable agent class can be defined further.

Axiom ::='*EnumeratedAgentClass*(' *agentClassID*['*deprecated*']{*annotation*} {*individualID*}')'

When ontology collaborates with agents, the*EnumeratedAgentClass* defined above can form many clusters managed by ontology. As for these defined clusters, they cannot cross with each other where no common individual agent for any two clusters exists. To *OWL DL*, ontology should be allowed to manage these clustered agents by means of a layered way. We add further the following definitions.

Axiom ::='*disjointAgentClass*(' {*description*}')' |'*EquivalentAgentClass* (' {*description*}')'
|'*SubClassofAgentClass*(' {*description*}')'

This kind of extension helps to achieve layered management in a larger applicational field; the description of the axiom above can be defined as follows:

Description: *: = agentClassID* | *restriction* | *'unionOf('* {*description*} ')'
| *'intersectionOf* (' {*description*} ')'
| *'complementOf('* *description* ')' | {*agentID*} ')'

Restriction:: = '*Restriction* (' *Agent-Link* {*agentRestrictionComponent*}')'

agentRestrictionComponent:: =
'*allValueFrom('Description*')'|'*someValuesFrom('Description*')'.

The *IOWL*, which is constructed in this section, has solved the restricted ontological administrable problem with agents from the point of syntactic component views.

74.5 Some Important Characteristics

Because *IOWL* is totally confined to the range of *OWL-DL*, next we will discuss the theory supported by *OWL-DL*. First, we will introduce some relevant basic signs.

Suppose that K_1 is an ontology constructed over an application field based on *OWL-DL*, K_2 is another ontology based on *IOWL-DL* and K_1, and V_1, V_2 is the thesaurus corresponding to K_1, K_2 respectively, then there is a supplementary set of V_1 in the V_2, mark it V_O. Apparently $K_1 \subseteq K_2$, $V_1 \subseteq V_2$, at the same time $V_1 \subseteq K_2$. if we use *I* as an interpretation of description logics, the interpretation to V_1 and V_2 should be written down as $I(V_1)I(V_2)$ respectively.

Theorem 1 *V0 has been an independent thesaurus restricted to agents*

Prove: Each agent has its own independent formalism. Every ontology includes axioms and facts in the abstract syntactic meaning. V_O is a supplementary set of V_1 in the V_2. It is made of some agent individual *agentID* and *Agentclass* name according the meaning of *datastruct*, whose field is confined to the typed agent category that includes cluster (class), property, objects. V_O observes the abstract *OWL* syntax and each individual agent belongs to its own class. So, V_0 has been a thesaurus restricted to agents.

Theorem 2 *$I(V_O)$ shows user application context in the V_2 thesaurus.*

Prove: Each agent has its own semantic context, According to the *IOWL* made in the fourth section, *agentID*, as an individual identification based on semantic context which users employed, is the source of some new knowledge appeared in V_2, by theorem 1, V_0 is an independent thesaurus restricted to agents, not controlled by V_1. On the contrary, V_2 controlled by V_0, the reason is that new knowledge comes from every *agentID* that is marked in the V_0. So, $I(V_0)$, interpretation to V_0 based on the description logic, shows user application context in the V_2 thesaurus on the semantic maining.

Definition 1 If the construction of an ontology *K* accords with *OWL-DL* standard, then we will call the ontology *K* supports *OWL-DL*.

Theorem 3 *OWL-DL supports the ontology K_2 based on IOWL-DL.*

Prove: The ontology K_2 can be divided into two relatively independent components: K_0, K_1. Among them K_1 is an ontology in the application field, it is constructed in the abstract syntax of *OWL-DL*, obviously K_1 supports *OWL-DL*; Considering the *IOWL* and section three, each agent is registered in the mutual ontology, and merges into its own cluster (class) according to his semantic data structure, and belongs to an enumerated *agentID* set under an *agent-linkID* chain. *AgentClass* is a kind of abstract agent type based on *datastruct*, and *Agent-Link* is further an abstract that is based on class *AgentClass*, that is to say, independent thesaurus V_0 itself has been a construction based on *OWL-DL*. On this meaning, the *IOWL-DL* definition in this chapter submits itself to *OWL-DL*, so *OWL-DL* supports the ontology K_2 based on *IOWL-DL*.

Further, consider the reasoning of *OWL-DL* can do it too in K_2. Then theorem 4 can be concluded.

Theorem 4 *The ontology* K_2 *based on IOWL-DL can support the reasoning of OWL-DL.*

74.6 Conclusions

One key issue in multi-agent systems is to develop a mechanism of being restricted administrably and how to interact and exchange information autonomously across applications among multi-agents. A new restrictive administrable approach is proposed in this chapter, which differs from the current various methods, and the standard *OWL* is extended by adding agent-link in the form of axiom that includes the link itself, its property, data structure, trigger sign, and so on. This *IOWL* itself constructs a administrable restricted model, in which a clustered structure is constructed according to the relation of various subdomains. In the fifth section, the author has also come to the conclusion that an ontology based on *IOWL-DL* can run the reasoning of the *IOWL-DL*. Therefore, this extension has solved the ontological administrable issues with agents from the ontology itself, and is a feasible solution to reflect the heterogeneous context. This chapter comes from our current work, and certainly there are still many problems to be solved which we want to study further.

References

1. W3C Recommendation 10 February 2004,"OWL web ontology language semantics and abstract syntax", http://www.w3.org/TR/2004
2. FIPA - Foundation for Intelligent Physical Agents website.http://www.fipa.org/
3. Wenyu Z, Lanfen L, Jiong Q, Ruofeng T, and Jinxiang D (2006) An Ontology-based functional modeling approach for multi-agent distributed design on the semantic web, CSCWD, LNCS 9865:334–343
4. Reference "description of the DAML + OIL ontology markup language". http://www.daml.org
5. DAML - Darpa agent markup language website.http://www.daml.org
6. Marek O, Vladimir M (2009) Ontologies for multi-agent systems in manufacturing domain". DEXA'09, 1529-4188 /02,2009
7. Zhang L, Zhang W, Wang Q (2007) An Ontology-based collaborative reasoning strategy for multidisciplinary design in the semantic grid. CSCWD 2006, LNCS 4402:419–427

8. Rodriguez S, et al. (2011) Agent-based virtual organization architecture. Engineering Applications of Artificial Intelligence, doi:10.1016/j.engappai.(2011)
9. Yang SY, Chang YY (2011) An active and intelligent network management system with ontology-based and multi-agent techniques. Expert Syst Appl 38:10320–10342
10. Sharp B, Atkins AS, Kothari H (2011) An ontology based multi-agent system to support HABIO outsourcing framework. Expert Syst Appl 38:6949–6956
11. Afacan Y, Demirkan H (2011) An ontology-based universal design knowledge support system. Knowl-Based Syst 24:530–541
12. Jung JJ (2011) Exploiting multi-agent platform for indirect alignment between multilingual ontologies: A case study on tourism business. Expert Syst Appl 38:5774–5780 5775
13. Babitski G, Bergweiler S, Hoffmann J (2009) Ontology-based integration of sensor web services in disaster management. GeoS 2009,LNCS 5892:103–121
14. Lezcano L, Sicilia MA, Rodr C, guez-Solano (2011) Integrating reasoning and clinical archetypes using OWL ontologies and SWRL rules. J Biomed Inform 44:343–353

Chapter 75
Research on the Access Control Model under Grid Environment

Zenan Chu, Xinzhi Guo and Xinfa Wang

Abstract The protection of a more reasonable and safer means of grid authorization, which guarantees the efficient implementation of grid tasks, is required by the application and development of grid technology. Aiming at some difficult problems caused by Role-Based Access Control (RBAC), introducing the concept of context, tasks and conditions as well as the addition of monitoring devices, this paper proposes a Multi-restriction Access Control model on grounds of RBAC with reference to the existing traditional access control methods, meets the minimum privileges principle, and achieves the purpose of dynamic grid authorization, conditional sharing and secure interoperability among cross-domain resources. Experiments show that this approach can effectively avoid the insecurity of static authorization holders, and has superiority over other approaches in security.

Keywords Grid security · Access control · Role · RBAC · MRAC

75.1 Introduction

Grid computing has become a high-performance computing method based on the Internet, but a lot of academic and business applications, and grid security problems have always restricted rapid grid development [1]. Grid safety requires using

Z. Chu (✉) · X. Guo
Anyang Institute Of Technology,
Huanghe Rd. 73, Henan Anyang 455000, China
e-mail: chuzenan@yahoo.com.cn

X. Wang
Henan Institute of Science and Technology,
Henan Xinxiang 453000, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_75,

large amounts of resources, dynamic application and release resources [2], establishing credibility and realizing the dynamic relationship, implementating the safety interoperability between different resource management domains and trust domains.

At present, the access control model under grid environment protects resources mostly from angle system, which cannot solve the above requirements. Traditional access controls are based on access matrix, against system extension and manage maintenance. Role-based Access Control (RBAC) uses the main body of static authorization, no logo in object class jurisdiction as stipulated by the application range, which lack sufficient flexibility and dynamics [3]. Previous studies have many solutions to these questions, but the period of validity of the control, the dynamic authorized, multiple management domain, the context awareness etc. problems are without any well-integrated solutions [4]. To address this, this paper proposes the multi-restriction access control model, making it not only associated authorized with role but also considered in context, concrete task request and to condition the influence, thus realizing cross-realm resources conditions sharing, and security interoperability, according to the context, the task execution state, user attributes and object attribute changes to the dynamic authorization.

75.2 Principle Basis

75.2.1 Traditional Access Control Principle

Traditional access control based primarily on main attribute, object attribute and access control to authorize, mainly means independent access control and mandatory access control. Traditional access control model based access matrix to authorized, when the subject and object numbers increase, the growing of matrix becomes very large, against system expansion, and is hard to manage and maintain.

75.2.2 Role-Based Access Control Principle

RBAC Model contains user sets, role sets, access authority, conversation and restriction, access authority into operation sets and object sets. Introducing the role concept with permissions will have users connected effectively. The core, to define multiple roles and set the corresponding user access, was awarded the role, to get access. RBAC can be based on practical role setting, authorized more agile and can realize the fine grain access control. But, role based on subject attributes and object attribute by the static authorization way, has still some problems [5]: (1) in RBAC Static authorization, to perform tasks, the subject has prior permissions, task execution or executed will still have access, and makes the system safe from hidden trouble, there is a large against management authority. (2) RBAC Authorized process only considers roles that did not consider the task, to a different task as long as the same role can get the same permissions, not satisfying the authorized “minimum privileges” principle. (3) as RBAC Authorized in advance, the cause of RBAC cannot dynamically consider the changing of subject’s or object’s attributes.

75.3 Multi-Restriction Access Control model

The groups authorized technologies based on role, although it overcame the disadvantages of enabling access control list independently authorizing the increases indefinitely, while it could not solve the problems existing in the role authorized itself well. For grid needs, we proposed the Multi-Restriction access control model. Because this model considers the multiple restriction conditions, the model is called Multi-Restriction Access Control (MRAC) [6], as shown in Fig. 75.1.

Based on task MRAC, it can realize the fine grain authorized and meet the needs of this mission. Permissions can have timely recovery, avoiding abuse of the authority. Through introducing the monitoring function, preventing achieving dynamic authorized by the user attribute, the execution environment changes but appears as over-rides operation. At the same time, since condition testing can implement environmental or other factors, performance condition could not eliminate the phenomenon of operating authority.

75.3.1 Model Composition

In Multi-Restriction Access Control, the elements are defined as follows:

Users: operating object entity. Attributes including identity, safety, level, credibility, account information, etc.
Objects: user operation objects. Object attribute including security level, credibility, etc.
Role: a group of users executing operating set within the organization, means that the user functions within the organization.
Task: a group of operations that the user completes implements a requestion.
Conditions: use authorized rules authorized, let users access objects prior to inspection decision factors set. Conditions are environmental or decision factors for system.
Context: might trigger a role and permissions changing factors set, including the dynamic Context environment as well as the dynamic produced sensitive information.

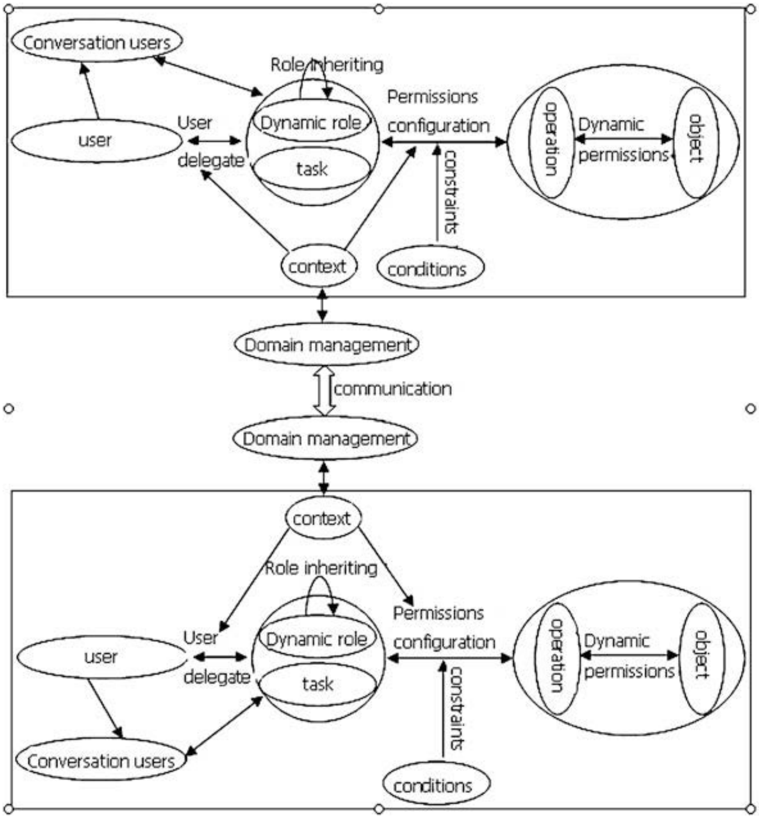


Fig. 75.1 MRAC access control model

Active Role: Dynamic Assigned roles (u:Users, c:Cont) → 2^{Roles}, Users and corresponding permissions changes with context.
Active Permissions: Dynamic Assigned permissions (r:Roles, c:Cont) → 2^{Perms}, Users and corresponding role changes with the context.
Domain Management: Used for maintenance domain access control strategy.
Communication: Used for cross-realm access control requests, including domain between dynamic role mapping.

75.3.2 Domain Model Definition

MRAC domain model expressed as a six-member group (U, O, R, T, C, P). Among them, U stands for users, O for object, R for role, T for task, C for conditions, P for permissions. ATT (U) represent user attributes, ATT (O) the object attribute.

RPA stands for roles and permissions relations, RPA $\subseteq$ Permission $\times$ Role, with RPA (p, r) as the role r with permissions p.
The URA is the users and role relationship, URA $\subseteq$ Subject $\times$ Role, use URA (u, r) represent user u belonging to role r.
TPA is tasks and permissions relationship, TPA $\subseteq$ Permission $\times$ Task, with TPA (p, t) as task t requirements operating permissions p.
UPA says a user has permissions, UPA $\subseteq$ Permission $\times$ Subject, UPA (u, o, t, p) said the mission t, users u for object o have permissions p.
Policy (attributes): Attribute set of attributes according to users is the role of decision this user belongs to.

Definition 1 task execution of prior authorization:

(1) URA (u, r) = policy (ATT(u))
(2) UPA (u, o, t, p) = RPA (p, r) $\cap$ TPA (p, t) Among them, the P_1 said the permissions owned role r, P_2 says task t required privileges, P says mission t users u for object o have to permissions.
(3) Allowed (u, o, t, p) = preA (UPA (u, o, t, p), c)

PreA says authorized before the start mission, satisfy condition c, complete a task t user u for object o allowed the permissions are p.
In advance license, according to user attribute decision and role, which will take intersects with the users under the permissions and task required access, satisfy conditions. If context does not conform, it cannot complete authorized.

Definition 2 task execution Authority:

(1) Because ATT (u) $\neq$ Update (ATT(u)) so URA (Update(ATT(u)), r$'$) = policy (Update(ATT(u)))

Update (ATT (u)) shows that dynamic Update user attributes. r$'$ says users attribute changes, users u belong to the new role.

(2) UPA (u, o, t, p$'$) = RPA ($p_1{}'$, r$'$) $\cap$ TPA (p_2, t), Among them, the $p_1{}'$ says new role r' have operating privileges, p' says user role changes, in order to mission t user u for object o have jurisdiction.
(3) Because preA (UPA (u, o, t, p), c) $\subset$ onA (UPA (u, o, t, p), c), so allowed (u, o, t, p$'$) = onA (UPA(u, o, t, p$'$), c$'$), Because preA(UPA(u, o, t, p), c) $\supset$ onA (UPA(u, o, t, p$'$), c$'$), so stop. Among them, the onA said authorized in task execution, c$'$ says users attribute or execution environment changes condition.

MRAC increased in monitoring function, can realize dynamic authorized. Authorized users attribute change will trigger, judge whether or not to change role or conditions, and obtain new meet access, if new access contains old access, continues to perform tasks; Otherwise stop the task.

75.3.3 Domain Between Dynamic Role Mapping

75.3.3.1 Management Domain and its Tectonic Definition

Definition 1 Domain $D_i = (U_i, R_i, P_i, UA_i, PA_i, RH_i)$,Among them, i ∈ N, Says the first *i* domain; U_i says users set; R_i said role sets (if Guest $\notin R_i$, then Ri = Ri ∪ {Guest}); P_i says permissions set; UA_i says users to role assignment; PA_i says permissions to role assignment; RH_i says role level (Partial order relation $RH_i \subseteq R_i \times R_i$).

Definition 2 Grid system, Grid = {Di|Di = (Ui, Ri, Pi, UAi, PAi, RHi), i ∈ [1, n]}, Among them, n represent management domain number. D $_i$∈ Grid, According to their own domain setup define the corresponding role sets R_i and role hierarchy RH_i. R_i roles in only D_i meaningful, once losing context, can't get its semantic analytical. Therefore, to realize interoperability, each management domain in Grid must first consultation each other's role semantics. In D_i when user $user_{im} \in U_i$, and role $r_{in} \in R_i$ visit D_j, D_j according to the role r_{in} semantic in r_{in} and role sets R_j build a dynamic maps, get a combination role hierarchical relationships, provide decision-making basis for the implementation of the access control.

Definition 3 Domain between dynamic role mapping: $R_i$said role sets in D_i, RH_i says role level in D_i. If $(r_{ik}, r_{il}) \in RH_i$, Among them $r_{ik}, r_{il} \in R_i$, show $r_{ik} > r_{il}$. From role r_{ik} in R_i to role sets R_j in D_j role mapping is $f_{Rj(rjk)} = \{(r_{ik}, r_{jl})|r_{jl} \in R_j i \neq j\}$ If $(r_{ik}, r_{jl}) \in fR_j\ (r_{ik})$, recorded $r_{ik} \rightarrow r_{jl}$, said $r_{ik} \geq r_{jl}$. D_i role sets R_i to D_j role sets R_j all role mapping set: $R_iR_j = \bigcup_{k=1}^{|R_{i|}} f_{R_j}(r_{ik})$.

R_i role r_{ik} in R_iR_j mapping head set: $H_{R_iR_j}(r_{ik}) = \{r|(r_{ik}, r) \in R_iR_j\}$
R_i role r_{ik} in R_iR_j mapping tail set: $T_{R_jR_i}(r_{ik}) = \{r|(r_{ik}, r) \in R_jR_i\}$
R_i role r_{ik} ancestors role set: $Ancestor(r_{ik}) = \{r_{il}|r_{il} \in R_i r_{il} \geq r_{ik}, k \neq 1\}$
R_i role r_{ik} sons role set: $Child(r_{ik}) = \{r_{il}|r_{il} \in R_i r_{ik} \geq r_{il}, k \neq 1\}$

75.3.3.2 Domain Mapping Rules Between Dynamic Roles

D_i role sets R_i and D_j role sets R_j (i ≠ j) established dynamic role mapping, can adopt different rules. Define four rules:

(1) The default rules: $\forall r_{ik} \in R_i, f_{R_j}(r_{ik}) = \{(r_{ik}, Guestj)\}$ Among them Guest ∈ R_j
(2) Specifies rules: $\forall r_{ik} \in R_i, f_{R_j}(r_{ik}) = \{(r_{ik}, r_{jl})|r_{jl} \in R_j\}$
(3) Some designated rules: $\exists\, r_{ik} \in R_j, f_{R_j}(r_{ik}) \neq \Phi$, If $\exists\, r_{il} \in R_i,\ r_{il} \geq r_{ik}, f_{R_j}(r_{il}) \neq \Phi$, then $f_{R_j}(r_{il}) = f_{R_j}(r_{ik})$

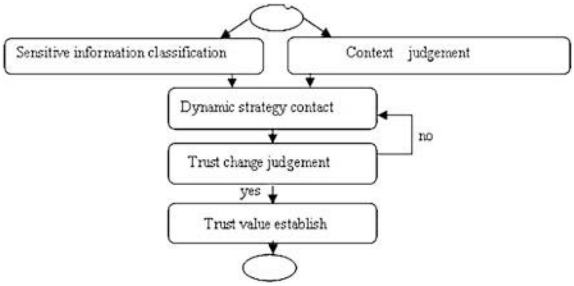


Fig. 75.2 dynamic security engine

(4) Temporary specified rules: When domain D_j cannot establish appropriate for the r_{ik} role mapping, first set up temporary roles tr_{jl}, According to r_{ik} semantics tr_{jl} distribution permissions, make $f_{R_j}(r_{ik}) = \{(r_{ik}, tr_{jl})\}$.

The rules (1) mapps all roles in R_i to role $GUEST_j$ in R_j, providing basic interoperability, but the lack of flexibility and of different roles without discrimination. Rules (2) for each role in R_i explicitly indicates the role mapping in R_j, although it can differentiate between different roles, but it also brought heavy role mapping burden. Rules (3) is the rule (1) and rules (2) compromise, do not have to compromise in each role for R_i explicitly indicate that in the role of R_j mapping, no explicit indicated mapping role through its role hierarchical relationships can be indirect gain role mapping. The rules (4) used not to establish appropriate role mapping, is to rules (2) and rules (3) supplement.

75.3.4 *Based on the Context of the Dynamic Security Engine*

In grid environment, the need for dynamics produces sensitive data security protection, taking into account the context of the proposed changes, causing the concept of dynamic security engine. The process of dynamic security engine role is shown in Fig. 75.2.

75.4 To Achieve the MRAC in Grid Platform

75.4.1 *MMRAC Implementation Process*

MRAC implementation is shown in Fig. 75.3. Users in a domain login, through the logon authentications, access and authorize users according to related domain between the domain mapping rules set authorized users to delegate context agent role to users host environment and initiate relevant role state machine, context agent control state machine to dynamic adjustment activities role. Once allowing resources and services, it will provide service into the grid startup corresponding

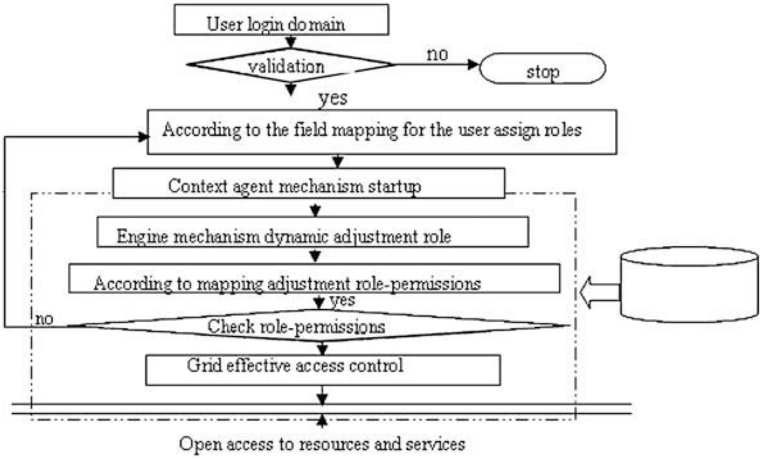

User login domain
validation
no
stop
yes
According to the field mapping for the user assign roles
Context agent mechanism startup
Engine mechanism dynamic adjustment role
According to mapping adjustment role-permissions
yes
no
Check role-permissions
Grid effective access control
Open access to resources and services

Fig. 75.3 MRAC basic processes

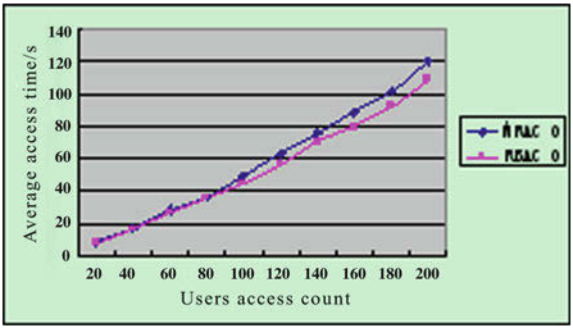


Fig. 75.4 more users to access the results

role permissions state machine. A character might be related to multiple resources. According to the resources and services and other circumstances dynamic adjustment role permissions are set. Permissions state machine visit a moment at the moment the resources mean in this context, in which the user belongs to certain roles of operating in a certain context.

75.4.2 Results Analysis

In order to test the function of the system, we deployed three nodes. Nodes A, B and C use Linux platform, installed JAVA environment, GT4, Tomcat and PostgreSQL database system. The user-role relationship table in Authorized servers are influenced by the number of users. Single user access to resources, MRAC and RBAC distinction is not obvious. In many tests users access to authorize server, by writing a script producing multiple, simultaneous operations of the visit statements. Many users to access test results are as shown in Fig. 75.4.

After verification, the model experiments deployment MRAC grid environment based on multiple successfully realized restriction conditions of dynamic access control. Compared with RBAC, when there are 200 users sending resources requests, the system response time is higher by 11%, a proof of this model about decision-making that higher efficiency can increase in the case of loss of grid, thus greatly enhancing security with higher practicability.

75.5 Conclusion

The application and development of grid technology requires a more reasonable, safe grid accredit way as guarantee, to ensure effective implementation of grid task. Based on the existing access control methods, RBAC, based MRAC model, the introduction of context, tasks, and conditions of the concept, joining monitoring function, realized the dynamic grid environment, guaranteed the authorized principle of least privilege, and avoided static hold authority because of users and appearing all sorts of unsafe phenomenon. At present, this is the study model of the stage, there are still some problems needing further in-depth study.

References

1. Zhou W, You J, He J (2006) Design and implementation: a model of privilege management infrastructure based on RBAC. Microcomput Inf 5(3):3–36
2. Joshy J, Craig F (2005) Grid computing [M]. Tsinghua University Press, Beijing
3. Humphrey M, Thompson MR, Jackson KR (2005) Security for grids [J]. Proc IEEE 93(3):644–652
4. Qiang W, Hai Jin, Shi X (2005) RB-GACA: a RBAC based grid access control architecture [J]. Int J Grid Util Comput 1(1):61–70
5. Yao H, Hu H, Huang B et al (2005) Dynamic role and context-based access control for grid applications [C]. In: Proceedings of the 6th International conference on parallel and distributed computing, applications and technologies. [S.I.], IEEE Press, pp 404–406
6. Alfieri R, Cecchini R, Ciaschini V (2005) CAS, an authorization system for virtual organization [C]. In: Proceeding in CAS conference. Springer, Berlin

Chapter 76
The Illegal Ex-connection Monitoring System of Electric Power Network

Junhui Fu, Yanpei Liu, Yafeng Han and Xueyong Li

Abstract The monitoring module of the illegal Ex-connection monitoring system in this chapter has the strict judgment logic and thorough detailed judgment process, which overcomes the network ways that traditional monitoring model can not monitor. The three-way handshake mechanism is adopted in the communication protocol, which can prevent a malicious program posing as server to the client, sends a message. After testing, the monitoring model demonstrates strong robustness, adaptability and security environment.

Keywords Electric power network · Illegal Ex-connection · Monitoring · Process hide

The monitoring system of Electric power and data scheduling network as a critical infrastructure of electric power system, the security issue has been the focus of state departments. The intranet power network has done well in resisting the attacks of hackers, viruses, malicious code from outside network, especially the attacks against the group, so as to ensure real-time closed-loop power control system and the security of data scheduling network. However, if a host connects to the external network or if in such a closed-loop scheduling data network, it will lead to, adding a physical link without firewall protection between the internal and external network, so as to damage the security of the existing internal network, hackers are likely to access the internal network by the host, by the means of "overlapping" to "bug" the transmission power control information, then to collect and "tamper" the internal key information or sensitive information by sniffing, cracking password, etc., to attack other hosts on the intranet by the host as

J. Fu (✉) · Y. Liu · Y. Han · X. Li
School of Information Engineer, Henan Institute of Science and Technology,
453003 Xinxiang, China
e-mail: 599246483@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_76,

a "springboard" [1–3]. Therefore, it has become an important research direction to solve illegal Ex-connection problem of electric power network [4].

76.1 The Study Status of Host Within Intranet Illegal Ex-connection Monitoring Technique

Currently there are two main methods about monitoring illegal Ex-connection system, one is based on detection of internal and external network monitoring program, referred to as dual-architecture generally used in large-scale local area network; the other is based on monitoring the device driver Agent monitor program, namely the Client/Server (C/S) architecture, two architectures have their own advantages and disadvantages. In local isolated networks generally use the C/S architecture. C/S structure that is illegal Ex-connection monitoring system based on monitoring agent, is an embedded program in host procedures, embed the monitor program in device drivers by the development of bottom system, through monitoring illegal foreign Communication with the device to achieve monitoring the illegal acts of Ex-connection. Based on the traditional C/S structure the monitoring system consists of two parts: the monitoring agent and monitoring alarm center [5, 6].

The task of monitoring alarm center is to take charge monitoring policy setting and issue monitoring agent software, and to process the real-time feedback of the monitoring agent dial-up information, when monitored host violate the security policy to dial-up, to generate the real-time alarm information [7].

Monitoring agent is to take charge of monitoring the host's dial-up behavior, cut off the dial-up or alarm, meanwhile store the relevant information in the log to prepare for queries.

C/S structure of the illegal Ex-connection has advantage of not only monitoring the online hosts but also off-line hosts, and can disconnect the user's illegal Ex-connection is not be achieved in the double structure. Even if the host is offline, the monitoring agency on it will monitor the user's dial-up behavior as the same, once connect, monitoring agent will immediately establish contact with the monitoring alarm center, and will report host offline user behavior during this time. But C/S structure is likely to restrict on cross-boundary local area network, and the protective measures between monitoring agents and the alarm center will result in omission, so will these deficiencies [8, 9]. The existence of these problems can be solved using a variety of ways. In this paper, combined with the specific needs of electric power company about internal network illegal Ex-connection based on C/S structure, gives an improved surveillance detection logic model [10].

76.2 Improved Monitoring System Architecture

The basic function of The Illegal Ex-connection Monitoring System of Electric Power Network is real-time monitoring of each host in the intranet, cut off the dial-up connection of the host once the host tries an illegal Ex-connection. In addition, it also has functions: alarm, logging, real-time viewing, process-hidden features and other extension functions generally used in network environment and single host environment [11]. The software adopts C/S architecture, including server subsystem and client terminal subsystem.

76.2.1 The Server System

This subsystem works as a master control program. It runs on stable server in local area network, receives signals sent by each host and detects active host in network currently, and provides management and configuration interface to the administrator. The subsystem consists of three modules: the control center module, the communication management module and the log management module.

The control center complete system setting functions and initializes the system according to the settings (mainly by communication managers and log managers); communication managers supervise all the network traffic, show the data receiving on real time monitoring UI, and generate server logs by log manager; The log manager is responsible for formatting information, provides log view, search and other function for the log UI.

76.2.2 The Client System

The client system consists of six modules: check module, protection module, communication module, situation processing module, log management module and information processing module [12].

From the point of implementation, the client subsystem is mainly composed of the following four threads: the main thread, which mostly accomplishes the function of information processing module; the check thread, which completes the function of inspection module; the protection thread, which completes the main function of protection module; the communication thread, which accomplishes the function of communication module. The main thread generates other sub-threads, control and abstract information from sub-threads, produces corresponding action. The relationship between the various threads and each modules is shown in Fig. 76.1.

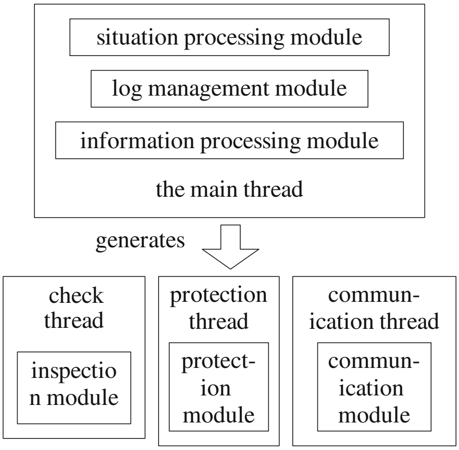


Fig. 76.1 The client of implementation module

76.3 Design and Implementation of Key Modules in System

76.3.1 Design of Monitoring Detection Model

The traditional control model could find all the ordinary dial-up to access Internet, but could not judge the ADSL, wireless Internet access and other Internet services. This model could not only judge the general features of the Internet, but also be able to estimate the features shown by ASDL, wireless Internet and other way access to internet.

(1) Check the existence of PPP virtual network in host, if exists, can judge illegal network connection, if not, go to (2).
(2) Check whether there are multiple network cards in host, if there are multiple network cards, then try and establish connection to the outside world (by means of establishing the connection with the well-known website, the process to establish the connection with the Internet well-known websites is as follows: first, resolve the domain name, if successful, then connect to the other host on port 80, if the connection has been built successfully, it illustrates that watched host has connected with Internet), if the connection is successful, it can be judged that there is illegal network connection; if the connection fails, no illicit network connections exist. If there are multiple network cards, go to (3).
(3) Check whether the server transfers the legal gateway list, and if so, check whether the current gateway exists in legal list (the legal list is provided by administrator), if it exists there is no illegal network connections, if it does not exist, then to try connect the outside world; if the server does not pass legal gateway list, go to (4).
(4) Try to connect external.

From the above strategy, criterion most time is whether try to connect external, in order to ensure reliable, a few well-known Web site domain names are added. Only two exceptions, one is finding the PPP protocol existing that can immediately determine the illegal Ex-connection Internet; the other one is the current gateway existing in the current legal list of the gateway, that can be determined by no illegal network Ex-connection.

When illegal network connection has been judged to exist, the program will cut off the connection work. To do: (1) to identify the items about server connected in routing table, and retain these. (2) delete the other items in routing table. This will not only remove the illegal connections, but also retain the normal connect between the client and service.

76.3.2 The Implementation of Client Process Hide Technology

Since the client is installed on the monitored host, so as to detect whether the monitored computer acts against the requirements, and report the behavior information to the server, so it requires the good counter-surveillance feature, which demands the client hidden to run in the monitored host, so the client can not be found by users, and can not be stopped even if found or deleted [13].

In the VC, in addition to function HOOK, DLL in Windows is also the executable file. DLL file does not have program logic, that is made up of a number of performance function, which can not run independently, generally called by processes. Because the DLL file can not be run independently, so the process list does not appear in DLL, therefore we have written DLL into the client process, and run it through processes, so the processes will occur in the process list, DLL will not appear, thus completely solving the shortage of the hidden process with HOOK functions blocking all the system messages.

76.3.3 The Implementation of Server and Client Communicate

Server and client communication protocol uses three-way handshake mechanism:

(1) server first sends a packet containing number A composed of random 128-bit to client;
(2) client returns B of 128-bit, which is a certain algorithm encryption of A (algorithm has been agreed);
(3) server encrypt number A using algorithm, then compare with the number of B, if equal server sends a 128-bit C, which is the encryption of B; or else disconnect;
(4) client encrypt B with algorithm, then compare with the number of C, if equal client startup a normal communication; or else disconnect.

In this way the connection is established, and it can prevent malicious programs imitating server to send messages to the client, such as stopping client command, to ensure the secure connection.

In addition to three-way handshake, server and client communication is also in accordance with a communication protocol, including:

(1) client regularly sends online declarations;
(2) server periodically sends polling packets;
(3) client accepts server's control commands [13]. ① terminate client command A_CMD_STOP. ② log extract command A_CMD_QUERY. ③ suspend monitoring command A_CMD_PAUSE;
(4) client transfers data to server: ① two types of alarming information; methods of Ex-connection dial-up, ISDN, ADSL. ② log search information.

Because the "flow" has no boundaries, so programmer determine the boundaries of the data, using three-step approach, the fixed head length of 8 bytes, (the boundary alignment), process of data is: first read head Information, then redistribution the buffer, finally read the data, and interpret according to type code. Type code is defined by the macro.

76.3.4 Implementation of Server Monitoring Host Online

Server sends probe packets to host on internal network to detect whether the host is online and its status. In order to ensure the accuracy of detection, the system uses the ARP protocol. ARP protocol is used for transmitting the hardware address between the LAN hosts, only for local area network. Most of the firewall does not filter ARP protocol, because once filtered will result in obstacle communicated with. In the windows system, cache of the machine can be seen with the arp—a command, including the LAN IP address and the hardware address of each host. In other words, as long as the arp cache of the local host saves records of the host, identifying that the host must be boot-strap. Using this idea, we adopt the following detection mechanism [14, 15]: (1) clear arp cache records of local host; (2) send ping packets to the host within the LAN; (3) Check arp cache of local host to find the IP address records whose hardware address not be 00-00-00-00-00 - 00, mark it as online.

76.4 Test and Debug

In the test environment, mainly do three important test of the system including management test, function test and log audit test. Test items of Management test consist of system installation, management manner, administrator identification, the division of authority, etc., afterward testing in several aspects, the results meet the demand. Test items of function test involve fetching the monitored host information, dial-up, event detection, the security of controlled host and so on, after conducting the tests, results are satisfactory, and show superiority particularly in event detection and security of controlled host. Log audit test is comprised of alarm information, system configuration management logs and record logs and so on. The results have demonstrated the completeness and superiority of logs alarm information.

76.5 Conclusion

Analyzing the current status of illegal Ex-connection monitoring system, bring forward a timely technical scheme to cut off the illegal Ex-connection computer programs, test results show that the system has the following features: real-time, automatically, accuracy, availability, efficient and compatibility. Here cutting off means illegal connections between the computer and external, rather than connection between computers within the network.

References

1. Yong S, Yixian Y (2004) Design and implement of illegal foreign Ex-connection monitoring system [J]. Netw Secur Technol Appl (12):21–22
2. Junwei W, Tao W (2003) Analysis the illegal connection monitoring system technology [J]. Control Aircr (9):17–18
3. Guanghan S (2005) Research and implement of illegal monitoring system [D]. Jilin University, Jilin, pp 134–137
4. Xiaoqing Y (2000) Network security defense physical isolation technique [J]. China Electron Publ (6):59
5. Yingliang W, Gang W (2000) Network intrusion and security countermeasures [J]. Appl Res Comput (11):37–39
6. Guohui H (2001) Dial-up network security solutions [J]. Comput Mod (6):137–139
7. Yuejin Z (2003) Network security and computer crime investigation technology [M]. Tsinghua University Press, Beijing, pp 201–221
8. Yingying W, Shengwei M, Yanbin M, Feng L (2010) Vulnerability assessment of power grid with distributed generation based on complex network theory. J Syst Sci Math Sci 06: 015–021
9. Ming TF (2010) Development of driver and storage procedures for power network meters. Program Controller Fact Autom 10:69–72
10. Xiao-lu D (2010) SSL-based security model of electric power network communication. Comput Knowl Technol 36:020–024
11. Yijia C, Guangzeng W, Zhejing B, Chuangxin G (2009) Temporal and spatial evolvement model of power grid. Electr Power Autom Equip 01:003–006
12. Xiuxia Y, Yi Z (2009) Study of large isolated power network topology structure identification and fault flow calculation. J Wuhan Univ Technol (Transp Sci Eng) 01:024–028
13. Li-jie D, Yi-jia C, Mei-jun L (2008) Dynamic modeling and analysis on cascading failure of complex power grids. J Zhejiang Univ (Eng Sci) 04:020–025

14. Sheng-wei M, Xiang-ping N (2008) Several new progresses on the complexity research of the interconnected power network. J Changsha Univ Sci Technol (Nat Sci) 5(2):103–106
15. Ming-min W, Jian-quan W (2008) Power network node ordering technology basd on particle swarm optimization. Mech Electr Eng Mag 25(8):87–89

Chapter 77
Design and Implementation of Temperature Programmed Controller by Fuzzy PID Based on SOPC

Suying Yang, He Zhang, Jianying Lin and Miaomiao Gao

Abstract According to the designing requirement, this paper puts forward a new way of combining the fuzzy PID control IP core with programmed temperature control of the Nios II soft processor by SOPC technology. The bottom layer of the fuzzy PID controller in IP core is achieved by the fuzzy rule table and the traditional PID algorithm. The output of the fuzzy rule table is the increment of PID parameters, which can be obtained from the deviation and the deviation variance ratio of the control system, and data transfer between the top layer and the Nios II processor is fulfilled through the parallel registers and Avalon interfaces. This controller is tested in programmed temperature control in which its object is of the micro reactor with features of first-order inertia and pure delay. The results show that the control can realize good traceability, zero steady-state error, non-overshoot, and strong anti- interference effect. Its time delay is no more than 10 s.

Keywords Temperature programmed control · Fuzzy PID · SOPC · IP core

77.1 Introduction

Traditional PID control is still the mainstream in chemical production. The PID controller can apply to different control plants, and obtain better control performance by actual parameters adjustment except for dynamic performance deteriorates [1]. With the rapid development of computer technology, some new modern control theories are well used in industrial control applications [2], such as fuzzy PID control algorithm which is the combination of PID control algorithm and fuzzy algorithm.

S. Yang (✉) · H. Zhang · J. Lin · M. Gao
Dalian University Technology, Dalian 116024, China
e-mail: rr319@dlut.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_77,

Specific to the features of micro-reactor, a control system which possesses obvious characters for time-lag, time varying, and nonlinear models is required, fuzzy control can meet the above conditions. The fuzzy PID control can better achieve temperature control of the reactor. Fuzzy reasoning rule makes online adjustments to the three parameters of PID controller [3]. The innovation of this article achieves customizing fuzzy PID IP core by hardware description language, and then sets the communication and control interface to combine with the embedded Nios II soft processor, finally the design of programmed temperature controller by fuzzy PID based on SOPC is accomplished, the fuzzy PID controller with Nios II soft core has the advantages of both the micro processor and FPGA system.

77.2 Design Model of Fuzzy PID Programmed Temperature Controller

Fuzzy PID programmed temperature controller is made up of fuzzy PID IP core and Nios II soft core processor system which is with the function of programmed temperature control.

The core mission of Nios II processor is the temperature programmed control which includes the response of key interrupt, temperature parameters setting, programmed temperature control, and the schedule of different operating condition. In addition, Nios II soft core processor is also responsible for the schedule and the operation of function modules, such as data acquisition, communication, and PWM control.

The core task of controller IP core is to realize the fuzzy PID controller. The development environment is Quartus II and SOPC Builder. The structure of the design model with fuzzy PID IP core can be divided into two layers. The bottom layer is PID controller, which realizes the combination of PID control algorithm and fuzzy rule table by introducing hardware description language on parallel FPGA, and achieves the basic function of fuzzy PID control. The top layer achieves data exchange between IP core and Nios II processor system. The module makes the micro processor access to the hardware language on FPGA seamlessly by the driver of IP core.

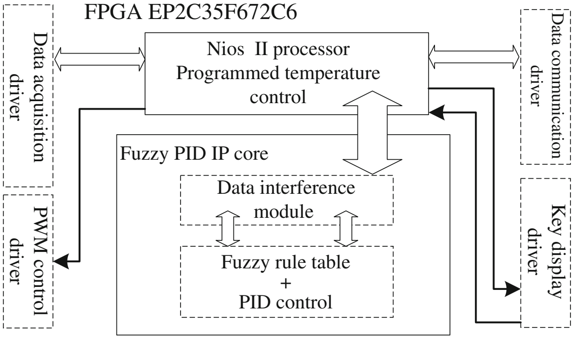
The design structure of fuzzy PID programmed temperature controller is shown in Fig. 77.1.

77.3 Fuzzy PID Controller IP Core

FPGA design of fuzzy PID. Reference [4] implements the PID control algorithm by parallel Verilog HDL logical language, and achieves the FPGA design of PID control. The structure of the fuzzy PID controller is shown in Fig. 77.2.

The simplest implementation of fuzzy controller is that it makes a series of fuzzy control rules transform into a query table which is stored in the controller for

Fig. 77.1 Design structure of fuzzy PID IP core



querying. The fuzzy control is the most basic form for simple construct and convenient use [5]; the first step is to confirm the structure of the fuzzy controller, and take ΔKp, ΔKi, and ΔKd as the output of fuzzy reasoning, taking e and Δe as input. The second step is to create fuzzy control rule in which e and Δe are regarded as the deviation and the deviation variance ratio of the input temperature, ΔKp, ΔKi, and ΔKd are regarded as output. The third step is to confirm the assignment table of fuzzy variable and the membership function of fuzzy language variable, and then confirm the degree of membership of elements in the domain for fuzzy language variable [6]. The last step is to establish fuzzy control table, the parameters are simulated by Matlab through fuzzy reasoning, and then export table is reasoned by invoking the toolbox.

IP core data interaction. There are two types of data interactions, IP core parallel register set and Avalon bus interface. The internal register of IP core can be addressed through Avalon interface “base address + offset”.

The function of Avalon bus realizes the internal linkage between master and slave components of SOPC system. Custom fuzzy PID module must provide some signals as a subordinate peripheral of the SOPC system. The signals make the address, data, and control of the Avalon connecting with the module, each signal needs to be assigned an effective signal type of Avalon.

In general, application program connects the basic equipment with Hardware abstraction layer (HAL) system library other than accessing directly.

77.4 Design of Programmed Temperature Control

The temperature control program is accomplished in the Nios II processor. In general, there are two main tasks in the Nios II soft processor, one is the setting of temperature segment, time segment, and PID parameters, the other is to realize programmed temperature control by invoking fuzzy PID IP core.

Design of key interrupt. The function of the key in the design is to set the parameters of temperature segment and time segment in programmed temperature control. The function prototype of key interrupt is:

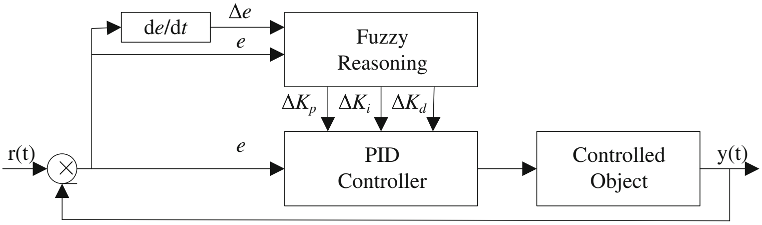
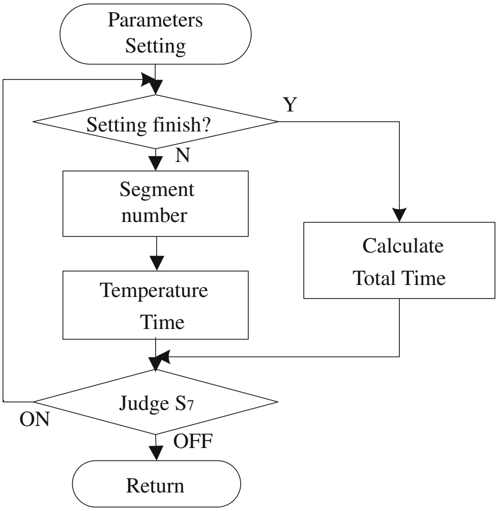


Fig. 77.2 Structure of the fuzzy PID controller

Fig. 77.3 Flow chart of parameters setting



void button_interrupts (void *edge_capture, int id);

Variable edge_capture represents the value of integer point. The id represents the interrupt number. The edge_capture is assigned with an integer value according to different keys when interrupt occurs.

Temperature parameters setting. Programmed temperature control can be described by three parameters including segment temperature ci, segment time ti, and segment number n (0 < i < n). The parameters setting flow of Nios II soft processor controller is shown in Fig. 77.3.

Programmed temperature control. The value of the temperature is measured by the control program in every control circle when the processor is working at the running state of the control. The program calculates the heating rate and current temperature expectation according to the segment temperature and segment time, and then writes the current temperature expectation and the measured value by invoking fuzzy PID controller IP core, the control value calculated by the IP core is transferred to the processor, solid-state relay which is in charge of temperature control obtains the value by the output of PWM control.

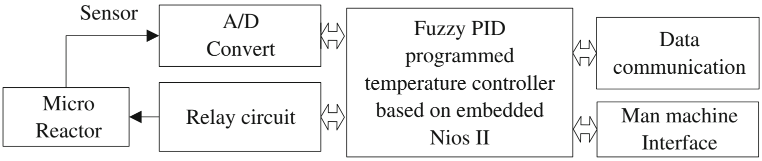


Fig. 77.4 Blocks of temperature programmed control system

77.5 Verification

Programmed temperature control can apply to the catalytic engine and chemical control field. A programmed temperature control system verifies the functions of fuzzy PID programmed temperature controller IP core. The core of control system is embedded Nios II fuzzy PID programmed temperature controller. The blocks of temperature programmed control system is shown in Fig. 77.4.

The sensor of the control system is a K-type thermocouple with the measured range from 0 to 1300°C. The analog–digital conversion (A/DC) with twelve bits converts analog voltage into digital data. The output of the controller is the Pulse Width Modulation (PWM) control IP core, which is in charge of the on or off of the solid-state relay. Seven switches and three keys are used for the state control and parameters setting of the temperature programmed controller. Data communication module realizes the communication between the controller and the computer by RS232 serial circuit.

Micro-reactor is the temperature controlled object for testing the control effects of the design. Its mathematical model of the micro reactor can be described by

$$\mathrm{W}(s) = \frac{1.21}{1480s+1} \cdot e^{-8.64s} \tag{77.1}$$

where the pure delay time is 8.64 s, the proportional coefficient is 1.21, and the time constant is 1480. The controlled object is first order inertial system with pure time delay, and the heating speed of the object is fast. The temperature curves are adjusted to a steady state and the remaining stable according to fuzzy PID control and the adjustment of overshoot. Six segments effect chart of programmed temperature control is shown in Fig. 77.5.

The initial temperature of the micro-reactor is 35°C, and the setting value is 40°C. The first segment temperature is 100°C. From the chart, it can be observed that the curve quickly catches up with small fluctuations and non-overshoot. The second segment is 100°C constantly, the measured temperature is 100°C, the steady-state error is zero, and the curve is smooth. The temperature in the third segment is up to 150°C, when the measured value is 150°C, the curvet is of non-overshoot and the relay time is less than 10 s. The fourth segment is also constant with the setting temperature of 150°C, the measured value is 150°C with good stability. The temperature in the fifth segment is set up to 200°C, the measured value is 200°C with non-overshoot, and relay time is less than 10 s. The last

Fig. 77.5 Six segments effect chart of temperature control

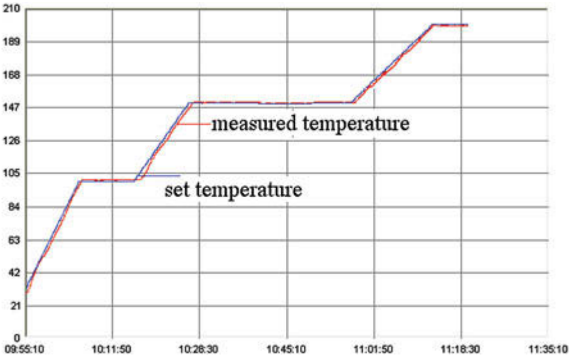


Fig. 77.6 Test for anti-interference performance

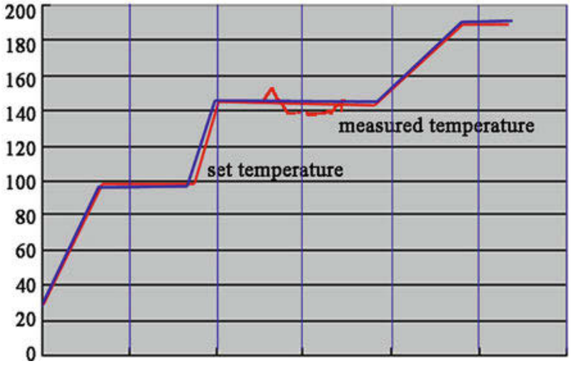
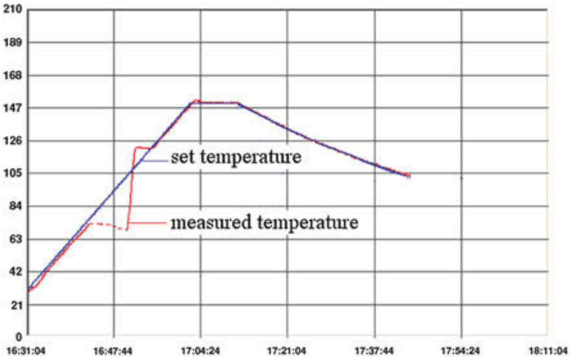


Fig. 77.7 Test for anti-interference performance



segment is constant with the setting temperature of 200°C, the measured value is 200°C without steady-state error.

The controller always suffers the interference from the environment. The interference tests of the controller are made to verify the performance of anti-interference. One effect of the test is shown in Fig. 77.6. When the temperature is stable at 150°C, take the thermocouple outside for a moment, then put it back after 10 s. The other test is shown in Fig. 77.7. When the temperature is 90°C, cut off the relay making the object cool down for a moment, and then turn on the power to continue.

The results show that the programmed temperature controller tracks the temperature quickly with non-overshoot, the steady-state error is zero and the recovery time after the disturbance is less than 6 min. The performance of the fuzzy PID programmed temperature controller is adequate for the micro reactor.

77.6 Conclusions

The key idea of the design based on SOPC in the paper is the combination of the fuzzy PID control by IP core with programmed temperature control of the Nios II soft processor, to realize the programmed temperature control of the fuzzy PID controller. The FPGA IP core describes the fuzzy PID algorithm and fuzzy rule table with Verilog HDL language. The controller adopting the fuzzy PID control algorithm achieves a better effect than traditional PID control. The design considers the micro reactor as a controlled object. Fuzzy PID based on SOPC controls the temperature up to the set value, the controller can achieve good tracking performance, the temperature rise curve is smooth without overshoot and steady-state error, lagging time is less than 10 s. The curve can keep steady after disturbance. For constituting the control system conveniently, the design provides man–machine interactive interface, including the keys and switches to set parameters. The controller linked to the computer observes the control effect directly. The design can bring about industrialization of SOPC technology and contribute to good practical application.

References

1. Fons F, Fons M, Canto E (2006) Custom-made design of a Digital PID Control System[C]. International conference on acoustics, speech and signal processing, Toulouse, 3:III–III.
2. Hu B, Ying H (2001) Review of fuzzy PID control techniques and some important issues[J]. Acta Autom Sinica 27(4):567–584
3. Kung Y-S (2009) FPGA realization of an adaptive fuzzy controller for PMLSM drive[C]. IEEE Trans Ind Electron 56(8):2923–2932
4. Joao L, Ricardo M, Cardoso Joao MP et al (2006) A methodology to design FPGA-based PID controllers[C]. IEEE international conference on systems, man and cybernetics, Taipei, vol 3, pp 2577–2583
5. Qi Y, Guo B (2008) Adaptive generalized common model control based on Fuzzy Logic Control[J]. Autom Instrum (4):5–8
6. Su D, Ren K, Luo J (2010) Programme and simulation of the Fuzzy Control List in Fuzzy Control[C]. The eighth world congress on intelligent control and automation, Jinan, pp 1935–1940

Chapter 78
Research on the Dynamic Strength Tester Based on Servo Control System

Xiaoguang Xu, Lijuan Yin and Feng Luo

Abstract Based on the requirement of dynamic strength in free-wheeling toys at the toy safety standard, we designed and developed the practical automatic dynamic strength tester. The tester, taking Siemens S7-224 PLC to control the motion of Minas A4 Series AC Servo Actuator, uses Pulse Train Output generator issued variable frequency high-speed pulse to control Servo Actuator, applies internal high-speed Counter to receive and count the feedback signal coding of Servo Actuator and applies Pulse Width Modulation generator to realize speed synchronization following function. Using this tester can not only be realized as the automatic test of dynamic strength in free-wheeling toys and greatly improved the precision of the test but also can enhance the level of whole system's automation.

Keywords PLC · Servo control · Dynamic strength test · High-speed pulse · Toy safety test

X. Xu (✉) · L. Yin
Shenzhen Entry-Exit Inspection and Quarantine Bureau, CIQ Building,
No.10 Airport Road, Shenzhen, Guangdong, China
e-mail: xu.xg@163.com

L. Yin
e-mail: yinlj411@163.com

F. Luo
Shenzhen University, Nanhai Ave 3688, Shenzhen, Guangdong, China
e-mail: llf@szu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_78,

78.1 Introduction

In order to ensure the health and safety of children, the children's playing toy must undergo a series of testing according to toy safety standard. Dynamic strength test is the key phase in mechanical and physical properties of free-wheeling toys. For example, the requirement of Dynamic strength test in the Nation toy safety standard (GB6675) and European Standard EN71-1 is: Toys propelled by a child or by other means and intended to bear the mass of a child, e.g.: roller-skates, inline skates, tricycles and hand carts shall conform to the requirements of the strength and stability of the toys. In strength and stability testing, we must load the toy on its sitting or standing surface with the appropriate mass at position that corresponds approximately to the normal use of the toy, accelerate smoothly, drive the toy three times at steady speed of (2 ± 0.2) m/s perpendicularly into a non-resilient step with height of (50 ± 2) mm [1].

At present, it has some questions in dynamic strength test: (1) unable to reach the accurate impact speed; (2) unable to ensure if the toys bear the outside force in the testing process; (3) difficult to fix up the toys; (4) difficult to load mass in testing process; (5) safety factor is low in case the testing failed; (6) cannot do fast testing because of the low testing efficiency.

With the development of computer-controlled technology, programable logic controller (PLC) has characteristics of modular structure, high anti-jamming I/O processing components, flexible hardware configuration, expansible and stability, which provides a stable platform in different application. It has been widely used in the field of automation control device [2]. With the application of AC Servo system, the AC Servo Actuator usually has been used in high accuracy and capability control system.

The advanced PLC control technology, the digital AC Servo system and high-performance and fine-segmentation driver control technology is used to design and develop an automation tester of dynamic strength. The tester, which has high precision, high efficiency and great stability, can meet with the related clause about strength testing requirement of Europe standard toy safety EN71-1 and national standard toy safety GB6675 and provide guarantee for the test of toy safety.

78.2 Working Process

The automation tester of toy dynamic strength is frame form configuration. The tester is composed of five parts: human–computer interaction part, control part, drive part, steel-frame part and load-mass part.

Human–Computer interaction part: Use the F920 operation panel of Japan's Mitsubishi and the operational button to set the various operational functions. Test parameters and test results can display on the screen. The operations are simple and convenient. The results are displayed timely and correctly.

Control part: Controller with high precision, respond rapidly and reliable stability characteristic is the S7-224 PLC of Siemens Co. Its primary function is to accept input signals, determine and process data in according with signals, output control signals to the driver of the tester. In this part, it is composed of two pair of time sensors and a pair of position sensors. The time sensors are used to record the time of the tested toy past by. In according to the setting distance of the sensors, we can calculate the tester's speed. The position sensors are used to check the impact position in order to send pulse signal to stop the Servo Actuator.

Drive part: Including an AC Servo Actuator, a Servo Driver and a synchronization strap wheel. The MDDDT540003 Servo Driver of Panasonic MINAS A4 Series is used to receive pulse signals. It controls the rotation direction and degree of Servo Actuator and provides all kinds of driving service in the test [5]. Output signals of Servo Actuator's rotary encoder form A and B phase orthogonal signals through Servo Driver, then the orthogonal signals become the signal source of the PLC counter by signal transform and come into being closed loop system controlled by processor of PLC. The drive part also includes synchronization strap wheel and a pair of feed screw nut.

Framed structure part: Mechanical part of the automatic tester mainly consists of the beams, inside and outside assistant clamp, non-resilient step, orbit, fixing screws and so on.

Load fixing part: In the test, it is necessary to simulate the real environment and apply the load which simulated the weight of children on the small vehicles, so this part mainly consists of the clamps of tester, load, wire rope, stationary rings, etc.

The mechanical structure of the automation tester of dynamic strength is shown in Fig. 78.1.

The automation tester of dynamic strength gets smooth speed with the Servo Actuator to ensure that the toy can reach the required speed evenly.

Framed structure of the tester is designed to ensure that the toy can drive perpendicularly to the 50 mm non-resilient step. On the upper part of the frame structure, the stationary ring connects with the load using a steel wire, so that the toy will not damage while the load falls after the toy impact. On the lower part of the frame structure there are three pairs of LED sensor on each side symmetrically. At the bottom of frame structure, four orbits are used to ensure that the toy moves perpendicular to the stage with linear motion. They are assistant clamps too. With these clamps the usual skateboards can be guided and moved. We can also use the specific clamps to make toys such as roller-skates or others to go along the orbits smoothly.

The Servo Actuator is fixed in the same plate, with the synchronization strap wheel installed on its output axis. When the output axis rotates and pulls the cord, the motor below it will drive the plate and the Servo Actuator perpendicular to the direction of the cord's movement. So the cord in the Servo Actuator's output axis rotates in spiral convolution. The purpose of this design is to make the cord's speed uniform, so that the test can be done smoothly with the uniform speed.

On the control panel, the test speed can be set and the torque curve of the gearbox axis can be drawn automatically. The status of sensor and trigger time point can be recorded.

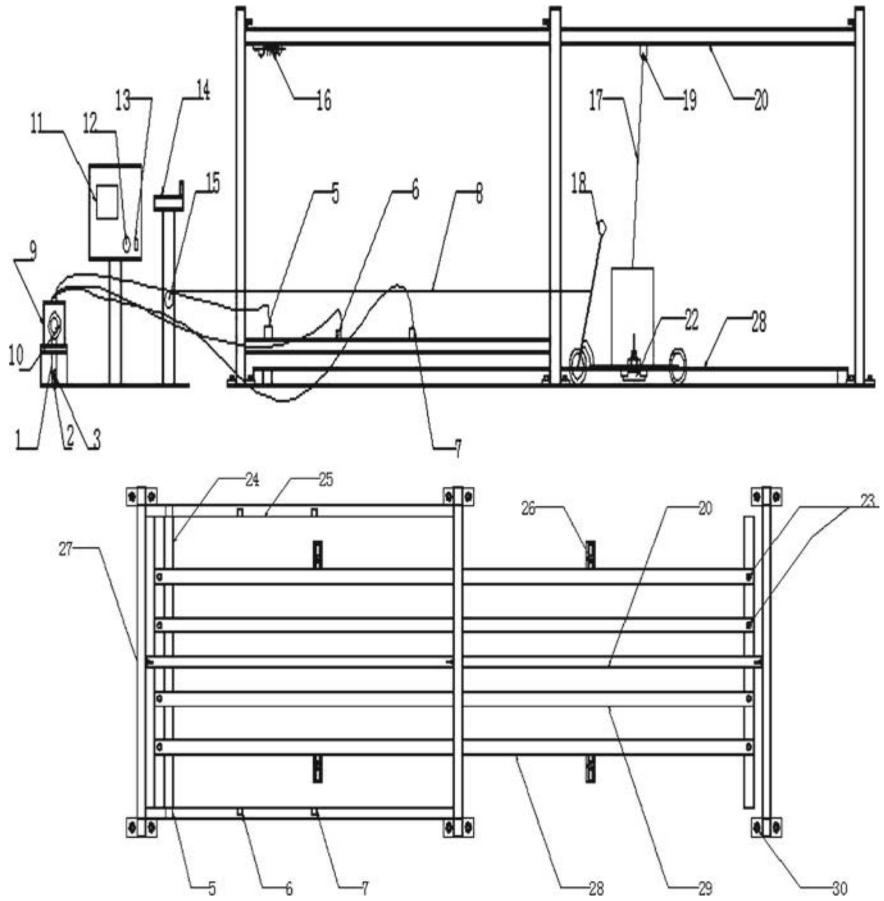


Fig. 78.1 Mechanical structure of the dynamic strength tester. *1* Synchronization strap wheel, *2* Gear wheel, *3* Screw thread, *4* Wire, *5* Position sensors, *6* Time sensors, *7* Time sensors, *8* Flexible rope, *9* Servo actuator, *10* Output axis, *11* Control panel, *12* Stop button, *13* Power switch, *14* Adjust-high wheel, *15* Chain wheel, *16* Spring, *17* Steel wire, *18* Test sample, *19* Stationary ring, *20* Orbit, *21* Load mass, *22* Clamp, *23* Fastness bolt, *24* Non-resilient step, *25* Placed sensors, *26* Foots bolt, *27* Framework, *28* Outside assistant clamp, *29* Inside assistant clamp, *30* Foot bolt

78.3 Working Principle

To meet the requirement of the precise control of the toys' speed, we choose the AC servo speed governing system as a transmission device. The rotary encoder at the end of the axis ensures the high accuracy of the AC Servo Actuator. Panasonic MINAS A4 series general-purpose AC Servo Actuator uses 17 bits encoder, the drive rotates one round when it receives 217 = 131,072 pulses, which means the pulse equivalent is 360/131,072 = 9.89 s. It is 1/655 of the pulse equivalent of step motor which the step angle is 1.8°. For example, MDMA152P1G has rated power of 200 W, and rated speed of 2000 RPM. It takes only a few milliseconds to accelerate from standstill to its rated speed 2000 RPM, and has strong overload ability [3].

78.3.1 Principle and Actualization of the Integrated Pulse Output Function of PLC

Each CPU of Siemens S7-224 PLC has two PWM/PTO generators, which are allocated to digital output ports Q0.0 and Q0.1, in order to create square wave of high-speed pulse train output and pulse width modulation.

Pulse Train Output Function. The PTO generator creates square wave of specific pulse numbers (50% duty cycle pulse), which controls cycles and pulse numbers. The equipment uses this function to control the AC Servo Actuator to realize orientation function and accelerate control function.

Pulse Width Modulation Function. The PWM generator is used to provide continuous and variable duty cycle pulses output, and provide the users the cycles of control and the pulse width. The equipment uses this function to control the AC Servo Actuator to realize speed synchronization following function.

78.3.2 High-Speed Pulse Output Control the AC Servo Actuator to Realize Speed Following and Precision Orientation Function

AC Servo Actuator usually has two control modes, which are position and speed controls. This equipment sets it to speed control mode. When the equipment is running, it receives high frequency pulses based on the high-speed pulse generator in PLC emitting. The pulse numbers determine the rotation angle of the Servo Actuator and the pulse frequency determines the rotation speed of the Servo Actuator. The rotation direction is controlled by varying the pulse output ports [6].

In Servo Actuator PLC controlling, S7-224 output high-speed pulses and direction signals to control the rotation of the Servo Actuator, and the same time receives the feedback high-speed pulse signals of the Servo Actuator and realizes closed loop control system.

S7-224 has two PTO/PWM generators to establish high-speed pulse train output or pulse width modulation waveform. A generator is assigned to digital outputs Q0.0, another generator assigned to the digital outputs Q0.1. PTO feature provides 50% duty cycle square wave output or a specified number of pulses and a specified period. PWM feature set available with variable accounting for a fixed period than the output. Each PTO/PWM generator has an 8-bit control byte, an unsigned 16-bit period value, an unsigned 16-bit pulse width value and an unsigned 32-bit pulse

Fig. 78.2 Following speed pulse output frequency–time curve

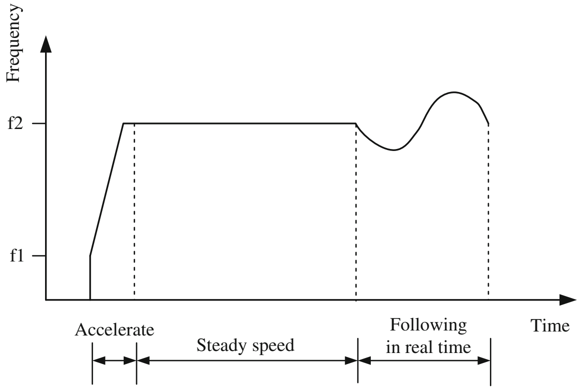
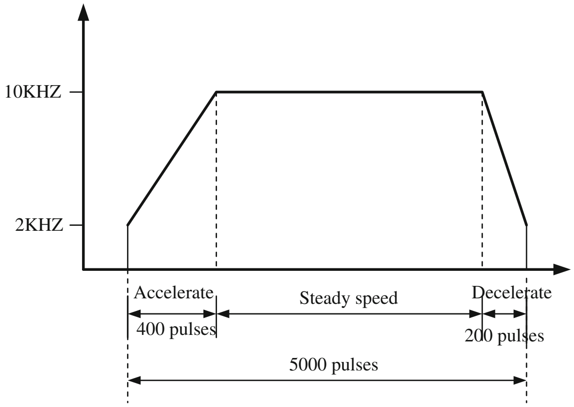


Fig. 78.3 Precision orientation pulse output frequency–time curve



value [4]. These values are all stored in special memory region specified location SM, once set these special memory bit, select the desired operation,and the implementation of pulse output commands starts operation.

The speed synchronization following function. Rationale for speed synchronization following is let the actuator accelerate to the speed currently or larger a bit, and then follow it in real time. The first and second sections adopt PTO output; the third section adopts PWM function and realizes speed following at real time, shown in Fig. 78.2. The following in real time according to the collected corresponding frequency pulse with the speed output. It receives the feedback high-speed pulse of the Servo Actuator and realizes the following in real time. In this equipment because the diameter of the output shaft is 80 mm, for ensuring the tested sample keep 2 m/s speed running, and calculate the output angle speed $\omega = V/2\pi r = 8$ rps $= 480$ rpm, so it just needs the synchronized following actuator output speed of 480 rpm.

The precision orientation function. The control of precision orientation is actually one application of the PLC pulse output function in many segments of

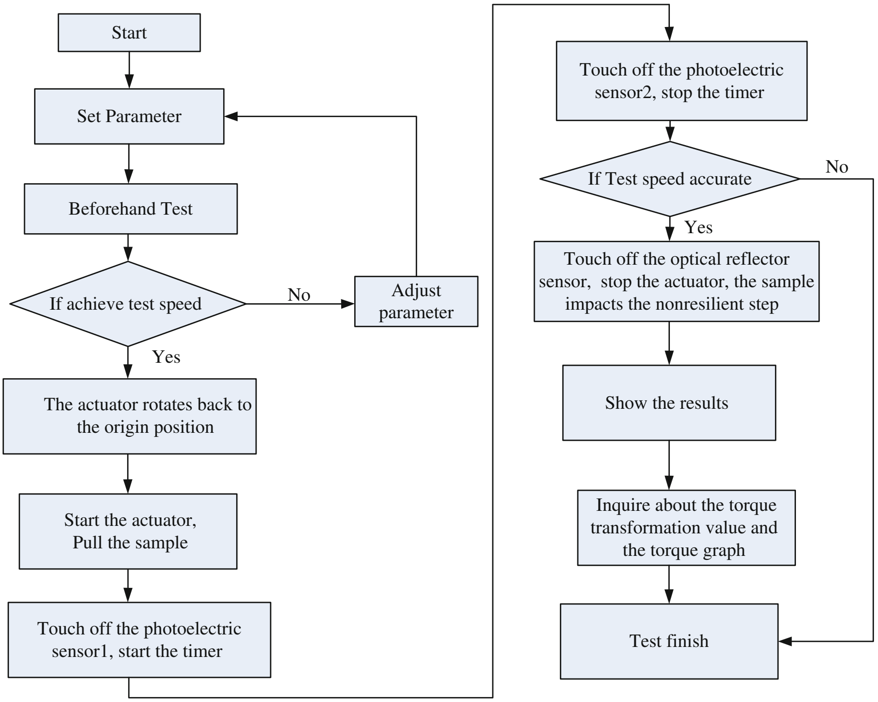

Start
Set Parameter
Beforehand Test
If achieve test speed
No
Adjust parameter
Yes
The actuator rotates back to the origin position
Start the actuator, Pull the sample
Touch off the photoelectric sensor1, start the timer
Touch off the photoelectric sensor2, stop the timer
If Test speed accurate
No
Yes
Touch off the optical reflector sensor, stop the actuator, the sample impacts the nonresilient step
Show the results
Inquire about the torque transformation value and the torque graph
Test finish

Fig. 78.4 Control program of the dynamic strength tester

PTO output [7]. Using this function, the equipment can easily control Servo Actuator realize accelerating smoothly to the steady speed, and decelerating to stop in the end. So it can orient to certain position. The main procedure realizes the frequency–time curve which is shown in Fig. 78.3. According to the test requirement, the test of dynamic strength in free-wheeling toys needs to quickly accelerate to the test speed and run in steady speed for a certain time. In the test speed do the collision test and stop the sample quickly.

78.4 Software Design

The I/O variables of the tester of dynamic strength are composed of digital input signals, digital output signals and intermediate variables. Digital input signals: running signal, start timing signal, stop timing signal, stop signal, begin position signal, stop position signal, clockwise rotate signal and counter-clockwise rotate signal; digital output signals: running control signal; intermediate variables: distance setting, speed setting, accelerate time and accelerate distance setting.

According to the received digital input signals and intermediate variables, PLC starts and controls the tester running. The control program is shown in Fig. 78.4.

First set the parameter, second take the test sample to do the first test. After that make sure whether the speed achieves the standard requirement or not, take the actuator back to the origin position and then start the actuator which will pull the sample. The sample first touches off the photoelectric sensor 1 and at the same time starts the timer, second it touches off the photoelectric sensor 2 and stops the timer, finally it touches off the optical reflector sensor and the actuator stops running, so that the sample impacts the non-resilient step. The actual impact speed shows in the screen and can inquire about the torque transformation value and the torque graph. And then the test finishes.

78.5 Conclusions

The tester of dynamic strength uses the Siemens S7-224 PLC controller and Japan's Panasonic advanced MINAS A4 series MDDDT540003 Servo Actuator driver to ensure stable operation, rapid response and high accuracy. Using this tester can not only reduce the work labor intensity but also improve the test efficiency and accuracy, and advance the level of automation in the test work as well. This tester has broad application prospect in the test of toy safety.

References

1. BS EN71-1:2006 Safety of toys-Part1: mechanical and physical properties [S]
2. Zhao Y (1998) The application of programmable logical controller (M). Chengdou University of Electronic Science and Technology Press, Beijing
3. Li Q (1994) Servo system and machine's electrical control. China Machine Press, Beijing, p 7
4. Yan Z (1985) Automatic control theory [M]. Metallurgical Industry Press, Beijing
5. Siemens Programmable Logic Controller Programmer Manual
6. Minas A4 series AC Servo Actuator driver technology information
7. Xi X, Wang W (2003) The application of PLC's integrated pulse output functions. Mech Electr Eng Tech 32(1):61–63nk

Chapter 79
Grey Sliding Mode Control for Autonomous Underwater Vehicle

Furong Liu and Dalin Zhu

Abstract The autonomous underwater vehicle has been attracting increasing interests in various areas like ocean resource exploitation. It is well known that the controller design of an underwater vehicle is complex for various reasons. The grey predicting theory is introduced into sliding mode control and a grey sliding mode controller is designed. The grey estimation can forecast and reject disturbances and parameter variations. The simulation results show that the application of grey estimation can improve the accuracy of the system and the application of grey compensation is effective to overcome uncertain disturbances. This system has a fast response and a good disturbance rejection capability.

Keywords Autonomous underwater vehicle · Grey sliding mode control · Grey estimation · Simulation

79.1 Introduction

The autonomous underwater vehicle (AUV) is a free swimming marine robot that requires little or no human intervention. In fact, AUV motion control in the unstructured underwater environment is a research hotspot. Till now, many motion control algorithms have applied in AUV motion control such as PID, Fuzzy

F. Liu (✉) · D. Zhu
College of Mechanical and Material Engineering,
China Three Gorges University, 443002, Yichang, Hubei, China
e-mail: chineselotus@163.com

D. Zhu
e-mail: dlzhu@ctgu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_79,

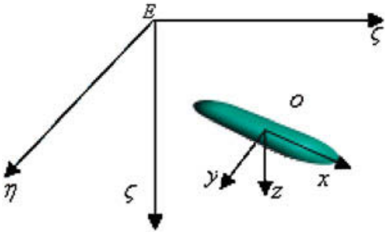


Fig. 79.1 The coordinate systems for AUV

method, neural network, self-adaptive method. However, the developing trend in control schemes is the integrated control fusing more than one control algorithm to compensate the respective restrictions of each algorithm.

One of the popular methods of robust control is the so-called sliding mode control (SMC). It has been proven as an effective and robust control technology. The sliding mode control can offer fast dynamic response, insensitivity to parameter variations, and external disturbances rejection. However, this theory still has the problem of state catching and chattering. The Grey system theory was first introduced in early 1980s by Professor Deng Ju-long [1]. The theory has since then become quite popular with its ability to deal with the systems that have partially unknown parameters. The grey prediction method only requires a few sampled data to develop the grey model and to forecast the future.

In this chapter, a grey predictor is used to forecast the values about uncertain outside disturbance parameters for use in the SMC. AUV modeling, including reference coordinates and rigid body dynamic, are discussed. SMC and grey predictor model are developed, respectively. The simulation results are presented to show the effectiveness of the proposed controller for AUV motion control with uncertain disturbances. Finally, conclusions are presented.

79.2 Modeling of AUV

The six degrees-of-freedom nonlinear equations of motion of AUV are defined with respect to two coordinate systems as shown in Fig. 79.1.

The AUV coordinates system $(o - xyz)$ has six velocity components of motion (surge, sway, heave, roll, pitch, and yaw). The velocity vector in the vehicle coordinate system is expressed as $v = [u, v, w, p, q, r]^T$. The global coordinate system $(E - \xi\eta\zeta)$ is a fixed coordinate system. Translational and rotational movements in the global reference frame are represented by $\eta = [x, y, z, \varphi, \theta, \psi]^T$ that includes earth fixed positions and Euler angles.

The equations of motion for AUV without manipulators can be written as follows [2]:

$$M(v)\dot{v} + C_D(v)v + g(\eta) + d = \tau \qquad \dot{\eta} = J(\eta)v \tag{79.1}$$

where $M(v) \in \Re^{6\times 6}$ is a 6 × 6 inertia matrix as a sum of the rigid body inertia matrix and the hydrodynamic virtual inertia (added mass); $C_D(v) \in \Re^{6\times 6}$ is a 6 × 6 Coriolis, centripetal and damping matrix; $g(\eta) \in \Re^6$ is a 6 × 1 vector containing the restoring terms formed by the AUV's buoyancy and gravitational terms; d is a 6 × 1 disturbance vector representing the environmental forces and moments(e.g. current); τ is a 6 × 1 vector including the control forces and moments; $J(\eta)$ is a 6 × 6 velocity transformation matrix that transforms velocities of the vehicle-fixed to the earth-fixed reference frame.

The expansion equations of the motion for 6-DOF of AUV based on rigid-body dynamics are written as follows [3]:

$$\begin{cases} m\cdot[(\dot u - vr + wq) - x_G\cdot(q^2 + r^2) + y_G(pq - \dot r) + z_G(pr + \dot q)] = X \\ m\cdot[(\dot v - wp + ur) - y_G\cdot(p^2 + r^2) + z_G(qr - \dot p) + x_G(pq + \dot r)] = Y \\ m\cdot[(\dot w - up + vp) - z_G\cdot(q^2 + p^2) + x_G(pr - \dot q) + y_G(qr + \dot p)] = Z \\ I_x\dot p + \left(I_z - I_y\right)qr + m\cdot[y_G(\dot w + pv - qu) + z_G(\dot v + ru - pw)] = K \\ I_y\dot q + (I_z - I_z)pr + m\cdot[z_G(\dot u + wq - vr) + x_G(\dot w + pv - uq)] = M \\ I_z\dot r + \left(I_y - I_x\right)qp + m\cdot[x_G(\dot v + ur - pw) + y_G(\dot u + qw - vr)] = N \end{cases} \tag{79.2}$$

79.3 Design of the Grey Sliding Mode Controller

79.3.1 System Description

Assume $D(x, k)$ is a disturbance additive to input $u(k)$ and the sampling time constant is T, discrete-time state-space model is presented as follows:

$$x(k+1) = Ax(k) + Bu(k) + BD(x, k) \tag{79.3}$$

where $D(x, k)$ denotes system uncertainty, including model parameter uncertainty and external disturbance. $D(x, k)$ can be a linear combination of $x(k)$:

$$D(x, k) = V_1x_1(k) + V_2x_2(k) + \cdots + V_nx_n(k) + d(k) \tag{79.4}$$

where V_i and $d(k)$ are disturbance parameters.

79.3.2 Sliding Mode Control

The design procedure of sliding mode control methodology consists of two main steps: First, a sliding surface that models the desired closed-loop performance is chosen, and then, the control law, such that the system state trajectories are forced toward the sliding surface is derived [3].

In this chapter, SMC is used to track the reference yaw angle trajectory. Hence, the switching function s is defined as

$$s = C[R(k) - x(k)] = 0 \tag{79.5}$$

where $R(k)$ is reference yaw angle trajectory. Vector C satisfies stability condition of sliding motion, $C = [C_1 C_2 \cdots C_n], C_n = 1$

According to the definition of switching function $s(k) = C(R(k) - x(k))$, the following formulation can be gained

$$\begin{aligned} s(k+1) &= C(R(k+1) - x(k+1)) \\ &= C(R(k+1) - Ax(k) - Bu(k)) \end{aligned} \tag{79.6}$$

Then the control law is presented:

$$u(k) = (CB)^{-1}(CR(k+1) - CAx(k) - s(k+1)) \tag{79.7}$$

where $s(k+1) = s(k) + (-\varepsilon T \mathrm{sgn}(s(k)) - qTs(k))$ is discrete-time exponential convergence law.

Substituting (79.6) into (79.7) gives discrete-time control law based on exponential convergence law

$$u_s(k) = (CB)^{-1}[C(R(k+1) - R(k) - C(A-I)x(k) - ds(k))] \tag{79.8}$$

where $ds(k) = -\varepsilon T \mathrm{sgn}(s(k)) - qTs(k), \varepsilon > 0, q > 0, 1 - qT > 0$

79.3.3 Grey Estimation

Grey system theory-based approaches can achieve good performance characteristics when applied to real-time systems, since grey predictors adapt their parameters to new conditions as new outputs become available. Because of this reason, grey controllers are more robust with respect to noise, lack of modeling information, and to other disturbances. Let $x^{(0)}$ be the original discrete-time data sequence

$$x^{(0)} = \left(x^{(0)}(1) x^{(0)}(2) \cdots x^{(0)}(n)\right) \tag{79.9}$$

where n is the sampling size of the recorded data.

In order to smooth the randomness, the primitive data obtained from the system is subjected to an operator, named accumulating generation operation (AGO) [4]]. When this sequence is subjected to the AGO, the following sequence $x^{(1)}(k_1)$ is obtained.

$$x^{(1)}(k_1) = \sum_{m=1}^{k} x^{(0)}(m) \tag{79.10}$$

where $i = 1, 2, \ldots, n$, $k = 1, 2, \ldots, N$, $k_1 = 1, 2, \ldots, N - 2$.

The following equation can be derived from (79.3)

$$D(x, k) = (B)^{-1}(x(k+1) - Ax(k) - Bu(k)) \tag{79.11}$$

where $u(k)$ is sliding mode control law based on exponential convergence law, i.e. $u(k) = u_s(k)$.

Using Eq. 79.11, discrete-time data sequence $D^{(0)}(k)$ corresponding with $D(x, k)$ can be achieved. Consequently, $D^{(1)}(k_1)$ is made by accumulating, i.e. $D^{(1)}(k_1) = \sum\limits_{m=1}^{k} D^{(0)}(m)$.

According to the least squares method, if BB^TBB has inverse, grey system identification results in

$$\hat{V}^T = \left(BB^TBB\right)^{-1}BB^TD^{(1)} \tag{79.12}$$

where $\hat{V} = \left(\hat{V}_1\, \hat{V}_2\, \cdots \hat{V}_n\, \hat{d}\right)$, $BB = \begin{bmatrix} x_1^{(1)}(2) & \cdots & x_n^{(1)}(2) & 1 \\ x_1^{(1)}(3) & \cdots & x_n^{(1)}(3) & 2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(1)}(N) & \cdots & x_n^{(1)}(N) & N-2 \end{bmatrix}$

79.3.4 Grey Sliding Mode Compensation

To integrate grey prediction into the proposed SMC, an effective disturbance compensation utilizing predicted disturbance parameters is given. Grey compensation controller is defined as [5]

$$u_c = -\left(\sum_{i=1}^{n} \hat{V}_i x_i + \hat{d}\right) \tag{79.13}$$

The total grey sliding mode control input u therefore can be defined as $u = u_s + u_c$.

79.4 Simulation Results

In this section, the newly proposed control scheme was numerically evaluated on a simulation example of an AUV with some parameters as follows [6]: Mass: $m = 800$ kg; Length: $L = 2.5$ m; Diameter: $D = 1.2$ m; Volume: $V = 2.5\text{m}^3$; Water density: ρwater = 1000 kg/m 3; AUV hull density: ρAUV = 320 kg/m3; Only yaw angle of AUV in the horizontal surface is given simulation results.

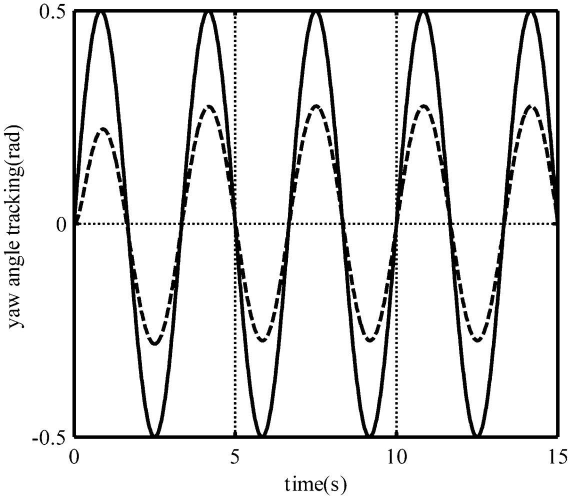


Fig. 79.2 Tracking without grey compensation

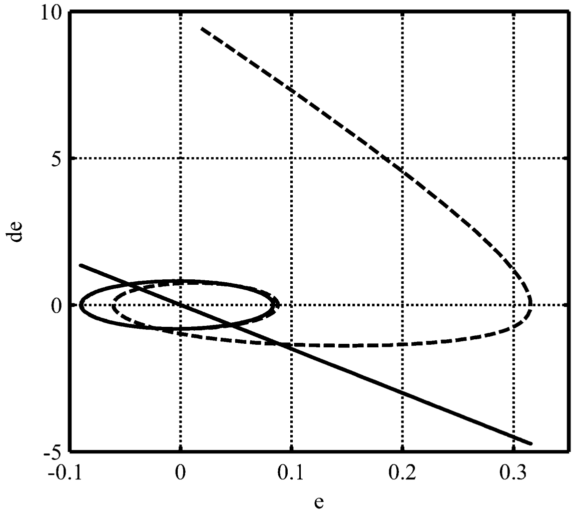


Fig. 79.3 Phase trajectory without grey compensation

The reference signal is chosen to sinusoidal signal. The simulation comprises two partitions:

79.4.1 Grey Prediction

Making use of the original data sequence at time $k = 1, 2, \ldots, N$, the accumulating data sequence at time $k_1 = 1, 2, \ldots, N - 2$ can be found; sequentially matrix *BB*

Fig. 79.4 Tracking with grey compensation

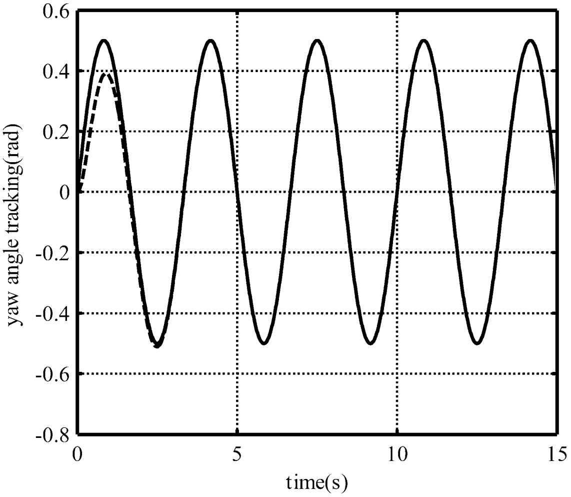
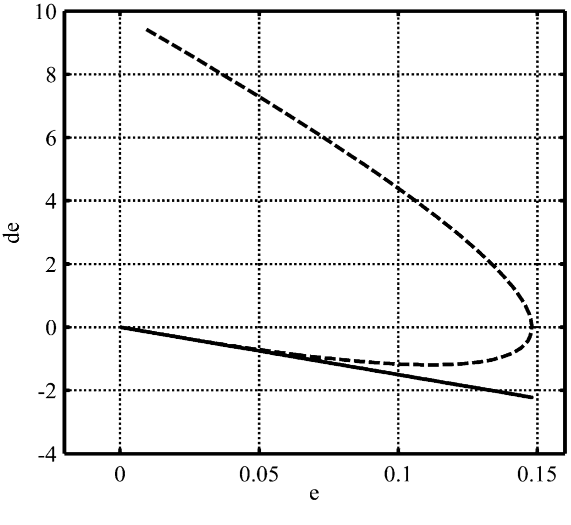


Fig. 79.5 Phase trajectory with grey compensation



also can be gained. The estimation values of disturbance parameters are $\hat{V} = [5.0000 \ -3.0000 \ 0.5000]$.

79.4.2 Adopting Grey Prediction Compensation

The simulation results are showed in Figs. 79.1 and 79.2 without grey prediction compensation. The results demonstrate that the only SMC controller cannot remove the effect of the disturbance. With grey prediction compensation, the controller can effectively overcome uncertain disturbance, as shown in Figs. 79.3, 79.4 and 79.5.

79.5 Summary

In this study, an SMC and an SMC with a grey predictor for AUV motion control have been proposed, and their performances have been compared both by simulation studies. According to various simulation results, the most attractive characteristic of the proposed controller is the robustness in the presence of the uncertainties in the system, such as noisy measurements or disturbances. The proposed grey controller has the ability to handle these difficulties.

References

1. Deng JL (1982) Control problems of grey system. Syst Control Lett 1:288–294
2. Li J, Lee P, Jun B (2004) Application of a robust adaptive controller to autonomous diving control of an AUV. In:The 30th annual conference of the IEEE industrial electronics society, Busan, pp 419–424
3. Wai R, Chang L (2006) Adaptive stabilizing and tracking control for a nonlinear inverted-pendulum system via sliding-mode technique. IEEE Trans Ind Electron 53(2):674–692
4. Deng JL (1989) Introduction to grey system theory. J Grey Syst 1:1–24
5. Lu HC (2004) Grey prediction approach for designing grey sliding mode controller. In: IEEE international conference on systems, man and cybernetics, The Netherlands. pp 403–408
6. Zhang Z (2005) Research on the method of motion control for AUV. Harbin Engineering University, Harbin

Part VIII
Green Computing

Chapter 80
Dynamics Analysis for a Generalized Approximate 3*x* + 1 Function

Shuai Liu, Weina Fu, Xiangjiu Che and Zhengxuan Wang

Abstract To study character of the generalized $3x + 1$ function is a difficult problem in fractal. At first, we put forward a generalized approximate $3x + 1$ function $D(z)$ and point out fractal character of $D(z)$ is similar to generalized $3x + 1$ function. Secondly we execute dynamics analysis of $D(z)$ and find the fixed point and periodic point. Thirdly, we point out all fixed points are at real axis of complex plane. We proved there is no other attract fixed point except zero and find the distribution of periodic point in complex plane. Then we proved that iteration of $D(z)$ is not divergence at whole number. Finally, we draw fractal figures to validate the dynamics character of $D(z)$.

Keywords Fractal · Generalized $3x + 1$ function · Periodic point · Fixed point · Complex plane

80.1 Introduction

Collatz puts forward the $3n + 1$ problem in 1950. In 60 years, many scholars studied $3n + 1$ problem [1–4] and put forward a $3n + 1$ conjecture. This conjecture is shown below.

S. Liu · X. Che · Z. Wang
College of Computer Science and Technology, Jilin University,
B234, The Computer Building, No. 2699, Qianjin Street,
Changchun 130021, China

S. Liu (✉) · W. Fu
Software College, Changchun Institute of Technology,
Room 205, The 9th Teaching Building, No. 2494,
Hongqi Street,
Changchun 130012, China
e-mail: ls_25210114@sohu.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_80,

For a natural number n, to convert with function $F(n)$.

$$F(n) = \begin{cases} \frac{n}{2} & n \equiv 0 \bmod 2 \\ \frac{3n+1}{2} & n \equiv 0 \bmod 2 \end{cases}$$

Then we trust that there exists integer $k \geq 0$ to make $F^k\ (n) = 1$ for each n. As usual, $F^k(n) = F \cdot \left(F^{k-1}(n)\right)$. Many researchers study this conjecture for long time and find many conclusions [5, 6]. But with more study, the scholars find that there is a long way to solve it.

For a long time, this conjecture could hardly find a way to solve until Mandelbrot created Mandelbrot Set by computer [7]. Soon Mandelbrot Set became a symbol of chaos dynamic which is called fractal. Because fractal figures are created by iteration of initial function, it is possible to study $3n + 1$ conjecture by fractal technology. Firstly, Pe and Dumont create $3x + 1$ piecewise function and give dynamic analysis of this function by fractal. The created fractal figures are good visual effect [8]. It is the first study of $3n + 1$ conjecture by fractal. In fact, the piecewise function is hard to solve with the existing formula . So the generalized $3x + 1$ function is created. Of course, it is a continuous function. Many scholars do dynamic analysis of it [9–12].

Nowadays, the study of generalized $3x + 1$ functions are more in $T(x) = \frac{1}{2}\left[x3^{\sin^2\left(\frac{\pi x}{2}\right)} + \sin^2\left(\frac{\pi x}{2}\right)\right]$ and fewer in $C(x) = x - \frac{x}{2}\cos \pi x + \frac{1-\cos \pi x}{4}$. So we create a approximate generalized $3x + 1$ function $D(x) = x - \frac{x \cos \pi x}{2}$. In fact, $D(x) = \frac{3x}{2}$ when x is odd, and $D(x) = x/2$ when x is even. So it is similar to generalized $3x + 1$ function. When we study more about $D(x)$, we will find that $D(x) \leq C(x) \leq D(x) + 0.5$. Thus, characters of $D(x)$ is similar to $C(x)$ too.

So we do dynamic analysis of $D(x)$ and find its fixed points and periodic points in complex plane. Then we study the convergence and divergence of whole number of $D(x)$ and prove that iteration with whole number is not divergence. It extended the $3n + 1$ conjecture.

80.2 Fixed Point and Periodic Point of D(z)

80.2.1 Fixed Point of D(z)

When extended $D(x)$ to complex plane, we call it $D(z)$ instead. We find distribution and attraction of $D(z)$'s fixed point by Theorem 1 and 2.

Theorem 1 *There exists fixed points of $D(z)$. All fixed points are at real axis. All fixed points are $n + 0.5$ $(n \in Z)$ except* 0.

Proof Set $D(z) - z = -(z\cos\pi z)/2 = 0$, we can easily find that $z = 0$ is a root of this equation. When $z \neq 0$, the equation changed to $\cos\pi z = 0$. To solve it, we get the roots are $z = n + 0.5(n \in Z)$. Theorem is proved.

Theorem 2 *The only attractive fixed points of D(z) is zero. All other fixed points are repulsive.*

Proof We know that $D'(z) = 1 - \frac{\cos \pi z}{2} + \frac{\pi z \sin \pi z}{2}$. To set $z = 0$ into it, we find $D'(x) = 0.5 < 1$. So zero is attractive. When $z = n + 0.5$, we know that $\cos\pi z = 0, \sin\pi z = \pm 1$ and $D'(x) = 1 \pm \pi z/2$. To solve inequality, we find $-4\pi/ < \pm z < 0$. We know there are fixed points -0.5 and 0.5 at this bound by Theorem 1. To set them in $D'(z)$, we get $D'(0.5) = \mathrm{D}'(-0.5) = 1 + \pi/4 > 1$. So we know that they are all repulsive fixed points.

To consider all above, Theorem 2 is proved.

So we know that $D(z)$'s fixed points are all at real axis and only zero is attractive. To extend conclusion from fixed point to periodic point, we gain the conclusion of periodic points.

80.2.2 Periodic Point of D(z)

We use Theorem 3 to explain the distribution of periodic points at real axis of $D(z)$. To prove Theorem 3, we have to prove Lemma 1 and 2 to gain the relation of i-periodic point and $i + 1$ periodic points of two next fixed points. We define $D^i(z) = D(D^{i-1}(z))$ and $D^1(z) = D(z)$ in calculation.

Lemma 1 *There exists $i + 1$ periodic points in $(n - 0.5, n + 0.5)$ when there exists z_0 to make $D^i(z_0) = n + 1$ in $(n-0.5, n + 0.5)$ $(n > 1, n \in Z)$.*

Proof We know $n - 0.5$ and $n + 0.5$ are all fixed points of $D(z)$, so there exists x_1 make $D(x_1) = n$. When $n > 1$ and $2|n$, we find that $D^{i+1}(x_1) = n/2 < n-0.5 < x_1$ and $D^{i+1}(z_0) = 3(n + 1)/2 > n + 0.5 > z_0$. So there exists k make $D^{i+1}(k) = k$ between z_0 and x_1. In opposite, when n is odd, we know that $D^{i+1}(x_1) > x_1$ and $D^{i+1}(z_0) < z_0$. It is to say that $i + 1$ periodic point is exist between x_1 and z_0. So Lemma 1 is proved.

Lemma 2 *There exists $i + 1$ periodic points in $(n - 0.5, n + 0.5)$ when there exists z_0 to make $D^i(z_0) = n - 1$ in $(n - 0.5, n + 0.5)$ $(n > 3, n \in Z)$.*

Proof The proof is similar to Lemma 1, ellipsis.

We can find the distribution of $D(z)$'s periodic points. Then we gain Theorem 3 to explain it.

Theorem 3 *There exists periodic points of every period in $(n - 0.5, n + 0.5)$ when $n > 0$, $n \in Z$ and $n \neq \pm 2$. There is not any other periodic point except zero in $(-0.5, 0.5)$.*

Proof We know $x\sin\pi x \geq 0$ in $[-0.5, 0.5]$. So $D'(z) = 1 - \frac{\cos \pi z}{2} + \frac{\pi z \sin \pi z}{2} > 0$. It is to say that $D(x)$ is monotone increasing. It is easy to find that $D(z) - z = -\frac{z \cos \pi z}{2}$ is not negative in $(-0.5, 0]$ and not positive in $[0, 0.5)$. So we find that zero is the only fixed point of $D(z)$ in this bound. Thus, we gain that $D(z) > z$ when z in $(-0.5, 0)$ and

$D(z) < z$ when z in (0, 0.5). So we know that $|Dk(z)| < |Dk - 1(z)| < \cdots < |z|$. It is to say that all points in this bound is attracted by zero. Then, it is not any other periodic point.

Proof finished.

Then we start to prove the conclusion in (n−0.5, n + 0.5).

Because $D(z)$ is an odd function, we assume that $n \geq 1$.

Of course, n is in (n − 0.5, n + 0.5). So $|D(n) - n|$ is $n/2$. (whether $D(n) = n/2$ or $3n/2$). When $n > 1$ and n is odd, there exists z_0 to make $D(z_0) = n + 1$. So we know that there exist 2-periodic point from Lemma 1 and so on. Then $D(z)$ has periodic points with every period in this bound. When n is even, there exists z_0 to make $D(z_0) = n - 1$. The theorem is proved when $n > 3$ by Lemma 2.

To consider all above, Theorem 3 is proved.

Then we can extend the conclusion to complex plane to gain the similar conclusion of $D(z)$.

80.2.3 Periodic Point of **D(z)** *in Complex Plane*

We use Theorem 4 to gain the distribution of $D(z)$'s periodic point in complex plane.

To set $z = a + bi$, then we gain formulas below.

$$D(z) = z - \frac{z\cos \pi z}{2} = a + bi - \frac{1}{2}(a + bi)\cos(\pi a + \pi bi) \tag{80.1}$$

To use definition of complex function, we gain formula (80.2).

$$\cos(\pi a + \pi bi) = \frac{e^{\pi b} + e^{-\pi b}}{2}\cos \pi a - i\frac{e^{\pi b} - e^{-\pi b}}{2}\sin \pi a \tag{80.2}$$

To set $D(z) = A + Bi$, $u = \frac{e^{\pi b}+e^{-\pi b}}{2}\cos \pi a$ and $v = \frac{e^{\pi b}-e^{-\pi b}}{2}\sin \pi a$, we put formula (80.2) into (80.1).

$$\begin{cases} A = a - \frac{ua}{2} - \frac{vb}{2} \\ B = b - \frac{ub}{2} + \frac{va}{2} \end{cases}$$

From the definition of u and v we know that they increase exponentially when $|b|$ increase and they increases periodically with period 2 increase. A and B have similar characters because A and B are linear calculations by u and v. So we know that when $|b|$ is large enough, $D(z)$ has no periodic point. In fact, $|D(z)| = \sqrt{A^2 + B^2} = |z|\sqrt{1 + \frac{u^2+v^2}{4} - u}$, so when we solve $H = 1 + \frac{u^2+v^2}{4} - u$, we can know the convergence and divergence of $D(z)$. Otherwise, we know that $H = 1 + \frac{u^2+v^2-4u}{4} > (\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos \pi a)^2$, so $|D(z)|/|\mathrm{z}| = \sqrt{1 + \frac{u^2+v^2}{4} - u} \geq |\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos \pi a|$.

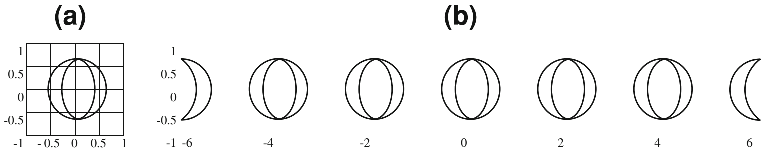


Fig. 80.1 **a** Is root region of (−1, 1) × (−1, 1) and **b** is root region in complex plane

It is to say that $|D(z)| > |z|$ when $\left|\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos\pi a\right| > 1$. As we know, $D(z)$ has no fixed point in this condition. To extend it, we gain Theorem 4.

Theorem 4 *To set $z = a + bi$, z is not a $n + 1$ periodic point when* $\left|\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos\pi a\right| > 2^n$.

Proof To consider with $|Dn + 1(z)| = |z| \cdot \prod_{i=0}^{n}\left|1 - \frac{\cos(\pi D^i(z))}{2}\right| \geq (1/2)n \cdot |z|$ $\sqrt{1 + \frac{u^2+v^2}{4} - u}$, we can know that Theorem 4 is proved.

We know that $D(z)$ has no $n + 1$ periodic point when $b > \ln\,(2^{n\,+\,2} + 4)^{1/\pi}$ by Theorem 4. To think from the opposite, we gain Inference 1.

Inference 1 z is not a $n + 1$ periodic point of $D(z)$ when $\left|\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos\pi a\right| + |\sin\pi a| < (2/3)^n$.

Especially, we gain region with center $a = k$ (k is even) by solving $\left|\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos\pi a\right| + |\sin\pi a| < 1$. It is similar to the region with center $a = 0$. The fixed point is outside the region. Then we gain another region by solving $\left|\frac{e^{\pi b}+e^{-\pi b}}{4} - \cos\pi a\right| > 1$ and the fixed point is inside it. So the fixed point is in the two root regions. Just like Fig. 80.1 shown. It is validate Theorem 1 and 4.

80.3 Dynamic Character and Fractal Figure of D(z)

We gain Theorem 5 to explain the symmetry of $D(z)$.

Theorem 5 *$D(z)$'s fractal figure is symmetry with both real axis and image axis.*

Proof Because $D(\bar{z}) = \bar{z} - \frac{\bar{z}\cos\pi\bar{z}}{2} = \bar{z} - \frac{\bar{z}\overline{\cos\pi z}}{2} = z - \frac{\overline{z\cos\pi z}}{2} = \overline{D(z)}$, $D(z)$ is symmetry with image axis. Otherwise, we know that $D(-\mathrm{z}) = -z\left(1 - \frac{\cos\pi(-z)}{2}\right) = -D(z)$, so $D(z)$ is symmetry with real axis. Theorem 5 is proved.

When we try to use generalized $3x + 1$ function to solve $3n + 1$ conjecture, we must consider about the whole number. It is hard to prove. But we can easily gain Theorem 6 to prove that the iteration of $D(z)$ is not divergence.

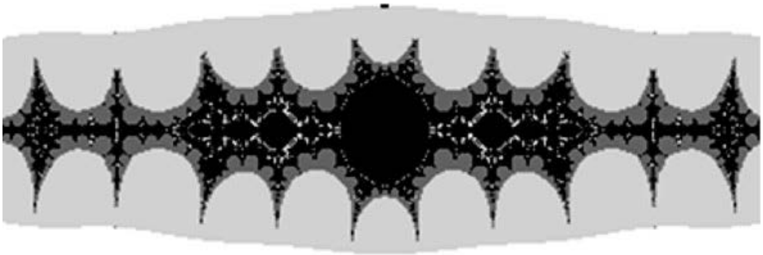


Fig. 80.2 $D(z)$'s fractal figure at (0, 0), $(-8, 8) \times (-1.5, 1.5)$

Theorem 6 *Set all integer* $n = 2^i n_1$ (n_1 *is odd*) *except* 0, $D^k(n) = \frac{3}{2}n_1$ *when* $k > i$.

Proof We know that $D^i(n) = n_1$ and $D^{k-i}(n_1) = D(n_1) = \frac{3}{2}n_1$ is a fixed point. So Theorem 6 is proved.

Because $D(z)$ is not divergent at integer and $n + 0.5$, so $D(z)$ can divide to connectivity regions with single point connect by the points which are iterated to fixed point. Then we can find the similarity by regions which converge to the same region.

As we know, $D(z)$ has no $n + 1$ periodic point when $b > \ln\,(2^{n+2} + 4)^{1/\pi}$. So the main fractal character of $D(z)$ is near real axis (For example: there is no 2 periodic point when $b > 0.7910$, and no 3 periodic point when $b > 0.9536\ldots$). We know that $D(z)$ has periodic point with every period between next foxed point bound in real axis except $(-2.5, -1.5)$ and $(1.5, 2.5)$ from Theorem 3. So we solve equation $\prod_{i=0}^{1}\left(1 - \frac{\cos(\pi D^i(z))}{2}\right) = 1$ to gain attractive 2 periodic point (1.2161, 1.6893) in (1.5, 2.5). So we guess that there is no attractive 3 periodic point of $D(z)$ in real axis by chaos formula.

By concluding all the discussion of $D(z)$, we can find the character of $D(z)$'s fractal figures.

(i) There are endless points to make $D^n(z) \to \infty$ at real axis.
(ii) $D^n(z)$'s attractive regions are attract by zero or 2-periodic orbits (1.2161, 1.6893) and $(-1.2161, -1.6893)$.
(iii) All integer at real axis is in Julia set of $D(z)$.

Then we use escape time algorithm to draw $D(z)$'s fractal to validate the above conclusions.

The max iteration number is 50 in all figures. The threshold is 1,000. The displayed area is in each figure.

We can easily validate Theorem 5 from Fig. 80.2. Then we can simply to find that all fractal figures is between $y = i$ and $y = -i$. To study with figures at real axis we find the discontinuous. It is to say that there are endless z to make $|D(\mathrm{z})| \to \infty$ from Fig. 80.3. We can find the self-similar by Fig. 80.4. Figures' displayed region in Fig. 80.4 are shown.

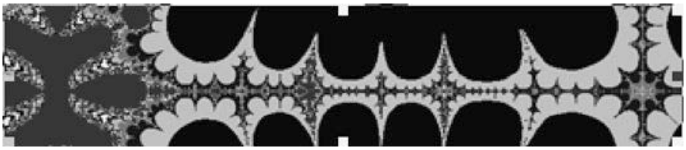


Fig. 80.3 (2.9999, 3.0001) × (−0.000075, 0.000075)

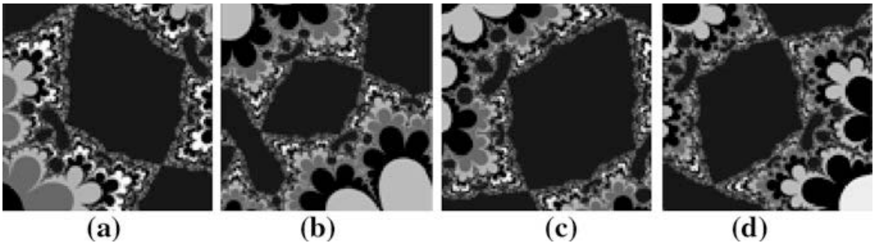


Fig. 80.4 The self-similar of $D(z)$'s fractal. **a** (9.9950758, 9.9950766) × (0.4687087, 0.4687095), **b** (−1.464046945, −1.464006945) × (−0.22868174, −0.22864174), **c** (−1.657693754, −1.657643754) × (0.32355618, 0.32360618), **d** (−1.213206622, −1.213006622) × (−0.672442411, −0.672242411)

The center of each figures satisfy $D3(a) = D2(b) = D(c) = d$ and finally reach to orbit (1.2161, 1.6893).

80.4 Conclusion

We created and analyzed dynamic an approximate generalized $3x + 1$ function D(z). Then we created fractal figures of $D(z)$ to validate the conclusions. Otherwise, we found the convergence is well from Fig. 80.2. It is to say that $D(z)$ can be used as a basic function of fractal algorithm.

As we know, $D(z) = z - \frac{z \cos \pi z}{2}$. Then we found that it is same to $3n + 1$ conjecture when z is even and difference is 0.5 when z is odd. We found that the iteration of $D(z)$ is not divergence at integer. It is a new way to solve $3n + 1$ conjecture.

References

1. Alves JF, Graca MM, Dias MES et al (2005) A linear algebra approach to the conjecture of Collatz. Lin Alg Appl 394(1):277–289
2. Lagarias JC (1985) The $3x + 1$ problem and its generalizations. Am Math Mon 92(1):3–23

3. Wirsehing GJ (1998) The dynamical system generated by the $3n$ +1 function. Lect Notes Math 1681:153–159
4. Wu J, Hao S (2003) On equality of the adequate stopping time and the coefficient stopping time of n in the $3\ N + 1$ conjecture. J Huazhong Univ Sci Tech (Nature Science Edition) 31(5):114–116
5. Belaga E, Mignotte M (1998) Embedding the $3x + 1$ Conjecture in a $3x + d$ context. Exp Math 7(2):145–151
6. Simons J, de Weger B (2005) Theoretical and computational bounds for m-cycles of the $3n + 1$ problem. Acta Arith 117(1):51–70
7. Mandelbrot BB (1982) The fractal geometry of nature. Freeman W H, San Francisco, pp 1–122
8. Pe JL (2004) The $3x + 1$ fractal. Comput Graph 25(3):431–435
9. Dumont JP, Reiter CA (2001) Visualizing generalized $3x + 1$ function dynamics. Comput Graph 25(5):553–595
10. Liu S, Wang Z (2009) Fixed point and fractal images for a generalized approximate $3x + 1$ function. J Comput Aided Des Comput Graph 21(12):1740–1744
11. Liu S et al (2011) The existence of fixed point for a generalized 3x + 1 function. Appl Mech Mater 55–57:1341–1345
12. Liu S et al (2011) periodic point at real axis for a generalized $3x + 1$ function. Appl Mech Mater 55–57:1670–1674

Chapter 81
Attraction in Positive Direction of Real Axis of Two Generalized 3*x* + 1 Functions

Liu Shuai, Fu Weina, Ke Hongchang and Wang Xin

Abstract To study attractive character of two generalized $3x + 1$ functions $T(z)$ and $C(z)$ at positive direction of real axis, at first, we define attractive domain and series. Secondly, we analyze attractive domain and series of $T(z)$ and $C(z)$ and point out the structure of attractive domains surrounding the same attractive series that are similar to each other. Finally, we draw the fractal figures of the two functions by escape time algorithm and validate the structure rule of the attractive domain and series of $T(z)$ and $C(z)$ from the figures.

Keywords Fractal · Generalized $3x + 1$ function · Dynamic system · Attractive series · Escape time

81.1 Introduction

The $3n + 1$ conjecture was born in the 1950s. For 60 years, many scholars studied the $3n + 1$ problem, but none could solve it. This conjecture is shown as below.

For a natural number n, to convert with function $F(n)$.

$$F(n) = \begin{cases} \dfrac{n}{2} & n \equiv 0 \bmod 2 \\ \dfrac{3n+1}{2} & n \equiv 1 \bmod 2 \end{cases}$$

L. Shuai (✉) · F. Weina
Software College, Changchun Institute of Technology, Room 205,
The 9th Teaching Building, No. 2494, Hongqi Street, Changchun 130012, China
e-mail: ls_25210114@sohu.com

L. Shuai · K. Hongchang · W. Xin
College of Computer Science and Technology, Jilin University, B234,
The Computer Building, No. 2699, Qianjin Street, Changchun 130021, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_81,

Then there exists integer $k \geq 0$ to make $F^k(n) = 1$ for each n. As usual, $F^k(n) = F\text{o}(F^{k-1}(n))$ and $F^1(n) = F(n)$. Many researchers have studied this conjecture for a long time [1, 2]. The experiments are from 1 to 2^{40} and every positive integer fits this conjecture. But the proof cannot be shown right now.

Many scholars trust that it cannot be proved in the current mathematic domain. So when Mandelbrot [3] created Mandelbrot Set on the computer, scholars tried to study the $3n + 1$ conjecture by fractal technology. In fact, fractal figures of functions are similar to its dynamic. Convergence of fractal is its stable region. First, Pe and Dumont created $3x + 1$ function and drew its dynamic system by fractal. The created fractal figures were good visual effects [4]. It was the first study of $3n + 1$ conjecture by fractal. Soon they created another function called generalized $3x + 1$ function [5]. Of course, this is a continuous function. Many scholars do dynamic analysis of it [6–9].

Nowadays, the study of generalized $3x + 1$ functions are mostly for $T(x) = \frac{1}{2}\left[x3^{\sin^2\left(\frac{\pi x}{2}\right)} + \sin^2\left(\frac{\pi x}{2}\right)\right]$ and $C(x) = x - \frac{x}{2}\cos \pi x + \frac{1 - \cos \pi x}{4}$. However, dynamic characters are hard to find because of their complexity. The study is an all near real axis [9] or for approximate function [7]. So after summarize studies heretofore, we compare $T(x)$ and $C(x)$ near real axis. We define attractive domain and series and find these characters of $T(x)$ and $C(x)$. In fact attractive domain and series of $T(x)$ and $C(x)$ are similar to each other. Then we draw their fractal figures to validate the similarity.

Our contribution is to show the similarity and self-similarity of $C(x)$ and $T(x)$'s attractive domain and series. So we put forward a conjecture to guess that all series are attracted by zero and orbit (1, 2).

81.2 Attractive Domain and Series

81.2.1 Divergence Domain, Attractive Domain and Attractive Series

To iterate $C(x)$ and $T(x)$ as basic function and set $C^n(z) = C(C^{n-1}(z))$, $T^n(z) = T(T^{n-1}(z))$, We conclude the iteration result of random point z in complex plane to three conditions.

A. When result $= \infty$, we call z divergence points and call set assemble by all these points divergence domain.

B. When there exists positive integer k make $F^k(z) = z_1$ and z_1 is a fixed point or z_1 is a periodic point of m-periodic orbit, the series assemble by $\{z,\ C(z),\ldots,\ F^{k-1}(z)\}$ is called z_1's series. In this case, we call the series attractive series.

C. When points neither fit case A nor B, we call these points wandering-domain. It is obvious that $\lim\limits_{k\to\infty} F^k(z) \neq \infty$ for any point z in wandering-domain and $F^m(z) \neq F^n(z)$ for any positive integer m and n when $m \neq n$.

In this chapter, we study attractive point and series mainly. It is easy to divide them into two kinds. One is to attract fixed points, the other is to attract periodic orbits. To set z_1 as a point attract to attractive point and series, in other words, $F^k(z_1) = z$ when z is an attractive n-periodic point, we can see that $|(F^n(z))'| < 1$. Thus, $|(F^{n+k}(z_1))'| = |(F^n(F^k(z_1)))'|\cdot|(F^k(z_1))'| < 1$. So we can say that the domain near z_1 is attractive domain.

81.2.2 Attractive Domain and Expansion Coefficient at Positive Integer

When we study characters as positive integers, we know that $\lim\limits_{k\to\infty} C^k(n) = \lim\limits_{k\to\infty} T^k(n)$ because the results are the same as the $3n+1$ conjecture. So we set expansion coefficient as k times derivative of function F by period k. Then we get Theorem 1.

Theorem 1 *When n is integer, $C^k(n) = T^k(n), C^{k'}(n) = T^{k`}(n)$ $(k = 1\ldots\infty)$.*

Then we solve equation $C(x) = T(x)$ in real domain and find theorem 2.

Theorem 2 *When $x \in R$, the solution of $C(x) = T(x)$ is $x \in Z$.*

Proof The equation $C(x) = T(x)$ is $\frac{1}{2}\left[x3^{\sin^2(\frac{\pi x}{2})} + \sin^2\left(\frac{\pi x}{2}\right)\right] = x - \frac{x}{2}\cos\pi x + \frac{1-\cos\pi x}{4}$. To simplify it we know that

$$x(3^{\sin^2(\frac{\pi x}{2})} - 2\sin^2(\frac{\pi x}{2}) - 1) = 0. \tag{81.1}$$

Obviously $x = 0$ is a root of Eq. (81.1). When $x \neq 0$, to set $\sin^2(\frac{\pi x}{2}) = X$, we can simplify Eq. (81.1) to $3^X - 2X - 1 = 0$. To solve it, we know it has two roots $X_1 = 0$ and $X_2 = 1$. Then we solve $\sin^2\left(\frac{\pi x}{2}\right) = 0, 1$ and find that the solutions are all integers.

To conclude the above, the theorem is proved.

Because both $C(x)$ and $T(x)$ are continuously derivable, we know that $C(x)$ and $T(x)$ have similar attractive domains near positive integers. It is known as theorems 3 and 4.

Theorem 3 *When $n \in Z$, $n_1 = C(n)$, Q is a continuous closed domain that contains n, $C^n(z) \neq \infty$ for all points z $(n \to \infty)$, Closed curve R is Q's boundary, R_1 is*

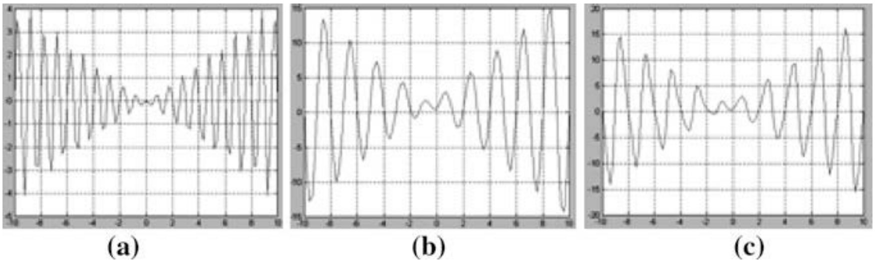


Fig. 81.1 **a** $K(x)$ **b** $C'(x)$ **c** $T'(x)$

Closed curve C(R), Q$_{1}$ *is domain enclosed by R*$_{1}$*, domain (n*$_{1}$*,ε) is similar to (n, ε/2) of C(x)'s dynamic system. T(x) has the same conclusion.*

Proof $C^{'}(n) = 1 - \frac{\cos(\pi n)}{2} + \frac{\pi n}{2}\sin(\pi n) + \frac{\pi \sin(\pi n)}{4}.C^{'}(n) = 0.5$ when n is even.

To use Taylor formula, we know $C(n + \varepsilon) = C(n) + \varepsilon \cdot C'(n) + O(\varepsilon^2)$. When ε is small enough, we know $C(n + \varepsilon) = n_1 + \varepsilon/2 + O(\varepsilon^2)$. It is to say that n with neighborhood ε is similar to n_1 with neighborhood ε/2.

As the same proof, we know that $T(x)$ has similar characters.

The theorem is proved.

Theorem 4 *When n is odd, n*$_{1}$ *= C(n), domain (n*$_{1}$*,ε) is similar to (n, 3ε/2) of C(x)'s dynamic system. T(x) has the same conclusion.*

Proof The proof is similar to theorem 3, ellipsis.

So we get inference 1 from theorems 3 and 4.

Inference 1 When $x \in R$, $x_1 = C(x)$, domain $(x_1{'}\varepsilon)$ is similar to $(x, \varepsilon C'(x))$ of $C(x)$'s dynamic system. $T(x)$ has the same conclusion.

So we know that characters of dynamic system near real axis of $T(x)$ and $C(x)$ is depend on derivative of them. As we known, $C^{'}(x) = 1 - \dfrac{\cos(\pi x)}{2} + \dfrac{\pi x}{2}\sin(\pi x) + \dfrac{\pi \sin(\pi x)}{4}$, $T^{'}(x) = \dfrac{1}{2} \cdot 3^{\sin^2(\frac{\pi x}{2})}\left(1 + \dfrac{\pi \ln 3}{2} x \cdot \sin(\pi x)\right) + \dfrac{\pi}{4}\sin(\pi x)$. Then, we define $K(x) = T^{'}(x) - C^{'}(x) = \frac{1}{2} \cdot \left(3^{\sin^2(\frac{\pi x}{2})} - 2\sin^2\left(\frac{\pi x}{2}\right) - 1\right) + \dfrac{\pi x}{2}\sin(\pi x)\left(\dfrac{\ln 3}{2} \cdot 3^{\sin^2(\frac{\pi x}{2})} - 1\right)$. To analyze $K(x)$ we find that $K(x)$ cross real axis by turns. To compare with $K(x)$, $C'(x)$ and $T'(x)$, we gain Fig. 81.1. We can see $C'(x)$ and $T'(x)$ are all zero near integer and $K'(x)$ is two multiple frequency than $C'(x)$ and $T'(x)$. In other words, it is to say that $C(x)$ and $T(x)$ are all attract at integer point. But $K(x)$ is attract at all points like n and $n + 0.5$. So we know that dynamic system of $T(x)$ and $C(x)$ are similar to each other at integer point from Theorems 1–4, inference 1 and Fig. 81.1. Now we extend the conclusion to real axis.

81.2.3 Compare Near Real Axis

We know that $T(x) \in$ R by $x \in$ R. Then if we set $x_1 = T(x)$, we confirm that dynamic systems near x_1 and x of $T(x)$ are similar to each other. As the same prove, we can use it to $C(x)$. It is that if we set $x_2 = C(x)$, dynamic systems near x_2 and x of $C(x)$ are similar to each other. Because $T(x)$ and $C(x)$ are similar, dynamic system near x_1 of $T(x)$ and x_2 of $C(x)$ are similar.

When we set complex number $z = a + bi$, because increment is too fast of $|T(z)|$ and $|C(z)|$ since $b > 1$, we define that $b < 1$ in this chapter. In fact, when $b > 1$, we can not use escape time algorithm to draw their fractal because both threshold and computation are too large.

So we analyze them with $b < 1$ in complex plane, to combine the past conclusions in this chapter, we gain conjecture 1 and 2.

Conjecture 1 *When $b < 1$ and $z_1 = T(z^*)$, the stable-domain created by z_1 is similar to the stable-domain created by z^*. So as $C(z)$.*

Conjecture 2 *When $b < 1$ and $z_1 = T(z^*),\ldots, z_n = T^n(z^*)$, the stable-domain created by $z_1, z_2,\ldots,z_n$ and z^* are similar to each other. So as $C(z)$.*

Then when we consider that $T(z)$ and $C(z)$ are all transcendental holomorphic functions, we confirm that there exist endless roots of equation $T(z) = z_0$ and $C(z) = z_0$ by argument z for every complex number z_0. To define roots as $\{T^{-1}(z)\}$ or $\{C^{-1}(z)\}$, we trust that there exist similar fractals to z_0.

81.3 Fractals of *C*(*z*) and *T*(*z*)

81.3.1 Fractals Near Real Axis

We use escape time algorithm to draw fractals of $T(z)$ and $C(z)$ in Fig. 81.2. We set the max iteration time as 100 and the threshold as 1,000. We iterate all points in display region since points iteration values are larger than threshold or iteration times reaching max times. We color points with different iteration times by different colors. We use black color to color the convergence region.

We draw some square borders in Fig. 81.2. From fractals of the two dynamic systems we find that the basic characters are similar to each other, though $T(z)$'s fractal is more complex than $C(z)$.

We get some random points z^* and $z_1 = T(z^*)$, $z_2 = C(z^*)$ to observe in complex plane. The created figures is Fig. 81.3. Multiple in Fig. 81.3 means the display region is the neighbor-domain $(-2/n,2/n) \times (-1.5/n,1.5/n)$ of z^*.

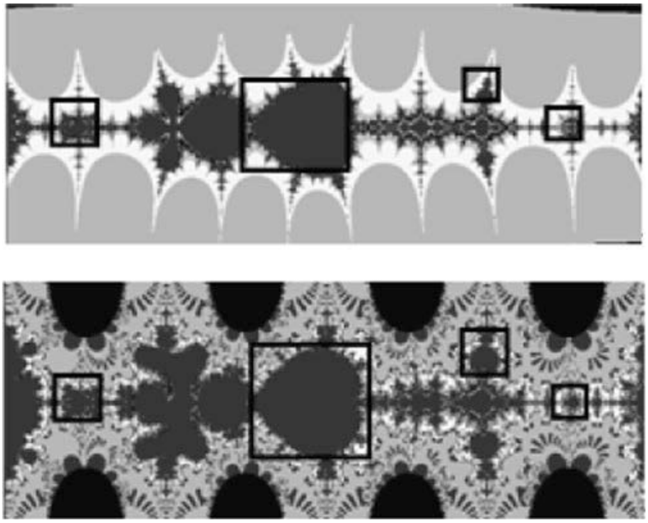


Fig. 81.2 Fractals at (0, 0) of $C(z)$ and $T(z)$. Display region is $(-4,4) \times (-1.5,1.5)$

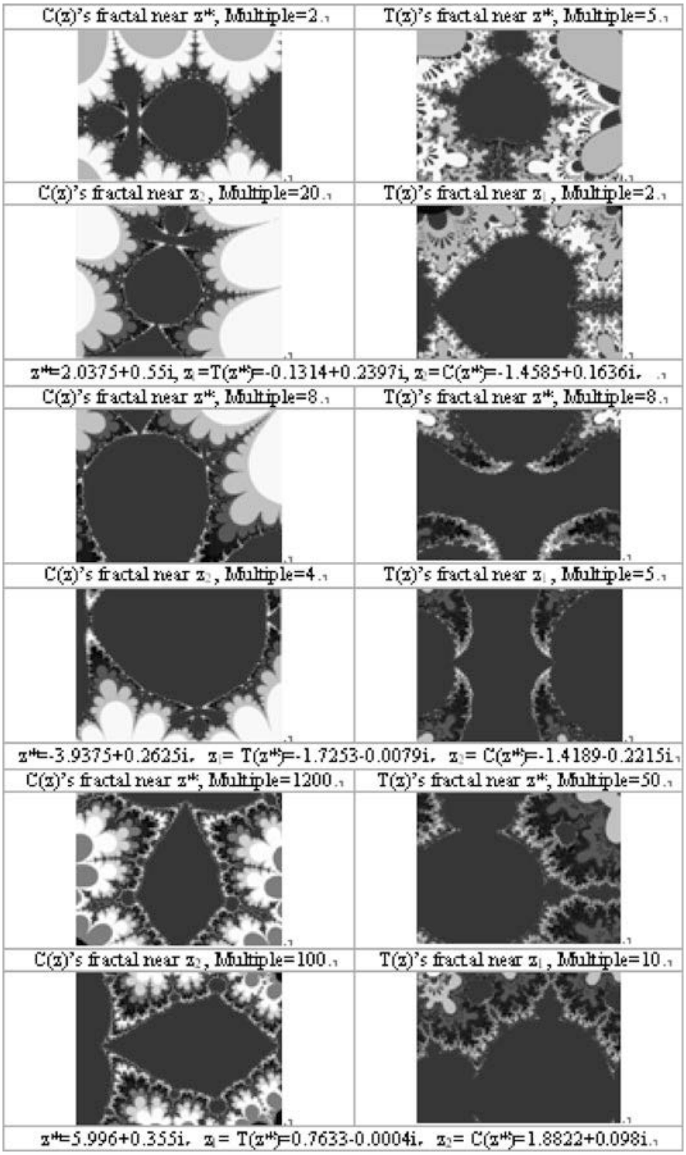
We can validate conjectures 1 and 2 from Fig. 81.3. In fact, we can get many similar sets by different attractive orbits z^*, $z_1^*,\ldots, z_n^*$. Otherwise, these similar sets construct fractals of generalized $3x + 1$ functions.

81.3.2 Fractals at Positive Integer

In order to observe rules at integer point of $T(z)$ and $C(z)$, we create an integer chain of generalized $3x + 1$ functions by $3n + 1$ conjecture and draw fractals of $T(n)$ and $C(n)$ by using integer chain in Fig. 81.4.

We can see that when transformation is $z/2$ except circle 1–2, just as a horizontal arrow, the convergence region changed nearly 0.5 times. When transformation is to $3z + 1$, just as vertical arrow, the convergence region changed nearly 1.5 times. It validates theorems 3 and 4, and also validates inference 1.

We observe that fractal near points 1, 2, 5 and 8 do not fit this rule. Moreover, we call points ‘well points’ when they fit theorems 3 and 4. Conversely, we call points ‘ill points’ when they do not fit theorems 3 and 4. Then we can find that all ill points are fit form $n^* = 2 + 3k$ except number 1. To observe deeply, we find that these points fit both theorems 3 and 4. In other words, there are two arrows pointing at these points. So we know that these points are restricted by both theorems 3 and 4 and they changed to ill points. Then we get



C(z)'s fractal near z*, Multiple=2

T(z)'s fractal near z*, Multiple=5

C(z)'s fractal near z_2, Multiple=20

T(z)'s fractal near z_1, Multiple=2

z*=2.0375+0.55i, z_1=T(z*)=-0.1314+0.2397i, z_2=C(z*)=-1.4585+0.1636i.

C(z)'s fractal near z*, Multiple=8

T(z)'s fractal near z*, Multiple=8

C(z)'s fractal near z_2, Multiple=4

T(z)'s fractal near z_1, Multiple=5

z*=-3.9375+0.2625i. z_1= T(z*)=-1.7253-0.0079i. z_2= C(z*)=-1.4189-0.2215i

C(z)'s fractal near z*, Multiple=1200

T(z)'s fractal near z*, Multiple=50

C(z)'s fractal near z_2, Multiple=100

T(z)'s fractal near z_1, Multiple=10

z*=5.996+0.355i. z_1= T(z*)=0.7633-0.0004i. z_2= C(z*)=1.8822+0.098i.

Fig. 81.3 Fractal at some points of *C*(*z*) and *T*(*z*)

conjecture 3. Of course, conjecture 3 can explain orbit 1–2 because they restrict each other.

Conjecture 3 *When n* fit form 2 + 3k (k ∈ N), fractal at n* are restricted by both 6k + 4 and 2k + 1*

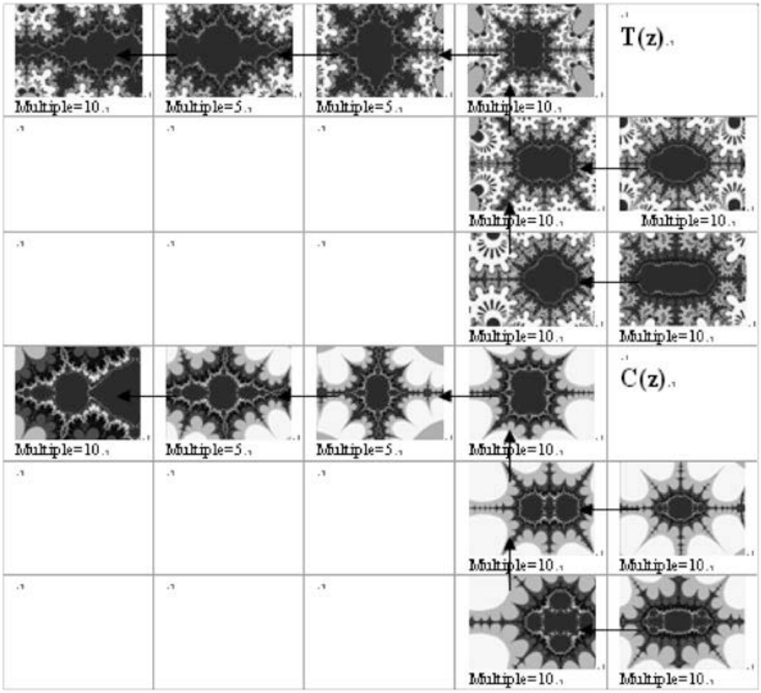


Fig. 81.4 Fractal at integer points of $C(z)$ and $T(z)$

81.4 Conclusion

We compared the dynamic systems of two generalized $3x + 1$ functions $C(z)$ and $T(z)$. We find that they are similar to each other, especially at integer points. That is to say, we can describe it with IFS if we can get the factors.

From this chapter we find two results. One is that characters of dynamic systems of C(z) and $T(z)$ are similar in complex plane. Moreover, the character of one point is similar to the character of its iteration. The other is that we find integer points of $C(z)$ and $T(z)$ have the same results as chains created by $3n + 1$ conjecture. To consider these two functions having the same iteration as $3n + 1$ conjecture with integer points, we can study $3n + 1$ conjectures deeply based on the conclusions of this chapter.

References

1. Belaga E, Mignotte M (1998) Embedding the 3x + 1 Conjecture in a 3x + d Context. Exp Math 7(2):145–151
2. Simons J, de Weger B (2005) Theoretical and computational bounds for m-cycles of the 3n + 1 problem. Acta Arith 117(1):51–70

3. Mandelbrot BB (1982) The fractal geometry of nature. Freeman W H, San Fransisco, pp 1–122
4. Pe JL (2004) The 3x + 1 fractal. Comput Graph 25(3):431–435
5. Dumont JP, Reiter CA (2001) Visualizing generalized 3x + 1 function dynamics. Comput Graph 25(5):553–595
6. Li X, Wang Z (2007) Generalized M-set produced by generalized 3x + 1 function and the artistic fractal images. J Comput-Aided Design Comput Graph 19(4):419–424 (in Chinese with English Abstract)
7. Liu S, Wang Z (2009) Fixed point and fractal images for a generalized approximate 3x + 1 function. J Comput-Aided Design Comput Graph 21(12):1740–1744
8. Liu S et al (2011) The existence of fixed point for a generalized 3x + 1 Function. Appl Mech Mater 55(57):1341–1345
9. Liu S et al (2011) Periodic point at real axis for a generalized 3x + 1 function. Appl Mech Mater 55(57):1670–1674

Chapter 82
The Initial and Neumann Boundary Value Problem for A Class Parabolic Monge–Ampère Equation

Juan Wang, Huizhao Liu and Jinlin Yang

Abstract Monge–Ampère equation is a typical fully nonlinear non-uniformly equation. The study of MA is motivated by the following two problems: Minkowski problem and Weyl problem. We consider the existence and uniqueness of a classical solution to the initial and Neumann boundary value problem for a class nonlinear parabolic equation of Monge–Ampère type. We show that such a solution exists for all times and is unique.

Keywords Parabolic Monge–Ampère equation · Neumann · Boundary value

82.1 Introduction

Monge–Ampère equation is a typical fully nonlinear non-uniformly equation. The study of MA is motivated by the following two problems: Minkowski problem and Weyl problem. One prescribes curvature type and the other is of embedding type. Monge–Ampère has many applications. In recent years new applications have been found in affine geometry and optical transportation problems. The otp was

J. Wang (✉) · J. Yang
School of Mathematics,Physics and Biological Engineering, Inner Mongolia
University of Science and Technology, Baotou 014010, China
e-mail: tlwangjuan@yahoo.com.cn

J. Yang
e-mail: yjl57715@163.com

H. Liu
Mathematical Institute Hebei, University of Technology, Tianjin 300130, China
e-mail: hz_liu@hebut.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_82,

proposed by Monge in 1781, and the main breakthrough was made by Kantorovich in 1940. Kantorovich introduced a very useful linear dual functional. The optimal mapping can be determined by the potential functions, namely maximizers of the dual functional, under certain conditions on the cost functions. The potential functions satisfy the second boundary condition of a Monge–Ampère type equation.

In this paper, we consider the existence and uniqueness of a classical solution to the initial and Neumann boundary value problem for a class parabolic equation of Monge–Ampère type :

$$\begin{cases} \dot{u} = \det^{\frac{1}{n}}(D_x^2 u) - g(x,u) & in\,\Omega \times (0,T] \\ u_\nu = \varphi(x,u) & on\,\partial\Omega \times [0,T] \\ u\big|_{t=0} = u_0 & in\,\Omega \end{cases} \tag{82.1}$$

where $u = \frac{\partial u}{\partial t}$, Ω is a bounded, uniformly convex domain in$R^n$with the boundary $\partial\Omega \in C^{4+\alpha}$ ν denotes the unit inner normal on $\partial\Omega$ which has been extended on $\overline{Q_T}$ to become a properly smooth vector field independent of t. The function $g \in C^{2+\alpha,2+\alpha}(\overline{\Omega} \times R)$, $\varphi \in C^{3+\alpha,3+\alpha}(\overline{\Omega} \times R)$ and the initial value $u_0 \in C^{4+\alpha}(\overline{\Omega})$, is a strictly convex function on $\overline{\Omega}$. In the sequel we assume for simplicity $0 \in \Omega$.

To guarantee the existence of the classical solutions for (82.1) and convergence to a solution with prescribed curvature, we have to assume several structure conditions analogous to [1]. These are

$$\varphi_z \equiv \frac{\partial \varphi(x,z)}{\partial z} \geq c_\varphi > 0 \tag{82.2}$$

$$g > 0\,\text{and}\,g_z \equiv \frac{\partial g(x,z)}{\partial z} \geq 0 \tag{82.3}$$

$$\det^{\frac{1}{n}}(D_x^2 u_0) - g(x,u_0) \geq 0. \tag{82.4}$$

Moreover, we will always assume the following compatibility conditions to be fulfilled on $\partial\Omega \times \{\text{t} = 0\}$

$$(u_0)_\nu = \varphi(x,u_0) \tag{82.5}$$

$$\left(\det^{\frac{1}{n}}(D_x^2 u_0\right) - g(x,u_0))_\nu \geq \varphi_z(x,u_0)\left(\det^{\frac{1}{n}}(D_x^2 u_0) - g(x,u_0)\right). \tag{82.6}$$

For elliptic equations of Monge–Ampère type have been explored by using the continuity method. Some of the techniques used there will be applied in our paper as well. For the parabolic case, Oliver and Knut Smoczyk [1] consider the flow of a strictly convex hypersurface driven by the Gauss curvature. For the Neumann boundary value problem and for the second boundary value problem they show that such a flow exists for all times and converges eventually to a solution of the prescribed Gauss curvature equation.

In Sect. 82.1, we shall obtain the uniqueness of the strictly convex classical solutions by the comparison principle. In Sect. 82.2, we shall prove uniform estimates for $|\dot{u}|$. This will be used in Sect. 82.3 to derive C^0- estimates. C^1- estimates then follow from [2–4, 12]. In Sect. 82.4, we shall derive C^2- estimates and the $C^{2+\beta,1+\frac{\beta}{2}}$ estimates. In Sect. 82.5, we will give the proof of Theorem 0.1.

Our main result is as follows.

Theorem 0.1 *Assume that Ω is a bounded, uniformly convex domain in R^n with the boundary $\partial\Omega \in C^{4+\alpha}$ ν denotes the unit inner normal on $\partial\Omega$ which has been extended on $\overline{Q_T}$ to become a properly smooth vector field independent of t. Let $g \in C^{2+\alpha,2+\alpha}(\overline{\Omega} \times R)$ and $\varphi \in C^{3+\alpha,3+\alpha}(\overline{\Omega} \times R)$ that satisfy* (82.2)–(82.3). *Let $u_0 \in C^{4+\alpha}(\overline{\Omega})$, be a strictly convex function that satisfies* (82.4). *Moreover, the compatibility conditions* (82.5–82.6) *are fulfilled. Then there exists a unique strictly convex solution of* (82.1) *in $K^{4+\alpha}$ for some $\alpha \in (0,1)$, where*

$$K^{4+\alpha}\!:= \{v(,t)|v(x,t) \in C^{2,1}(Q_T) \cap C^{1,0}(\overline{Q_T})\, and\, v(\cdot,t)\, is\, strictly\, convex\, for$$

every $t \in [0,T]\} \cap C^{4+\alpha,2+\frac{\alpha}{2}}(\overline{Q_T})$, $Q_T = \Omega \times (0,T], \nu$ is the inward ponting unit normal of $\partial\Omega$.

Proof Uniqueness of the strictly convex classical solution is given by Theorem 1.2. From the estimates obtained in Sects. 82.2–82.4, we get the existence of the classical solution in Sect. 82.5.

82.2 Comparison Principle and Uniqueness

Lemma 1.1 *Assume $u, v \in C^{2,1}(\overline{Q_T})$ and $u(\cdot,t), v(\cdot,t)$ are all convex for every time $t \in (0,T]$. Let $g \in C^{2,2}(\overline{\Omega} \times R)$ and $g_z = \frac{\partial g(x,z)}{\partial z} \geq 0$. Moreover, assume that*

$$-\dot{u} + \det^{\frac{1}{n}}(D_x^2 u) - g(x,u) \geq -\dot{v} + \det^{\frac{1}{n}}(D_x^2 v) - g(x,v)\; in\; \Omega \times (0,T];$$

if $u > v$, then $u_\nu > v_\nu$ *on* $\partial\Omega \times [0,T]$;

$$u \leq v\; on\; \Omega \times \{t = 0\};$$

where ν is the inward pointing unit normal of $\partial\Omega$, then $u \leq v$ in $\overline{Q}_T$.

Proof Using the given condition and the weak parabolic maximum principle.

Theorem 1.2 *Under the assumptions of theorem 0.1, there exists a unique classical solution of* (82.1).

Proof Using lemma 1.1.

82.3 u̇-Estimates

The proof of the u̇estimates can be carried out as in [1]. For a constant λ we define the function $r = e^{\lambda t}(\dot{\mathrm{u}})^2$ thus Eq. (82.1) implies the following evolution equation for r

$$\dot{r} = \frac{1}{n}\det^{\frac{1}{n}}(D_x^2 u)u^{ij} r_{ij} - \frac{2}{n} e^{\lambda t} \det^{\frac{1}{n}}(D_x^2 u)u^{ij}\, \dot{u}_i\, \dot{u}_j + (\lambda - 2g_z)r \tag{82.7}$$

Theorem 2.1 *As long as a strictly convex solution of* (82.1) *exists we obtain the estimates* $\left|\dot{u}\right|_{0,\overline{Q_T}} \leq \overline{M}$,

where $\overline{M}$ *is a controllable constant.*

Lemma 2.2 *If* $0 \leq \dot{u}(x,0) \neq 0$ *for* $t = 0$, *then a solution of* (82.1) *satisfies* $\dot{u} > 0$ *or equivalently* $\det^{\frac{1}{n}}(D_x^2 u) - g(x,u) > 0$ *for* $t > 0$..

82.4 C⁰- and C¹-Estimates

In this section, we derive the C^0- and C^1-estimates of the solution to problem (82.1).

Theorem 3.1 *Let Ω be a bounded, uniformly convex domain in R^n.$u \in C^{2,1}(Q_T) \cap C^{1,0}(\overline{Q_T})$, is a strictly convex solution of* (82.1). *Then there exists a controllable constant M_0, such that $|u|_{0,\overline{Q}_T} \leq M_0$.*

Proof. Proof u is uniformly a priori bounded from below and above.

Theorem 3.2 *Let Ω be a bounded, uniformly convex domain in R^n and $u \in C^{4,2}(Q_T) \cap C^{1,0}(\overline{Q_T})$ is a strictly convex solution of* (82.1). *Then we have*

$$\sup_{Q_T}|Du| \leq M^*$$

where M^**is a controllable constant.*

82.5 C²- and $C^{2+\beta,1+\frac{\beta}{2}}$-Estimates

Theorem 5.1 *Assume that Ω is a C^4 bounded, uniformly convex domain in $R^n$and $u \in C^{4,2}(\overline{Q_T})$ is a strictly convex solution of* (82.1). *Let $g \in C^{2,2}(\overline{\Omega} \times R)$ and $\varphi \in C^{3,3}(\overline{\Omega} \times R)$. Then we have*

$$\sup_{Q_T}\left|D_x^2u\right| \leq M''.$$

where M'' *is a controllable constant.*

From the uniform c^0 -estimates, $\dot{u}$ -estimates and the assumptions on g, we can conclude that $F(D^2u)$ has a priori positive bound from below. And using the uniform C^2-estimates for u, we obtain that the equation of (82.1) is uniformly parabolic. So we can apply the method of [5] to obtain the $C^{2+\beta,1+\frac{\beta}{2}}$ interior estimates and the estimates near the bottom. Using the estimates near the side in [6–8], we can get the Hölder semi-norm estimates for u and D_x^2u. Thus we have the $C^{2+\beta,1+\frac{\beta}{2}}$-estimates.

82.6 The Proof of Theorem 0.1

In Sect.82.2 we prove the uniqueness of the strictly convex solution for (82.1). The existence of the strictly convex solution for (82.1) is obtained by using the continuity method. Applying Theorem 5.3 in [9–11], the implicit function theorem and the Arzela–Ascoli theorem, we can get the desired result. Then the standard regularity of parabolic equation implies $u \in C^{4+\beta,2+\frac{\beta}{2}}$. Since there are sufficient a priori estimates,we can extend a solution of (82.1) on a time interval [0, T] to $[0, T+\varepsilon)$ for a small $\varepsilon > 0$. In this way we obtain existence for all $t \geq 0$ from the a priori estimates.

Now we complete the proof of Theorem 0.1.

References

1. Schnurer OC, Smoczyk K (2003) Neumann and second boundary value problems for Hessian and Gauss Curvature flows. Ann I H Poincar′e-AN 20(6):1043–1073
2. Lions PL, Trudinger NS, Urbas JIE (1986) The Neumann problem for equations of Monge–Amp‘ere type. Comm Pure Appl Math 39:539–563
3. Lions PL, Sznitman AS (1984) Stochastic differential equations with reflecting boundary conditions. Commun Pur Appl Math 37(4):511–537
4. Alvino A, Trombetti G, Diaz JI, Lions PL (1996) Elliptic equations and Steiner symmetrization. Commun Pur Appl Math 49(3):217–236
5. CHEN YZ (1986) Krylov's a priori estimates methods on fully nonlinear equation. Adv Math(China) 15(1):63–101
6. Dong GC (1998) Initial and nonlinear oblique boundary value problem for fully nonlinear partial equations. J PDE 2 Series A 1:12–42
7. Dong GC (1994) Elliptic and parabolic equations partial differential equations in China mathematics and its applications. Kluwer, Dordrecht, pp 30–41
8. Dong GC, Mu CL (1998) The measurable viscosity solutions for fully nonlinear elliptic equations. Nonlinear Anal 33(4):401–412
9. Ladyzenskaja A, Solonnikov VA, Ural'zeva NN (1967) Linear and quasilinear equations of parabolic type. (Russian) Translated from the Russian by S.Smith.Translations of

mathematical monographs. vol 23 American Mathematical Society, Providence, R.I. xi + 648

10. Temam R (1982) Behaviour at time t = 0 of the solutions of semilinear evolution equations. J Differ Equ 43(1):73–92
11. Thomee V, Wahlbin L (1974) Convergence rates of parabolic difference schemes for non-smooth data. Math Comp 28:1–13
12. Bahri A, Lions PL (1988) Morse index of some min–max critical points I Application to multiplicity results. Commun Pur Appl Math 41(8):1027–1037

Chapter 83
Method of Lower and Upper Solutions for Fourth-Order Multi-Point Boundary Value Problem with p-Laplacian Operator

Xianrui Meng, Yuxia Tong and Shujun Wang

Abstract In this chapter, the following fourth-order three-point boundary value problem is studied:

$$\begin{cases} \left(\phi_p(u''(t))\right)'' = f(t, u(t), u''(t)) \\ u(0) = au(\xi), u(1) = bu(\xi) \\ u''(0) = cu''(\eta), u''(1) = du''(\eta) \end{cases}$$

with the condition that the nonlinear term f is monotone, it is proved that there exists at least one solution to the above the fourth-order three-point boundary value problem by using the upper and lower solutions method and fixed point theorem.

Keywords Lower and upper solutions · p-Laplacian operator · Fixed point theorem

Project supported by Natural Science Foundation of Hebei Province of China (A2010000910) and Science Research Foundation of Hebei Polytechnic University (No. z200916).

X. Meng (✉) · Y. Tong
College of Science, Hebei United University, Tangshan,
People's Republic of China
e-mail: xianruimeng@yahoo.com.cn

Y. Tong
e-mail: tongyuxia@126.com

S. Wang
Tangshan College, Tangshan, People's Republic of China
e-mail: shujunwang88@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_83,

83.1 Introduction

Boundary value problems for ordinary differential equations play a very important role in both theory and applications. They describe a large number of physical,biological and chemical phenomena. Multi-point boundary value problems (BVPs)for ordinary differential equations arise in a variety of areas of applied mathematics and physics [1]; also, many problems in the theory of elastic stability can be handled by multi-point problems in [2].

In recent years, boundary value problems for even order differential equations can arise,especially for the fourth-order equations. The existence of solutions of second-order multi-point boundary value problems with p-Laplacian operator has been studied by many authors using the nonlinear alternative of Leray–Schauder, coincidence degree theory and fixed point theorem in cones [3–8]. There are very few works on the multi-point boundary value problem for higher order ordinary differential equation with p-Laplacian operator.

Recently, the existence of solutions of fourth-order three-point boundary value problems with p-Laplacian has been studied by Ref. [9] using the Leggett–Williams fixed point Theorem

$$\begin{cases} \left(\phi_p(u''(t))\right)''= a(t)f(u(t)), t \in (0,1) \\ u(0) = \xi u(1), u'(1) = \eta u'(0) \\ u''(0) = \alpha_1 u''(\delta), u''(1) = \beta_1 u''(\delta) \end{cases}$$

Motivated by the chapter [9], we are going to investigate fourth-order four-point boundary value problem of type p-Laplacian with the following four-point boundary value conditions

$$\begin{cases} \left(\phi_p(u''(t))\right)''= f(t, u(t), u''(t)) \\ u(0) = au(\xi), u(1) = bu(\xi) \\ u''(0) = cu''(\eta), u''(1) = du''(\eta) \end{cases} \tag{83.1}$$

where $\varphi_p(u) = |u|^{p-2}u, 1 < p < \infty,\ 0 < \xi, \eta < 1,\ 0 \leq a, b, c, d < 1, f : [0,1] \times R^2 \to R$ is continuous, it is easy to verify that the inverse function of φ_p is φ_q, where $\frac{1}{p} + \frac{1}{q} = 1$. Moreover, $\varphi_p(u)$ and $\varphi_q(u)$ are increasing functions with respect to $u \in (-\infty, +\infty)$, and they are odd functions.

The way of processing in this chapter is different from paper [9]. Under the conditions of existence of a upper solution−lower solution pair, we proved that the boundary value problem (83.1) has at least one solution.

83.2 Preliminaries and Lemmas

Denote $B = \left\{u \in C^2[0,1] : \varphi_p(u'') \in C^2[0,1]\right\}$, which are equipped with the norm $||u|| = \max_{t\in[0,1]} |u(t)|$, then $(B, ||\cdot||)$ is a Banach space.

Definition 1 Letting $\beta \in B$, we say β is an upper solution for the problem (83.1) if β satisfies

$$\begin{cases} \left(\phi_p(\beta''(t))\right)'' \geq f(t,\beta(t),\beta''(t)), & \text{for } t \in (0,1) \\ \beta(0) \geq a\beta(\xi), & \beta(1) \geq b\beta(\xi) \\ \beta''(0)\beta''(\eta), & \beta''(1) \leq d\beta''(\eta) \end{cases} \tag{83.2}$$

Definition 2 Letting $\alpha \in B$, we say α is a lower solution for the problem (83.1) if α satisfies

$$\begin{cases} \left(\phi_p(\alpha''(t))\right)'' \leq f(t,\alpha(t),\alpha''(t)), & \text{for } t \in (0,1) \\ \alpha(0) \leq a\alpha(\xi), & \alpha(1) \leq b\alpha(\xi) \\ \alpha''(0) \geq c\alpha''(\eta), & \alpha''(1) \geq d\alpha''(\eta) \end{cases} \tag{83.3}$$

Definition 3 For $u,\ \mathrm{v} \in B,$ we say $u \leq v,$ if and only if $u(t) \leq v(t), u''(t) \geq v''(t),$ $t \in [0,1]$

Lemma 1 [9] *If $f \in C(R,R), M_1 = 1 - a - (b-a)\xi \neq 0$, then the unique solution of the following second-order three-point boundary value problem*

$$\begin{cases} -u'' = f(t), t \in (0,1) \\ u(0) = au(\xi), u(1) = bu(\xi) \end{cases} \tag{83.4}$$

is

$$u(t) = \int_0^1 G(t,s)f(s)ds \tag{83.5}$$

where

$$H(t,s) = \frac{1}{M_1}\begin{cases} s(1-t) + bs(t-\xi), & 0 \leq s \leq\ < \xi < 1 \text{ or} \\ & 0 \leq s \leq \xi \leq t \leq 1 \\ t(1-s) + bt(s-\xi) + a(1-\xi)(s-t), & 0 \leq t \leq s \leq \xi < 1 \\ s(1-t) + b\xi(t-s) + a(1-t)(\xi-s), & 0 \leq \xi \leq s \leq t \leq 1 \\ (1-s)(t-at+a\xi), & 0 < \xi \leq t \leq s \leq 1 \text{ or} \\ & 0 \leq t < \xi \leq s \leq 1 \end{cases} \tag{83.6}$$

Now let us consider the following linear boundary value problem

$$\begin{cases} \left(\phi_p(u''(t))\right)''= y(t) \\ u(0) = au(\xi), u(1) = bu(\xi) \\ u''(0) = cu''(\eta), u''(1) = du''(\eta) \end{cases} \tag{83.7}$$

For BVP (83.7), *we have the following lemma which is a direct conclusion of Lemma* 1.

Lemma 2 *Let* $0<\xi,\eta<1, 0\leq a,b,c,d<1$, $M_1 = 1 - a - (b - a)\xi \neq 0$, $M_2 = 1 - c_1 - (d_1 - c_1)\eta \neq 0$, *and* $c_1 = \varphi_p(c), d_1 = \varphi_p(d)$, *if* $y \in C[0, 1]$, *then the BVP* (83.7) *has a unique solution*

$$u(t) = \int\limits_0^1 G(t,s)\phi_q\left(\int\limits_0^1 H(s,\tau)y(t)d\tau\right)ds \tag{83.8}$$

where

$$H(t,s) = \frac{1}{M_2}\begin{cases} s(1-t) + d_1 s(t-\eta), & 0\leq s\leq t<\eta<1 \text{ or} \\ & 0\leq s\leq \eta\leq t\leq 1 \\ t(1-s) + d_1 t(s-\eta) + c_1(1-\eta)(s-t), & 0\leq t\leq s\leq \eta<1 \\ s(1-t) + d_1\eta(t-s) + c_1(1-t)(\eta-s), & 0\leq \eta\leq s\leq t\leq 1 \\ (1-s)(t-c_1 t + c_1\eta), & 0<\eta\leq t\leq s\leq 1 \text{ or} \\ & 0\leq t<\eta\leq s\leq 1 \end{cases} \tag{83.9}$$

For any $u \in B$, *the operator K is defined as following*

$$(Ku)(t) = \int\limits_0^1 G(t,\ s)\cdot\phi_q\left(\int\limits_0^1 H(s,\ \tau)f(\tau,\mathrm{u}(\tau),\mathrm{u}''(\tau))d\tau\right)ds, \tag{83.10}$$

It is easy to say that, K is a completely continuous operator. It is obvious that the function u is the solution to p-Laplacian BVP (83.1) *if and only if* $u \in B$ *and* $u = Ku$, *that means u is the fixed point of operator K.*

83.3 Main Result

In this section,we will obtain existence of solution for p-Laplacian BVPs (83.1) by the upper and lower solutions method.

Theorem *Assume that function f is increasing in u,decreasing in* u'', *if there exist lower solution* $\alpha(t)$ *and upper solution* $\beta(t)$ *of BVP* (83.1), *then p-Laplacian BVP* (83.1) *have a solution* $u \in B$ *such that* $\alpha\leq u\leq \beta$.

Proof Define $\Omega = \{u \in B : \alpha \leq u \leq \beta\}$. We show that $K\Omega \subseteq \Omega$

In fact, for any $u \in \Omega$, denote $Ku = \omega$, let $x(t) = \varphi_p(\beta''(t)) - \varphi_p(\omega''(t))$, from (83.2) and (83.10), and f is increasing in u, decreasing in u'', we can get

$$\begin{cases} x''(t) = \big(\varphi_p(\beta''(t))\big)'' - \big(\varphi_p(\omega''(t))\big)'' \geq f\big(t, \beta(t), \beta(t)''\big) - f\big(t, u(t), u(t)''\big) \geq 0, \\ x(0) - c_1 x(\eta) = \varphi_p(\beta''(0)) - \varphi_p(\omega''(0)) - c_1\left[\varphi_p(\beta''(\eta)) - \varphi_p(\omega''(\eta))\right] \leq 0 \\ x(1) - d_1 x(\eta) = \varphi_p(\beta''(1)) - \varphi_p(\omega''(1)) - d_1\left[\varphi_p(\beta''(\eta)) - \varphi_p(\omega''(\eta))\right] \leq 0 \end{cases} \tag{83.11}$$

This means

$$\begin{cases} x''(t) \geq 0, t \in (0,1) \\ x(0) - c_1 x(\eta) \leq 0, (1) - d_1 x(\eta) \leq 0 \end{cases} \tag{83.12}$$

Let $x''(t) = g(t) \geq 0, x(0) - c_1 x(\eta) = A \leq 0, x(1) - d_1 x(\eta) = B \leq 0$, then $x(t)$ satisfy the following conditions

$$\begin{cases} x''(t) = g(t), t \in (0,1) \\ x(0) - c_1 x(\eta) = A, x(1) - d_1 x(\eta) = B \end{cases} \tag{83.13}$$

Then the unique solution of (83.13) is

$$x(t) = h(t) - \int_0^1 H(t,s) g(s) \mathrm{d}s \tag{83.14}$$

where

$$h(t) = \frac{(1 - d_1\eta - t + d_1 t)A + \left[c_1\eta + (1 - c_1)t\right] B}{(1 - c_1)(1 - d_1\eta) + c_1\eta(1 - d_1)} \leq 0 \tag{83.15}$$

because $H(t,s) \geq 0, g(t) \geq 0$, we know that $x(t) \leq 0$, which implies that $\varphi_p(\beta''(t)) - \varphi_p(\omega''(t)) \leq 0$, since φ_p is monotone increasing, we have

$$\beta''(t) \leq \omega''(t), \quad t \in (0,1) \tag{83.16}$$

Let $y(t) = \beta(t) - \omega(t)$, from (83.16) and (83.2) we can get

$$\begin{cases} y''(t) \leq 0 \\ y(0) - ay(\xi) \geq 0, y(1) - by(\xi) \geq 0 \end{cases} \tag{83.17}$$

Thence $y(t) \geq 0$, which implies

$$\beta(t) \geq \omega(t) \tag{83.18}$$

from (83.16) and (83.18), we can get

$$\omega \leq \beta \tag{83.19}$$

In the similar way, we can obtain that

$$\alpha \leq \omega \tag{83.20}$$

From (83.19) and (83.20), we know that

$$\alpha \leq Ku \leq \beta \tag{83.21}$$

It is true that $K\Omega \subseteq \Omega$. Therefore, by Schauder fixed point theorem, there exists a fixed point $u \in B$ such that $u = Ku$. It is obvious that $\alpha \leq u \leq \beta$. The proof of theorem is completed.

References

1. Moshiinsky M (1950) Sobre los problems de condiciones a la frontiera en una dimension de caracteristicas discontinuas. Bol Soc Mat Mexicana 7:1–25
2. Timoshenko S (1961) Theory of elastic stability. McGraw-Hill, New York
3. Wang J, Gao W, Lin Z (1995) Boundary value problems for general second order equations and similarity solutions to the Rayleigh problem. Tohoku Math J 47:327–344
4. Wang JY, Gao WJ (1996) A singular boundary value problem for the one-dimensional p-Laplacian[J]. J Math Anal Appl 201:851–866
5. Wong FH (1999) Existence of positive solutions for p-Laplacian boundary value problems[J]. Appl Math Lett 12:11–17
6. Wang JY (1997) The existence of positive solutions for the one dimensional p-Laplacian[J]. Proc Amermath Soc 125(8):2275–2283
7. Liu B, Yu JS (2001) Multiple positive solutions of singular boundary value problems with p-Laplacian[J]. Chin Ann Math 22A(6):721–728
8. Liu B (2005) The existence of positive solutions of singular boundary value systems with p-Laplacian[J]. Acta Math Sinica 48(1):35–50
9. Ma DX, Tian Y, Ge WG (2006) Existence theorems of positive solutions for a fourth-order three-point boundary value problem[J]. Taiwan J Math 10(6):1557–1573

Chapter 84
The Existence of Solutions to Fourth-Order Boundary Value Problems for Impulsive Differential Equations

Xianrui Meng, Liping Du and Nana Li

Abstract In this chapter, the existence of solutions to fourth-order three-point boundary value problem for impulsive differential equations is studied. With the condition that nonlinear term *f* is monotone and the existence of an upper–lower pair, it is proved that there exists at least one solution to fourth-order three-point boundary value problem with impulsive effects by using the upper and lower solutions method and monotone iterative theorem.

Keywords Lower and upper solutions · p-Laplacian operator · Impulsive

84.1 Introduction

During many evolution processes, evolution is subjected to a rapid change, that is, a jump in their states. This phenomenon is described by impulsive differential equations in mathematics. For the background, theory and applications of impulsive

Project supported by Natural Science Foundation of Hebei Province of China (A2010000910) and Science Research Foundation of Hebei Polytechnic University (No. z200916).

X. Meng (✉) · L. Du
College of Science, Hebei United University, Tangshan, People's Republic of China
e-mail: xianruimeng@yahoo.com.cn

L. Du
e-mail: liping.du@163.com

N. Li
Tangshan College, Tangshan, People's Republic of China
e-mail: lalazinana@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_84,

differential equations, we refer the readers to the monographs and some recent contributions as [1, 2].

Recently, the existence of positive solutions for Sturm–Liouville impulsive problem has been studied by Ref. [3] using the critical point theory and variational methods.

Motivated by the paper [3], we are going to investigate fourth-order boundary value problem for impulsive differential equations

$$\begin{cases} \left(\phi_p(x''(t))\right)''=f(t,x(t),x''(t)), \quad t_i<t<t_{i+1} \\ a_1x(0)-b_1x'(0)=0,\ c_1x(1)+d_1x'(1)=0 \\ x''(0)-c_2x''(\eta)=0,\ x''(1)-d_2x''(\eta)=0, \\ \Delta x(t_i)=\alpha_1(i), \qquad\qquad\qquad\qquad i=0,1,2\ldots k \\ \Delta x'(t_i)=\alpha_2(i,x(t_i),x''(t_i)) \\ \Delta\phi_p(x''(t_i))=\alpha_3(i) \\ \Delta\left(\phi_p(x''(t_i))\right)'=\alpha_4(i,x(t_i),x''(t_i)) \end{cases} \tag{84.1}$$

where $0=t_0<t_1<\cdots<t_k<t_{k+1}=1, 0\le a_1,b_1,c_1,\quad d_1,c,d<1, 0<\eta<1, \eta\neq t_k$; $\varphi_p(u)=|u|^{p-2}u, 1<p<\infty, \alpha_1(i)$ and $\alpha_3(i)$ is constant, $\alpha_l: R^2\to R, l=2,4$ is continuous, $1\le i\le k$; $\Delta x(t)=x(t^+)-x(t^-)$ and $x(t_i)=x\left(t_i^-\right), i=1,2\ldots k$. Moreover, $\varphi_p(u)$ is increasing functions with respect to $u\in(-\infty,+\infty)$, and it is odd function.

Under the conditions of existence of a upper solution–lower solution pair, we prove that the boundary value problem (84.1) has at least one solution.

84.2 Preliminaries and Lemmas

Let $PC^2[0,1]$ be the set of piecewise continuous functions defined [0,1] which satisfy $x^{(j)}$ is piecewise continuous on [0,1], $j=0,1,2$. $x\in PC^2[0,1]$ means that x is defined on [0,1] and $x\in C^2(t_i,\ t_{i+1}]$, $x^{(j)}(t_i^-)$ exists, $i=0,1,\ldots k,\ j=0,1,2$.

Denote $B=\{x\in PC^2[0,1]: x\in C^2(t_i,\ t_{i+1}], i=0,1,\ldots k\}$, which equipped with the norm $|x|=\max_{i=0,1,\ldots k}\{|x|_i,\ |x'|_i,\ |x''|_i\}$, and $|x|_i=\sup_{t_i\le t\le t_{i+1}}|x(t)|$. then $(B,||g||)$ is a Banach space.

Definition 1 Letting $\beta\in B(\alpha\in B)$, we say $\beta(\alpha)$ is a upper(lower) solution for the problem (84.1) if $\beta(\alpha)$ satisfies

$$\begin{cases} \left(\phi_p(\beta''(t))\right)''\ge(\le)f(t,\beta(t),\beta''(t)), \quad t_i<t<t_{i+1} \\ a_1\beta(0)-b_1\beta'(0)=0, c_1\beta(1)+d_1\beta'(1)=0 \\ \beta''(0)-c_2\beta''(\eta)=0, \beta''(1)-d_2\beta''(\eta)=0, \\ \Delta\beta(t_i)=\alpha_1(i), \quad i=0,1,2\ldots k \\ \Delta\beta'(t_i)\le(\ge)\alpha_2(i,\beta(t_i),\beta''(t_i)) \\ \Delta\phi_p(\beta''(t_i))=\alpha_3(i) \\ \Delta\left(\phi_p(\beta''(t_i))\right)'\ge(\le)\alpha_4(i,\beta(t_i),\beta''(t_i)) \end{cases} \tag{84.2}$$

Definition 2 For $u, v \in B$, we say $u \le v$ if and only if $u(t) \le v(t), u''(t) \ge v''(t)$, $t \in (t_i,\ t_{i+1}],\ i = 0, 1, \ldots k$

Lemma 1 *if* $f \in C[0,1], M_2 = 1 - \varphi_p(c_2) - \big(\varphi_p(d_2) - \varphi_p(c_2)\big)\eta \neq 0$, *Then the unique solution of the following second-order three-point boundary value problem*

$$\begin{cases} y''(t) = f(t), \quad t_i < t < t_{i+1} \\ y(0) - \varphi_p(c_2)y(\eta) = 0, y(1) - \varphi_p(d_2)y(\eta) = 0, \quad i = 0, 1, \ldots k \\ \Delta y(t_i) = \alpha_3(i), \Delta y'(t_i) = \alpha_4(i, x(t_i), x''(t_i)) \end{cases} \tag{84.3}$$

is

$$y(t) = I_2(t; \alpha_3, \alpha_4) + \int\limits_0^1 H(t,s)f(s)\mathrm{d}s \tag{84.4}$$

where

$$I_2(t; \alpha_3, \alpha_4) = \sum_{i=1}^{k} I_2(i, t; \alpha_3(i), \alpha_4(i)) \tag{84.5}$$

$$I_2(i, t; \alpha_3(i), \alpha_4(i)) = \begin{cases} t(-\alpha_3(i) - (1 - t_i)\alpha_4(i)), 0 \le t \le t_i \\ (1 - t)(\alpha_3(i) - t_i\alpha_4(i)), t_i \le t \le 1, \end{cases}, i = 1, 2 \ldots k \tag{84.6}$$

$$H(t,s) = \frac{-1}{M_2}\begin{cases} s(1-t) + \varphi_p(d_2)s(t-\eta), & \begin{array}{l} 0 \le s \le t \le \eta < 1\, or \\ 0 \le s \le \eta \le t \le 1 \end{array} \\ t(1-s) + \varphi_p(d_2)t(s-\eta) + \varphi_p(c_2)(1-\eta)(s-t), & 0 \le t \le s \le \eta < 1 \\ s(1-t) + \varphi_p(d_2)\eta(t-s) + \varphi_p(c_2)(1-t)(\eta-s), & 0 \le \eta \le s \le t \le 1 \\ (1-s)\big(t - \varphi_p(c_2)t + \varphi_p(c_2)\eta\big), & \begin{array}{l} 0 \le \eta \le t \le s \le 1\, or \\ 0 \le t \le \eta \le s \le 1 \end{array} \end{cases} \tag{84.7}$$

Proof From Ref. [4, 5] and superposition principle, the conclusion is obvious.

Now let us consider the following linear boundary value problem

$$\begin{cases} \phi_p(x''(t)) = y(t), \quad t_i < t < t_{i+1} \\ a_1x(0) - b_1x'(0) = 0, c_1x(1) + d_1x'(1) = 0, \quad i = 0, 1, \ldots k \\ \Delta x(t_i) = \alpha_1(i), \Delta x'(t_i) = \alpha_2(i, x(t_i), x''(t_i)) \end{cases} \tag{84.8}$$

For BVP (84.8), we have the following lemma which is a direct conclusion of lemma 1.

Lemma 2 *Let* $\rho_1 = a_1d_1 + a_1c_1 + b_1c_1 > 0,\ M_2 = 1 - \varphi_p(c_2) - \big(\varphi_p(d_2) - \varphi_p(c_2)\big)\eta \neq 0$, *if* $y \in C[0,1]$,

then the BVP (84.8) *has a unique solution*

$$x(t) = I_2(t; \alpha_1, \alpha_2) + \int_0^1 G(t,s)\phi_p^{-1}(y(s))\mathrm{d}s \tag{84.9}$$

where

$$G(t,s) = \frac{-1}{\rho_1}\begin{cases}(a_1 t + b_1)(c_1(1-s)+d_1), 0 \le t \le s \le 1\\(a_1 s + b_1)(c_1(1-t)+d_1), 0 \le s \le t \le 1\end{cases} \tag{84.10}$$

For any $x \in B$, the operator T is defined as following

$$(Tx)(t) = I_2(t;\ \alpha_1,\ \alpha_2(x,\ x'')) + (Kx)(t),\ t \in\ [0,\ 1], \tag{84.11}$$

$$(Kx)(t) = \int_0^1 G(t,\ s)\cdot\phi_q\left(I_2(s;\ \alpha_3,\ \alpha_4(x,\ x'')) + \int_0^1 H(s,\ \tau)f(\tau, x(\tau), x''(\tau))d\tau\right)\mathrm{d}s \tag{84.12}$$

It is easy to see that, T is a completely continuous operator

It is obvious that the function u is the solution to P-Laplacian BVP (84.1) if and only if $u \in B$ and $u = Ku$, that means u is the fixed point of operator T.

84.3 Main Result

In this section, we will obtain existence of solution for p-Laplacian BVPs (84.1) by the monotone iterative method.

Theorem *Assume that function f and α_4 are increasing in x, decreasing in x'', function α_2 is decreasing in x, increasing in x''. If there exists lower solution $\alpha_0(t)$ and upper solution $\beta_0(t)$ of BVP* (84.1)*, and $\alpha_0 \le \beta_0$, then p-Laplacian BVP* (84.1) *has a solution $u \in B$.*

Proof We divide our proof into three steps.

Step1: We show that T is a monotone operator, i.e. if $x \le y$, then $Tx \le Ty$.

$\forall x, y \in B, x \le y$, i.e. $x \le y, -x''(t) \le -y''(t) \quad t \in (t_i, t_{i+1}], i = 0, 1, \ldots k$, from (84.5) and (84.6), we can get

$$I_2(t; \alpha_1, \alpha_2(x, x'')) \le I_2(t; \alpha_1, \alpha_2(y, y'')); I_2(t; \alpha_3, \alpha_4(x, x'')) \ge I_2(t; \alpha_3, \alpha_4(y, y'')) \tag{84.13}$$

Note that $G(t,s) \le 0,\ H(t,s) \le 0$, function f is increasing in x, decreasing in x'', we get

$$\begin{aligned} & I_2(t;\alpha_3,\alpha_4(x,x'')) + \int\limits_0^1 H(t,s)f(s,x(s),x''(s))\mathrm{d}s \\ & \geq I_2(t;\alpha_3,\alpha_4(y,y'')) + \int\limits_0^1 H(t,s)f(s,y(s),y''(s))\mathrm{d}s \end{aligned} \tag{84.14}$$

Since $\varphi_q(u)$ are increasing, then from (84.13) and (84.14), we get

$$(Tx)(t) \leq (Ty)(t) \tag{84.15}$$

Moreover

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}(Tx)(t) &= \phi_q\left(I_2(t;\alpha_3,\alpha_4(x,x'')) + \int\limits_0^1 H(t,\tau)f(\tau,x,x'')\mathrm{d}\tau \right) \\ & \geq \phi_q\left(I_2(t;\alpha_3,\alpha_4(y,y'')) + \int\limits_0^1 H(t,\tau)f(\tau,y,y'')\mathrm{d}\tau \right) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}(Ty)(t) \end{aligned} \tag{84.16}$$

Then from (84.15) and (84.16), we can know

$$Tx \leq Ty \tag{84.17}$$

Step 2: We show that $T\alpha \geq \alpha,\ \beta \geq T\beta$

$$\begin{aligned} & (T\beta)(t) = I_2(t;\alpha_1,\alpha_2(\beta,\beta'')) + \int\limits_0^1 G(t,s)\phi_q\left(I_2(s;\alpha_3,\alpha_4) + \int\limits_0^1 H(s,\tau)f(\tau,\beta,\beta'')\mathrm{d}\tau \right)\mathrm{d}s \\ & \leq I_2(t;\Delta\beta,\Delta\beta') + \int\limits_0^1 G(t,s)\phi_q\left(I_2(s;\Delta\phi_p(\beta''),\Delta(\phi_p(\beta''))') + \int\limits_0^1 H(s,\tau)\big(\phi_p(\beta'')\big)''\mathrm{d}\tau \right)\mathrm{d}s \\ & = \beta(t). \end{aligned} \tag{84.18}$$

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}(T\beta)(t) &= \phi_q\left(I_2(t;\alpha_3,\alpha_4) + \int\limits_0^1 H(t,\tau)f(\tau,\beta,\beta'')\mathrm{d}\tau \right) \\ & \geq \phi_q\left(I_2\Big(t;\Delta\phi_p(\beta''),\Delta\big(\phi_p(\beta'')\big)'\Big) + \int\limits_0^1 G(t,\tau)\big(\phi_p(\beta''(\tau))\big)''\mathrm{d}\tau \right) \\ & = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\beta(t) \end{aligned} \tag{84.19}$$

Thence $\beta \geq T\beta$. In the similar way, we can obtain that $T\alpha \geq \alpha$

Step 3: We will obtain existence of solution for p-Laplacian BVPs (84.1) by the monotone iterative method.

Denote the sequences $\{\beta_n\}$ and $\{\alpha_n\}$ by $\beta_{n+1} = T\beta_{n,}\alpha_{n+1} = T\alpha_n, (n = 0, 1, 2\ldots)$, then from step 1 and step 2 we can know that

$$\alpha_0 \leq T\alpha_0 = \alpha_1 \leq \beta_1 = T\beta_0 \leq \beta_0 \tag{84.20}$$

By an induction argument, we have $\alpha_n \leq \alpha_{n+1} \leq \beta_{n+1} \leq \beta_n, (n \geq 1)$, therefore, there exist two functions $\alpha(t), \beta(t)$ in B such that $\beta_n \to \beta(n \to \infty), \alpha_n \to \alpha(n \to \infty)$, and $\alpha \leq \beta$. In order to explain this, it needs to note that $\beta_n^{(j)}\left(t_i^-\right)$ exists $j = 0, 1, 2, n \geq 1$, for $\{\alpha_n\}, \left\{\alpha_n''\right\}$ it is also true. therefore $\{\alpha_n\}, \left\{\alpha_n''\right\}, \{\beta_n\}, \left\{\beta_n''\right\}$ converge uniformly on the intervals $[t_i, t_{i+1}], i = 0, 1, 2\ldots k$ So the above conclusions are true.

Since α and β are the fixed point to T, then they are the solutions of BVPs (84.1).

References

1. Chu J, Nieto J (2008) Impulsive periodic solutions of first-order singular differential equations. Bull Lond Math Soc 40:143–150
2. Li J, Nietio J, Shen J (2007) Impulsive periodic boundary value problems of first-order differential equations. J Math Anal Appl 325:226–236
3. Tian Y, Ge W (2010) Variational methods to Sturm_Liouville boundary value problem for impulsive differential equations. Nonlinear Anal 72:277–287
4. Eloe P, Islam M (2001) Monotone methods and fourth order Lidstone boundary value problems with impulsive effects. Comm Appl Anal 5:113–120
5. Ma D, Tian Y, Ge W (2006) Existence theorems of positive solutions for a fourth-order three-point boundary value problem. Taiwan J Math 10:1557–1573

Chapter 85
Some Discussions on the Uniform Continuity Function

Bin Ran

Abstract Uniform continuity of the function is an important theoretical mathematical analysis course. In this study we begin with concept for the continuity of the function, discuss uniform continuity of the function, give proposition for uniform continuity of the criterion function, and take examples of its application, so that we can have a more comprehensive understanding of meaning and understanding for consistent with the continuity of the function.

Keywords Function · Uniformly continuous · Non-uniformly continuous

85.1 Introduction

Uniform continuity of the function is very common, important and abstract mathematical concept in application of mathematical analysis. It reflects overall nature of the function in an interval [1]. It is the basis of calculus and plays a key role in follow-up courses of the study.

We know that function is the mathematical study of the course, continuity is a function of a form of state. The function $f(x)$ in an interval is continuous, defined as the function $f(x)$ continuous at every point within the range, which reflects the local nature of the function $f(x)$ in the vicinal interval, but uniformly continuously reflects the overall nature of the function $f(x)$ in the interval. Therefore, the understanding and application of uniform continuity for function $f(x)$ and the

B. Ran (✉)
College of Mathematics and Computer Science,
Changjiang Normal University, Chongqing 408000, China
e-mail: rbfl@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_85,

mastery of the criterion for uniform continuity of function $f(x)$ will help to study the trend and nature for the function.

85.2 Some Definitions on Consistency

Uniform continuity is a very important concept, often used in calculus and other subjects, and for the function sequence, uniform convergence and uniform continuity has a close relationship [2]. In the research for convergence of function sequence, the convergence between the function sequence and functions, the relationship between uniform continuity and uniform convergence of are often used. We will discuss the function of the uniform continuity as follows:

We know that the following definition:

Definition 1 Suppose that the function $f(x)$ is defined in the interval I, if, $\exists\delta>0$, $\forall \mathrm{x}1, x2 \in I : |x1 - x2| < \delta$, then $^{|}\ f(x_1)$- $f(x_2)^{|} < \varepsilon$, so say function uniform continuous in the interval I.

Obviously, if function $f(x)$ in the interval I meet Lipschitz condition: $|f(x_1) - f(x_2)| \leq \mathrm{L}|x_1 - x_2|$, then function must be uniformly continuous in the interval I.

The definition of uniform continuity of functions makes us consider the issue from the perspective of the issue into the overall localization issue such as: bounded continuous function on the closed interval must be uniformly continuous. In the research of the overall nature of function, the uniform continuity seems to be particularly useful. For example: Riemann integral exists, there is such a proposition as follows [3, 4]:

Proposition, if the function $f(x)$ in the interval [a, b] is bounded on, and at most a finite number of discontinuity points, then the function $f(x)$ in the interval [a, b] can be integrated. In particular, the function $f(x)$ in the interval [a, b] is continuous, then the function $f(x)$ in the interval [a, b] can be integrated. In the proof of this proposition, the uniform continuity of the function plays a very important role. In the real variable functions, the famous Luzin theorem, mention continuous functions and the relationship between measurable functions, show that a continuous function approaching. Measurable function, using the more familiar continuous function, so that grasp the more abstract measurable function, and in some cases, we may appropriately change measurable functions into continuous functions [5]. Thus, many equivalent propositions and theorems for continuity and uniform continuity provide the basis for practical problems in study. The uniform continuity for the function sequence $\{f_n(x)\}$ is more complex than uniform continuity of function $f(x)$ because it involves not only a function, but relate to a sequence function of $\{f_n(x)\}$, which is defined as follows:

Definition 2 Suppose that the function $\{f_n(x)\}$ is defined in the interval E, if for any given $\varepsilon > 0$, always exist $\delta > 0$, makes that when X1, X2∈E, and $^{|}$X1-X2$^{|}$<δ, for any n∈N, exist $\{f_n(x_1)\} - \{f_n(x_2)\}| < \varepsilon$, then called function sequence $\{f_n(x)\}$ is uniformly continuous in interval E.

Definition 3 Function $f(x)$is continuous in the interval I means that function $f(x)$is continuous in each point of the interval I, if I include the left point, then the left point right continuous; if I include the right end point, then the left point left continuous [6, 7].

Definition 4 Function $f(x)$ is continuous at X0, means that if,$\exists\delta > 0$, $\forall$x1: $|\mathrm{X} -\ \mathrm{X0}| < \delta$, then $|f(x) - f(x_0)| < \varepsilon$.

To grasp uniform continuity of the function concept, should pay attention to the following three aspects:

Note the difference and contact between continuity and uniform continuity for function in the interval.

Compare the definition of continuity with uniform continuity for function in the interval: the former δ is not only related to ε, but also related to the point x0, that is, for different x0, δ is different in general. It suggests that as long as the function is continuous at every point in the interval I, the function continuous in the interval; the latter δ only related to ε, has nothing to do with the x0, that is for different x0, δ is the same. This indicates that the continuity of function in the interval, not only requires the function are continuous at each point in this interval, but also requires a continuous function on the interval is "consistent".

The essence of uniform continuous for function is that when the full value of the difference between any two points close to each other in the interval, in absolute terms, can be arbitrarily small, that is if $\forall x_1, x_2 \in I : |x_1 - x_2|\langle\delta$, then ${}^{|}f(x_1)\text{-}f(x_2)^{|} < \varepsilon$.

Pay attention to the negative narrative for uniformly continuous of function.

The negative of uniform continuity is just non-uniform continuity that is set the function $f(x)$ is defined in the interval I, $\exists\varepsilon > 0, \forall\delta > 0, \exists x_1, x_2 \in I{:}|x_1 - x_2|\langle\delta, |f(x_1) - f(x_2)| \geq \varepsilon_0$,then called function $f(x)$is non-uniformly continuous in the interval.

In general, the continuity of the function reflects the local nature of the function, and the uniform continuity of function reflects the overall nature of function in the interval. They are not only different but also related to each other[8].

85.3 Criterion for Uniform Continuity

Theorem 1 (*G • Cantor theorem*) *If the function $f(x)$ in the interval $[a,\ b]$ is continuous, the function $f(x)$ in the interval $[a,\ b]$ is also uniformly continuous.*

Proof (by contradiction) Assume that function in [a, b) is non-uniformly continuous, take $\delta = 1/\mathrm{n}$, (n = 1,2,……), then there in [a, b) exist two point x1(n)→x2(n)(n = 1,2,……), there I x1(n)-x2(n)I < 1/n, but $|f(x_1^{(\mathrm{n})}) - f(x_2^{(\mathrm{n})})| \geq \varepsilon0$

According to Weil Stella theorem, in a borded series {x1(n)} there is a convergence subsequence:

x1(n) → x0(k → ∞), where x0 ∈ [a,b].

Also because |x1(n)-x2(n)| < 1/nk, i.e. x1(n)-x2(n) → 0(k → ∞)

For x1(n) → x0(k → ∞), then x2(n) → x0(k → ∞) and $|f(x_1^{(n)}) - f(x_2^{(n)})| \geq \varepsilon 0$ are established for all k. According to the relationship between function and sequence limit, there $\lim_{k\to x_0^-} f(x_1^{(n)}) = f(x_0), lim_{k\to x_0^-} f(\mathrm{x}_2^{(n)}) = f(x_0)$, then $lim_{k\to x_0} f(x_1^{(n)}) - \lim^{-}_{k\to x_0} f(\mathrm{x}_2^{(n)}) = 0$ and$|f_{(x1(n))} - f_{(x2(n))|\geq\varepsilon 0}$ proof over.

Theorem 2 *If the function $f(x)$ satisfy the Lipschitz condition in the interval I, that is, $\forall \mathrm{x}, y \in I$, $|f(x) - f(y)| < k\ |x - y|$, where k is a constant, then the function $f(x)$ in the interval I is uniform continuous.*

Proof As the function $f(x)$ satisfy the Lipschitz condition in the interval I, that is, $\forall \mathrm{x}, y \in I, |f(x) - f(y)| > \mathrm{k}|x - y|$, then $\forall \varepsilon > 0$, since $|f(x) - f(y)|_< k|x - y| < \varepsilon$, then $|x - y| < \varepsilon/\mathrm{k}$,Take $\delta = \varepsilon/\mathrm{k} > 0$, and δ is not related to x, y, then $\forall \varepsilon > 0, \exists \delta = \frac{\varepsilon}{\mathrm{k}} > 0, \forall \mathrm{x}, y \in I, |x_1 - x_2| < \delta.]] >$There $|f(x_1) - f(x_2)| < \varepsilon$, so function $f(x)$ is uniform continuous in the interval.

Theorem 3 *If the function $f(x)$ in the interval I(finite or infinite) exist a bounded derivative function, i.e $\exists \mathrm{M} > 0$, $\forall x \geq I$, $|f'(x)| \leq M$, then function $f(x)$ is uniform continuous.*

Proof Since function $f(x)$ in the interval I exist derivate, so function $f(x)$ is continuous in the interval I. Also because $\forall x_1, \forall x_2 \in I$, function$f(x)$ in the interval meet the conditions of Lagrange's theorem. That is, there is a point ξ in [x_1,x_2] makes $|f(x_1) - f(x_2)| = |f'(x)| \leq M|x_1 - x_2|$, according to Theorem 3, function $f(x)$ in the interval I is uniform continuous.

Theorem 4 *The function $f(x)$ in the open interval (a,b) is uniform continuous ⇔ The function $f(x)$ in the open interval (a,b) is continuous and $f(\mathrm{a} + 0), f(\mathrm{b} - 0)$ both exist and limited.*

Proof Sufficiency suppose that $f(\mathrm{a} + 0), f(\mathrm{b} - 0)$ both exist and limited, proceed $f(x)$ for continuous development.

Define: when $x = \mathrm{a}, f(x) = f(\mathrm{a} + 0)$; when x ∈ (a,b), F(x) = $f(x)$; when $x = b$, $f(x) = f(\mathrm{b} - 0)$, easy to know F(x) in [a, b] is continuous, then uniform continuous in [a, b]. So function $f(x)$ is uniform continuous in [a, b].

Necessity: If $f(x)$ in (a, b) is uniform continuous, i.e. $\forall \varepsilon > 0$, $\exists \delta > 0(\delta < \mathrm{b} - \mathrm{a})$, when x_1,x_2 ∈ (a, b), and $|x_1 - x_2| < \delta$, there $|f(x_1) - f(x_2)| < \varepsilon$, take x_1,x_2 in (a, a + δ) or (b–δ,b), then according to Cauchy convergence criteria, $f(\mathrm{a} + 0)$, $f(\mathrm{b} - 0)$ both exist and limited.

Theorem 5 *If function $f(x)$ is continuous in $[a, +\infty)$, and $lim_{x\to\infty} f(x)$ exist and limited, then $f(x)$ is uniformly continuous in $[a, +\infty)$.*

Proof Suppose that $\lim_{x\to\infty} f(x) = \mathrm{A}$, i.e. $\forall \varepsilon > 0, \exists \mathrm{n} > 0, \forall x_1, x_2 \in (N, +\infty)$, there $|f(x_1) - f(x_2)| < \varepsilon$, so $f(x)$ is continuous in [a, $+\infty$) and also because $f(x)$ is continuous in [a, N], then $f(x)$ is continuous in [a, N], so $f(x)$ is uniformly continuous.

Theorem 6 *If* $f(x)$ *is continuous in* $[a, +\infty)$ *and* $lim_{\mathrm{x}\to+\infty}(\mathrm{b}x - f(x)) = 0$, *there b is non-zero constant, then* $f(x)$ *is uniformly continuous.*

Proof Suppose F(x) = $f(x)$-(ax + b). According above supposition, $f(x)$ is continuous in [a,+∞), and a$x + b$ is linear function, continuous in [a,+∞), thereby the difference between the continuous function is a continuous function, and thus the function F (x) in [a, + ∞) is continuous, according to above meaning of subject, $\lim_{\mathrm{x}\to+\infty} f(x) = 0$, so by Theorem 5, F(x) is uniformly continuous in [a, + ∞) .

Theorem 7 *If the monotone bounded function in a limited or infinite interval I is continuous, then the function in the interval I is uniformly continuous.*

Proof Suppose that I = (a, b). Since $f(x)$ is monotone and bounded, then $f(x)$ is also bounded in the interval (a, b), so limit $\lim_{\mathrm{x}\to a^+} f(x)$ and $\lim_{\mathrm{x}\to b^-} f(x)$ both exist.

According to Theorem 7, function $f(x)$ is uniformly continuous.

85.4 Application Examples

Example 1 Determine uniform continuity for the function $f(x) = \sqrt{\mathrm{x}}\ln x^2 + x$

Proof Take function X for comparison function, there $\lim_{x\to+\infty} \frac{\sqrt{\mathrm{x}}Inx^2+x}{x} = 1$, Function $g(x) = x$ is uniformly continuous in the interval (0, + ∞). According to Theorem 6, function $f(x) = \sqrt{\mathrm{x}}Inx^2 + x$ is uniformly continuous in (0, + ∞).

Example 2 Function$f(x)$ = x3, x ∈ (0,1) is uniformly continuous or not.

Proof Obviously, $f(x)$ = x3 is continuous in the interval (0,1), and $\lim_{\mathrm{x}\to0^+}\mathrm{x}^3 = 0, \lim_{\mathrm{x}\to0^-}\mathrm{x}^3 = 1$,Ie,$\lim_{\mathrm{x}\to0^+}\mathrm{x}^3 = 0$ and $\lim_{\mathrm{x}\to0^-}\mathrm{x}^3 = 1$ both exist, so function $f(x) = \sqrt{\mathrm{x}}\mathrm{ln}x^2 + x$ is uniformly continuous in the interval (0,1).

Example 3 $f(x) = 1/(1 + x2), x \in (0, +\infty)$ uniformly continuous or not

Proof Obviously, $f(x)$=1/(1 + x2) is continuous in the interval (0,1), and $\lim_{\mathrm{x}\to0^+}\mathrm{f(x)} = \lim_{\mathrm{x}\to0^+}[1/(1 + x^2)] = 1, \lim_{\mathrm{x}\to0^-}\mathrm{f(x)} = \lim_{\mathrm{x}\to0^-}[1/(1 + x^2)] = 0$, so $f(x) = 1/(1 + x2), \mathrm{x} \in (0, +\infty)$ is uniformly continuous.

Example 4 Determine uniform continuity for the function $f(x)$ = $f(x) = \sqrt{\mathrm{x}^4 + \mathrm{x}^2 + \sin x}$ in [0, + ∞)

Proof Construct $g(x) = \mathrm{x}2$, then $\lim_{\mathrm{x}\to+\infty} \frac{f(x)}{\mathrm{g}(x)} = \lim_{\mathrm{x}\to+\infty} \frac{\sqrt{\mathrm{x}^4+\mathrm{x}^2+\sin x}}{\mathrm{X}^4} = 1$, there $g(x) = x2$ is non-uniformly continuous in the interval.

According to Theorem 1, $f(x) = \sqrt{\mathrm{x}^4 + \mathrm{x}^2 + \sin x}$ in $[0, +\infty)$ is uniformly continuous.

Example 5 Determine uniform continuity for the function $f(x) = (\mathrm{x} + 2)$ $e^{\frac{1}{x}}$, $\mathrm{x} \in f(x)[1 + \infty)$.

Proof According to Theorem 1, suppose $g(x) = x + 3$, $g'(x)$ is limited in the interval $[1, +\infty)$, so $\mathrm{g}(x) = x + 3$ is uniformly continuous. But $f(x) = (\mathrm{x} + 2)e^{\frac{1}{x}}$ is continuous in the interval $[1, +\infty)$, and $\lim_{\mathrm{x}\to+\infty}[\mathrm{g}(x) - f(x)] = \lim_{\mathrm{x}\to+\infty}\left[(x + 3) - (\mathrm{x} + 2)e^{\frac{1}{x}}\right] = 0$, Then, $f(x)$ is uniformly continuous in the interval $[1, +\infty)$.

85.5 Conclusion

Uniform continuity of functions in the mathematical analysis is an important concept, it is not only the basic theory for continuous function Riemann in the closed interval, but also are closely related to concept about the subsequent integral with parameters, series of functions and so on. So to determine the uniform continuity of the function is an important part of mathematical analysis. This paper summarizes the definition of uniform continuity of functions, give general theorems for uniform continuity of function in the different kinds of interval and take examples of its application.

References

1. Xiangdong W (1994) Concepts and methods of mathematical analysis [M]. Shanghai Science and Technology Press, Shanghai
2. Peiwen L (2005) Mathematical analysis of typical problems and methods [M]. Higher Education Press, Beijing
3. Yulian L (1995) Number of notes [M]. Higher Education Press, Beijing
4. Wenjiu Q (1991) Mathematical analysis of the basic concepts and methods [M]. Higher Education Press, Beijing
5. Zhongnan Y (1997) Function uniform continuity at infinity [J]. Jimei Univ 2(1):70–75
6. Chen H, He C (2006) Revisited same function continuity at infinity [J]. Yichun Univ 2:45–46
7. Xinhua F (2004) Discriminant function in several ways consistent with continuous [J]. Changzhou Inst Technol (4):021–024
8. Yuanhua L (2004) Some of the discussion of the Uniform Continuity Function [J]. Hechi Teachers Coll (1):36–40

Chapter 86
A Popularized Inequality About Hausdorff Measure

Shanhui Sun, Jing Liu and Shaoyuan Xu

Abstract In this article, using the Helly theorem and other skills, we get a new inequality about Hausdorff measure. The result popularizes the relevant result of article (Falconer (1990) Fractal geometry—mathematical foundation and application. Wiley, New York).

Keywords Hausdorff measure · Hausdorff dimension · Helly theorem

86.1 Introduction

The Hausdorff dimension and Hausdorff measure are the most basic concepts in the study of fractals. To calculate or estimate their values for fractals is one of the most important problems. But in most cases, to calculate them, especially to calculate the Hausdorff measure, is very difficult. Till date, the computation of Hausdorff measure, even for this simplest class of fractals, is still difficult (see [1–3]). Especially for fractals with Hausdorff dimensions larger than 1, how to compute Hausdorff measure remains an open problem. Nevertheless, efforts have been made to estimate

S. Sun (✉) · J. Liu
College of Mathematics and Statistics, Suzhou University,
Suzhou 234000, Anhui, People's Republic of China
e-mail: sshh10304@163.com

J. Liu
e-mail: liujingfulliji@126.com

S. Xu
College of Mathematics, Gannan Normal University,
Ganzhou 341000, Jiangxi, People's Republic of China
e-mail: xushaoyuan@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_86,

the lower and upper bounds of their Hausdorff measure. Because of the above reasons, many authors devote themselves to study inequality about Hausdorff measure and gain fruitful achievements.

Falconer [4, 1] concluded some useful inequalities about Hausdorff measure. Zhou [5] gained an inequality named partial estimation principle. Luo, Zhou [6] gained an isodiametric inequality about Hausdorff measure. In this article, we study a popularized inequality about Hausdorff measure [7, 8]. First, the Hausdorff measure and dimension of the set $E \subset R^3$ are defined either by arbitrary cover or other types of cover of E. Then, using the famous Helly theorem and other skills, we gain the main Theorem about Hausdorff measure as follows:

Theorem 0.1 *Let E be a subset of R^3, then the following inequality*

$$H^s(E) \le B^s(E) \le \left(\sqrt{6}\Big/2\right)^s H^s(E) \tag{86.1}$$

had been obtained, where $s = \dim_H S$, $B^s(E)$ is defined in (86.8).

Finally, we prove that the constant $\sqrt{6}/2$ is the minimum and the best one in the inequality (86.1) about Hausdorff measure.

86.2 Hausdorff Measure and Hausdorff Dimension

Let U is a nonempty subset of R^n [9]. We define the diameter of U as

$$|U| = \sup\{\ |x-y| : x, y \in U\ \}. \tag{86.2}$$

If $E \subset \cup_i U_i$ and $0 < |U_i| \le \delta$ for each i, we say that $\{U_i\}$ is a $\delta-$ cover of E. Let E be a subset of R^n and let s be a non-negative number. For $\delta > 0$ define

$$H^s_\delta(E) = \inf\left\{\sum_{i=1}^{\infty} |U_i|^s :\ \{U_i\} \text{ is a cover of } E\right\}.\quad \delta \tag{86.3}$$

It is clear that $H^s_\delta(E)$ is increasing as δ decreases. Let $\delta \to 0$, mark [10]

$$H^s(E) = \lim_{\delta \to 0} H^s_\delta(E). \tag{86.4}$$

The limit may be 0 and ∞. We call $H^s(E)$ is the s dimensional Hausdorff measure.

For any subset E of R^n and $\delta < 1$, it is clear that H^s_δ is non-increasing as s increase. So H^s is non-increasing as s increases [11]. In fact, we can get a better conclusion:

$$\sum |U_i|^t \le \delta^{t-s} \sum |U_i|^s. \tag{86.5}$$

By the definition of Hausdorff measure, we get $H^s_\delta \le \delta^{t-s} H^s_\delta(E)$. Thus, there is a unique value, $\dim E$, called the Hausdorrf dimension such that

$$\begin{cases} H^s = \infty, & 0 \le s < \dim E; \\ H^s(E) = 0, & \dim E < s < \infty. \end{cases} \tag{86.6}$$

Mark

$$\dim_H E = \inf\{s: \ H^s(E) = 0\} = \sup\{s: \ H^s(E) = \infty\}. \tag{86.7}$$

If $s < \dim_H E$, then $H^s(E) < \infty$. If $s > \dim_H E$, then $H^s(E) = 0$. Mark $s = \dim_H E$, then $H^s(E)$ may be 0, ∞, or $0 < H^s(E) < \infty$.

Hausdorrf dimension has the following properties:

If $E \subset R^n$ is open set, then $\dim_H E = n$.

If $E \subset F$, then $\dim_H E = \dim_H F$.

If $F_1, F_2, F_3, \ldots$ is a list of sets, then $\dim_H \cup_{i=1}^{\infty} F_i = \sup_{1 \le i \le \infty}\{\dim_H F_i\}$.

If E is countable, then $\dim_H E = 0$.

We can define the Hausdorff measure and dimension of the set $E \subset R^3$ either by arbitrary cover or other types of cover of E. For example, we can define

$$B_\delta^s(E) = \inf\left\{\sum_{i=1}^{\infty} |B_i|^s \text{: } \{B_i\} \text{ is a ball of } E\right\}. \quad \delta \tag{86.8}$$

Then, we get a measure $B^s(E) = \lim_{\delta \to 0} B_\delta^s(E)$ and a dimension.

86.3 Lemma and Notation

A subset C of R^n is called convex, if $x \in C$, $y \in C$ and $0 < \lambda < 1, (1 - \lambda)$ $x + \lambda y \in C$.

Let C be a non-empty convex set of R^n, if C contains all the semi-straight lines which stars from points in C and has the direction D, we call C is forth and back in direction D. In other words, C is forth and back in direction $y(y \neq 0)$ if and only if for every $\lambda \ge 0$ and $x \in C$, $x + \lambda y \in C$.

Lemma 2.1 (Helly Theorem) [1] *Let* $\{C_i : i \in I\}$ *is a family nonempty closed convex set of* R^n, *where I is a arbitrary indexed set. Assume* C_i *have no common forth and back direction. If the subfamily constituded with* $n + 1$ (*no less than* $n + 1$) *selements of* $\{C_i : i \in I\}$ *has the nonempty intersection, then* $\cap_{i \in I} C_i$ *is nonempty.*

Remark 2.1 If one or more of C_i are bounded, the hypothesis of Helly theorem is satisfied.

In order to prove the follow Lemmas, we denote that

Symbol $O(O, 2r)$ expresses a closed circle with center O and radius r.

Symbol $\Omega(O, 2r)$ express a closed ball with center O and radius r.

Lemma 2.2 [3] *Assume AB is the longest side of triangle ABC, the length of AB is d. Then triangle ABC can be covered by a circle, the radius of which is* $\sqrt{3}/3d$.

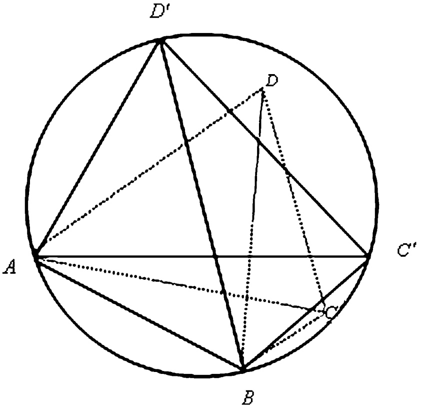


Fig. 86.1 Tetrahedron graph

Lemma 2.3 *Assume AB is the longest side of tetrahedron ABCD, the length of AB is d. Then tetrahedron ABCD can be covered by a ball, the radius of which is* $\sqrt{6}/4d$.

Proof Since AB is the longest side of tetrahedron $ABCD$, let C', D' and C, D be at the same side of AB, forming the regular tetrahedron $ABC'D'$ by using base triangle ABC'. Let Ω be the circumscried ball of regular tetrahedron $ABC'D'$. Note that $\angle AOB \geq \pi - \arccos 1/3$. Hence, when $\angle AOB > \pi - \arccos 1/3$, point C and D are in the interior of ball Fig. 86.1 Ω.

When $\angle AOB = \pi - \arccos 1/3$, point C and $D$are on the spherical surface of ball Ω.

It is easy to see that, the radius of ball Ω satisfies:

$$r^2 + r^2 - 2 \cdot r^2 \cos \angle AOB = d^2. \tag{86.9}$$

So tetrahedron $ABCD$ can be covered by a ball, the radius of which is $\sqrt{6}/4d$.

Proposition 2.4 *Assume U is a subset of* R^3, *the diameter of U is d. Then U can be cobered by a ball, the diameter of which is* $\sqrt{6}/2d$.

Proof It is easy to see that, U can be covered by a ball $\Omega(O, \sqrt{6}/2d)$, if and only if for every $P \in U$, there exists a common point $O \in \cap_{P \in U} \Omega(O, \sqrt{6}/2d)$.

For $\forall P_1, P_2, P_3, P_4 \in U$, since $|U| = d$, $|P_1P_2| \leq d$, $|P_2P_3| \leq d$, $|P_3P_4| \leq d$, $|P_{41}P_1| \leq d$. By the Lemma 5.3, there exists a ball $\Omega(O, 2r)$ such that $\forall P_1, P_2, P_3, P_4 \in \Omega(O, \sqrt{6}/2d)$, i.e., there exists a point O such that

$$|P_1O| \leq \frac{\sqrt{6}}{4} d |P_2O| \leq \frac{\sqrt{6}}{4} d |P_3O| \leq \frac{\sqrt{6}}{4} d |P_4O| \leq \frac{\sqrt{6}}{4} d \tag{86.10}$$

So, $O \in \Omega(P_1, \sqrt{6}/2d) \cap \cdots \cap \Omega(P_4, \sqrt{6}/2d)$.

For $\forall P_1, P_2, P_3 \in U$, since $|U| = d$, $|P_1P_2| \le d$, $|P_2P_3| \le d$, $|P_3P_1| \le d$, and $2/3\sqrt{3}d \le \sqrt{6}/2d$. By the Lemma 5.2, there exists a circle $O\left(O, 2/3\sqrt{3}d\right) \subset \Omega\left(O, \sqrt{6}/2d\right)$ such that

$$P_1, P_2, P_3 \in \mathrm{O}\left(O, \frac{2}{3}\sqrt{3}\,d\right) \subset \Omega\left(O, \frac{\sqrt{6}}{2}d\right) \tag{86.11}$$

i.e., there is exists a point O such that

$$|P_1O| \le \frac{\sqrt{3}}{3}d \le \frac{\sqrt{6}}{4}\ d|P_2O| \le \frac{\sqrt{3}}{3}d \le \frac{\sqrt{6}}{4}\ d|P_3O| \le \frac{\sqrt{3}}{3}d \le \frac{\sqrt{6}}{4}d \tag{86.12}$$

So, $O \in \Omega\left(P_1, \sqrt{6}/2d\right) \cap \cdots \cap \Omega\left(P_3, \sqrt{6}/2d\right)$.

For $\forall P_1, P_2 \in U$, since $d < 2/3\sqrt{3}d \le \sqrt{6}/2d$, so $\Omega\left(P_1, \sqrt{6}/2d\right) \cap \Omega\left(P_2, \sqrt{6}/2d\right) \neq \phi$.

By Helly theorem, there exists a common point $O \in \cap_{P\in U}\Omega\left(P, \sqrt{6}/2d\right)$. So, U can be cobered by a ball, the diameter of which is $\sqrt{6}/2d$.

Remark 2.2 When U is a regular tetrahedron with the side d, the diameter of the regular tetrhedron's circumscribed ball will be $\sqrt{6}/2d$. So the constant $\sqrt{6}/2d$ is the best.

86.4 Proof of the Theorem

Proof Since $\delta-$ball cover of E is also the cover in the definition of $H^s_\delta(E)$ [12], so

$$H^s_\delta(E) \le B^s_\delta(E). \tag{86.13}$$

Let $\delta \to 0$, we have

$$H^s(E) \le B^s(E). \tag{86.14}$$

If $\{U_i\}$ is a $\delta-$cover of E, by the Proposition 5.4, there exists a ball cover $\{B_i\}$ of E such that

$$|B_i| \le \frac{\sqrt{6}}{2}|U_i| \le \frac{\sqrt{6}}{2}\delta \tag{86.15}$$

where B_i is a ball which contains U_i to every i. So,

$$\sum |B_i|^s \le \sum \left(\frac{\sqrt{6}}{2}|U_i|\right)^s = \left(\frac{\sqrt{6}}{2}\right)^s \sum |U_i|^s. \tag{86.16}$$

By the definition of $H^s_\delta(E)$ and $B^s_\delta(E)$, we have

$$B^s_{\frac{\sqrt{6}}{2}\delta}(E) \le \left(\frac{\sqrt{6}}{2}\right)^s H^s_\delta(E). \tag{86.17}$$

Let $\delta \to 0$, we have

$$B^s(E) \le \left(\frac{\sqrt{6}}{2}\right)^s H^s(E). \tag{86.18}$$

Remark 3.1 From the above Psroposition, we can see that the dimensions defined by the two measures in (86.3) and (86.8) are the same.

Remark 3.2 Proposition 5.5 popularizes the relevant result of pages 32 and 33 in article [4].

Remark 3.3 From Remark 2, we can see that the constant $\sqrt{6}/2d$ is the best.

This article is Sponsored by the natural science research foundation of Suzhou University (No: 2009yzk26).

References

1. Falconer KJ (1997) Techniques in fractal geometry. Wiley, Chichester
2. Zhou Z, Feng L (2004) Twelve open problems on the exact value of the Hausdorff measure and on topological entropy: a brief survey of recent results. Nonlinearity 17:493–502
3. Zhou Z, Feng L (2009) Some problems on fractal geometry and topological dynamical systems. Ann Theory Appl 25:5–15
4. Falconer KJ (1990) Fractal geometry—mathematical foundation and application. Wiley, New York
5. Zhou Z (1998) Hausdorffmeasure of self-similar set the Koch curve. Sci China Ser A 41: 723–728
6. He W, Luo J, Zhou Z (2005) Hausdorff measure and isodiametric inequalities. Acta Math Sinica 48:939–946
7. Zhou Z, Feng L (2000) A new estimate of the Hausdorff measure of the Sierpinski gasket. Nonlinearity 13:479–491
8. Zhou Z, Feng L, Zhai C (2008) The structure of the self-similar sets—Hausdorff measure and theupper convex density. Science Press, Beijing
9. Jia B, Zhou Z, Zhu Z (2002) A lower estimate for the Hausdorff measure of the Sierpinski gasket. Nonlinearity 15:393–404
10. Zhai X, Jia B (2006) An inequality about the Hausdorff measure. Math Pract Theory (in Chinese) 36:379–382
11. Jia B (2007) Bounds of the Hausdorff measure of Sierpinski carpet. Anal Theory Appl 23: 1–15
12. Lay SR (1982) Convex sets and their applications [J]. Wiley, New York

Chapter 87
Multivariate Statistical Evaluation of Heavy Metals in Pond Sediments from a Rural Area in Sixian County, Northern Anhui Province, China

Linhua Sun and Dongsheng Xu

Abstract Heavy metal concentrations from pond sediments in a rural area in Sixian County, Northern Anhui Province, China had been determined by using X-ray fluorescence, and the calculation of enrich factor and index of geo-accumulation, as well as multivariate statistical analysis (including principle component analysis and cluster analysis) had been brought out to light. The results indicate that V, Cr, Cu and Zn of pond sediments are unpolluted, whereas copper is light pollution when normalize to soil environmental background value of China. Two sources of heavy metals have been obtained by using multivariate statistical analysis, including lithogenic (V and Cr) and anthropogenic (Cu), whereas Zn and Pb are mainly contributed from anthropogenic activities, with little contribution from the background. The distribution of Cu, Zn and Pb concentrations are similar to the distribution of towns and roads, also indicate an anthropogenic origin.

Keywords Heavy metals · Pond sediments · Correlation matrix · Principle component analysis · Cluster analysis

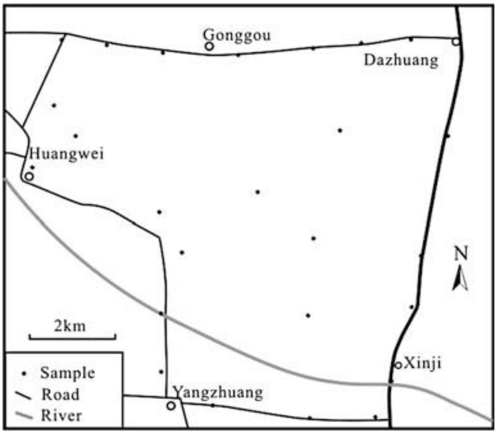
87.1 Introduction

Heavy metals have long been recognized for their hazards in human health, including their occurrence in water, air, dusts, soils and sediments, although they have only trace levels. With the evolution of technology (especially the method related to trace element analysis), as well as economics of the society, more and

L. Sun (✉) · D. Xu
School of Earth Science and Engineering,
Suzhou University, Suzhou 234000, China
e-mail: sunlinh@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_87,

Fig. 87.1 Sample location of study area (*long dash line* showing the inner area, whereas the *short dash line* showing the area affected by town or village)



more researches have been applied to heavy metals in cities or city-related environments recently. For instance, many studies have been taken place in city dusts [1], soils [2] and river sediments near the city [3]. However, the similar consideration paid toward rural area is still far less, with only few documents had been reported [4].

Pond water in rural area is important because it is always considered not only for farmland irrigation and laundry, but also drinking in some areas, especially in many rural areas in North China. It is thus pollution of pond water is not only dangerous for agriculture, but also potential risk for the health of residents.

Because sediments are good indicators for water safety [5], in this paper, we report the concentrations of heavy metals (V, Cr, Cu, Zn and Pb) in pond sediments in a typical rural area in Sixian County, Northern Anhui Province, China. We tend to (1) determine the degrees of heavy metal pollution and (2) discriminate the source of them by using multivariate statistical analysis (e.g. correlation, principle analysis and luster analysis).

87.2 Research Background and Methods

The study area is located in the north part of Sixian County in Northern Anhui Province, China, with its longitudes range from 117°45′00″ to 117°52′00″, and latitudes range from 33°37′00″ to 33°43′00″. There are four major towns in the area, including Huangwei, Yangzhuang, Xinji and Dazhuang, and a large village (Gonggou) located in the north (Fig. 87.1). The study area is surrounded by roads between four major towns with heavy traffic. There was a lot of industry in Huangwei, including iron, copper and chemical works, etc. but now some of them had been banned. It is also known that Huangwei is a high incidence area of cancer with two to three persons died per year.

Table 87.1 Heavy metal concentrations (mg/kg) of pond sediments in the study area

	N	Min	Max	Mean	Std. dev	BC	AEF	I_{geo-1}	I_{geo-2}
V	25	62	101	80.0	9.0	82.4	0.97	−0.63	−0.22
Cr	25	40	67	55.6	7.7	61.0	0.91	−0.72	−0.11
Cu	25	24	74	39.5	9.6	22.6	1.75	0.22	0.13
Zn	25	44	107	69.0	15.9	74.2	0.93	−0.69	0.06
Pb	25	20	34	25.6	4.9	26.0	0.98	−0.61	−0.23

Note BC means soil environmental background values of China [6], Igeo-1 means normalized to BC, whereas Igeo-2 means normalized to minimum in the Table

Only the surface sediments (0–5 cm) had been collected by using a homemade sediment sampler. A total of twenty-five pond sediment samples had been collected in the study area (Fig. 87.1). Sample collection was performed in December, 2010. Samples were firstly air-dried in natural condition, and the debris of animals and plants had been removed by hands. Then the samples were powdered to 200 meshes after parching for 24 h with 80°C in dryer.

Samples were condensed to be tablets by using a 30t condenser, and then analyzed by X-ray fluorescence (Explorer 9000SDD) in the Engineering Research Center of Coal Exploration, Anhui Province. National standard sediment sample (GBW07307) is analyzed simultaneously for calibration, and the relative standard derivation is less than 10%. Statistical analysis is performed by SPSS (version 11).

87.3 Results

87.3.1 Descriptive Statistics

Concentrations of heavy metals of pond sediments in the study area are listed in Table 87.1. As can be seen in the table, the mean values of the heavy metal contents arranged in the following decreasing order: V > Zn > Cr > Cu > Pb. Therefore, micronutrients such as V and Zn present higher levels in pond sediments, whereas Pb presents the lowest values. In comparision with the soil environmental background values of China (Table 87.1) [6], Cu of the pond sediments are only light pollution because they are little higher than the background values with AEF values (Average Enrich Factor = AC/BC, where AC and BC are the average concentration of sample and background, respectively) <2 [7], whereas V, Cr, Zn and Pb are considered as unpolluted as their concentrations are lower than or equivalent to the background values.

87.3.2 Correlation Between Variables

Table 87.2 is the correlation matrix of the six heavy metals. Because there are 25 samples in this study in total, the critical value of correlation coefficient can be

Table 87.2 Correlation matrix of heavy metals of pond sediments in the study area

	V	Cr	Cu	Zn	Pb
V	1				
Cr	0.71^{**}	1			
Cu	0.20	−0.01	1		
Zn	0.49^{*}	0.21	0.47^{*}	1	
Pb	0.43^{*}	0.49^{*}	0.38	0.76^{**}	1

Note ** and * mean correlation are significant at the 0.01 and 0.05 level, respectively

Table 87.3 Total variance explained and component matrixes for heavy metal contents

Component	Extraction sums of squared loadings			Rotation sums of squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
PC1	2.7	53.8	53.8	2.0	39.7	39.7
PC2	1.2	24.7	78.5	1.9	38.8	78.5
Component matrix				*Rotated component matrix*		
	PC1	PC2			PC1	PC2
V	0.77	−0.43		V	0.26	0.85
Cr	0.66	−0.67		Cr	0.01	0.94
Cu	0.51	0.66		Cu	0.83	−0.12
Zn	0.82	0.36		Zn	0.84	0.31
Pb	0.85	0.16		Pb	0.72	0.47

calculated as 0.40. It can be seen from the table that some of the heavy metals such as V, Cr, Zn and Pb are closely correlated, especially the relationship between V and Cr ($r = 0.71$), Zn and Pb ($r = 0.76$).

87.3.3 Principle Component Analysis

In order to obtain detailed statistical information, more robust statistical methods are applied to the data set to shed some light on the origin of the elements in this study. Varimax rotation is applied to minimize the number of variables and facilitate the interpretation of results.

The results of PCA for heavy metal concentrations are presented in Table 87.3. According to these results, the eigenvalues of the two extracted components are >1 both before and after the matrix rotation. As a consequence, heavy metals can be grouped into a two-component model that accounts for 78.5% of all the data variation. PC1 is responsible for 39.7% of the total variance and is represented by Cu, Zn and Pb, whereas PC2 explains 38.8% of the total variance and is mainly participated by V and Cr (Table 87.3, Fig. 87.2a).

Figure 87.2b shows the score plot of the two PCs, it can be concluded that the pond sediment samples can be subdivided into two groups: Group 1 is characterized by high PC2 scores, whereas Group 2 is characterized by low PC2 scores.

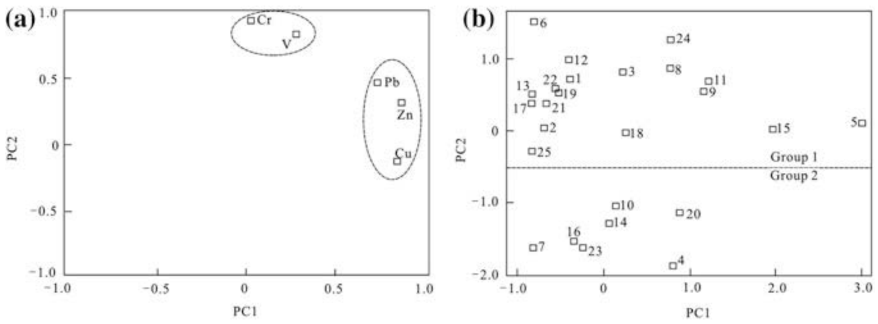


Fig. 87.2 **a** Principle component analysis loadings for the two rotated components. **b** Principle component scores for the pond sediments

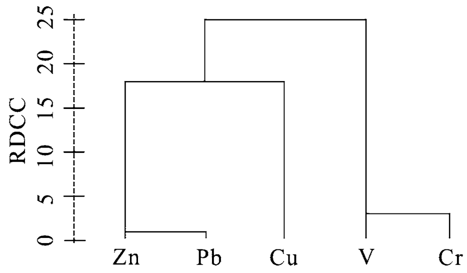


Fig. 87.3 Dendrogram of the hierarchical cluster analysis of variables (RDCC-Rescaled distance cluster combine)

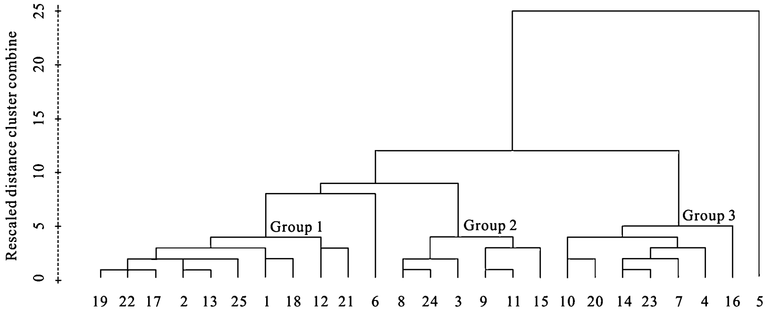


Fig. 87.4 Dendrogram of the hierarchical cluster analysis of cases

Moreover, samples (e.g. 5, 8, 9, 11, 15 and 24) show high PC1 scores in Fig. 87.2b are characterized by high lead concentrations (>30 mg/kg).

87.3.4 Cluster Analysis

The results of the hierarchical cluster analysis of variables and cases are given as dendrorgams in Figs. 87.3 and 87.4, respectively. As can be seen from Fig. 87.3, the variables are separated into two groups, V–Cr and Cu–Pb–Zn, the result is similar to the principle component analysis (Fig. 87.2a). Additionally, cluster analysis also shows that all of the samples can be subdivided into three groups (Group 1, high V and Cr; Group 2, high Cu, Zn and Pb; Group 3, low concentration of metals), similar but more significant than the result obtained from principle component analysis.

87.4 Discussions

87.4.1 Source of Heavy Metals

Principle analysis indicates that the sources of heavy metals in this study can be subdivided into two kinds, including lithogenic and anthropogenic. PC2 is considered to be a lithogenic component because the variability of the elements has similar concentrations with their background values (Table 87.1). The parent rock factor explains 38.8% of the total variance. This result suggests that the distribution of V and Cr has a lithogenic control because they have high positive loadings in PC2 (0.85 and 0.94, respectively). Additionally, part of Zn and Pb also has a lithogenic contribution because their loadings over PC2 are 0.31 and 0.47, respectively.

However, PC1 are considered to be an anthropogenic component because these metals (including Cu, Zn and Pb), especially their maximum concentrations are much higher than those of background values (Table 87.1). The anthropogenic factor explains 39.7% of the total variance. It can be seen from Table 87.3 that Cu, Zn and Pb have high positive loadings over PC1 (0.83, 0.84 and 0.72, respectively), indicating that they have an anthropogenic control. It is worth noting that Cu has high positive loading over PC1, but small negative loading over PC2, which indicates that copper sediments are anthropogenic in origin without lithogenic contribution, which is different from Zn and Pb, although they are mainly controlled by anthropogenic sources, the contribution of lithogenic origin is also observed in Table 87.3.

87.4.2 Environmental Risk Evaluation

The index of geo-accumulation (Igeo) enables the assessment of contamination by comparing the current and pre-industrial concentrations originally used with bottom sediments [8], and is computed using Igeo = log2C/(1.5*B) (where C is the measured concentration of the element in the sediment and B is the geochemical background value). The measurement of Igeo can be subdivided into five degrees (<0, unpolluted; 0–1, light pollution; 1–3, moderate pollution; 3–5, heavy pollution; >5, serious pollution). The calculated Igeo results are listed in Table 87.1. As seen from the Table, the average Igeo data normalized to soil environmental background values of China [6] indicating that V, Cr, Zn and Pb are unpolluted (Igeo-1 < 0) with only copper is light pollution (Igeo-1 $= 0.22$).

Although it is instructive to normalize metal concentrations of the sediments in this study to the soil environmental background values of China [6] for purposes of comparison with earlier studies. However, because different areas have different background values, here we choose the lowest value of metals in the study area as background values for calculation (Table 87.1). The results of average Igeo-2 indicate that the Cu and Zn are light pollution. However, the Igeo-2 of V, Cr, Cu, Zn and Pb are 0.12, 0.16, 1.04, 0.70 and 0.18, respectively, indicating that copper in some of the study area is moderate pollution, whereas other metals are light pollution with different degrees.

In summary, the pond sediments in the study area can be considered as unpolluted or only light polluted by heavy metals such as V, Cr, Zn and Pb, with the exception of Cu. However, as shown by the Igeo data normalized to the minimum concentration of metals in the study area, Pb, Zn and Cu pollution should be noticed because these metals are always considered to be related to motor vehicles or industries. Although the pollution of these metals is limited at the moment, they will become more serious with migration of industries into the rural areas, as well as increasing of motor vehicles in the areas.

87.5 Conclusions

Heavy metal concentrations of pond sediments in a rural area in Sixian County, Northern Anhui Province, China had been determined, and a series conclusions can be obtained as follows:

Calculation of enrich factor and Igeo based on soil environmental background of China indicate that V, Cr, Cu and Zn of pond sediments are unpolluted, whereas copper is light pollution;

Multivariate statistical analysis indicates that these metals have two sources: lithogenic (V and Cr) and anthropogenic (Cu), whereas Zn and Pb are mainly contributed from anthropogenic activities, to a lesser extent, the background;

Loading score of principle component analysis, as well as cluster analysis can subdivide all samples into three groups: high V and Cr, high Cu, Zn and Pb and low concentration of metals, the distribution of Cu, Zn and Pb concentrations are similar to those of towns and roads.

Acknowledgments This work was financially supported by the Key Project of Natural Science of the Department of Education, Anhui Province (KJ2011A261).

References

1. Al-Khlaifat A Temperature distribution in a microwave heated gas chromatographic packed column. Journal of Porous Media 38:6803–6812 (in press)
2. Zheng Y, Chen T, He J (2008) Abundance and community composition of Methanotrophs in a Chinese paddy soil under long-term fertilization practices. J Soils Sediments 8:51–58
3. Li LF, Zeng XB, Li GX, Mei XR (2007) Bending capability of foam aluminum sandwich beams. Acta Scien Circum 27:289–297
4. Pagotto C, Remy N, Legret M, Cloirec PL (2001) Pond sediments as indicator of traffic related lead pollution in rural area. Environ Technol 22:307–319
5. Wang J, Chen ZL, Wang C, Ye MW, Shen J, Nie ZL (2007) Heavy metals accumulation in river sediments of Chongming Island, Shanghai City, and its environmental risk. Chin J Appl Ecol 18:1518–1522
6. CEPA (Chinese Environmental Protection Administration) (1990) Elemental background values of soils in China. Environmental Science Press of China, Beijing
7. Hakanson L (1980) An ecological risk index for aquatic pollution control - a sedimentological approach. Water Res 14:975–1001
8. Muller G (1969) Subaerial cementation and subsequent dolomitization of lacustrine carbonate muds and sands from Paleo-Tuz Gölü ("Salt Lake"). J Geol 2:108–118

Chapter 88
Minimum Mean Square Error Estimator for Shearlet Coefficients Reconstruction

Chengzhi Deng

Abstract In this paper, a minimum mean square error (MMSE) estimator for shearlet coefficient reconstruction is proposed. The MMSE estimator, which is subband-adaptive, is inspired by a recent wavelet domain method ProbShink. Experimental results show the new estimator achieves better performance than several published method such as BLS-GSM and ProbShrink that is a state-of-the-art denoising technique.

Keywords Image denoising · Shearlet · MMSE estimator · Bayesian estimation

88.1 Introduction

During acquisition and transmission, images are often corrupted by additive noise that can always be modeled as Gaussian. The main goal of image denoising is to reduce the noise, while preserving the image features. Partial differential equations (PDE) and computational harmonic analysis (CHA) are two widely used classes of methods to achieve this goal.

Recently, Bayesian methods for nonlinear wavelet thresholding and nonlinear wavelet shrinkage estimators are widely studied by researchers. The Bayesian methods impose a prior distribution on wavelet coefficients and use a suitable Bayesian rule to derive a nonlinear mapping function for processing the noisy wavelet coefficients. The prior distribution which captures the distribution of

C. Deng (✉)
School of Information Engineering, Nanchang Institute
of Technology, Nanchang, China
e-mail: dengchengzhi@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_88,

wavelet coefficients is vital in Bayesian methods. Research results show that the distribution of wavelet coefficients for natural images is heavy-tailed. The popular prior for wavelet coefficient is a scale mixture of two Gaussian distribution [1], one Gaussian distribution and a point mass at zero [2], or one Laplacian distribution and a point mass at zero [3]. Various other probability modes were proposed in the literature with better capability in modeling the distribution of wavelet coefficients, e.g., the generalized Gaussian distribution [4], the Gaussian scale mixture model [5], α-stable distribution [6], Bessel K form model [7], elliptically contoured exponential mixture model [8], and mixture of Laplacian distribution [9].

However, the conventional wavelets cannot provide a 'sparse' representation for line or curve singularities. That is why they ignore the geometric properties of objects and do not exploit the regularity of edge curves. When wavelet is used to image denoising, it will lead to oscillatory artifacts along the edges. Within recent years, Demetrio and his collaborator developed a new geometric multiscale transform, named shearlet transform [10, 11]. The shearlet transform breaks the limitation of the wavelet transform and provides spares representation for the objects with singularities. In this paper, we proposed minimum mean squared error (MMSE) estimator for shearlet coefficients reconstruction. The core of the estimator is estimation of the probability that a given coefficient contains a significant noise-free component. In the end, the experiment is used to evaluate the performance of the estimator.

88.2 Shearlet Transform

The shearlet transform is a multiresolution representation with basis functions well-localized in space, frequency and orientation. It is generated by one single function which is dilated by a parabolic scaling and a shear matrix and translated in the time domain. The shearlet mother function is a composite wavelet that satisfies appropriate admissibility conditions [11].

The composite wavelet, recently introduced in [10], exhibits the geometric and multiscale properties by taking advantage of classical theory of affine systems. In dimension $n = 2$, the affine systems with composite dilations are defined as follows.

$$\Psi_{AB}(\psi) = \left\{ \psi_{f,k,l}(\xi) = |\det A|^{j/2} \psi(B^l A^j \xi - k) : j, l \in Z, k \in Z^2 \right\} \qquad (88.1)$$

where $\Psi \in L^2(R^2), A, B,$ are both 2×2 invertible matrices, $|\det B| = 1$. The elements of this system are called wavelet if Ψ_{AB} forms a Parseval frame for $L^2(R^2)$.

The dilations matrices A^j and B^l are associated with scale transformations and area-preserving geometric transformations, respectively. The above framework can be used to construct Parseval frames whose elements, in addition to ranging at various scale and location, also range at various orientations.

The shearlet is a special Parseval frame of composite wavelets in $L^2(R^2)$. These are collections of the form $\Psi_{AB}(\psi)$ where $A = A_0$ is the anisotropic dilation matrix and $B = B_0$ is the shear matrix, which are given by

$$A_0 = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

As described in [11], the shearlets provide a non-uniform angular covering of the frequency plane when restricted to the finite discrete setting for implementation. Thus, it is preferred to reformulate the shearlet transform with restrictions supported in the regions given by $D_0 = \left\{(\xi_1, \xi_2) \in \hat{R}^2 : |\xi_1| \geq 1/8, |\xi_2/\xi_1| \leq 1\right\}$ and $D_0 = \left\{(\xi_1, \xi_2) \in \hat{R}^2 : |\xi_2| \geq 1/8, |\xi_1/\xi_2| \leq 1\right\}$. For any $\xi = (\xi_1, \xi_2) \in \hat{R}^2$, $\xi_1 \neq 0$, define $\psi^{(0)}$ as

$$\hat{\psi}^{(0)}(\xi) = \hat{\psi}^{(0)}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_1)\hat{\psi}_2(\xi_2/\xi_1) \tag{88.2}$$

where $\hat{\psi}_1, \hat{\psi}_2 \in C^\infty(\hat{R})$, supp $\hat{\psi}_1 \subset [-1/2, 1/6] \cup [1/6, 1/2]$, and supp $\hat{\psi}_2 \subset [-1, 1]$. This implies that $\hat{\psi}^{(0)}$ is C^∞ and compactly supported with supp $\hat{\psi}^{(0)} \subset [-1/2, 1/2]^2$. In addition, assume $\sum_{j\geq 1} \left|\hat{\psi}_1(2^{-2j}\omega)\right|^2 = 1$ for $|\omega| \geq 1/8$. And for each $j \geq 0$, assume $\sum_{l=2^j}^{2^j} \left|\hat{\psi}_2(2^j\omega) + l\right|^2 = 1$ for $|\omega| \leq 1$. This implies

$$\sum_{j\geq 0} \sum_{l=-2^j}^{2^j} \left|\hat{\psi}^{(0)}(\xi A_0^{-j} B_0^{-l})\right|^2 + \sum_{j\geq 0} \sum_{l=-2^j}^{2^j} \left|\hat{\psi}_1(2^{-2j}\xi_1)\right|^2 \left|\hat{\psi}_2(2^j\xi_2/\xi_1 + l)\right|^2 = 1 \tag{88.3}$$

for $(\xi_1, \xi_2) \in D_0$. That is, the function $\{\hat{\psi}^{(0)}(\xi A_0^{-j} B_0^{-l})\}$ forms a tilling of D_0, and the collection

$$\left\{\psi^{(0)}_{j,k,l}(\xi) = 2^{3j/2}\psi^{(0)}(B_0^l A_0^j \xi - k) : j \geq 0, -2^j \leq l \leq 2^j, k \in Z^2\right\}$$

is a Parseval frame for $L^2(D_0)^\vee = \{f \in L^2(R^2) : \operatorname{supp} \hat{f} \subset D_0\}$. Similarly we can construct a Parseval frame for $L^2(D_0)^\vee$. Let

$$A_0 = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}, \quad B_0 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

and φ^1 be given by $\hat{\psi}^{(1)}(\xi) = \hat{\psi}^{(1)}(\xi_1, \xi_2) = \hat{\psi}_1(\xi_2)\hat{\psi}_2(\xi_1/\xi_2)$.

Then the collection $\left\{\psi^{(0)}_{j,k,l}(\xi) = 2^{3j/2}\psi^{(1)}(B_1^l A_1^j \xi - k) : j \geq 0, -2^j \leq l \leq 2^j - 1,\right.$ $\left. k \in Z^2\right\}$ is a Parseval frame for $L^2(D_1)^\vee$. Let $\varphi \in L^2(R^2)$ satisfy

$$|\hat{\varphi}(\xi)|^2 + \sum_{j\geq 0} \sum_{l=-2^j}^{2^j} \left|\hat{\psi}^{(0)}(\xi A_0^{-j} B_0^{-l})\right|^2 + \sum_{j\geq 0} \sum_{l=-2^j}^{2^j} \left|\hat{\psi}^{(1)}(\xi A_1^{-j} B_1^{-l})\right|^2 = 1$$

for $\xi \in \hat{R}^2$. Thus, we have the following:

Theorem 1 Let $\varphi_k(x) = \varphi(x-k)$ and $\psi^{(d)}_{j,k,l}(\xi) = 2^{3j/2}\psi^{(d)}\left(B^l_d A^j_d \xi - k\right)$, where φ, ψ are given as above. Then the collection of shearlet: $\{\varphi_k : k \in Z^2\} \cup \left\{\psi^{(d)}_{j,k,l} : j \geq 0, -2^j \leq l \leq 2^j - 1, k \in Z^2, d = 0,1\right\}$ is a Parseval frame for $L^2(R^2)$.

For each $f \in L^2(R^2)$, the shearlet transform is the mapping on $L^2(R^2)$ defined by

$$SH : f \to SH_\psi f(j,k,l) = \left\langle f, \psi^{(d)}_{j,k,l} \right\rangle,\ d = 0,1 \tag{88.4}$$

88.3 Proposed MMSE Estimator

Supposing we observe noisy coefficient $g = f + n$ where n is independent, zero-mean Gaussian noise. The aim of denoising is to estimate the noise-free coefficient f as accurately as possible according to some criteria. After shearlet transform, the problem can be formulated as $y = x + n$ where y is the noisy shearlet coefficients, x is the noise-free shearlet coefficient and n is noise.

For a special threshold $\tau \in R^2$, define the shrinkage function $S_\tau(x)$ to be x if $|x| \geq \tau$ and zero otherwise. Let H_1 denote signal of interest, which is equivalent to $|x| \geq \tau$. Otherwise, H_0 denotes no signal of interest. The MMSE estimate of x is given as follows.

$$\hat{x} = E(x|y) = E(x|y,H_1)P(H_1|y) + E(x|y,H_0)P(H_0|y) \tag{88.5}$$

Since H_0 refers to the absence of a signal of interest, we can set $E(x|y,H_0) = 0$. We further assume we can approximate $E(x|y,H_1)$ by y. These assumptions lead to

$$\hat{x} = P(H_1|y)y \tag{88.6}$$

Using Bayes rule, Eq. 88.6 can be rewritten as

$$\hat{x} = P(H_1|y)y = \mu\eta y/(1 + \mu h) \tag{88.7}$$

where $\mu = P(H_1)/P(H_0), \eta = p(y/H_1)/p(y/H_0)$, and the product $\mu\eta$ is called the generalized likelihood ratio.

For the assumed additive noise model where the noise coefficients and the noise-free coefficients are ,respectively, realizations of two stochastic processes that are statistically independent, the conditional probability $p(y/H_1)$ and $p(y/H_0)$ can be resulted from the following convolutions.

$$p(y/H_1) = \int_{-\infty}^{\infty} \phi(y - x; \sigma_n)p(x|H_1)\mathrm{d}x,$$

$$p(y/H_0) = \int_{-\infty}^{\infty} \phi(y - x; \sigma_n) p(x|H_0) \mathrm{d}x,$$

where $\phi(x; \sigma_n)$ denotes the zero-mean Gaussian density with the standard deviation σ_n. In this paper, we assume that the prior probability of true image shearlet coefficients is a generalized Laplacian distribution.

$P(H_1)$ and $P(H_0)$ are probabilities of the important signal and unimportant signal. Given the threshold τ, they can be estimated by $P(H_1) = 1 - \int_{-\tau}^{\tau} f(x)\mathrm{d}x$ and $P(H_0) = \int_{-\tau}^{\tau} f(x)\mathrm{d}x$. The factor μ can be derived as

$$\begin{aligned} \mu &= P(H_1)/P(H_0) = \left(1 - \int_{-\tau}^{\tau} f(x)\mathrm{d}x\right) \Bigg/ \int_{-\tau}^{\tau} f(x)\mathrm{d}x \\ &= \left(\Gamma\left(\frac{1}{v}\right) - \int_{0}^{(\lambda\tau)^v} t^{\frac{1}{v}-1} \mathrm{e}^{-t} \mathrm{d}t\right) \Bigg/ \int_{0}^{(\lambda\tau)^v} t^{\frac{1}{v}-1} \mathrm{e}^{-t} \mathrm{d}t \end{aligned} \qquad (88.8)$$

In this paper, the choice of the threshold τ is important. The threshold determines the actual signal of interest. The signal of interest should be chosen to minimize mean square error of the denoised image. Traditionally, the threshold is chosen according to the deviation σ_n of the noise, i.e., $\tau = \sigma_n$. Unfortunately, the shearlet transform is not normal preserving and, therefore, the variance of the noisy shearlet coefficients will depend on the shearlet index λ. In this paper, we set $\tau = c\sigma_n$, in which c is determined according to shearlet index λ. Letting S denote the discrete shearlet transform matrix, when the AWGN $n \sim N(0, \sigma_n^2)$, we have $S(n) \sim N(0, SS^T \sigma_n^2)$. We calculate an approximate value σ_λ^2 of the individual variance using Monte Carlo simulations where the diagonal elements of SS^T are simply estimated by evaluating the shearlet transform of a few standard white noise images.

88.4 Experimental Results

We now test our method on a set of 8-bit gray-scale test images (Lena and Barbara), of size 512 × 512 pixels, each contaminated with computer-generated additive Gaussian white noise at different variances. In the experiments, we compare it to other recently published denoising methods. The peak signal-to-noise ratio (PSNR) and visual quality are used as the assessing criteria of performance. Three other denoising algorithms are considered. In shearlet domain, the universal hard threshold (SHeT-H) [11] is considered. In wavelet domain, the new SURE approach (SURE) and Scale Mixtures of Gaussian (BLS-GSM) [5] are

Table 88.1 Comparison of PSNR values for different algorithms

	SHeT-H	BLS-GSM	SURE	Proposed method
Lena				
10	35.06	34.7	34.4	35.39
20	31.92	31.4	31.1	32.18
30	30.09	29.5	29.2	30.38
Barbara				
10	34.07	32.7	32.2	34.28
20	30.11	28.6	27.9	30.68
30	27.89	26.5	25.9	28.61

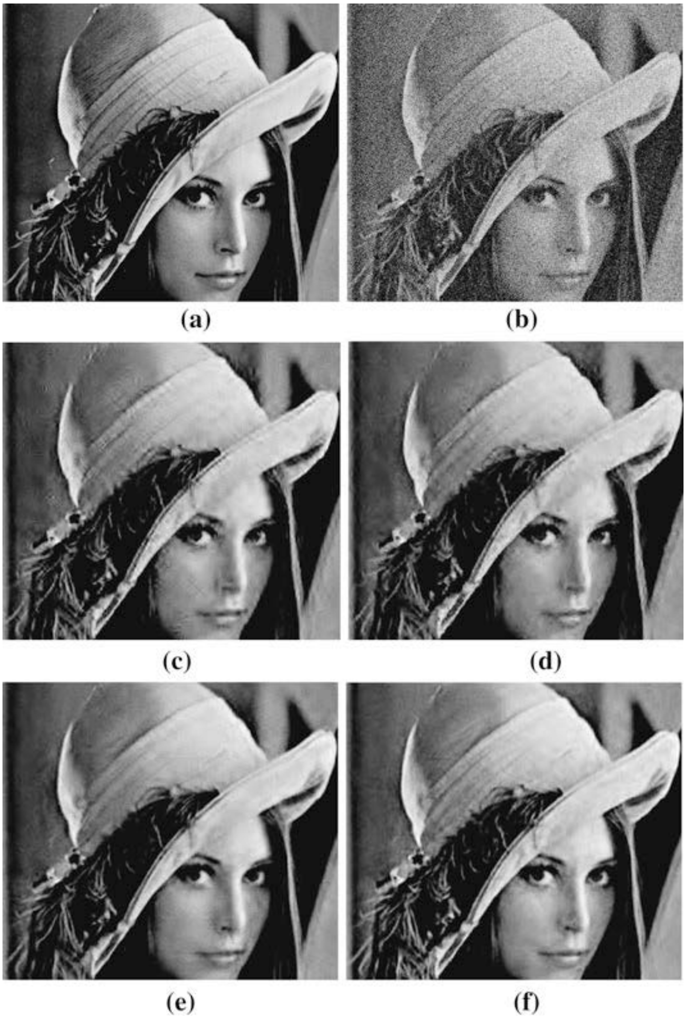


Fig. 88.1 Detail of the reconstruction results of Lena **a** Original, **b** Noisy image, **c** BLS-GSM, **d** SURE, **e** SHeT-H, **f** Proposed method

compared. In wavelet domain, a four-scale orthonormal wavelet with the Symlet-8 filter is used.

Table 88.1 shows the PSNR values (in dB) for the denoised images obtained by the three methods considered in this paper for three different noise standard deviations, 10, 20, and 30. In this table, the highest PSNR value is bolded. From Table 88.1, we can find that the proposed method performances better than almost all other methods. For example, at noise level 10, the proposed performance 0.7 dB better for Lena, 1.5 dB for Barbara than BLS-GSM. While at noise level 20, the proposed performance 0.8 dB better for Lena, 1.6 dB for Barbara than BLS-GSM.

Figure 88.1 shows the estimated images for each denoising method for Lena image with noise level 25. From the results we can find that the new proposed method yields the best denoising results. Due to the sparse representation of shearlet transform for curve singularities, the denoising methods in shearlet domain (SHeT-H and Proposed) show good performance for the edge preserving denoising. In all the figures, the proposed method achieves the best visual quality.

88.5 Conclusions

In this paper, a MMSE estimator for shearlet coefficients reconstruction is developed. The generalized Laplacian density is used to model the noise-free shearlet coefficients. The hard threshold rule is used to determine the interest of signal. The threshold is adaptively calculated using Monte Carlo simulations in shearlet domain. Experimental results have shown that the proposed method is superior to the other methods in terms of the PSNR as well as visual quality.

References

1. Crouse MS, Nowak RD, Baraniuk RG (1998) Wavelet-based statistical signal processing using hidden Markov models. IEEE Trans Signal Process 46:886–902
2. Abramovich F, Sapatinas T, Silverman B (1998) Wavelet thresholding via a Bayesian approach. J R Stat Soc B 60:725–749
3. Chang S, Yu B, Vetterli M (2000) Spatially adaptive wavelet thresholding with context modeling for image denoising. IEEE Trans Image Process 9:1532–1546
4. Pizurica A, Philips W (2006) Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising. IEEE Trans Image Process 15:654–665
5. Portilla J, Strela V, Wainwright M, Simoncelli EP (2003) Image denoising using scale mixtures of Gaussians in the wavelet domain. IEEE Trans Image Process 12:1338–1351
6. Boubchir L, Fadili JM (2006) A closed-form nonparametric Bayesian estimator in the wavelet domain of images using an approximate α-stable prior. Pattern Recognit Lett 17:1370–1382
7. Fadili JM, Boubchir L (2005) Analytical form for a Bayesian wavelet estimator of images using the Bessel k forms densities. IEEE Trans Image Process 14:323–329

8. Shi F, Selesnick IW (2007) An elliptically contoured exponential mixture model for wavelet based image denoising. Appl Comput Harmon Anal 23:131–151
9. Rabbani H, Vafadust M (2008) Image/video denoising based on a mixture of Laplace distributions with local parameters in multidimensional complex wavelet domain. Signal Process 88:158–173
10. Guo K, Lim W, Labate D, Weiss G, Wilson E (2006) Wavelets with composite dilations and their MRA properties. Appl Comput Harmon Anal 20:231–249
11. Easley G, Labate D, Lim W-Q (2008) Sparse directional image representations using the discrete shearlet transform. Appl Comput Harmon Anal 5:25–46

Chapter 89
Research on the Self-Assessment of English Writing Skills by Chinese University Students

Chunling Sun, Huilan Li, Guoping Feng and Lei Zhou

Abstract This paper attempted to study the reliability and validity of self-assessment used in evaluating the English writing skills of Chinese university students. Eighty one Chinese university students participated in this study. The results showed that the self-assessment of the English writings skills by Chinese university students was affected by their English proficiency; that is, students in the low-leveled group were not able to self-evaluate their writing skills appropriately, and they often overestimated their own and their classmates' writing ability, and the scores from self-assessment and peer-assessment of the high-leveled group were similar to those of teachers'. Hence, self-assessment and peer-assessment, as testing devices, were not suitable for students in the low-leveled group.

Keywords Self-assessment · Peer-assessment · Teacher-assessment

89.1 Introduction

From the perspective of students, foreign language test falls into two basic types: (1) The test is conducted by ways of self-reporting and self-assessment. This kind of test is a student-centered inner activity. (2) The test is conducted by examinations. This kind of test is an examiner-centered outside activity. At present, the latter was used widely, while the former was used and less studied .

C. Sun (✉) · H. Li · L. Zhou
Hebei United University, Tangshan 063000, Hebei, China
e-mail: sunchunling1969@163.com

G. Feng
No.1 Middle School of Kailuan, Tangshan 063000, Hebei, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_89,

With the increasing study of self-assessment, people have found that it has more advantages. Oscarson [1] listed the advantages of self-assessment in detail: (1) To promote learning. (2) To increase the degree of self-awareness of students. (3) To understand better of the purpose of their own learning. (4) To expand the scope of testing. (5) To reduce the burden on teachers. (6) To benefit self-learning of students after school. Some studies have also shown the effectiveness of self-assessment test and there existed significant correlation between self-assessment test, objective test and teacher-assessment test [2–4]. Of course, some studies have also showed its existing problems. For example, some studies found that there was no significant correlation between self-assessment scoring and tests; and students may not accurately self-evaluate their writing abilities. Low-leveled students tended to overestimate their language skills, while high-leveled students tended to underestimate their abilities.

The results of different studies made people use self-assessment test with care. Although studies have showed that students could accurately assess their English writing skills, but these studies were based on integrating writing and other language skills. There was less specific research on self-assessment of writing ability. The paper aimed to study the reliability and validity of self-assessment used in evaluating the English writing skills of the Chinese university students. This paper examined the following questions:

Was the students' self-assessment of their English writing ability valid and credible?

Whether students' self-assessment of their English writing ability was influenced by their English level?

Which was more reliable and effective, students' self-assessment or peer-assessment?

89.2 Research Design

Subjects. In this study, 81 students, who were English-majored sophomores from the Foreign Language Department (39) and law-majored sophomores from the Law Department (42), of which there were 47 boys and 34 girls,who took part in this study. English-majored students were classified as high-leveled group; while law-majored students as low-leveled group.

Research materials. High-leveled students were required to write an argumentative composition on the topic of "What, do you think, are the qualities contributing to success?", while low-leveled students were required to write a composition on the topic of "My view on Failure". Two groups of students must complete this task in 1 week.

Research procedure. After the students handed in their writing assignments, the teachers would make three copies of each student composition. First, one copy was sent back to the student himself, asking him to score his own composition. Before this scoring, every student was given a detailed scoring criteria, which was based primarily on three aspects: content, language and structure. To each part, the highest score was 5, the lowest was 0, and 0.5 as a rate of scoring. Each student's score was the sum of the scores of the above three aspects, the highest was 15 points. Students had 1 day to make this assessment. The second copy was given to their peer students at random. In this process, the teachers would check the assigned compositions in case the students should assess their own ones. Similarly, each student had 1 day to complete this assessment with the same criteria as mentioned before. The third copy was left to the teachers to make assessment. The criteria was also the same as before. High-leveled students' compositions were scored by two professional English teachers in teaching writing course, while low-leveled students' compositions were scored by two college English teachers. The final score of each student was the average of the total scores from the two teachers. At the same time, in order to further illustrate the reliability and validity of teachers' scoring, the scores from teachers would be compared with the writing scores and the total paper scores got by students in their final examination (scored by other teachers) respectively. Finally, the scoring results would be treated with statistical analysis in terms of reliability and validity.

89.3 Results and Discussions

Students' self-assessment scores, peer-assessment scores and teacher-assessment scores were all put into computer, and all the data were analyzed by SPSS software.

89.3.1 Difference Analysis to Indications in Low-Leveled Group

Table 89.1 shows the results of average points and variance analysis of indications in low-leveled group. It could be seen that self-assessment average points in low-leveled group were lower than that of the peer-assessment, higher than that of teacher-assessment; peer-assessment average points were higher than that of peer-assessment and teacher-assessment, respectively; while teacher-assessment points were lower than that of self-assessment and peer-assessment. Analysis of variance showed that there existed significant difference between self-assessment, peer-assessment and teacher-assessment in low-leveled group in terms of content, language and structure scoring, respectively. Scheffe test (Table 89.2) showed that there existed no significant difference between self-assessment and peer-assessment in low-leveled group in terms of content and language scoring, but there existed significant difference in structure scoring; no significant difference between self-assessment and teacher-assessment in low-leveled group in terms of language and structure scoring was seen, but there existed significant difference in content scoring; significant difference between peer-assessment and teacher-assessment in

Table 89.1 Results of average points and variance analysis of indications in low-leveled group

	N	Mean	Std. deviation	Sig.
Content				
Self-assessment	42	3.7976	0.9043	0.000
Peer-assessment	42	3.8333	0.7938	
Teacher-assessment	42	2.9048	0.6648	
Total point in content	126	3.5119	0.6240	
Language				
Self-assessment	42	3.2857	0.9183	0.000
Peer-assessment	42	3.4524	0.7715	
Teacher-assessment	42	3.0833	0.6336	
Total point in language	126	3.2738	0.5568	
Structure				
Self-assessment	42	3.1667	0.8239	0.000
Peer-assessment	42	3.5238	0.8900	
Teacher-assessment	42	2.9643	0.7358	
Total point in structure	126	3.2183	0.5539	
Total point				
Self-assessment	42	10.250	2.0099	0.000
Peer-assessment	42	10.809	2.0392	
Teacher-assessment	42	8.9524	1.5919	
Total point	126	10.033	1.5036	

Table 89.2 Scheffe test of indications in low-leveled group

Variable	(*I*)	(*J*)	Mean (*I* − *J*)	Std. deviation	Sig.
Content	Self-assessment	Peer-assessment	−0.036	0.702	0.743
	Self-assessment	Teacher-assessment	0.893	0.901	0.000
	Peer-assessment	Teacher-assessment	0.929	0.928	0.000
Language	Self-assessment	Peer-assessment	−0.167	0.908	0.241
	Self-assessment	Teacher-assessment	0.2027	1.05	0.218
	Peer-assessment	Teacher-assessment	0.369	0.898	0.011
Structure	Self-assessment	Peer-assessment	−0.357	0.983	0.023
	Self-assessment	Teacher-assessment	0.202	1.11	0.244
	Peer-assessment	Teacher-assessment	0.559	1.04	0.001
Total point	Self-assessment	Peer-assessment	0.238	0.418	0.002
	Self-assessment	Teacher-assessment	0.294	0.583	0.002
	Peer-assessment	Teacher-assessment	0.056	0.443	0.021

low-leveled group in terms of content, language and structure scoring was observed. As for the total points, there existed significant difference between peer-assessment, peer-assessment and teacher-assessment in low-leveled group respectively, which suggested self-assessment and peer-assessment in low-leveled group were not accurate, and low-leveled students tended to overestimate their own writing ability, and a great gap existed between their assessment and

Table 89.3 Results of average points and variance analysis of indications in high-leveled group

	N	Mean	Std. deviation	Sig.
Content				
Self-assessment	39	3.6026	0.6302	0.008
Peer-assessment	39	4.1538	0.7178	
Teacher-assessment	39	3.1667	0.7285	
Total point in content	117	3.6410	0.4511	
Language				
Self-assessment	39	3.4359	0.6196	0.000
Peer-assessment	39	3.9231	0.6543	
Teacher-assessment	39	3.3205	0.6929	
Total point in language	117	3.5598	0.3927	
Structure				
Self-assessment	39	3.5000	0.5257	0.002
Peer-assessment	39	3.9103	0.6676	
Teacher-assessment	39	3.3333	0.8138	
Total points in structure	117	3.5812	0.4238	
Total point				
Self-assessment	39	10.5385	1.1378	0.000
Peer-assessment	39	11.9872	1.24334	
Teacher-assessment	39	9.8205	1.5196	
Total point	117	10.782	0.8491	

teacher-assessment. In summary, students in low-leveled group could not make accurate judgments on their writing ability.

89.3.2 *Difference Analysis to Indications in High-Leveled Group*

Table 89.3 shows the results of average points and variance analysis of indications in high-leveled group. It could be seen that self-assessment average points in high-leveled group were lower than that of peer-assessment, higher than that of teacher-assessment; peer-assessment average points were higher than that of peer-assessment and teacher-assessment respectively; while teacher-assessment points were lower that that of self-assessment and peer-assessment. Analysis of variance showed that there existed significant difference between self-assessment, peer-assessment and teacher-assessment in high-leveled groups in terms of content, language and structure scoring. Scheffe test (Table 89.4) showed that there existed significant difference between self-assessment and peer-assessment in high-leveled group in terms of content, language and structure scoring; There were no significant differences between self-assessment and teacher-assessment in high-leveled group in terms of language, and structure scoring, but there existed significant difference in content scoring; also, significant difference was seen between peer-assessment and teacher-assessment in high-leveled group in terms of content, language and structure

Table 89.4 Scheffe test of indications in high-leveled group

Variable	(*I*)	(*J*)	Mean (*I* − *J*)	Std. deviation	Sig.
Content	Self-assessment	Peer-assessment	−0.551	0.826	0.000
	Self-assessment	Teacher-assessment	0.436	0.968	0.008
	Peer-assessment	Teacher-assessment	0.987	0.9428	0.000
Language	Self-assessment	Peer-assessment	−0.487	0.623	0.000
	Self-assessment	Teacher-assessment	0.115	1.042	0.493
	Peer-assessment	Teacher-assessment	0.603	1.008	0.001
Structure	Self-assessment	Peer-assessment	−0.411	0.794	0.003
	Self-assessment	Teacher-assessment	0.167	1.009	0.309
	Peer-assessment	Teacher-assessment	0.577	0.943	0.000
Total point	Self-assessment	Peer-assessment	0.081	0.633	0.428
	Self-assessment	Teacher-assessment	0.059	0.491	0.451
	Peer-assessment	Teacher-assessment	-0.021	0.497	0.790

scoring, respectively. As for the total points, there existed no significant difference between peer-assessment, peer-assessment and teacher-assessment in high-leveled group, which suggested the scores from self-assessment and peer-assessment in high-leveled group were similar to those of teacher-assessment.

89.4 Conclusions

The results showed that students' self-assessment of their English writing ability was influenced by their English proficiency. Students in low-leveled group could not make accurate assessment to their own and their classmates' writing, and they often overestimated their own writing ability. While students in high-leveled group could make accurate assessment in language and structure aspects, but could not do so in the content aspect. But from the perspective of total points, students' scoring in high-leveled groups was similar to that of teacher-assessment.

Thus, self-assessment and peer-assessment as testing devices should be used cautiously, and they could be properly used as an aid to teaching writing. If students are allowed to access their own composition before the teachers' correcting, they will have the opportunity to compare their own with the other's compositions so as to promote their learning.

References

1. Oscarson M (1989) Self-assessment of language proficiency, rationale and applications. Lang Test 6(1):1–13
2. Blanche P (1990) Using standardized achievement and oral proficiency tests for self-assessment purposes, the ELITLC study. Lang Test 7(2):202–229

3. Blue GM (2000) Self-assessment and defining learners' needs. In: Blue GM, Milton J, Saville J (eds) Assessing english for academic purposes. Peter Lang, Oxford, pp 237–255
4. Freeman MI (2000) A Self-evaluation instrument for the measurement of student proficiency levels. In: Blue GM, Milton J, Saville J (eds) Assessing english for academic purposes. Peter Lang, Oxford, pp 219–236

Chapter 90
Some Properties of *A*H* Over Weak Hopf Algebras

Yan Yan and Chunfeng Liu Bing Han

Abstract Let H be a weak Hopf algebra and A an H-module algebra. Using the properties of the trace function we describe some properties of $A*H$ over Weak Hopf Algebras.

Keywords Right twisted smash product algebra · Weak Hopf algebra · Trace function

90.1 Introduction

Weak Hopf algebras have been proposed by Bohm and Nill as a generalization of ordinary Hopf algebras in the following sense: the defining axioms are the same, but the multiplicativity of the counit and the comultiplicativity of the unit are replaced by weaker axions. Perhaps the easiest example of a weak Hopf algebra is a groupoids algebra, other examples are face algebras, quantum groupoids and generalized Kac algebra.

The initial motivation to study weak Hopf algebras was their connection with the theory of algebra extension, and another important application of weak Hopf algebras is that they provide a natural framework for the study of dynamical twists in Hopf algebras.

It turns out that many important properties of ordinary Hopf algebras have "weak" analogues. For example, using the theory of integrals for weak Hopf algebras developed in [1], which is essentially parallel to that of ordinary Hopf

Y. Yan (✉) · C. L. B. Han
College of Sciences, Hebei United University, Tangshan, Hebei, China
e-mail: yanjxky@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_90,

algebras, one can prove an analogue of Maschke's theorem [2, 3] for weak Hopf algebras and show that semisimple weak Hopf algebras are finite dimensional. But the structure of weak Hopf algebras is much more complicated than that of ordinary Hopf algebras, even in the semisimple case. For example, the antipode of a semisimple weak Hopf algebra over the field of complex number may have an infinite order.

In this paper, we mainly study the concept of the right twisted smash products over weak Hopf algebras and investigate their properties.

90.2 Preliminaries

For the foundations of weak Hopf algebra theory we refer the reader to [1]. Throughout this paper, k denotes a field. We use Sweedler's notation for the comultiplication: $\Delta(x) = x_1 \otimes x_2$.

Definition 2.1 Let K be a field and $H = <\mu, \eta, V, \varepsilon>$ be both an associative algebra over K and a coalgebra over K [4, 5]. If H satisfies the following conditions (1)–(3), it is caller a weak bialgebra. If it satisfies the following conditions (1)–(4), it is caller a weak Hopf algebra with a weak antipode S.

(1) $\Delta(xy) = \Delta(x) \otimes \Delta(y)$,

(2) $1_1 \otimes 1_2 \otimes 1_3 = 1_1 \otimes \widehat{1}_1 1_2 \otimes \overset{\frown}{1_2} = 1_1 \otimes 1_2 \widehat{1}_1 \otimes \overset{\frown}{1_2}$,

(3) $\varepsilon(xyz) = \varepsilon(xy_1)\varepsilon(y_2z) = \varepsilon(xy_2)\varepsilon(y_1z)$,

(4) $x_1S(x_2) = \varepsilon(1_1x)1_2,\ S(x_1)x_2 = 1_1\varepsilon(x1_2),\ S(x_1)x_2S(x_3) = S(x)$.

For any weak Hopf algebra H, the following conditions are equivalent:

(1) H is a Hopf algebra,

(2) $\Delta(1) = 1 \otimes 1$,

(3) $\varepsilon(xy) = \varepsilon(x)\varepsilon(y),\ \forall x, y \in H$

(4) $\varepsilon(1_1x)1_2 = 1_1\varepsilon(x1_2) = \varepsilon(x),\ \forall x \in H$.

Let H be a weak bialgebra, the linear maps $\Pi^L, \Pi^R : H \to H$ is defined by the formulas

$$\Pi^L(x) = \varepsilon(1_1x)1_2, \quad \Pi^R(x) = 1_1\varepsilon(x1_2)$$

and we denote their images by $H^L = \Pi^L(H)$ and $H^R = \Pi^R(H)$. Clealy, they are the subalgebras of H.

Bohm and Nill [1] proved that in a weak Hopf algebra H, for all $x, y \in H$, we have

$$\Pi^L(x) = \varepsilon(1_1x)1_2 = x_1S(x_2) = \varepsilon(S(x1_1))1_2 = S(1_1)\varepsilon(1_2x)$$

$$\Pi^R(x) = 1_1\varepsilon(x1_2) = S(x_1)x_2 = 1_1\varepsilon(1_2S(x)) = \varepsilon(x1_1)S(1_2)$$

And for all $x \in H$, we have the following relations

$$S\big(\Pi^L(x)\big) = \Pi^R(S(x)), \quad S(\Pi^R(x)) = \Pi^L(S(x))$$
$$x_1 \otimes x_2 S(x_3) = 1_1 x \otimes 1_2, \quad S(x_1)x_2 \otimes x_3 = 1_1 \otimes x 1_2$$
$$x_1 \otimes S(x_2)x_3 = x 1_1 \otimes S(1_2), \quad x_1 S(x_2) \otimes x_3 = S(1_1) \otimes 1_2 x$$

Moreover, the relation $x_1 S(x_2) x_3 = x$ also holds. In fact,

$$x_1 S(x_2) x_3 = \Pi^L(x_1) x_2 = \varepsilon(1_1 x_1) 1_2 x_2 = x$$

If H is a weak Hopf algebra, its weak antipode S is both an anti-multiplicative map and an anti-comultiplicative map, that is, for all $x, y \in H$ [6]

$$S(xy) = S(y)S(x),\ S(1) = 1,\ S(x)_1 \otimes S(x)_2 = S(x_2) \otimes S(x_1),\ \varepsilon(S(x)) = \varepsilon(x)$$

Definition 2.2 Let H be a weak Hopf algebra with antipode S, and A be an algebra. A is called an H-bimodule angebra if the following conditions hold:

A is an H-bimodule with the left H-module structure map "→" and with the right H-module structure map "←";

A is not only left H-module algebra with the left module action "→" but also right H-module algebra with the right module action "←"

Let H be a weak Hopf algebra. An algebra A is a (left) H-module algebra if A is a left

H-module via $x \otimes a \mapsto x \rightarrow a$ and

(1) $x \rightarrow \mathrm{ab} = (x_1 \rightarrow a)(x_2 \rightarrow b)$
(2) $x \rightarrow 1 = \Pi^L(x) \rightarrow 1$

An algebra A is a (left) H-comodule algebra if A is a left H-comodule via $\Delta_A : A \rightarrow H \otimes A$, $\Delta_A(a) = a_{-1} \otimes a_0$ and

(1) $\Delta_A(\mathrm{ab}) = a_{-1} b_{-1} \otimes a_0 b_0$
(2) $\Delta_A(1) = 1 \otimes 1$

A coalgebra A is a (left) H-module coalgebra if A is a left H-module via $x \otimes a \mapsto x \rightarrow a$ and

(1) $\Delta_A(x \rightarrow a) = (x_1 \rightarrow a) \otimes (x_2 \rightarrow a_2)$
(2) $\varepsilon_A(x \rightarrow a) = \varepsilon_H(x)\varepsilon_A(a)$

It is clear that a left H-module algebra is also a right H^*-comodule algebra if H is finite dimensional.

Let A be a left H-module algebra, set $A^H = \{a \in A \big| x \rightarrow a = \Pi^L(x) \rightarrow a, \forall x \in H.\}$, then A^H is a subalgebra of A, which is called the invariant subalgebra of H. In fact, for all $x \in H, s, t \in A^H$, we have

$$\begin{aligned}
x \rightarrow \mathrm{st} &= (x_1 \rightarrow s)(x_2 \rightarrow t)\\
&= (x_1 \rightarrow s)\left(\Pi^L(x_2) \rightarrow t\right)\\
&= (x_1 \rightarrow s)(\varepsilon(1_1 x_2)1_2 \rightarrow t)\\
&= \left(1^{'}_1 x_1 \rightarrow s\right)\left(\varepsilon\left(1_1 1^{'}_2 x_2\right)1_2 \rightarrow t\right)\\
&= (1_1 x_1 \rightarrow s)(\varepsilon(1_2 x_2)1_3 \rightarrow t)\\
&= (1_1 x_1 \rightarrow s)(1_2 \rightarrow t)\\
&= 1 \rightarrow ((x \rightarrow s)t)\\
&= (x \rightarrow s)t\\
&= \left(\Pi^L(x) \rightarrow s\right)t
\end{aligned}$$

While by the same method, we have

$$\begin{aligned}
\Pi^L(x) \rightarrow (\mathrm{st}) &= (\Pi^L(x)_1 \rightarrow s)(\Pi^L(x)_2 \rightarrow t)\\
&= (\Pi^L(x)_1 \rightarrow s)(\Pi^L\left(\Pi^L(x)_2\right) \rightarrow t)\\
&= (1_1 \Pi^L(x) \rightarrow s)(1_2 \rightarrow t)\\
&= 1 \rightarrow (= (\Pi^L(x) \rightarrow s)t)\\
&= (\Pi^L(x) \rightarrow s)t
\end{aligned}$$

So, $st \in A^H$ and A^H is a subalgera of H.

Now, we introduce the Sweedler's arrow notation. That is, for any finite dimensional weak Hopf algebra H, its dual vector space $H^* = \mathrm{Hom}_k(H, k)$ has also a weak Hopf algebra structure. And for $x \in H$ and $\phi \in H^*$, set

$$x \rightarrow \Phi = \Phi_1 < \Phi_2,\ x > ,\ \Phi \leftarrow x = < \Phi_1,\ x > \Phi_2$$
$$\Phi \rightarrow x = x_1 < \Phi,\ x_2 > ,\ x \leftarrow \Phi = < \Phi,\ x_1 > x_2$$

Then for all $y \in H$, we have $< x \rightarrow \Phi, y > \ = \ < \Phi, yx >$ and $< \Phi \leftarrow x, y > \ = \ < \Phi, xy >$.

Definition 2.3 Let A be an H-bimodule algebra. A right twisted weak smash product A*H is defined on the vector space $A \otimes H$. Define a multiplication $(a \otimes h)(b \otimes g) = a(h_1 \rightarrow b \leftarrow S(h_3)) \otimes h_2 g$ on tensor space $A \otimes H$, for all $a, b \in A, h, l \in H$. Let a*x denote the class of $a \otimes x$ in $A \otimes H$, the multiplication in A*H is given by the familiar formula$\overset{\wedge}{\underset{1}{1}} \rightarrow a \otimes \overset{\wedge}{\underset{2}{1}} h = a \otimes h, a \leftarrow S\left(\overset{\wedge}{\underset{2}{1}}\right) \otimes \overset{\wedge}{\underset{1}{1}} h = a \otimes h.$

90.3 References Some Properties of A*H

In this section, let H be a weak Hopf algebra with a bijective weak antipode S and A a H-module algebra [7, 8].

Definition 3.1 A left (right) integral in a weak Hopf algebra H is an element such that $l \in H (r \in H)$ such that $xl = \pi^L(x)l(rx = r\pi^R(x))$ for all $x \in H$. A left or right integral in a weak Hopf algebra H is called nondegenerate if it defines a non-degenerate functional on H^*. A left integral l is $\pi^L(l) = 1$ called normalized if $\pi^L(l) = 1$.

Similarly, a right integral r is normalized if $\pi^R(l) = 1$.

Lemma 3.2 *Let H be a finite dimensional weak Hopf algebra and l a non-zero left integral of H, and let A be a left H-module algebra. Then the map $\hat{t} : A \to A$ given by $\hat{t}(a) = t \to a$ is an A^H-bimodule map with values in A^H* [9].

Proof Let $a \in A, s \in A^H$, we have $\forall x \in H$

$$x \to (t \to a) = (\mathrm{xt}) \to a = (\Pi^L(x)t) \to a = \Pi^L(x) \to (t \to a)$$
$$\therefore \hat{t}(a) = t \to a \in A^H.$$

Next, we proof that $\hat{t} : A \to A$ is a bimodule map.

$$\because t \to (\mathrm{as}) = (t_1 \to a)(t_2 \to s) = (t_1 \to a)(\Pi^L(t_2) \to s)$$
$$= \hat{1} \to ((t \to a)s) = (t \to a)s$$
$$\therefore \hat{t}(\mathrm{as}) = \hat{t}(a)s$$
$$\because t \to (\mathrm{sa}) = (t_1 \to s)(t_2 \to a) = (\Pi^L(t_1) \to s)(\Pi^L(t_2) \to a)$$
$$= (\hat{1}_1 \to s)(\hat{1}_2 \to (t \to a)) = s(t \to a)$$
$$\therefore \hat{t}(\mathrm{sa}) = \mathrm{s}\hat{\mathrm{t}}(a)$$

Definition 3.3 The map $\hat{t} : A \to A^H$ as in Lemma 3.2 is called a (left) trace function for H on A [10].

Definition 3.4 A ring R is called semiprime if it has no non-zero nilpotent ideals.

Lemma 3.5 *Assume that A*H is semiprime, and choose $0 \neq t \in \int_H^l$. If I is any non-zero left or right H-stable ideal of A, then $\widehat{t}(I) \neq 0$.*

Proof If $\widehat{t}(I) = 0$, then tIt=0. Thus if I is a left ideal, then $J =$ It is a left ideal of $A*H$ such that $J^2 = 0$. Since $A*H$ is semiprime, $J = 0$ and thus $I = 0$ a contradiction. If I is a right ideal, the same argument works using $J = tI$.

Corollary 3.6 *Let H be a weak Hopf algebra with a bijective weak antipode S, and let A be a left H-module algebra. Assume there exists a normalized left (or right) integral r in H (i.e. H is semisimple artinian). Then if A is semisimple artinian, so is A*H.*

Theorem 3.7 *Let H be finite-dimensional acting on A and assume that $\widehat{t}$ is surjective. If A is left Noetherian, then so is A^H* [11].

Theorem 3.8

$$A^H \cong \mathrm{End}({}_{A*H}A)$$

Proof Suppose $\Phi{:}A^H \to \mathrm{End}({}_{A*H}A)$, and $\Phi(a) = a_r, a \in A^H, a_r$ is a left multiplicatine map of A, that is $\forall b \in A$, $a_r(b) = ba$. It is clear that Φ is a injection. $\forall \sigma \in \mathrm{End}({}_{A*H}A)$, $a \in A$, we have $(a \otimes \hat{1}) \to 1 = a\big(\hat{1}_1 \to 1 \leftarrow S(\hat{1}_2)\big) = a1_A = a$

So, $\sigma(a) = \sigma\big((a \otimes \hat{1}) \to 1\big) = \sigma(a \to 1) = a\sigma(1)$

That is, $\sigma = \sigma(1)$.

$\forall h \in H,\quad$ we have $h \to \sigma(1) = 1 \otimes h \to \quad \sigma(1) = \sigma(1 \otimes h \to 1) = \sigma(h \to 1) = \sigma(\pi^L(h) \to 1) = \pi^L(h)\sigma(1) \forall a, b \in A^H, \forall c \in A, \Phi(\mathrm{ab})(c) = c(\mathrm{ab}) = \mathrm{cab}, \Phi(b)\Phi(a)(c) = \Phi(b)(\mathrm{ca}) = \mathrm{cab}$

So, $\Phi(\mathrm{ab}) = \Phi(b)\Phi(a)$.

Theorem 3.9 *Let A be a left Noetherian algebra and H finite dimensional weak Hopf algebra. Assume that A is an H-module algebra such that the trace function* $\widehat{t} : A \to A^H$ *is surjective. If A is an k-affine algebra, then so is* A^H.

References

1. Bohm G, Nill F, Szlachanyi K (1999) Weak Hopf algebras I: integral theory and C* structure. J Algebra 221:385
2. Sweedler ME (1969) Hopf algebras. Berjamin, New York
3. Hirata K, Sugano K (1966) On semisimple extensions and separable extensions over non commutative rings. J Math Soc Japn 18(4):360–373
4. Cohen M, Fishman D (1986) Hopf algebra actions. J Algebra 10:363–379
5. Wang S, Li JQ (1998) On twisted smash products for bimodules algebras and the drinfel‘d double. Commun Algebra 26:2435–2444
6. Zheng N (2009) Smash biproduct over weak Hopf Algebras. Adv Math (05):003–006
7. Shi M (2008) The complexity of smash product. Adv Math (05):010–015
8. Jiao Z, Li J (2009) The L-R weak smash product over weak Hopf algebras. Journal of Henan Normal University(Natural Science) (01):002–006
9. Ling J (2009) Maschke-type therems for weak smash coproducts. J Math Res Expo (04): 028–033
10. Yin Y, Zhang M (2009) The structure theorem for weak Hopf Algebras. Adv Math (06): 011–016
11. Ju T (2010) Cocyclic module constructed by the right adjoint action of Hopf algebras. J Math (02):221–230

Chapter 91
MRM–BEM Method for Buckling Eigenvalue Problem and its Convergence Analysis

Zeng Fengxia, Cui Yuhuan and Wang Xinghua

Abstract The buckling eigenvalue problem as the background in this paper, we discussed the specific process of the numerical solution, which uses multiple reciprocity method and boundary element method to solve a class of elliptic boundary value problems. Firstly, express the integral expression of the boundary value problem in the closed area, and convert it into the corresponding MRM-boundary integral equation and MRM-boundary variational equation with the MRM–BEM method; followed with the corresponding error estimates; Finally, numerical example shows that the MRM–BEM method is quick and has fast convergence, high accuracy and so on.

Keywords Boundary value problem MRM–BEM · Boundary variational equation · Boundary integral equation

91.1 Introduction

Buckling eigenvalue problem in the analysis of the engineering structures stabilization has great significance and now becomes a universal attention subject in computational mechanics. In control equation of the problem contains Laplace operator and biharmonic operator, so we should introduce two series of the high-order fundamental solution sequences, to set up the replacement formula between each sequence and two sequences. Using replacement repeatedly, we can gain

Z. Fengxia (✉) · W. Xinghua
KaiLuan No.1 middle school, Tangshan, Hebei, China
e-mail: zeng_fengxia@126.com

C. Yuhuan
Qinggong College, Hebei United University, Tangshan, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_91,

MRM-boundary integral equation and MRM-boundary variational equation, and later the convergence of analysis of MRM method for buckling eigenvalue problem is given, so as to ensure the reliability of numerical solution method.

91.2 The Conventional Boundary Integral Equation

In this paper we consider the following boundary value problem

$$\begin{cases} \Delta^2 u - s\Delta u = 0 \quad x \in \Omega \cup \Omega' \\ u|_\Gamma = u_0, \quad \left.\dfrac{\partial u}{\partial n}\right|_\Gamma = g_0 \end{cases} \tag{91.1}$$

where $\Omega \in R^2$ is an opening region with smooth boundary, Γ is a piecewise smooth and closed curve, $\Omega' = R^2 \backslash \bar{\Omega}$, $s = k^2$, k is a real number. u_0, g_0 are known functions in the definition, n is the outer normal of Γ. For external problem, u at infinity must satisfy

$$u = \mathrm{O}(|x|^{-2}), \quad D^m u = \mathrm{O}(|x|^{-4}) \quad (m = 1, 2) \tag{91.2}$$

In the fixed boundary conditions, buckling eigenvalue problem can be stated as problem (91.1). By Ref. [1], we can conclude its fundamental solution

$$u^*(x,y) = 2\int_0^{|x-y|} \rho^{-1}\left(\int_0^{\rho} \tau K_0(k\tau)\mathrm{d}\tau\right)\mathrm{d}\rho$$

where $K_0(z)$ is an amendary Bessel function of the second kind with order zero $K_0(z) = \sum_{n=0}^{\infty} a_n z^{2n} \lg z + \sum_{n=1}^{\infty} b_n z^{2n}$, $a_0 = -1$, a_n, b_n are constants.

Theorem 1 *If* $u_0 \in H^{3/2}(\Gamma)$, $g_0 \in H^{1/2}(\Gamma)$, *then the internal and external boundary value problem of* (91.1) *has only solution in* $H^2(\Omega)$ *and* $W_0^2(\Omega')$ (*The proof of this theorem and the method of Ref.* [1] *are similar*).

Theorem 2 *The boundary integral expression of the solution in closed region* $\bar{\Omega}$ *of problem* (91.1) *is*

$$c(y)u(y) = -\int_\Gamma \left(u^*\frac{\partial \Delta u}{\partial n} - u\frac{\partial \Delta u^*}{\partial n} - \Delta u\frac{\partial u^*}{\partial n} + \Delta u^*\frac{\partial u}{\partial n}\right)\mathrm{d}s_x + s\int_\Gamma \left(u^*\frac{\partial u}{\partial n} - u\frac{\partial u^*}{\partial n}\right)\mathrm{d}s \tag{91.3}$$

Among them, when $y \in \Omega$, $c(y) = 1$, *expression* (91.3) *is Integral expression of the solution u in problem* (91.1). *When* $y \in \Gamma$, *if y is smooth point in the border,*

$c(y) = 1/2$; Else, $c(y) = \theta/2\pi$, $\theta = \theta(y)$ expresses outer angle of the left tangent and right tangent at point y, expression (91.3) is boundary integral equation of the problem (91.1). The proof of this theorem and the method of Ref. [2] are similar. First, multiplying $u^*(r)$ to both sides of the above formula, integrating on Ω, then using Green formula, it can be attested.

91.3 The Construction of MRM-Boundary Integral Equation

First, introducing two series of the high-order fundamental solution sequences

$$\begin{cases} u_j^*(r) = \dfrac{r^{2(2j+1)}}{2\pi 4^{2j+1}((2j+1)!)^2}\ln r - l(2j+1) \\ m_j^*(r) = \dfrac{r^{4j}}{2\pi 4^{2j}((2j)!)^2}\ln r - l(2j) \end{cases} \quad (j = 1, 2, 3, \ldots) \tag{91.4}$$

in which $r = |x - y|, l(0) = 0, l(m) = \sum_{k=1}^{m} 1/k$; It is very easy to verify that high-order fundamental solution sequences $u_j^*(r)$, $m_j^*(r)$ satisfy the following properties: $u_0^*(r) = r^2/8\pi \ln r$ and $m_0^*(r) = -1/2\pi \ln r$ respectively expresses the fundamental solution of biharmonic operator and Laplace operator, they satisfy

$$\Delta^2 u_0^*(x,y) = \delta(y-x), \quad \Delta m_0^*(x,y) = \delta(y-x)$$

High-order fundamental solution sequences $u_j^*(r)$, $m_j^*(r)$ satisfy the following reciprocity formula

$$\Delta^2 u_{j+1}^* = u_j^*, \Delta^2 m_{j+1}^* = m_j^*, \quad \Delta u_j^* = m_j^* (j = 0, 1, 2, \ldots) \tag{91.5}$$

By Ref. [3] the below theorem is existed:

Theorem 3 *Let Ω be a bounded region on plane and Γ be a smooth boundary curve, then the solution of problem (91.1) by MRM follows*

$$\begin{aligned} u(y) = & \sum_{j=0}^{\infty} s^{2j} \int_\Gamma -\left(u_j^*(r)\frac{\partial \Delta u(x)}{\partial n} - u(x)\frac{\partial \Delta u_j^*(r)}{\partial n} - \Delta u(x)\frac{\partial u_j^*(r)}{\partial n} + \Delta u_j^*(r)\frac{\partial u(x)}{\partial n}\right) \mathrm{d}s_x \\ & + \sum_{j=0}^{\infty} s^{2j+1} \int_\Gamma \left(u_j^*(r)\frac{\partial u(x)}{\partial n} - u(x)\frac{\partial u_j^*(r)}{\partial n}\right) \mathrm{d}s_x \\ & + \sum_{j=1}^{\infty} s^{2j-1} \int_\Gamma -\left(m_j^*(r)\frac{\partial \Delta u(x)}{\partial n} - u(x)\frac{\partial \Delta m_j^*(r)}{\partial n} - \Delta u(x)\frac{\partial m_j^*(r)}{\partial n} + \Delta m_j^*(r)\frac{\partial u(x)}{\partial n}\right) \mathrm{d}s_x \\ & + \sum_{j=1}^{\infty} s^{2j} \int_\Gamma \left(m_j^*(r)\frac{\partial u(x)}{\partial n} - u(x)\frac{\partial m_j^*(r)}{\partial n}\right) \mathrm{d}s_x \end{aligned} \tag{91.6}$$

The infinite term must be truncated in calculation, it yields the following error estimates of MRM–BEM iterative method [5].

Theorem 4 *Let Ω be bounded open region with smooth boundary and $u(y)$ be the solution of equation $\Delta^2 u - s\Delta u = 0$, mark the approximate solution by the truncation of n times as follows:*

$$\begin{aligned} u_n(y) = & \sum_{j=0}^{n} s^{2j} \int_\Gamma -\left(u_j^*(r) \frac{\partial \Delta u(x)}{\partial n} - u(x) \frac{\partial \Delta u_j^*(r)}{\partial n} - \Delta u(x) \frac{\partial u_j^*(r)}{\partial n} + \Delta u_j^*(r) \frac{\partial u(x)}{\partial n} \right) \mathrm{d}s_x \\ & + \sum_{j=0}^{n} s^{2j+1} \int_\Gamma \left(u_j^*(r) \frac{\partial u(x)}{\partial n} - u(x) \frac{\partial u_j^*(r)}{\partial n} \right) \mathrm{d}s_x \\ & + \sum_{j=1}^{n} s^{2j-1} \int_\Gamma -\left(m_j^*(r) \frac{\partial \Delta u(x)}{\partial n} - u(x) \frac{\partial \Delta m_j^*(r)}{\partial n} - \Delta u(x) \frac{\partial m_j^*(r)}{\partial n} + \Delta m_j^*(r) \frac{\partial u(x)}{\partial n} \right) \mathrm{d}s_x \\ & + \sum_{j=1}^{n} s^{2j} \int_\Gamma \left(m_j^*(r) \frac{\partial u(x)}{\partial n} - u(x) \frac{\partial m_j^*(r)}{\partial n} \right) \mathrm{d}s_x + s^{2n+1} \int_\Omega u(x) m_n^*(r) \mathrm{d}\Omega \end{aligned}$$

Then it has the following estimate formula

$$|u(y) - u_n(y)| \le \max_{y\in\Omega} |u(y)| \frac{s^{2n+1}}{2\pi e} \left(\frac{\mathrm{e}d^2}{4(2n)} \right)^{2n} \quad \forall y \in \Omega$$

The proof of the theorem can be in Ref. [4].

91.4 The Construction of MRM-Boundary Variational Equation

For convenience, let

$$q_j^*(r) = \begin{cases} s^{2j} u_j^*(r) + s^{2j-1} m_j^*(r) & j = 1, 2, \ldots \\ u_0^* & j = 0 \end{cases}$$

then expression (91.6) can be stated as follows:

$$\begin{aligned} u(y) = & \sum_{j=0}^{\infty} - \int_\Gamma \left(q_j^*(r) \frac{\partial \Delta u(x)}{\partial n} - u(x) \frac{\partial \Delta q_j^*(r)}{\partial n} - \Delta u(x) \frac{\partial q_j^*(r)}{\partial n} + \Delta q_j^*(r) \frac{\partial u(x)}{\partial n} \right) \mathrm{d}s_x \\ & + \sum_{j=0}^{\infty} s \int_\Gamma \left(q_j^*(r) \frac{\partial u(x)}{\partial n} - u(x) \frac{\partial q_j^*(r)}{\partial n} \right) \mathrm{d}s_x \end{aligned} \tag{91.7}$$

Theorem 5 *Let the boundary be smooth enough, if* $u \in c^{\infty}(\overline{\Omega}) \cap c^{\infty}(\Omega^{'}) \cap \Psi^{'}(R^2)$ *is the solution of the problem* (91.1), *u and* $\partial u/\partial n$ *are continuous in both side, then the integral expression on the whole plane of the Eq.* (91.1) *can be stated*:

$$u(y) = \sum_{j=0}^{\infty} \int_{\Gamma} -\left(q_j^*(r) \left[\frac{\partial \Delta u(x)}{\partial n} \right] - [u(x)] \frac{\partial \Delta q_j^*(r)}{\partial n} - [\Delta u(x)] \frac{\partial q_j^*(r)}{\partial n} + \Delta q_j^*(r) \left[\frac{\partial u(x)}{\partial n} \right] \right) \mathrm{d}s_x$$
$$+ \sum_{j=0}^{\infty} s \int_{\Gamma} \left(q_j^*(r) \left[\frac{\partial u(x)}{\partial n} \right] - [u(x)] \frac{\partial q_j^*(r)}{\partial n} \right) \mathrm{d}s_x$$

Let $\sigma(x) = [\partial \Delta u/\partial n]$, $\varphi(x) = [\Delta u]$, $\Psi^{'}(R^2)$ *is dropped function space. If* $y \in \Gamma$, *then the left stated by* $(\alpha u^+(y) + (2\pi - \alpha)u^-(y))/2\pi$, α *is an inner angle of the left tangent and right tangent on. If* Γ *is smooth, then* $\alpha = \pi$, u, $\partial u/\partial n$ *are continuous when across the border*, *namely* $u = \partial u/\partial n = 0$. *According to the above expression*, *the solution of the problem on the whole plane can be stated*:

$$u(y) = -\sum_{j=0}^{\infty} \int_{\Gamma} \left(\sigma(x) q_j^*(r) - \varphi(x) \frac{\partial q_j^*(r)}{\partial n_x} \right) \mathrm{d}s_x \quad y \in \Omega \cup \Omega' \tag{91.8}$$

Contacting the boundary conditions of the problem (91.1), we could gain the MRM—boundary integral equation:

$$u_0(y) = -\sum_{j=0}^{\infty} \int_{\Gamma} \left(\sigma(x) q_j^*(r) - \varphi(x) \frac{\partial q_j^*(r)}{\partial n_x} \right) \mathrm{d}s_x \quad \forall\, y \in \Gamma \tag{91.9}$$

$$g_0(y) = -\sum_{j=0}^{\infty} \int_{\Gamma} \left(\sigma(x) \frac{\partial q_j^*(r)}{\partial n_y} - \varphi(x) \frac{\partial^2 q_j^*(r)}{\partial n_y \partial n_x} \right) \mathrm{d}s_x \quad \forall\, y \in \Gamma \tag{91.10}$$

According to the Theorem 1, Theorem 5 and trace theorem, expression (91.9) and (91.10) define an isomorphism mapping on $V \to V'$ $(V = H^{3/2}(\Gamma) \times H^{1/2}(\Gamma),\ V' = H^{-3/2}(\Gamma) \times H^{-1/2}(\Gamma))$. Multiplying $\sigma'(y) \in H^{-3/2}(\Gamma)$ to both sides of the expression (91.9), and multiplying $\varphi'(y) \in H^{-1/2}(\Gamma)$ to both sides of the expression (91.10), integrating on Γ about y, then subtracting two expressions mutually, it can be gained:

$$\left.\begin{aligned}
&\begin{cases} \mathit{find}\ v=(\sigma,\phi)\in V,\ \mathit{satisfy}\\ b(v,v')=F(v'),\ \forall\, v'\in V,\ v'=(\sigma',\phi')\end{cases}\\
&b(v,v') = -\sum_{j=0}^{\infty}\left\{\int\limits_\Gamma\int\limits_\Gamma\left[\sigma(x)\sigma'(y)q_j^*(r)-\phi(x)\sigma'(y)\frac{\partial q_j^*(r)}{\partial n_x}\right]\mathrm{d}s_xs_y\right.\\
&\qquad\qquad\left.-\int\limits_\Gamma\int\limits_\Gamma\left[\sigma(x)\phi'(y)\frac{\partial q_j^*(r)}{\partial n_y}-\phi(x)\phi'(y)\frac{\partial^2 q_j^*(r)}{\partial n_y\partial n_x}\right]\mathrm{d}s_xs_y\right\}\\
&F(v')=\int\limits_\Gamma u_0\sigma'(y)\mathrm{d}s_y-\int\limits_\Gamma g_0\phi'(y)\mathrm{d}s_y
\end{aligned}\right\}\tag{91.11}$$

Theorem 6 *The variational equation* (91.11) in $H^{-3/2}(\Gamma)\times H^{-1/2}(\Gamma)$ *exists only one solution*

Proof We just need to prove compulsive of the bilinear forms b. The continuity, symmetry and the continuity of the linear functional in (91.11) are obvious.

$$\begin{aligned}
b((\sigma,\varphi),(\sigma,\varphi)) &= \int\limits_{R^2}\Delta u\Delta v\mathrm{d}\Omega + s\int\limits_{R^2}\nabla u\nabla v\mathrm{d}\Omega\\
&= |u|^2_{W_0^2(R^2)}+s|u|^2_{W_0^1(R^2)}\\
&\geq c\|u\|^2_{W_0^2(R^2)}\geq c\left(\|\sigma\|^2_{H^{-3/2},\Gamma}+\|\varphi\|^2_{H^{-1/2},\Gamma}\right)
\end{aligned}$$

The last inequality is decided by the continuity of linear functional $(\sigma,\varphi)\to u$, namely:

$$\left(\|\sigma\|^2_{H^{-3/2},\Gamma}+\|\varphi\|^2_{H^{-1/2},\Gamma}\right)^{1/2}=\sup_{v\in W_0^2(R^2)}\frac{\left|\int_{R^2}\Delta u\Delta v\mathrm{d}\Omega+s\int_{R^2}\nabla u\nabla v\mathrm{d}\Omega\right|}{\|v\|_{W_0^2(R^2)}}\leq M\|u\|_{W_0^2(R^2)}$$

Here M is a constant, according to the Lax–Milgram theorem, variational equation (91.11) has a unique solution on $H^{-3/2}(\Gamma)\times H^{-1/2}(\Gamma)$.

To sum up, the way to get the solving problem (91.1): by the variational equation (91.11) we can gain $(\sigma,\varphi)\in H^{-3/2}(\Gamma)\times H^{-1/2}(\Gamma)$, get it into expression (91.8), then $u\in M\subset W_0^2(R^2)$ is obtained.

Fig. 91.1 The value of w with $x = 0$

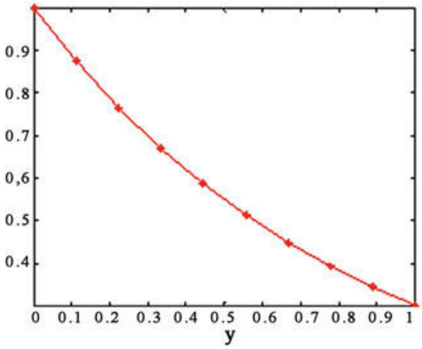
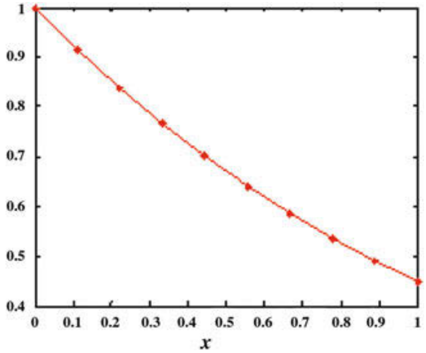


Fig. 91.2 The value of w with $y = 0$



91.5 Numerical Example

So as to validating the precision of the method in this paper, consider the following example:

$$\begin{cases} \Delta^2 w - 2\Delta w = 0 \quad \Omega = [0,1] \times [0,1] \\ w|_{\Gamma_1} = \mathrm{e}^{-x},\ w|_{\Gamma_2} = \mathrm{e}^{-(y+1)},\ w|_{\Gamma_3} = \mathrm{e}^{-(x+1)},\ w|_{\Gamma_4} = \mathrm{e}^{-y} \\ \left.\dfrac{\partial w}{\partial n}\right|_{\Gamma_1} = \mathrm{e}^{-x}, \quad \left.\dfrac{\partial w}{\partial n}\right|_{\Gamma_2} = -\mathrm{e}^{-(y+1)} \\ \left.\dfrac{\partial w}{\partial n}\right|_{\Gamma_3} = -\mathrm{e}^{-(x+1)}, \quad \left.\dfrac{\partial w}{\partial n}\right|_{\Gamma_4} = \mathrm{e}^{-y} \end{cases}$$

where

$$\Gamma_1 = \{(x,y) : x \in [0,1], y = 0\}, \quad \Gamma_2 = \{(x,y) : y \in [0,1], x = 1\}$$
$$\Gamma_3 = \{(x,y) : x \in [0,1], y = 1\}, \quad \Gamma_4 = \{(x,y) : y \in [0,1], x = 0\}$$

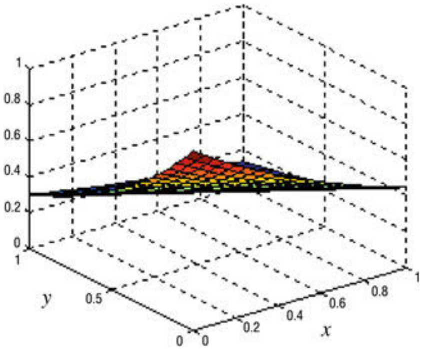


Fig. 91.3 The value of the whole area

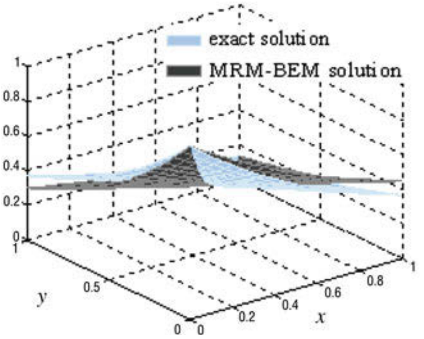


Fig. 91.4 Two values of the whole area

The accurate solution of the above problem is $w(x, y) = \mathrm{e}^{-(x+y)}$. We can gain the calculation results by MRM–BEM method. When $x = 0$, the change image of w with y is presented in Fig. 91.1; When $y = 0$, the change image of w with x is presented in Fig. 91.2; In Fig. 91.3 all values of the whole area $\Omega = [0, 1] \times [0, 1]$ are presented; Fig. 91.4 gives the results of the problem, one is gained by MRM–BEM method in this paper, the other is accurate solution, we can compare the two results effectively through the figure.

Acknowledgments This work is supported by the Scientific Research Fund Project (qz201006) of Qinggong College, Hebei United University.

References

1. Zhu J (1991) Boundary element analysis for elliptic boundary value problem[M]. Science Press, Beijing (in Chinese)
2. Ding FY, Zhang X, Ding R (1999) Boundary mixed variational inequality in friction problem. Appl Math Mech 20:201–210

3. Zeng F, Xu Y, Chen Y (2008) Multiple Reciprocity Method for Buckling Eigenvalue Problem and Its Convergence Analysis. 3rd International Conference on Innovative Computing, Information and Control, Dalian, China, (EI&ISTP search)
4. Ding R, Peng D, Wu Z (2004) Multiple reciprocity method for parabolic variational inequalities of second kind[J]. Chin Q Mech 25(2):239–247 (in Chinese)
5. Ding F, Ding R, Li B (2003) Multiple reciprocity method with two series of sequences of high-fundamental solution for thin plate bending [J]. Appl Math Mech 24(12):1267–1275 (in Chinese)

Chapter 92
Wavelet Boundary Element Method for Numerical Solution of Laplace Equation

Yuhuan Cui, Jingguo Qu, Yamian Peng and Qiuna Zhang

Abstract Using wavelet boundary element method to solve Laplace equation, the purpose is to solve the difficulties of singular integrals in natural boundary element method, reduce computation and improve accuracy. The basic idea is that the differential equation is matched by an equivalent variational problem after natural boundary element naturalization, then use wavelet interpolation method to discrete it, and obtain the stiffness matrix which has a unique advantage, so that we can greatly reduce the computation. In this paper, Shannon wavelet scaling functions are used as basis functions, applied to the natural boundary element method to solve the harmonic equation boundary value problems on the half-plane.

Keywords Natural boundary element method · Wavelet · Laplace equation

92.1 Natural Boundary Element Naturalization

Consider the upper half-plane area Ω with smooth boundary surface Γ, Laplace equation boundary value problems

$$\begin{cases} \Delta u = 0, & \text{in}\,\Omega \\ \frac{\partial u}{\partial n} = u_n, & \text{on}\,\Gamma \end{cases} \tag{92.1}$$

Y. Cui · J. Qu (✉) · Q. Zhang
Qinggong College, Hebei United University, Tangshan, China
e-mail: qujingguo@163.com

Y. Peng
College of Science, Hebei United University, Tangshan, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_92,

$\partial u/\partial n$ denotes the exterior normal derivative to boundary Γ, $u_n \in H^{-1/2}(\Gamma)$ is the given function for Γ. To ensure problem solvability, u_n must meet the compatibility condition

$$\int_{\Gamma} u_n \mathrm{d}s = 0$$

Let

$$\widehat{H}^{-1/2}(\Gamma) := \left\{ v_n \in H^{-1/2}(\Gamma) | \int_{\Gamma} v_n \mathrm{d}s = 0 \right\}$$

so $u_n \in \widehat{H}^{-1/2}(\Gamma)$.

The Laplace equation boundary value problems are naturalized as natural boundary integral equation

$$\frac{\partial u}{\partial n}(x,0) = -\frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{u(x',0)}{(x-x')^2} \mathrm{d}x'$$

That is

$$u_n(x) = \frac{\partial u}{\partial n}(x,0) = -\frac{1}{\pi} {}^{*} u_0(x) \tag{92.2}$$

* denotes convolution on the variable x, because the integral kernel of the convolution integral is strongly singular integral kernel, the integral should be understood in the sense of generalized functions for the Hadamard finite part integral.

Define the natural integral operator

$$\Re : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$$

The natural integral equation on the upper half-plane

$$\Re u_0(x) = u_n(x)$$

Defined bilinear form

$$D(u_0, v_0) = \int_R v_0 \Re u_0 \mathrm{d}x = -\iint_{R^2} \frac{v_0(x) u_0(y)}{\pi (x-y)^2} \mathrm{d}x \mathrm{d}y \tag{92.3}$$

Linear functional

$$F(v_0(x)) = \int_R v_0(x) u_n(x) \mathrm{d}x \tag{92.4}$$

Theorem 1 *On the half-plane bilinear form* $D(u_0, v_0)$ *derived by harmonic equations natural integral operator* $\Re$ *is positive definite continuous bilinear form of* $H^{1/2}(\Gamma)$.

From Theorem 1, it is easy to obtain the uniqueness of solution. Assumed that the u_1, u_2 are two solutions of the equation, we can see from the positive definite of $D(u_0, v_0)$:

$$D(u_1 - u_2, u_1 - u_2) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} |\xi||\hat{u}_1(\xi) - \hat{u}_2(\xi)|^2 \mathrm{d}\xi \geq 0$$

However, due to

$$D(u_1, u_1 - u_2) - D(u_2, u_1 - u_2) = 0$$

therefore, if the solution of equation exists, it will be unique.

92.2 Shannon Wavelet

Shannon wavelet [1–3] is a well-known wavelet image processing, having good Bureau of frequency. We first define the function$f$ of the Fourier transform [4]

$$\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(x)e^{-i\xi x}\mathrm{d}x$$

Shannon wavelet scaling function Fourier transform

$$\hat{\phi}(\xi) := \begin{cases} 1 & -\pi \leq \xi \leq \pi \\ 0 & \text{others} \end{cases}$$

That is

$$\varphi(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \hat{\varphi}(\xi)e^{-i\xi x}\mathrm{d}\xi = \frac{\sin(\pi t)}{\pi t}$$

So $\langle \varphi(t), \varphi(t-n) \rangle = \delta_{0n}$
Let $V_0 = \mathrm{Span}\{\varphi(t-n)\}_{n\in Z}$
This shows that $V_0 = \left\{ f \in L^2(R) \middle| \mathrm{supp}\hat{f} \subset [-\pi, \pi) \right\}$.

Theorem 2 (Shannon sampling theorem) *Let* $f \in L^2$, *and* $\mathrm{supp}\hat{f} \subset [-\pi, \pi)$, *then*

$$f(t) = \sum_n f(n)\frac{\sin[\pi(t-n)]}{\pi(t-n)}$$

Let

$$\varphi_{j,k}(t) := 2^{j/2}\varphi\left(2^{j}t-k\right), \quad j,k \in Z$$

$$V_j := \operatorname{span}\left\{\varphi_{j,k}(t)\right\}_{k\in Z} = \left\{f \in L^2 \ \operatorname{supp}\hat{f} \subset \left[-2^{j}\pi, 2^{j}\pi\right)\right\}$$

Then

$$\cdots \subseteq V_{-m} \subseteq \cdots \subseteq V_{-1} \subseteq V_0 \subseteq V_1 \subseteq \cdots \subseteq V_m \subseteq \cdots;$$

$\overline{\cup_m V_m} = L^2(R); \cap_m V_m = \{0\};$ $V_0 = \operatorname{span}\{\varphi(t-n)\}_{n\in Z}$, $\{\varphi(t-n)\}_{n\in Z}$ *is the standard orthogonal basis of* V_0.

So $\{V_m\}$ *is the multi-resolution analysis generated by scaling function* $\varphi(t)$.

92.3 Discretization

Suppose that $u_0^j(x)$ is approximation for the boundary value set $u_0(x)$ [5, 6], that

$$u_0(x) \approx u_0^j(x) = \sum_{k=-M}^{M-1} a_k\varphi_{j,k}(x) \tag{92.5}$$

Consider the approximate variational problem

$$D\left(u_0^j, v_0^j\right) = F\left(v_0^j\right) \tag{92.6}$$

Substitute (92.5) into (92.6), introduced

$$\sum_{k=-M}^{M-1} a_k D\left(\varphi_{j,k}(x), \varphi_{j,l}(x)\right) = F\left(\varphi_{j,l}(x)\right) \tag{92.7}$$

Among of them $-M \le l \le M$, let

$$A_j := \left(D\left(\varphi_{j,k}, \varphi_{j,l}\right)\right)_{2M\times 2M}$$

$$X := (a_{-M}, \ldots, a_{M-1})^T, \ b := (b_{-M}, \ldots, b_{M-1})^T$$

and

$$b_k := F\left(\varphi_{j,l}(x)\right)$$

Then (92.7) is written as $A_jX = b$ notice that

$$F\left[-\frac{1}{\pi x^2}\right] = |\xi|,$$

let

$$a^j_{k,l} = D\big(\varphi_{j,k}, \varphi_{j,l}\big) = -\iint\limits_{R^2} \frac{\varphi_{j,k}(y)\varphi_{j,l}(x)}{\pi(x-y)^2}\mathrm{d}x\mathrm{d}y$$

Among them

$$\varphi_{j,k}(y) = 2^{j/2}\varphi\big(2^j y - k\big)$$

$$\varphi_{j,l}(x) = 2^{j/2}\varphi\big(2^j x - l\big)$$

Let $2^j y = m$, $2^j x = n$, so

$$\begin{aligned} a^j_{k,l} &= -\iint\limits_{R^2} \frac{2^{j/2}\varphi(2^j y - k)2^{j/2}\varphi(2^j x - l)}{\pi\left(\frac{n}{2^j} - \frac{m}{2^j}\right)^2} 2^{-2j}\mathrm{d}x\mathrm{d}y \\ &= -2^j \iint\limits_{R^2} \frac{\varphi_{0,k}(m)\varphi_{0,l}(n)}{\pi(n-m)^2}\mathrm{d}m\mathrm{d}n = 2^j\left\langle \varphi_{0,l}, \varphi_{0,k} * \frac{-1}{\pi x^2}\right\rangle \end{aligned}$$

Fourier transform for $\varphi_{0,l}$, that $\hat{\varphi}_{0,l} = \int_R \varphi(n-l)e^{-i\omega n}\mathrm{d}n$

Let $p = n - l$, so $\hat{\varphi}_{0,l} = \int_R \varphi(p)e^{-i\omega p}e^{-i\omega l}\mathrm{d}p = \hat{\varphi}(\xi)e^{-i\omega l}$. The same reason $\hat{\varphi}_{0,k} = \overline{\hat{\varphi}(\xi)e^{-i\omega k}}$

The known $\langle f, g\rangle = \frac{1}{2\pi}\langle \hat{f}, \hat{g}\rangle$, $\langle f \hat{*} g\rangle = \hat{f}\cdot\hat{g}$ so

$$\begin{aligned} a^j_{k,l} &= \frac{2^j}{2\pi}\left\langle \hat{\varphi}_{0,l}, \hat{\varphi}_{0,k} * \frac{-1}{\pi x^2}\right\rangle = \frac{2^j}{2\pi}\left\langle \hat{\varphi}(\xi)e^{-i\omega l}, \overline{\hat{\varphi}(\xi)}e^{-i\omega k}\cdot|\xi|\right\rangle \\ &= \frac{2^j}{2\pi}\int\limits_{-\infty}^{+\infty} \hat{\varphi}(\xi)e^{-i\omega l}\cdot\overline{\hat{\varphi}(\xi)}e^{-i\omega k}\cdot|\xi|\mathrm{d}\xi = \frac{2^j}{2\pi}\int\limits_{-\pi}^{\pi} |\hat{\varphi}(\xi)|^2 e^{-i\omega(l-k)}\cdot|\xi|\mathrm{d}\xi \\ &= \frac{2^j}{2\pi}\left[\int\limits_0^{\pi} \xi e^{-i\omega(l-k)}\mathrm{d}\xi + \int\limits_0^{\pi} \xi e^{i\omega(l-k)}\mathrm{d}\xi\right] \\ &= \frac{2^j}{\pi}\int\limits_0^{\pi} \xi\cos(l-k)\mathrm{d}\xi \end{aligned}$$

when $k = l$, $a^j_{k,k} = 2^{j-1}\pi$. When $k \neq l$,

$$a^j_{k,l} = \frac{2^j}{\pi}\frac{(-1)^{k-l}-1}{(k-l)^2}$$

It is not difficult to find $a^j_{k,l} = a^j_{l,k}$, let $s = |k - l|$, so $a^j_s = a^j_{k,l}$, then

$$a_s^j = \begin{cases} 2^{j-1}\pi & s = 0 \\ \frac{2^j}{\pi}\frac{(-1)^s - 1}{s^2} & s = 1, 2, \ldots, 2M - 1 \end{cases}$$

Found from the above analysis, the singularity of the natural boundary integral equations has been eliminated, and there is a very simple and accurate formula for the element of stiffness matrix.

92.4 Numerical Examples

We use a group of discrete examples in the sub-base V_{2M}^j of space-scale space V_j to discretizate variational problem derived from the natural boundary integral equation, with the relaxation method for solving linear equations. The following table gives the error of exact solution and approximate solution L^2 [7].

Example 1 Neumann problems on the half-plane

$$\begin{cases} \Delta u = 0 \\ u_n = \frac{2x}{(1+x^2)^2} \end{cases}$$

Its exact solution is

$$u(x, y) = \frac{x}{x^2 + (y+1)^2}, \; u_0(x) = \frac{x}{x^2+1}.$$

Example 2 Neumann problems on the half-plane

$$\begin{cases} \Delta u = 0 \\ u_n = -\frac{1}{\pi(1+x^2)} + \delta(x) \end{cases}$$

Its exact solution is

$$u(x, y) = \frac{1}{4\pi}\ln\frac{x^2 + (y+1)^2}{x^2 + (y-1)^2} + y\delta(x), \; u_0(x) = 0.$$

Tables 92.1 and 92.2, respectively, are the calculations of Examples 1 and 2. There is no singularity interference in the calculation, which is consistent with the theoretical results, and there are no discrete calculation error and quadrature error, but only the iterative solution Error and bandwidth error work entirely decided by the condition number of coefficient matrix, bandwidth, and the nature of the function.

Table 92.1 Error analysis

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
Error	1.56334e-01	1.060848e-01	9.770571e-02	7.733752e-02

Table 92.2 Error analysis

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
Error	1.836327e-01	7.216365e-02	3.008517e-02	1.41527e-02

92.5 Conclusions

The Wavelet boundary element method effectively solves problems that occur when we solve the Laplace equation, through a specific example, validates the effective type of the method.

Acknowledgment This work is supported by the Scientific Research Fund Project (qz201006) of Qinggong College, Hebei United University.

References

1. Chen W, Zhang Y (2001) Shannon wavelet method of harmonic equation natural boundary element. Numer Math 2:125–134
2. Xu Y, Sun M (2008) Shannon wavelet chaotic neural network and its TSP problem. Control Theory Appl 25(3):574–577
3. Chen W (1998) Non-stationary wavelet with nature numerical solution of integral equations. Ph.D thesis, Zhongshan University
4. Tao R, Li BZ, Wang Y (2007) Spectral analysis and reconstruction for periodic nonuniformly sampled signals in frac-tional Fourier domain. IEEE Trans Signal Process 55(7):3541–3547
5. Qu J, Cui Y, Li B, Yang A, Zhang Y (2009) The research of fingerprint identification model. The third international workshop on matrix analysis and applications, vol 1, pp 75–81
6. Qu J, Cui Y, Yang A, Gong D, Feng L (2009) Computer simulation of fuzzy comprehensive evaluation and its application in water quality evaluation. 2009 international conference on test and measurement, vol 5–6, pp 342–345
7. Qu J, Cui Y, Wang X, Yang A (2009) Computer simulation of QR Algorithm and its application in the matrix Eigenvalue problem. 2009 international conference on test and measurement, vol 5–6, pp 338–341

Chapter 93
Asymptotically Almost Periodic Sequence Solutions for a Class of Difference Equations

Jingguo Qu, Yuhuan Cui, Aimin Yang and Huancheng Zhang

Abstract Differential equations is proposed when people study the variational equation of the orbits in discrete dynamical systems. It has important applications in many areas, including differential equations numerical methods, finite element method, control theory and computer science research. In recent years, many authors studied the properties of difference equations and applied them to discrete dynamical systems and differential equations with piecewise constant variables. This chapter uses the exponential dichotomy; the polytrophic exponent to give the sufficient conditions of differential equation asymptotically almost periodic sequence solution.

Keywords Differential equations · Asymptotically almost periodic sequence · Exponential dichotomy

93.1 Asymptotically Almost Periodic Sequence Solutions for a Class of Linear Difference Equations

In this chapter, we use $||\cdot||$ denote the Euclidean norm in the vector space R^d.

Consider the homogeneous linear difference equation

$$x(n+1) = A(n)x(n) \quad n \in Z \tag{93.1}$$

J. Qu (✉) · Y. Cui · H. Zhang
Qinggong College, Hebei United University, Tangshan, China
e-mail: qujingguo@163.com

A. Yang
College of Science, Hebei United University, Tangshan, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_93,

Here, for each one$n \in Z$, $A(n)$ is $d \times d$ order invertible matrix.

The non-homogeneous differential equation

$$x(n+1) = A(n)x(n) + h(n) \quad n \in Z \tag{93.2}$$

Here, for each one $n \in Z$, $A(n)$ is $d \times d$ order invertible matrix, $h(n) \in R^d$.

93.2 Introduction

We first give definitions of the exponential dichotomy and polytrophic exponent, and some lemmas needed to prove the main results

Definition 1 [1] Equation 93.1 is called the exponential dichotomy in $Z\,(Z^+, Z^-)$, when there is the projection $P\,(P^2 = P)$ and a constant $K > 0, \alpha > 0$ such that for any $m, n \in Z\,(Z^+, Z^-)$, there

$$|X(n)PX^{-1}(m)| \leq Ke^{-\alpha(n-m)}, \quad n \geq m$$

$$|X(n)(I-P)X^{-1}(m)| \leq Ke^{-\alpha(m-n)}, \quad m \geq n$$

Here, $X(n)$ is the based solution matrix of Eq. 93.1, which satisfy $X(0) = I$ (unit matrix).

Definition 2 [1] Equation 93.1 is called polytrophic exponent in Z, when it is that if there is projection P_1, P_2, P_3, $P_iP_j = 0, i \neq j$ and $P_1 + P_2 + P_3 = I$, as well as constant $K > 0, \alpha > 0$, such that for any $m, n \in Z$, there

$$|X(n)P_1X^{-1}(m)| \leq Ke^{-\alpha(n-m)}, \quad n \geq m$$

$$|X(n)P_2X^{-1}(m)| \leq Ke^{-\alpha(m-n)}, \quad m \geq n$$

and

$$|X(n)P_3X^{-1}(m)| \leq \begin{cases} Ke^{-\alpha(n-m)}, & 0 \leq m \leq n \\ Ke^{-\alpha(m-n)}, & n \leq m \leq 0 \end{cases}$$

Here, $X(n)$ is the based solution matrix of Eq. 93.1, which satisfy $X(0) = I$ (unit matrix).

Lemma 1 [1] *Suppose that linear differential equations (93.1) is the exponential dichotomy in Z, then there is no non-zero bounded solution of Eq. 93.1 in Z.*

If $\{h(n)\}_{n \in Z}$ is a bounded sequence, then the Eq. 93.2 has a unique bounded solution on $\{y(n)\}_{n \in Z}$, and

$$||y|| \leq K\frac{e^{\alpha}+1}{e^{\alpha}-1}||h|| \tag{93.3}$$

Lemma 2 [1] *Equation 93.1 is polytrophic exponent in* Z, *if and only if for every bounded sequence* $\{h(n)\}_{n\in Z}$, *Eq. 93.2 at least has one bounded solution.*

In fact, if the Green function

$$G(n,m) = \begin{cases} X(n)P_{-}X^{-1}(m), & m<0\leq n, \text{ or, } m<n<0 \\ -X(n)(I-P_{-})X^{-1}(m), & n<m<0 \\ X(n)P_{+}X^{-1}(m), & 0<m<n \\ -X(n)(I-P_{+})X^{-1}(m), & 0<n<m, \text{ or, } n<0<m \end{cases} \tag{93.4}$$

Among of them, $P_{+} = I - P_2,\ P_{-} = P$, so

$$y(n) = \sum_{s=-\infty}^{+\infty} G(n,s+1)h(s), \quad n \in Z \tag{93.5}$$

and

$$||y|| \leq 8K^2\frac{e^{\alpha}}{e^{\alpha}-1}||h|| \tag{93.6}$$

93.3 Main Results

Theorem 1 *If the homogeneous differential equation* (93.1) *is polytrophic exponent in* Z *corresponding the difference equation* (93.2), *then for any* $h = \{h(n)\}_{n\in Z} \in C_0(Z, R^d)$, *Eq. 93.2 has solutions* $y = \{y(n)\}_{n\in Z} \in C_0(Z, R^d)$, *in particular, if* $P_3 = 0$, *the solution* y *is the unique* [2].

Proof Let P_1, P_2, P_3 are the projections in the definition of polytrophic exponent, we define $P_{+} = I - P_2,\ P_{-} = P_1$ [3]. Take into account the defined Green functions in Eq. 93.4.

For a given $h = \{h(n)\}_{n\in Z} \in C_0(Z, R^d)$, defines the new sequence by (93.5), that is,

$$y(n) = \begin{cases} y_1(n), & n \in Z^{+} \\ y_2(n), & n \in Z^{-} \end{cases}$$

Here

$$\begin{aligned} y_1(n) = &\sum_{s=-\infty}^{-1} X(n)P_{-}X^{-1}(s+1)h(s) + \sum_{s=0}^{n-1} X(n)P_{+}X^{-1}(s+1)h(s) \\ &- \sum_{s=n}^{+\infty} X(n)(I-P_{+})X^{-1}(s+1)h(s) \end{aligned}$$

$$y_2(n) = \sum_{s=-\infty}^{n-1} X(n)P_{-}X^{-1}(s+1)h(s) + \sum_{s=n}^{-1} X(n)(I-P_{-})X^{-1}(s+1)h(s)$$
$$- \sum_{s=0}^{+\infty} X(n)(I-P_{+})X^{-1}(s+1)h(s)$$

First, $h = \{h(n)\}_{n \in Z} \in C_0(Z, R^d)$, then there is $0 < G < +\infty$, satisfies

$$||h(n)|| \leq G,\ \forall n \in Z \tag{93.7}$$

Thus, for $q \geq 0$, from (93.7)

$$\sum_{k=-q}^{q} ||h(k)|| e^{-\alpha|k|} \leq G \sum_{k=-q}^{q} e^{-\alpha|k|} = \frac{1+e^{\alpha}-2e^{-\alpha q}}{e^{\alpha}-1} G$$

Let $q \to +\infty$, denote $M = \frac{1+e^{\alpha}}{e^{\alpha}-1} G$, then

$$\sum_{k=-\infty}^{+\infty} ||h(k)|| e^{-\alpha|k|} = \lim_{q \to +\infty} \sum_{k=-q}^{q} ||h(k)|| e^{-\alpha|k|} \leq \frac{1+e^{\alpha}}{e^{\alpha}-1} G = M < +\infty \tag{93.8}$$

For a given $n \in Z^{+}$, from (93.8) and the property of polytrophic exponent,

$$\sum_{s=-\infty}^{-1} ||X(n)P_{-}X^{-1}(s+1)h(s)|| \leq K \sum_{s=-\infty}^{-1} ||h(s)|| e^{-\alpha(n-s-1)} \leq KMe^{-\alpha(n-1)} < +\infty$$

$$\sum_{s=n}^{+\infty} ||X(n)(I-P_{+})X^{-1}(s+1)h(s)|| \leq K \sum_{s=n}^{+\infty} ||h(s)|| e^{-\alpha(s+1-n)} \leq KMe^{\alpha(n-1)} < +\infty$$

This proved that $y_1(n)$ has the definition in Z^{+}. Similarly, for a given $n \in Z^{-}$, from (93.8) and the property of polytrophic exponent, we know that [4],

$$\sum_{s=-\infty}^{n-1} ||X(n)P_{-}X^{-1}(s+1)h(s)|| \leq K \sum_{s=-\infty}^{n-1} ||h(s)|| e^{-\alpha(n-s-1)} \leq KMe^{-\alpha(n-1)} < +\infty$$

$$\sum_{s=0}^{+\infty} ||X(n)(I-P_{+})X^{-1}(s+1)h(s)|| \leq K \sum_{s=0}^{+\infty} ||h(s)|| e^{-\alpha(s+1-n)} \leq KMe^{\alpha(n-1)} < +\infty$$

This proved that $y_2(n)$ has the definition in Z^{-}. So $y(n)$ has the definition in Z.
Easy to verify that y is the solution of differential equation (93.2).
The following proof $\lim_{|n| \to \infty} ||y(n)|| = 0$.
In fact, we prove respectively that $\lim_{n \to +\infty} ||y_1(n)|| = 0$, $\lim_{n \to -\infty} ||y_2(n)|| = 0$
First deal with the three items of $y_1(n)$. Noted that, here $n \in Z^{+}$.

The first item $P_- = P_1$, from (93.8) and the property of polytrophic exponent, then

$$\sum_{s=-\infty}^{-1} ||X(n)P_-X^{-1}(s+1)h(s)|| \leq K \sum_{s=-\infty}^{-1} e^{-\alpha(n-s-1)}||h(s)||$$
$$= Ke^{-\alpha(n-1)} \sum_{s=-\infty}^{-1} e^{-\alpha(-s)}||h(s)|| \leq KMe^{-\alpha(n-1)} \to 0, \quad (n \to +\infty)$$

The second item, $h \in C_0(Z)$, for $\forall \varepsilon > 0$, exists $N_1 > 0$, such that when $|n| \geq N_1$, there

$$||h(n)|| < \varepsilon \tag{93.9}$$

Because when $n \to +\infty$, $e^{-\alpha(n-1)} \to 0$, then for the above$\varepsilon > 0$, exists $N_2 > 0$, such that when $|n| \geq N_2$, there

$$e^{-\alpha|n-1|} < \varepsilon \tag{93.10}$$

$P_+ = I - P_2 = P_1 + P_3$, Take $N = \max\{N_1, N_2\}$, when $n \geq N$, from (93.7), (93.9), (93.10) and the property of polytrophic exponent, there

$$\sum_{s=0}^{n-1} ||X(n)P_+X^{-1}(s+1)h(s)|| \leq K \sum_{s=0}^{n-1} e^{-\alpha(n-s-1)}||h(s)||$$
$$= K\sum_{s=0}^{N-1} e^{-\alpha(n-s-1)}||h(s)|| + K\sum_{s=N}^{n-1} e^{-\alpha(n-s-1)}||h(s)|| = I_1 + I_2$$

Among them

$$I_1 = K\sum_{s=0}^{N-1} e^{-\alpha(n-s-1)}||h(s)|| \leq KGe^{-\alpha(n-1)} \sum_{s=0}^{N-1} e^{-\alpha(-s)} < \frac{KG(1-e^{\alpha N})}{1-e^{\alpha}}\varepsilon$$

$$I_2 = K\sum_{s=N}^{n-1} e^{-\alpha(n-s-1)}||h(s)|| \leq K\frac{1-e^{-\alpha(n-N)}}{1-e^{-\alpha}}||h(s)|| < \frac{K}{1-e^{-\alpha}}\varepsilon$$

This proves that

$$\sum_{s=0}^{n-1} ||X(n)P_+X^{-1}(s+1)h(s)|| \to 0, \quad (n \to +\infty)$$

the third item, $I - P_+ = P_2$, from (93.9) and the property of polytrophic exponent, when $n \geq N_1$, for $q \geq n$ there

$$\sum_{s=n}^{q} ||X(n)(I-P_+)X^{-1}(s+1)h(s)|| \leq K\sum_{s=n}^{q} e^{-\alpha(s+1-n)}||h(s)|| < K\frac{1-e^{-\alpha(q+1-n)}}{e^{\alpha}-1}\varepsilon$$

Let $q \to +\infty$, so

$$\sum_{s=n}^{+\infty} ||X(n)(I - P_+)X^{-1}(s+1)h(s)|| \leq \lim_{q \to +\infty} K \frac{1 - e^{-\alpha(q+1-n)}}{e^{\alpha} - 1} \varepsilon < \frac{K}{e^{\alpha} - 1} \varepsilon$$

Thus

$$\sum_{s=n}^{+\infty} ||X(n)(I - P_+)X^{-1}(s+1)h(s)|| \to 0, \quad (n \to +\infty)$$

So $\lim_{n \to +\infty} ||y_1(n)|| = 0$.

Secondly, deal with the three items of $y_2(n)$. Noted that, here $n \in Z^+$.

For any $\varepsilon > 0$, from Eq. 93.9, when $n \leq -N_1$, $||h(n)|| < \varepsilon$ and $P_- = P_1$, from the property of polytrophic exponent, when $n \leq -N_1$, $q < n$,

$$\sum_{s=q}^{n-1} ||X(n)P_- X^{-1}(s+1)h(s)|| \leq K \sum_{s=q}^{n-1} ||h(s)|| e^{-\alpha(n-s-1)} < K \frac{1 - e^{-\alpha(n-q)}}{1 - e^{-\alpha}} \varepsilon$$

Let $q \to -\infty$, so

$$\sum_{s=-\infty}^{n-1} ||X(n)P_- X^{-1}(s+1)h(s)|| \leq \lim_{q \to -\infty} K \frac{1 - e^{-\alpha(n-q)}}{1 - e^{-\alpha}} \varepsilon < \frac{K}{1 - e^{-\alpha}} \varepsilon$$

Thus

$$\sum_{s=-\infty}^{n-1} ||X(n)P_- X^{-1}(s+1)h(s)|| \to 0, \quad (n \to -\infty)$$

The second item, for $\forall \varepsilon > 0$, from Eq. 93.9, when $n \leq -N_1$, $||h(n)|| < \varepsilon$, from Eq. 93.10, when $n \leq -N_2$, $e^{-\alpha(1-n)} < \varepsilon$.

$I - P_- = P_2 - P_3$, let $N = \max\{N_1, N_2\}$, from (93.9) and the property of polytrophic exponent, when $n \leq -N$,

$$\sum_{s=n}^{-1} ||X(n)(I - P_-)X^{-1}(s+1)h(s)|| \leq K \sum_{s=n}^{-1} e^{-\alpha(s+1-n)} ||h(s)||$$
$$\leq K \sum_{s=n}^{-N-1} e^{-\alpha(s+1-n)} ||h(s)|| + K \sum_{s=-N}^{-1} e^{-\alpha(s+1-n)} ||h(s)|| = I_3 + I_4$$

Among them

$$I_3 = K \sum_{s=n}^{-N-1} e^{-\alpha(s+1-n)} ||h(s)|| \leq K \frac{e^{-\alpha} - e^{-\alpha(-N-n)}}{1 - e^{-\alpha}} \varepsilon \leq K \frac{e^{-\alpha}}{1 - e^{-\alpha}} \varepsilon$$

$$I_4 = K\sum_{s=-N}^{-1} e^{-\alpha(s+1-n)}||h(s)|| \le KGe^{-\alpha(1-n)}\sum_{s=-N}^{-1} e^{-\alpha s} < KG\frac{e^{\alpha}-e^{N\alpha}}{1-e^{\alpha}}\varepsilon$$

This proves that

$$\sum_{s=n}^{-1}||X(n)(I-P_-)X^{-1}(s+1)h(s)|| \to 0, \quad (n \to -\infty)$$

The third items, $I - P_+ = P_2$, from (93.8) and the property of polytrophic exponent, then

$$\sum_{s=0}^{+\infty}|X(n)(I-P_+)X^{-1}(s+1)h(s)| \le K\sum_{s=0}^{+\infty} e^{-\alpha(s+1-n)}||h(s)||$$
$$= Ke^{-\alpha(1-n)}\sum_{s=0}^{+\infty} e^{-\alpha s}||h(s)||$$

$$KMe^{-\alpha(1-n)} \to 0, \quad (n \to -\infty)$$

So $\lim_{n\to+\infty}||y_1(n)|| = 0$.

Finally, because $h \in C_0(Z, R^d)$, h is a bounded sequence, so inequality (93.6) was established, that y is bounded. If $P_3 = 0$, the differential equation is the exponential dichotomy, Lemma guarantees that y is the unique bounded solution of a differential equation (93.2), and the inequality (93.3) holds.

93.4 Conclusions

The chapter in the conditions of polytrophic exponent, gives the sufficient conditions of almost periodic sequence for a class of linear differences. Subsequently, in the conditions of exponential dichotomy, use the contraction mapping principle to discuss the problems of asymptotically almost periodic sequence solution existence corresponding nonlinear differential equations.

Acknowledgment This work is supported by the Scientific Research Fund Project (qy201004) of Qinggong College, Hebei United University.

References

1. Hong JL (1998) The almost periodic type difference equations [J]. Math Comput Model 28(12):21–31
2. Cui Y, Qu J, Yang A, Chang J, Li B (2009) Computer simulation of the grey system prediction method and its application in water quality prediction. 2009 international conference on test and measurement, vol 5–6, pp 323–326

3. Cui Y, Qu J, Chen W, Yang A (2009) Divide and conquer algorithm for computer simulation and application in the matrix eigenvalue problem. 2009 international conference on test and measurement, vol 5–6, pp 319–322
4. Cui Y, Qu J, Yang A, Li B (2009) The research of optimization model based on airplane seating problem. Fourth international conference on innovative computing, information and control

Chapter 94
Linear Method of the Power Spectrum Estimation Based on MATLAB Simulation

Zhongliang Sun, Xiaokan Wang and Sanci Guo

Abstract The power spectrum estimation is used to estimate power spectrum of the finite length data signal. It is very important for understanding a random signal or other applications which is one of the important research for digital signal processing. Study several typical methods of the power spectrum estimation and design a simulation signal of random sequence, and the spectrum is estimated with MATLAB. Simulation results show that if reasonable choosing the best power spectrum estimation function according to the actual work demand can reduce the spectrum analysis error of the actual signal. All that may provide a reference value for scientific research and engineering applications of power spectrum estimation based on MATLAB.

Keywords Power spectrum estimation · MATLAB · Correlation function · Periodogram

94.1 Introduction

In the practical work, the general processing signal can be divided into two kinds: regular signals (deterministic signals) and random signals (non-deterministic signals) [1]. Regular signal can be explicitly described by mathematic relationship,

Z. Sun (✉) · X. Wang · S. Guo
Henan Mechanical and Electrical Vocational Group,
Zhengzhou, Henan, China
e-mail: sunzhl2008@126.com

X. Wang
e-mail: wxkbbg@163.com

S. Guo
e-mail: guosanci@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_94,

while the random signal is generally impossible to describe with the clear mathematics relationship and also unable to forecast its future instantaneous precise value. So these randomness data can described with the probability and statistics average method, such as mean, standard deviation, variance, probability density function, probability distribution function, correlation function and power spectral density function.

The spectrum of signal is often much simpler expression than the original signal in the modern data analysis [2, 3]. Linear in the spectrum is easily expressed simply and can clearly see the changed characteristic of signal when interpreting the random signal processing or filtering effect. Therefore, for solving the random signal problem, we often need to do power spectral analysis and power spectrum estimation of random signal [4].

Because of the uncertainty of random signal, voltage spectrum is uncertain. But the stationary random signal with ergodic can get the determined correlation function, the dimension of its sequence is the power unit. The Fourier transform or z transform of correlation function is power spectral density function of random signal which is the abbreviation of power spectrum. Spectrum estimation is using the given N sample data to estimate power spectral density of the stationary random signal.

94.2 Common Power Spectral Estimation

94.2.1 Linear Estimation

Linear estimation method mainly includes: (1) the correlation function (BT) [5, 6]. This method could first get the correlation function of signal in the time domain, then it could calculate the Fourier transform of the correlation function with fast algorithm to get the estimation value of the power spectrum. When the time delay is much smaller than the data length, we can have a good estimation precision. (2) Periodogram. The method is a classical estimation method, the basic idea is that power spectrum can obtain directly from the Fourier transform of random signal series. The other two derivative estimation methods of periodogram are the average periodogram and the smoothing periodogram.

94.2.2 Nonlinear Estimation

Nonlinear estimation methods include: (1) maximum entropy estimation method (MEEM) [7]. It is an adaptive spectral analysis method, there is no fixed window function. When estimating the power spectrum of a frequency, it could automatically adjust the power spectrum of other frequencies to regulate the minimum disturbance. So it has the prediction error function of Wiener filter and has a higher resolution than the general linear spectral estimation method; (2) minimum cross-entropy method (MCEM) [8]. The resolution of this method is higher than MEEM, both of them have close relations. When the priori probability density function is uniform distribution, the minimum cross-entropy estimation is equivalent to the MEEM; (3) maximum likelihood method (MLM). The method is an unbiased minimum variance estimation method; (4) auto-regressive moving average model method (ARMA) [9]. The method is represented by zero and pole signal model parameters of spectral estimation methods with high resolution and greater flexibility.

94.3 Basic Principles of Linear Estimation on Power Spectrum

94.3.1 Correlation Function

The method consists of two steps: the first step, using N samples to estimate the value of random signal series's correlation function; the second step, calculating the estimation value of the power spectrum by the discrete Fourier transform of correlation function. Algorithm is as follows:

Correlation function valuations:

$$\widehat{r}_{xx}(m) = r_{Nx}(m) = \frac{1}{N}\sum_{n=0}^{N-|m|-1} x(n)x^*(n+m) \tag{94.1}$$

Power spectrum estimation:

$$\widehat{s}_x(\omega) = s_{Nx}(\omega) = \sum_{m=-|N-1|}^{N-1} r_{Nx}(m)e^{-jm\omega T} \tag{94.2}$$

Periodogram

The method could directly calculate the estimation value of the power spectrum from its own discrete Fourier transform of random signal samples. The algorithm is:

$$\widehat{s}_x(\omega) = s_{Nx}(\omega) = \frac{1}{N}\sum_{m=0}^{N-1} x(m)e^{-jm\omega T}\sum_{n=0}^{N-1} x(n)e^{-jn\omega T} = \frac{1}{N}X(\omega)X^*(\omega) \tag{94.3}$$

The above equation is called periodogram of the N length steady random signal series which is usually written as:

$$I_N(\omega) = \frac{1}{N}|X(\omega)|^2 \tag{94.4}$$

Theory has proved that the results of the periodogram spectrum estimation and the BT spectrum were the same, and belong to biased and non-uniform estimation.

94.3.2 Average Periodogram

The basic idea of the method is that random series *x* (*n*) is divided into *K* sections, and the length of each section is *L*, so *N* = *KL*. The *i* series can be expressed as:

$$x_i(n) = x(n + (i-1)L), \quad 0 \leq N \leq L-1,\ i = 1, 2, \ldots, K \tag{94.5}$$

Its periodogram is:

$$I_i(\omega) = \frac{1}{L}\left|\sum_{n=0}^{L-i} x_i(n)e^{-jn\omega T}\right|^2, \quad 1 \leq i \leq K \tag{94.6}$$

Therefore, the estimator of the power spectrum is defined as:

$$\hat{s}_x(\omega) = s_{Nx}(\omega) = \frac{1}{K}\sum_{i=1}^{K} I_i(\omega) \tag{94.7}$$

Theory study proved that the power spectrum estimation of average periodogram is biased and asymptotic consistent estimation.

94.3.3 The Smoothing Periodogram

The *N* length sequence of real stationary random signal *x* (*n*) is divided into *K* sections, the length of each section is *L*, so *N* = *KL*. Then,

$$x_i(n) = x(n + (i-1)L), \qquad 0 \leq N \leq L-1, \quad i = 1, 2, \ldots, K$$

We could calculate the periodogram of each segment sequence by using the following formula which has weighted with the window function *w* (*n*).

$$I_i(\omega) = \frac{1}{LU}\left|\sum_{n=0}^{L-i} x_i(n)\omega(n)e^{-jn\omega T}\right|^2, \quad 1 \leq i \leq K \tag{94.8}$$

where, *U* is the energy of the window sequence,

$$U = \frac{1}{L}\sum_{n=0}^{L-i} \omega^2(n) \tag{94.9}$$

The estimation value of the power spectrum is:

$$\widehat{s}_x(\omega) = s_{Nx}(\omega) = \frac{1}{K}\sum_{i=1}^{K} I_i(\omega) \tag{94.10}$$

94.4 Experimental Simulation

94.4.1 MATLAB Introduction

MATLAB was launched by the MathWorks company in 1982 which is a high-performance numerical computation and visualization software, so it was often called matrix laboratory. It integrated numerical analysis, matrix computation, signal processing and graphic display and it constitutes a convenient and friendly user-interface environment. The introduction of MATLAB was paid widespread attention by experts and scholars in various fields. Its powerful extension function expansion provided the foundation for all fields's application.

94.4.2 Simulation Results

In this simulation experiment, the different frequency of sinusoidal signals and Gaussian white noise composed of the random sequence. Figure 94.1 is the realization curves of the four simulation algorithms; Fig. 94.2 is the simulation curves of the different L values by the average periodogram and smoothing periodogram. The random sequence expression of Figs. 94.1 and 94.2 is:

$$\begin{aligned} x(n) = {} & \sqrt{20}\sin(0.1\times 2\pi n) + \sqrt{20}\sin(0.13\times 2\pi n) + \sqrt{5}\sin(0.3\times 2\pi n) \\ & + \sqrt{5}\sin(0.35\times 2\pi n) \end{aligned} \tag{94.11}$$

The normalized frequency of sinusoidal signal are 0.1, 0.13, 0.3 and 0.35, respectively, while the length of random sequence $N = 256$. In Fig. 94.1, the random sequence of the average periodogram and smoothing periodogram are divided into $K = 4$ sections, the length of each section is $L = 64$. Suppose $L = 32$ for the upper part and $L = 128$ for the lower part in the Fig. 94.2. The window function used is Hanning window in the smoothing periodogram.

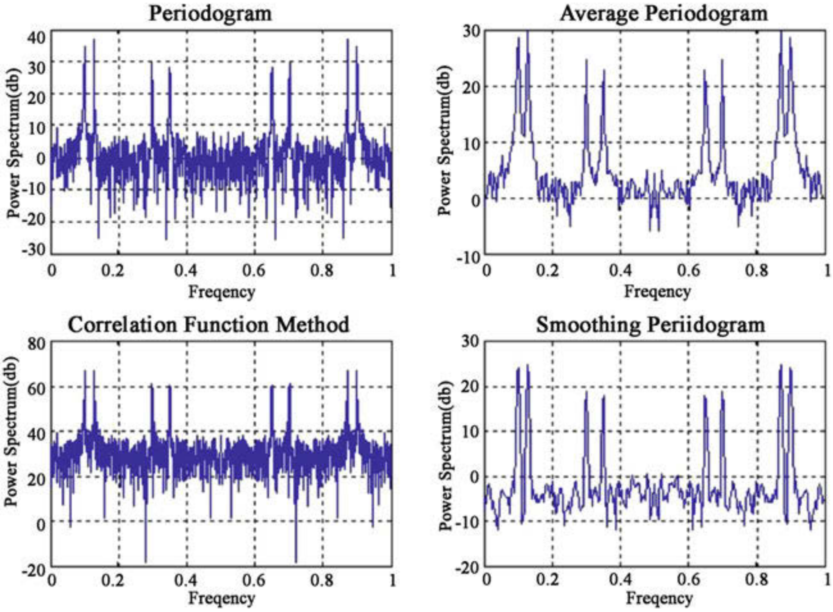


Fig. 94.1 Four typical spectrum estimation curves

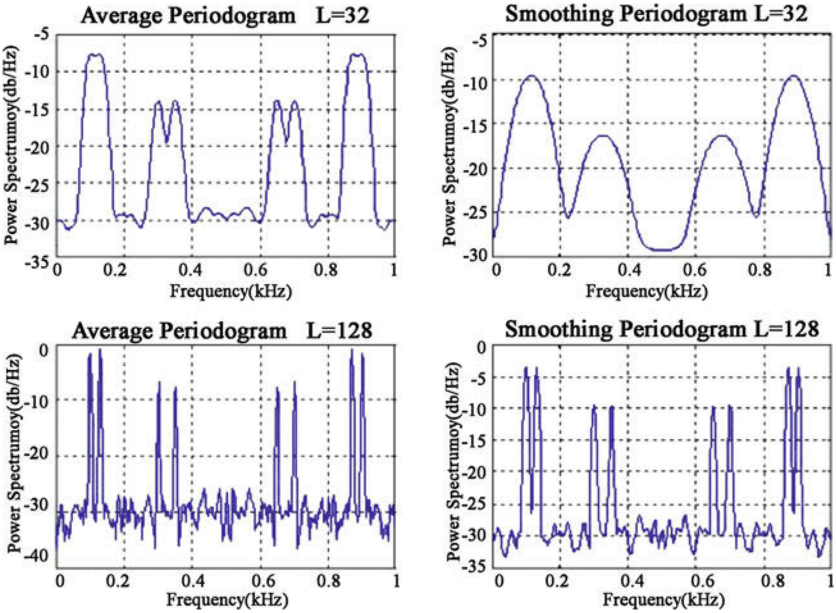


Fig. 94.2 The curves with different length of the average periodogram and the smoothing periodogram

94.5 Discussion

We could directly obtain the following properties from the Figs. 94.1 and 94.2.

The results which are estimated by power spectrum of the correlation function and periodogram are consistent, and have the characteristics of large discreteness, rough curve and the larger variance, but its resolution is higher than other methods.

The average periodogram and the smoothing periodogram have good convergence, smooth curves and small variance, but the main lobe of power spectrum is wide and the resolution is low. This is caused by Gibbs phenomenon which due to the sections processing of random sequences will come into the limited length.

The smoothing periodogram has relatively smoother spectrum valuation and lower resolution than the average periodogram. The main lobe of the power spectrum will become wider when it gets smoothing estimated results at the same time, by giving the appropriate window weighted to each sequence. So the resolution decreased.

Seen from the Fig. 94.2, the main lobe of the power spectrum will become narrower and improve the resolution if increasing the length L of the sections for a fixed random sequence with length N in the average periodogram and the smoothing periodogram. But the estimation variance will also increase because of the increasing length L. In actual application we should reasonably choose the length L to aviod the the contradiction of the resolution and the variance.

Acknowledgments The work was supported by the Innovation Fund for Small Technology-based Firms of China (No. 09C26214101775, No. 2009GJD00027), the Natural Science Foundation of Henan (102300410240) and Innovation Scientists and Technicians Troop Construction Projects of Zhengzhou City (10CXTD160).

References

1. Li J, Chen JB, Zhang LL (2010) Probability density function estimation of stochastic processes[J]. Chin J Appl Mech 03:53–57, 211–212
2. Feng Y, Trashi N (2010) Weighted noise power spectrum estimation based on the inter-frequency correlation[J]. Inf Electron Eng 04:63–67
3. Qu HY, Li L, Qian XL (2006) How to use MATLAB to optimize power spectral estimation of random signals treated by the basic period map method[J]. Comput Digit Eng 03:36–39
4. Yao WJ (2007) Research on AR model power spectrum estimation based on the algorithm and Burg algorithm[J]. Comput Digit Eng 10:8, 46–48
5. Zhang WW, Xing WQ (2008) Matlab realization of random signal spectrum estimation[J]. Mod Electron Tech 18:139–140
6. Yu XF, Ma DW, Wei L (2008) Simulation analysis of window function in power spectrum estimation based on modified periodogram[J]. Comput Simul 03:117–120
7. Wang FJ, Pan HX (2009) Analysis and selection of several power spectrum estimation functions in MATLAB. Electron Prod Reliab Environ Test 06:32–35
8. Wang FY, Zhang LL (2006) Power spectrum density estimation and the simulation in Matlab[J]. Microcomput Inf 11(31):294–296
9. Wei X, Zhang P (2005) The window function analysis of amended periodic table in power spectrum estimation[J]. Mod Electron Techn 03:20–21

Chapter 95
Research on Discrete Mathematical Model of Special Helical Surface

Lizhi Gu, Peng Lei and Qi Hong

Abstract Special helical surface is widely used in machinery manufacturing industry. This chapter discussed the effect of fundamental and important parameters on generic model of the helical surface, based on differential geometry, spatial engaging theory and method of Cartesian coordinate conversion. A 'guide helix line method' was proposed, and with this method a discrete mathematical model of non-cylinder helical surface was established. The obtained discrete mathematical model can readily have access to the coordinate information of each point and can conveniently design high-precision helical surface through programming and execute the software in the MATLAB environment. This mathematical model may find specific and significant application discretely, in the f computer-aided design for products with helical surface, especially for those with abnormal helical surface.

Keywords Special helix · Special helical surface · Guide helix · Discrete mathematical model

L. Gu (✉) · P. Lei · Q. Hong
College of Mechatronics Engineering and Automation,
Huaqiao University, Xiamen, People's Republic of China
e-mail: gulizhi888@163.com

P. Lei
e-mail: leipeng900@126.com

Q. Hong
e-mail: lusia_361005.student@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_95,

95.1 Introduction

Modeling design of special helical surface plays an important role in the modern machine manufacturing field, which is the link between product-designing and product-manufacturing. The non-cylinder helical surface is a widely used equipment in the machinery manufacturing industry [1], the development and application of special helical surface has guaranteed equipment's performance and service life. For example, the application of the variable pitch screw in large extrusion machines, the conical miller with equal helix angle and rotary bure in milling, and the variable pitch propeller installed on ships, play the key role in the mechanical equipment [2]. Hence, solving the problem of modeling technology for special helical surface has become an the important project in the machinery manufacturing industry.

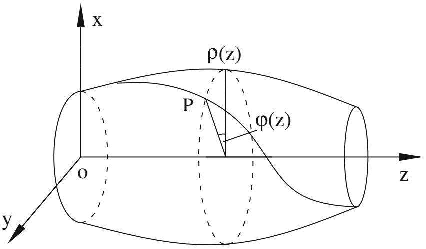
One point or a curve does a special screw motion around a central axis, it is the so-called special helical movement [3]. There are two ways for the helical movement: sliding along generatrix of rotate parts and revolving around a central axis. The sliding movement consists of axial movement and radial motion. So when one point or a curve does special screw motion, its orbit is a helix line or a helical surface. It was necessary to point out that solid of revolution, in this chapter, consists of cylinder and non-cylinder, but we mainly discussed the non-cylinder parts.

At present, there are few document resources about research on discrete mathematical model for special helical surface modeling [4, 5]. Most of the existing papers redefined some related concepts and summarized its generalized mathematical model briefly, there are serious limitations valuable for engineering applications. This chapter developed an approach to discrete mathematical model for special helical surface based on the screw theory, which was of great significance for the systematical research on helical surface.

95.2 Establishment of Mathematical Model for Guide Helix

To establish the mathematical model for special helical surface on rotational parts, first, there must to be defining of the space track of helical surface's truncate generatrix. The space track we defined it as 'guide helix', which was the oriented line for the helical surface truncate generatrix. The key problem was how to express the guide helix with precise mathematical model, because its spatial location changed, the mathematical model will change. So if we build a moving frame on the guide helix, the mathematical model of helical surface's truncate generatrix would become the only mathematical model. Then we scatter its truncate generatrix into finite coordinate points, alteration of the number of coordinate points can change the modeling accuracy. Each point following the guide helix forms helix trajectory, all of the helix trajectories fit into helical

Fig. 95.1 Special helix



surface. The following is the concrete step to establish the mathematical model of guide helix.

Take the central axis of non-cylinder revolution parts as Z axis, one end as the XOY plane to build system of rectangular coordinates O-XYZ, as shown in Fig. 95.1. For convenience of explanation, we chose the guide helix on the face of revolution solid. It was similar to not on the face of revolution solid. Equation of special helix can be obtained according to the forming principle of non-cylindrical revolution surface as mentioned above

$$r(z,\varphi) = [\rho\cos\varphi, \rho\sin\varphi, z]^T \tag{95.1}$$

where $\rho(z)$ means the non-cylinder revolution parts radius, and φ means the intersection angle formed between X axis and the non-cylinder revolution parts radius. When the functional relation about φ and z is ascertained, the guide helix on the surface of non-cylinder revolution parts can be expressed as

$$r(z) = r(z,\varphi(z)) = [\rho(z)\cos\varphi(z), \rho(z)\sin\varphi(z), z]^T \tag{95.2}$$

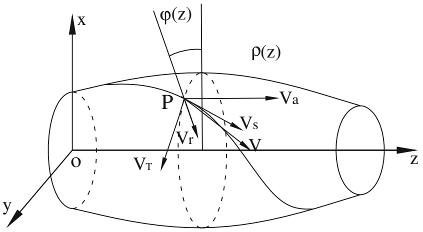
The key to establish guide helix mathematical model was the functional relation about φ and z according to the design requirements of helical surface. The traditional circular helix method cannot work well with the establishment of mathematical model for guide helix, hence, guide helix must be classified by its different characteristics of motion to establish mathematical model.

As shown in Fig. 95.2, $\overrightarrow{V}$ is velocity vector of point P, $\overrightarrow{V}_S$ stands for velocity vector along helical surface's truncate generatrix. $\overrightarrow{V}_T$ is tangent velocity vector of point P, $\overrightarrow{V}_a$, $\overrightarrow{V}_r$ are axial and radial velocity vector, their functional relationship is just expressed by $\overrightarrow{V}_S = \overrightarrow{V}_a + \overrightarrow{V}_r$. The helix angle is defined as the intersection angle formed between velocity vector $\overrightarrow{V}$ along helix and velocity vector $\overrightarrow{V}_S$ along helical surface's truncating generatrix [6].

$$\overrightarrow{V} = \overrightarrow{V}_S + \overrightarrow{V}_T = \overrightarrow{V}_a + \overrightarrow{V}_r + \overrightarrow{V}_T \tag{95.3}$$

The velocity of point $P\overrightarrow{V}$ can be broken into three vectors $\overrightarrow{V}_T$, $\overrightarrow{V}_a$, $\overrightarrow{V}_r$. The three values directly affect the shape of guide helix. To find out the value of $\overrightarrow{V}_S$, the angular velocity ω can be assumed to be constant, then the helix angle of point P can be described as follows

Fig. 95.2 The forming of helix vectogram



$$\tan\beta(z) = \frac{\left|\overrightarrow{V}_T\right|}{\left|\overrightarrow{V}_S\right|} = \frac{\omega r(z)}{V_S(z)} \tag{95.4}$$

For angle α is the angle between $\overrightarrow{V}_a$ and $\overrightarrow{V}_a$, according to functional expression $V_a = \mathrm{d}z/\mathrm{d}t,\ V_r = \mathrm{d}r/\mathrm{d}t$, Eq. 94.5 can be obtained in terms of

$$\cos\alpha = \frac{V_a}{V_S} = \frac{\mathrm{d}z/\mathrm{d}t}{\sqrt{(\mathrm{d}r/\mathrm{d}t)^2 + (\mathrm{d}z/\mathrm{d}t)^2}} = \frac{1}{\sqrt{1 + (\mathrm{d}r/\mathrm{d}t)^2}} \tag{94.5}$$

According to $\mathrm{d}\varphi = \omega \mathrm{d}t$, $\mathrm{d}z = \overrightarrow{V}_a \mathrm{d}t$, combination of Eqs. 95.4 and 95.5, the function with respect to $\varphi(z)$ is set as follows:

$$V_r = \frac{\mathrm{d}\varphi}{\mathrm{d}z} = \frac{\omega \mathrm{d}t}{\overrightarrow{V}_a \mathrm{d}t} = \frac{\tan\beta(z) V_S(z)}{r(z)\overrightarrow{V}_a \mathrm{d}t} = \frac{\tan\beta(z)}{r(z)}\sqrt{1 + \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2} \tag{95.6}$$

$$\varphi(z) = \varphi_0 + \int \frac{\tan\beta(z)}{r(z)}\sqrt{1 + \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2}\mathrm{d}z \tag{95.7}$$

Putting Eqs. 95.7 into 95.2, a common equation of guide helix according to spiral angle is obtained [7]. It can provide a mathematic model to calculate and optimize geometric model before meshing on a computer as a result of establishing the equation, where φ_0 can be obtained by the φ value of a certain given space position. When $\beta(z)$ is constant, the helix will be the equal spiral angle helix, as described in Eq. 95.7.

$$\varphi(z) = \varphi_0 + \tan\beta \int \frac{1}{r(z)}\sqrt{1 + \left(\frac{\mathrm{d}r}{\mathrm{d}t}\right)^2}\mathrm{d}z \tag{95.8}$$

Putting Eqs. 95.8 into 95.2, an equation about equal spiral angle helix is obtained. So the following conclusion can be drawn that equal spiral angle helix is a special case of special helix.

95.3 Establishment of Programmable Discrete Special Helical Surface Model

After common equation of guide helix has been confirmed, The next step is how to build the programmable discrete model for special helical surface. According to principle of "envelope forming", helical surface is deemed to be generated by its truncate generatrix, and that truncating generatrix can be divided into finite coordinate point. These points do spiral motion along guide helix, their trajectories are helices, through those helices, we can construct the helical surface model. To realize the data compression and smooth surface modeling within the user-defined precision, a proper increment of control points is necessary. Then Eq. 95.9 is a truncating generatrix expression of the special helical surface [8].

$$\begin{cases} x = R_{iz}(i,z)\cos\varphi_{iz}(i,z) \\ y = R_{iz}(i,z)\sin\varphi_{iz}(i,z) \\ z = z \end{cases} \tag{95.9}$$

where $R_{iz}(i,z)$, $\varphi_{iz}(i,z)$ stand for radius of gyration and gyrating angle respectively, while the point P is moved to the position $z = z$.

Figure 95.3 shows the section of revolution parts in the position $z = z$, the point P_i is a random point, r_i is its radius of gyration, its gyrate angle is φ_i. The point P_i can be described by the following equations.

$$r_i = \sqrt{x_i^2(i) + y_i^2(i)} \tag{95.10}$$

$$\mu = \arctan\left[\frac{y_i(i)}{x_i(i)}\right] \tag{95.11}$$

The moving frame was built on the guide helix, so the relative position was not changed between the origin of moving frame and the point of truncating generatrix. In the position $z = z$, r_{iz}, φ_i are its radius of gyration and gyrate angle, respectively. r_{ik} is the radius of guide helix in the position $z = z$. The expressions of correlation functions are as follows:

$$r_{iz} = \frac{r_1}{r_{ik}} \cdot r_z \tag{95.12}$$

$$\frac{\mathrm{d}r_{iz}}{\mathrm{d}z} = \frac{r_i}{r_{zk}} \cdot \frac{\mathrm{d}r_z}{\mathrm{d}z} \tag{95.13}$$

$$\frac{\mathrm{d}^2 r_{iz}}{\mathrm{d}z^2} = \frac{\mathrm{d}^2 r_z}{\mathrm{d}z^2} \tag{95.14}$$

$$\varphi_{iz} = \varphi_z + \mu_z = \varphi_z + \mu \tag{95.15}$$

According to expressions mentioned above, a discrete mathematical model can be obtained in the form of

Fig. 95.3 Cross-section

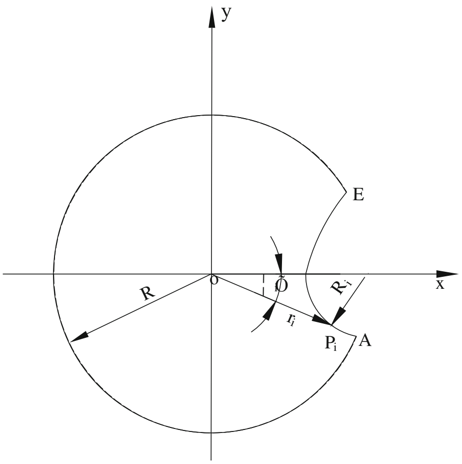
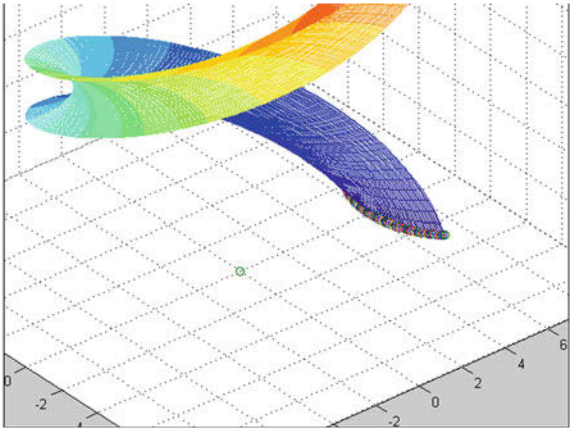


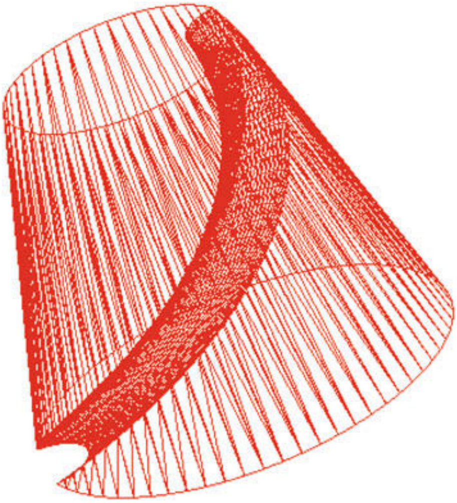
Fig. 95.4 Helical surface model



$$\begin{cases} x = \dfrac{\sqrt{x_1^2(i) + y_1^2(i)}}{r_{ik}} \cdot r_z \cos\left(\varphi(z) + \operatorname{arctg}\left[\dfrac{y_1(i)}{x_1(i)}\right]\right) \\ y = \dfrac{\sqrt{x_1^2(i) + y_1^2(i)}}{r_{ik}} \cdot r_z \sin\left(\varphi(z) + \operatorname{arctg}\left[\dfrac{y_1(i)}{x_1(i)}\right]\right) \\ z = z \end{cases} \tag{95.16}$$

The discrete mathematical model for helical surface has been obtained through programming with MATLAB, and it has been demonstrated that the method proposed and developed is used to build 3D digital model of helical surfaces in Figs. 95.4 and 95.5.

Fig. 95.5 3D model



95.4 Conclusion

Aiming at geometric modeling of special helical surface precisely to derive a general mathematical model for helical surface, a ‘guide helix line method’ was proposed, and with this method a discrete mathematical model for special helical surface was established. The obtained 3D digital model of helical surface can conveniently design high precision helical surface through programming, executing in the MATLAB environment and designing axial section of cutter. The method may find application in mechanical design with high efficiency and can be used as important reference for manufacturing such special helical surface of machine parts.

Acknowledgments The study is supported by Fujian Natural Science Fund (2010J01295) and by Xiamen Science and Technology Project Fund (3502Z20093031), Quanzhou Science and Technology Project Fund (2008G5), also by Heilongjiang Natural Science Fund (E201043).

References

1. Liu kexin BC (1998) Helicoid theory for design of cutting tools [M]. Publish House of Mechanical Industry, Beijing, pp 5–20
2. Li RX (2005) Theory and algorithm for design of cutting tools [M]. Publish House of Science and Technology in Jiangsu Province, Nanjing, pp 20–35
3. Wang YJ (2006) A theory of the helicoid surface of the spherical hob [J]. Chin J Mech Eng 27(4):16–21
4. Cui YQ (2008) Theory of helicoid of spherical hob [J]. Chin J Mech Eng 32(2):54–61
5. Gong ZH, Bin HZ (2009) Generalized definition of helical angle and its applications [J]. China Mech Eng 1:14–15

6. Tang YY (2010) General mathematical mode for grinding of precise helicoid [J]. Chin J Mech Eng 27(3):32–37
7. Shi PL, Wang W, Tang YY (1998) Research on mathematical mode in manufacturing of ball end mills [J]. Chin J Mech Eng 30(5):55–59
8. Xi W (2005) Design of spiral flute for form milling cutter with CAD [J]. Tool Eng 29(12):11–13

Part IX
Graphics and Visualizing

Chapter 96
PCNN-DDF Filter for Color Image

Beiji Zou, Haoyu Zhou, Hao Chen and Cao Shi

Abstract This chapter presents a new filtering approach capable of detecting and removing impulsive noise in multi-channel images. The filter detects noise pixels in the image by utilizing pulse-coupled neural networks specific feature that the fire of one neuron can capture firing of its adjacent neurons due to their spatial proximity and intensity similarity. Then it estimates the noise pixels by a DDF-likely vector filtering. Experimental results reported in this chapter indicate that the proposed filter has excellent performance, and is able to preserve fine details while suppressing impulsive noise.

Keywords Multi-channel image processing · Nonlinear vector filtering · PCNN

96.1 Introduction

The perception of color is important to humans and machine vision systems, since they use color information to sense the environment and recognize the objects on the scene. Because the acquisition or transmission of digital images through sensors or

B. Zou · H. Zhou (✉) · H. Chen · C. Shi
School of Information Science and Engineering,
Central South University, Changsha, China
e-mail: zhou_haoyu@163.com

B. Zou
e-mail: bjzou@vip.163.com

H. Chen
e-mail: xschenhao@hotmail.com

C. Shi
e-mail: plfdg@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_96,

communication channels is often affected by impulsive noise [1–4], the aim of preprocessing techniques is the noise filtering [5, 6] which enables communication in noisy environments [5] and processing of different kinds of multi-channel images (e.g., enhancement of cDNA microarray images [7, 8], digitized artwork images [9, 10], old movies [11–13], and images acquired by sensors [14, 15]).

It has been widely recognized [16–19] that the nonlinear vector processing of color images is the most effective way to filter out outliers. For this reason, a number of filtering approaches, such as those presented in [20–22], have been developed to extend the filtering efficiency of the standard filtering approaches. Vector median filter [16] is a typical example of such an extension, when the median defined over the gray-scale samples has been replaced with the lowest ranked multi-channel sample achieved by vector ordering [1]. This filter is very often used for the removal of impulsive noise in color images. On the other hand, the standard median filter [23] or its multi-channel extensions, i.e., the vector median filter [16] and the basic vector directional filter [18], are unable to adapt their behavior to varying noise and signal statistics related to the local image information of the samples inside a sliding filtering window. These filters performing the fixed amount of smoothing result in blurring of fine image details.

In the late 1980s, Eckhorn et al. discovered that the midbrain in an oscillating way created binary images that could extract different features from the visual impression when they studied the cat visual cortex [24]. Due to this discovery they developed a neural network, called Eckhorn's model, to simulate this behavior. Then Johnson et al. carried on a number of modifications and variations to improve its performance as image processing algorithms [25]. This modified neural model is called pulse-coupled neural networks (PCNN). As a new generation of neural network, the PCNN is good at digital image processing and applied in many fields like image segmentation, image enhancement, image fusion, object and edge detection, pattern recognition, etc. [26].

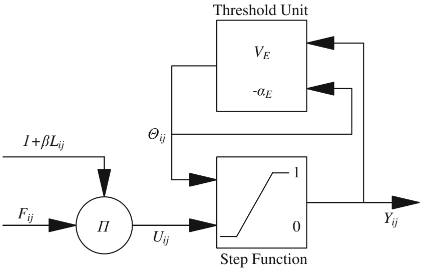
The remainder of this chapter is organized as follows. In Sect. 96.2, a brief overview of the PCNN model is presented. In Sect. 96.3, we propose a new color image filter based on PCNN. In Sect. 96.4, the proposed method is tested compared with some existing filters. Section 96.5 is the conclusion and this chapter ends with acknowledgements.

96.2 PCNN Model

As mentioned above, the PCNN is two-dimensional, single layered, laterally connected neural network of pulse-coupled neurons, which connect with image pixels one to one. Because each image pixel is associated with a neuron of the PCNN, processing the pixels can be translated into processing the corresponding neurons of the PCNN.

The PCNN neuron's structure is shown in Fig. 96.1. The neuron consists of an input part, linking part and a pulse generator. The neuron receives the input signals

Fig. 96.1 PCNN model



from feeding and linking inputs. Feeding input is the primary input from the neuron's receptive area. The neuron receptive area consists of the neighboring pixels of corresponding pixel in the input image. Linking input is the secondary input of lateral connections with neighboring neurons. The difference between these inputs is that the feeding connections have a slower characteristic response time constant than the linking connections. The standard PCNN model is described as iteration by the following equations:

$$F_{ij}[n] = e^{-\alpha_F} F[n-1] + V_F \sum_{kl} w_{ijkl} Y_{ij}[n-1] + I_{ij} \tag{96.1}$$

$$L_{ij}[n] = e^{-\alpha_L} L_{ij}[n-1] + V_L \sum_{kl} m_{ijkl} Y_{ij}[n-1] \tag{96.2}$$

$$U_{ij}[n] = F_{ij}[n](1 + \beta L_{ij}[n]) \tag{96.3}$$

$$Y_{ij}[n] = \mathrm{step}\big(U_{ij}[n] - E_{ij}[n-1]\big) \tag{96.4}$$

$$E_{ij}[n] = e^{-\alpha_E} E_{ij}[n-1] + V_E Y_{ij}[n] \tag{96.5}$$

In these equations, I_{ij} is the input stimulus such as the normalized gray level of image pixels in (i, j) position, $F_{ij}[n]$ is the feedback input of the neuron in (i, j), and $L_{ij}[n]$ is the linking item. $U_{ij}[n]$ is the internal activity of neuron, and $E_{ij}[n]$ is the dynamic threshold. $Y_{ij}[n]$ stands for the pulse output of neuron and it gets either the binary value 0 or 1. The input stimulus (the pixel intensity) is received by the feeding element and the internal activation element combines the feeding element with the linking element. The value of internal activation element is compared with a dynamic threshold that gradually decreases at iteration. The internal activation element accumulates the signals until it surpasses the dynamic threshold and then fires the output element and the dynamic threshold increases simultaneously strongly [18]. The output of the neuron is then iteratively fed back to the element with a delay of one iteration.

The inter-connections M and W are the constant synaptic weight matrices for the feeding and the linking inputs, respectively, which are dependant on the distance between neurons. Generally, M and W (normally $W = M$) refer to the Gaussian weight functions with the distance. β is the linking coefficient. α_F, α_L and α_E are the attenuation time constants of $F_{ij}[n]$, $L_{ij}[n]$ and $E_{ij}[n]$, respectively. V_F, V_L and V_E denote the inherent voltage potential of $F_{ij}[n]$, $L_{ij}[n]$ and $E_{ij}[n]$, respectively.

96.3 Introduction of Proposed Framework

96.3.1 Basic Idea

In case of impulsive noise, color image noise pixels corrupted each channel of the image independently [16–19]. Each channel has been corrupted in different positions, but they are all corrupted in the same characteristic and pattern. Impulsive noise is a very interested noise: in each channel of the image, the value of the corrupted pixel is quite different from the pixels that have not been affected nearby, which means that the neurons corresponding to the noise pixels will not fire synchronously with the neurons that have not been affected nearby. Thus the basic idea for adopting PCNN for noise detection and removal is to fix the stimulus of noise pixels to make them fire with the normal pixels synchronously.

Color image is a multi-channel image which is not likely a gray image. For a gray image, we can build a PCNN whose neurons are one by one corresponding to each pixel in the image. If we use PCNN for each channel separately and then combine the filtered channels together, because each channel is affected in a different place, unexpected new color will be introduced to the image and hue will change in the original noise pixels. To solve this problem, we have to translate the image from original R, G, B three channels into a new space which can amplify the difference between noise pixels with normal pixels.

Thus we proposed a noise detection and removal filter based on PCNN which only modifies the noise pixels in the new space. The filter only operates a neuron (i, j) when it fires; the algorithm is implemented in the following steps:

1. When neuron (i, j) fires within its nearby neurons already fired, it means that this neuron is not able to be captured to fire because of its nearby neurons firing, so the corresponding pixel is a noise pixel, and because its fire time is later than the nearby neurons, so this pixel is darker than the supposed normal value. It needs to be set brighter.
2. When neuron (i, j) fires with its nearby neurons do not fire yet, it means that this neuron's firing cannot capture its nearby neurons firing, so the corresponding pixel is a noise pixel, and because its fire time is earlier than the nearby neuron, so this pixel is brighter than supposed normal value. It needs to be set darker.
3. When neuron (i, j) fires with half of the nearby neurons already fired and half not fired yet, it means that this neuron is not a noise pixel and it should remain unchanged.
4. After this action, the algorithm will get a list of suspicious noise pixels in the image. For each noise pixel, we employ a DDF-likely algorithm for filtering: build a processing window which central point in the noise pixel, then use DDF

for the pixels in the processing window expecting the pixels in the suspicious noise pixels list to select the median value to replace the central pixel.

96.3.2 Noise Pixel Detection

RGB color space set up from primary color spectrum is quite suitable for hardware implementation, but not good for explanation in human vision system. In RGB color space, impulsive noise corrupts one or more channels, the probability of corruption in each channel is equal. So if we detect noise directly in RGB color space, we should consider the corruption situation in three channels equally, and the output image also cannot fix the characteristics of human vision system. Human vision system use hue, saturation and intensity to describe color and observe colorful objects. In HSI color space, Impulsive noise corruption focus more on intensity than hue and saturation. The following equations can translate image from RGB space into HSI space:

$$H = \begin{cases} \theta, & B \leq G \\ 2\pi - \theta, & B > G \end{cases}, \quad \theta = \arccos\left(\frac{2 \times R - G - B}{2 \times ((R-G)^2 + (R-B)(G-B))^{1/2}}\right) \tag{96.6}$$

$$S = 1 - 3 \times \min\frac{(R, G, B)}{(R + B + G)} \tag{96.7}$$

$$I = \frac{(R + G + B)}{3} \tag{96.8}$$

If the image is filtered in *H*, *S*, *I* channel separately, new color will be introduced to the image, and details and edges will also be destroyed. Thus we construct *m*:

$$m = (\eta + \cos H)^p S^q I^{(1-p-q)}, \quad \eta \geq 2,\ p \geq 0,\ q \geq 0,\ p + q \leq 1 \tag{96.9}$$

By constructing *m*, we amplify the difference between noise pixel and its nearby pixels, and consider affection to hue, saturation and intensity by the noise pixel. Thus we avoided the problem of handling the noise in one or more channels separately. Using this equation for every pixel in the image, we will get a matrix $M = \{m_1, m_2, \ldots, m_n\}$ which has the same size as the image. To detect the noise pixel, we use *M* to activate a PCNN which has the same size as*M*. Let the neuron network runs have all the neurons fired and record the firing moment of each neuron to the firing time map (FTM). For each pixel m_i in *M*, check the slide window which has this pixel as central point, remark the total pixel amount in the window as S_0, remark the amount of pixel firing before central point as S_1, remark the amount of pixel firing after central point as S_2, if $S_1 > S_0/2$ or $S_2 > S_0/2$ then the central point is corrupted by noise, otherwise it is not corrupted.

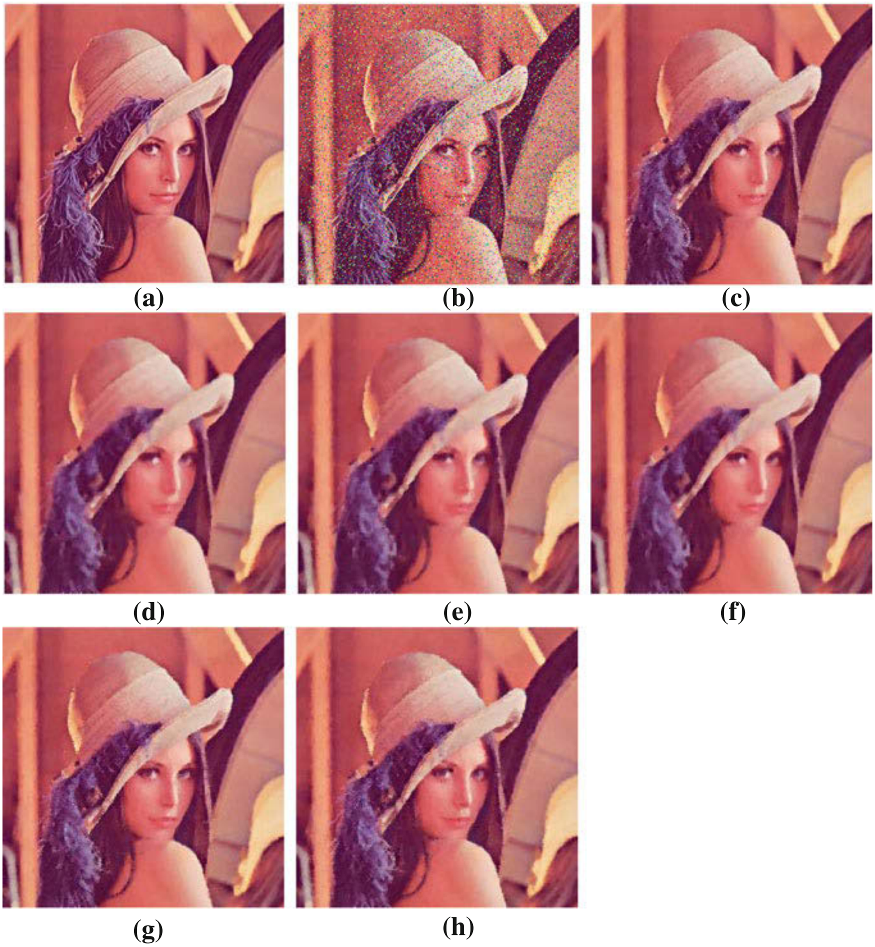


Fig. 96.2 Original Lena image, impulsive noise ($p_v = 0.1$) image and filtered outputs. **a** Original Lena image (24-Bit, 256 × 256); **b** Noise image; **c** PCNN-DDF; **d** VMF; **e** BVDF; **f** DDF; **g** Rank SVMF; **h** Mean SVMF

96.3.3 Noise Pixel Removal

If the pixel is not corrupted, leave it unchanged. If the pixel is corrupted, we proposed an auto parameter a DDF likely algorithm: set up a processing window which has the corrupted pixel as central point, if over half of the pixels in the window fire time are earlier than central point ($S_1 > S_0/2$), choose the pixels in this processing window and are not in the noise pixel list to build $\Gamma = \{x_1, x_2, \ldots, x_L\}$, if x_l and x_k are two pixels in Γ, the Euclidean distance $D(l,k)$ is: $D(l,k) = \|x_l - x_k\|_2$, angle distance $A(l,k)$ is:

Table 96.1 Comparison of the presented algorithms using impulsive noise corruption $p_v = 0.1$

Methods/criterions	MAE	MSE	NCD $\times$ 10^{-4}
PCNN-DDF	110.19	3.62	68.76
VMF	174.35	6.94	118.62
BVDF	183.65	7.21	128.16
DDF	161.20	6.37	106.93
Rank SVMF	121.35	4.01	75.12
Mean SVMF	138.75	4.89	93.51

Table 96.2 Comparison of the presented algorithms using impulsive noise corruption $p_v = 0.3$

Methods/criterions	MAE	MSE	NCD $\times$ 10^{-4}
PCNN-DDF	192.45	6.24	102.02
VMF	240.28	8.49	143.74
BVDF	264.31	8.92	160.23
DDF	231.76	7.78	136.87
Rank SVMF	206.23	6.92	112.21
Mean SVMF	216.32	7.56	136.97

$$A(l,k) = \arccos\left(\frac{x_l^T x_k}{\|x_l\|_2 \|x_k\|_2}\right),$$

then set central point as

$$\arg\min_{x_k \in \Gamma}\left\{\left[\sum_{l=1}^{L} D(l,k)\right]^{g}\left[\sum_{l=1}^{L} A(l,k)\right]^{1-g}\right\} \quad (0 \le g \le 1).$$

If over half pixels in the window fire later than central point($S_2 > S_0/2$), then choose the pixels in this processing window that are not in the noise pixel list to execute the above operations. By using this method, we can remove noise without affect image details and edges, and avoid disturbance from other noise pixels.

The parameters' value should consider the following points: (1) to make neurons in PCNN fire quickly because of E_{ij}'s value rise rapidly, V_E should be rather bigger; (2) for neuron connect with its nearby neurons, and locate noise point, the linking coefficient β should be bigger; (3) to set m's value, η is used to smooth effect of hue. Set value of p, q should consider impulsive affection on image focus on intensity and hue. In the experiments of of the following section, the parameters' values are set as follows:

$$\beta = 0.3,\ \ \alpha_E = 0.1,\ \ VE = 310,\ \ w = \begin{bmatrix} 0.2 & 1 & 0.2 \\ 1 & 0 & 0.2 \\ 0.2 & 1 & 0.2 \end{bmatrix};$$
$$\eta = 2, p = 0.5, q = 0.1, g = 0.25.$$

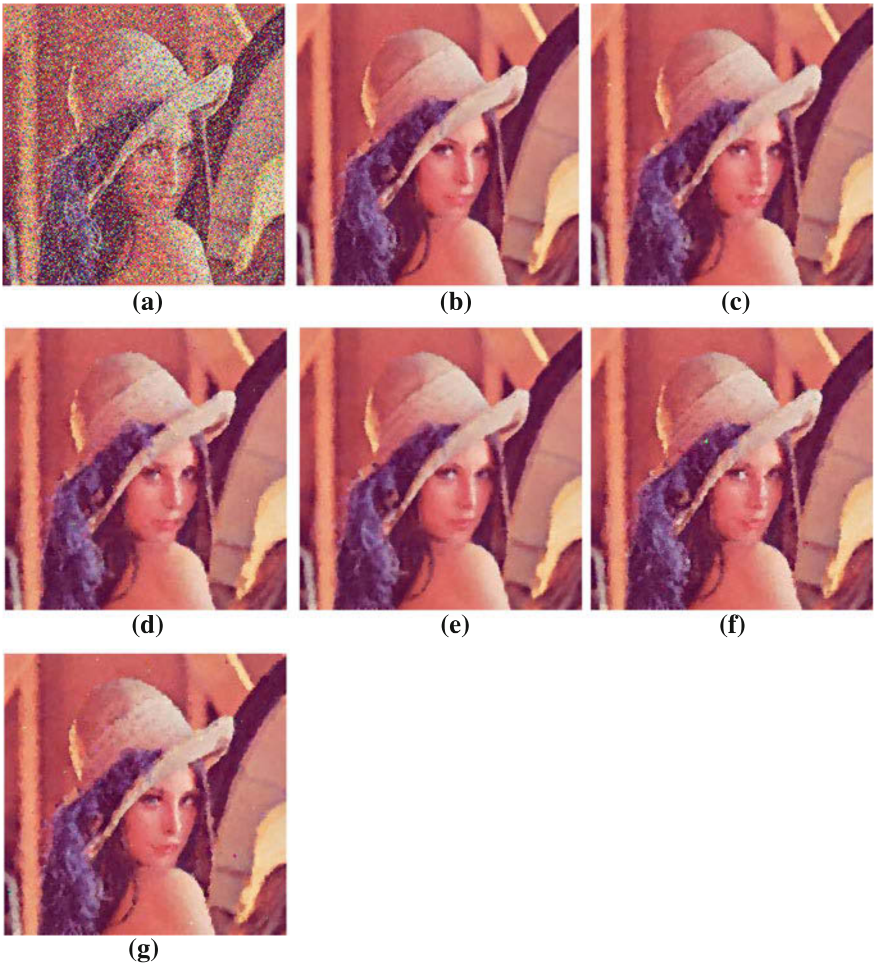


Fig. 96.3 Impulsive noise ($\underline{p_v} = 0.3$) image and filtered outputs. **a** Noise image; **b** PCNN-DDF; **c** VMF; **d** BVDF; **e** DDF; **f** Rank SVMF; **g** Mean SVMF

96.4 Experimental Results

The primary goal of all filtering algorithms presented in this chapter is to remove impulses and outliers from the image. This type of noise is often introduced through bit errors [5], especially during the scanning or transmission over the noisy information channel.

The achieved results were evaluated by the commonly used objective criteria [27], such as the mean absolute error (MAE), the mean square error (MSE), and the normalized color difference (NCD).

The methods were tested using test image Lena (Fig. 96.2a), whose size is 256 × 256. We added impulse noise with $p_v = 0.1$, $p_v = 0.3$. Then we use our PCNN-DDF, VMF, BVDF, DDF ($g = 0.25$), Rank SVMF and Mean SVMF on the image. Tables 96.1 and 96.2 and Figs. 96.2 and 96.3 show the experimental results.

The results show that PCNN is superior to the comparison algorithms in removal noise and image detail preservation. Especially when $p_v > 0.2$, vector filters bring distort of image, but PCNN not only preserves the details but also removes the noise.

96.5 Conclusion

In this chapter, a new color image filtering framework based on PCNN has been proposed.

The achieved results show excellent detection and image detail preservation capabilities of the new approach, while still holding the impulsive noise attenuation characteristics of standard vector filters. The new filters clearly outperform the standard vector filtering schemes as well as their adaptive modifications. In our experiments, the best results were achieved by PCNN-DDF scheme.

Acknowledgments We thank the associate editor, reviewers, and people, who helped us to improve this chapter, for their helpful and constructive suggestions. This chapter is jointly supported by the National Natural Science Foundation of China (Grant Nos. 60970098 and 60803024), the Major Research Plan of the National Natural Science Foundation of China (Grant No 90715043).

References

1. Lukac R, Smolka B, Martin K, Plataniotis KN, Venetsanopoulos AN (2005) Vector filtering for color imaging. IEEE Signal Process Mag (special issue on color image processing) 22:74–86
2. Plataniotis KN, Venetsanopoulos AN (2000) Color image processing and applications. Springer, Berlin
3. Pitas I, Tsakalides P (1991) Multivariate ordering in color image filtering. IEEE Trans Circuits Syst Video Technol 1:247–259
4. Smolka B, Chydzinski A, Wojciechowski K, Plataniotis KN, Venetsanopoulos AN (2001) On the reduction of impulsive noise in multi-channel image processing. Opt Eng 40:902–908
5. Astola J, Kuosmanen P (1997) Fundamentals of nonlinear digital filtering. CRC Press, Boca Raton
6. Peltonen S, Gabbouj M, Astola J (2001) Nonlinear filter design: methodologies and challenges. In: Proceedings of the IEEE region 8-EURASIP symposium on image and signal processing and analysis ISPA 01 in Pula. Croatia, pp 102–106
7. Lukac R, Plataniotis KN, Smolka B, Venetsanopoulos AN (2004) A multi-channel order-statistic technique for cDNA microarray image processing. IEEE Trans Nanobiosci 3:272–285
8. Lukac R, Plataniotis KN, Smolka B, Venetsanopoulos AN (2005) cDNA microarray image processing using fuzzy vector filtering framework. J Fuzzy Sets Syst (special issue on fuzzy sets and systems in bioinformatics) 152:17–35
9. Barni M, Bartolini F, Capellini V (2000) Image processing for virtual restoration of artworks. IEEE Multimed 7:34–37

10. Li X, Lu D, Pan Y (2001) Color restoration and image retrieval for Donhunag fresco preservation. IEEE Multimed 7:38–42
11. Kokaram AC (1998) Motion picture restoration. Springer, Berlin
12. Tenze L, Carrato S, Ramponi G (2002) An alignment algorithm for old motion pictures. IEEE Signal Process Lett 9:309–311
13. Barni M, Buti F, Bartolini F, Capellini V (2004) A quasi-Euclidean norm to speed up vector median filtering. IEEE Trans Image Process 9:1704–1709
14. Garber NJ, Hoel LA (1999) Traffic and highway engineering. Brooks/Cole, Pacific Grove
15. Morillas S, Gregori V, Peris-Fajarnés G, Latorre P (2005) A fast impulsive noise color image filter using fuzzy metrics. Real-Time Imaging 11(6):417–428
16. Astola J, Haavisto P, Neuvo Y (1990) Vector median filters. Proc IEEE 8:678–689
17. Lukac R (2003) Adaptive vector median filtering. Pattern Recogn Lett 24:1889–1899
18. Lukac R, Smolka Bogdan, Plataniotis KN, Venetsanopoulos AN (2006) Vector sigma filters for noise detection and removal in color images. J Vis Commun Image Represent 17:1–26
19. Plataniotis KN, Androutsos D, Venetsanopoulos AN (1999) Adaptive fuzzy systems for multi-channel signal processing. Proc IEEE 87:1601–1622
20. Lukac R, Smolka B, Plataniotis KN, Venetsanopulos AN (2004) Selection weighted vector directional filters. Comput Vision Image Understand (special issue on colour for image indexing and retrieval) 94:140–167
21. Lucat L, Siohan P, Barba D (2002) Adaptive and global optimization methods for weighted vector median filters. Signal Process Image Commun 17:509–524
22. Viero T, Oistamo K, Neuvo Y (1994) Three-dimensional median related filters for color image sequence filtering. IEEE Trans Circuits Syst Video Technol 4:129–142
23. Pitas I, Venetsanopoulos AN (1992) Order statistics in digital image processing. Proc IEEE 80:1892–1919
24. Eckhorn R, Reitboeck HJ, Arndt M, Dicke PW (1989) A neural network for feature linking via synchronous activity: results from cat visual cortex and from simulations. In: Models of brain function. Cambridge University Press, Cambridge, pp 255–272
25. Reitboeck HJ, Eckhorn R, Arndt M, Dicke P (1989) A model of feature linking via correlated neural activity. In: Synergistics of cognition. Springer, New York, pp 112–125
26. Wang Z, Ma Y, Cheng F, Yang L (2010) Review of pulse-coupled neural networks. Image Vis Comput 28:5–13
27. Lee JS (1983) Digital image smoothing and the sigma filter. Comput Vis Graph Image Process 24:255–269

Chapter 97
3D Dynamic Reconstruction of Rigid Object using Space–Time Correlation in Multi-View

Li Xiuxiu, Zheng Jiangbin, Zhang Yanning and Chen Ning

Abstract A three dimensional (3D) dynamic reconstruction algorithm for rigid objects in multiview is presented. In this algorithm, the space–time correlation is utilized to fill larger holes due to the insufficiency of viewpoints. For this purpose, firstly, an improved generalized voxel coloring GVC is used to reconstruct the initial 3D shape using the frame images in multiview at different time based on the color constraints, and a new voxel space can be obtained; secondly, the space–time correlation between obtained voxel spaces is calculated at different time using the features motion parameters and iterative closest points; finally, the dynamic reconstruction is implemented by combining the obtained voxel space of the previous frame and the current frame images using the space–time correlation. Owing to the space–time correlation, all information of frame images can be combined to reconstruct the 3D shape of an object, and most holes are filled. Experiments are given to demonstrate the efficiency of this reconstruction algorithm.

Keywords Dynamic 3D reconstruction · A multi-view system · Space–time correlation · Hole-filling

97.1 Introduction

The three dimensional (3D) shape reconstruction in a multiview is an important technique to get the 3D model of a realistic 3D object. It has been an important research topic in computer vision for long time [1], which has been applied in

L. Xiuxiu (✉) · Z. Jiangbin · Z. Yanning · C. Ning
School of Computer Science and Technology,
Northwestern Polytechnical University, Xi'an, China
e-mail: Lixiuxiu1013@sohu.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_97,

video conferencing, virtual reality, virtual museum and historical archiving demand etc. However, in the multi-view reconstruction, holes are existent in a reconstructed 3D model because of the insufficiency of viewpoints or self-occlusions, which would affect the usage of the reconstructed model and display effects.

To solve the problem caused by holes, various approaches are proposed to fill or avoid holes. According to the representation way of the reconstructed 3D object, different solutions are proposed. In the reconstruction based on topological representations (e.g. reconstruction based on mesh), the hole-filling is integrated into the reconstruction process commonly for the explicit connectivity [2–4]. In [3], the local radial basis function is used to automatically identify and interpolate hole regions in triangulated models. In [4], the holes are repaired in edge-constrained Delaunay triangulation. In the reconstruction based on point cloud or voxel, there are two classes of methods to fill holes. In the first class, the interpolation is implemented to fill holes [5–7], e.g. alpha shapes, crusts or balls. In [5], a signed distance field (SDF) is used to interpolate holes. In [6, 7], the point cloud model is triangulated and a moving least squares (MLS) approach is used to interpolate the hole regions. In the second class, the surface evolving is used to avoid holes [8, 9]. The class can be conducted by solving certain partial differential equations. In [8], the shape is repaired automatically based on context information only by solving a planar PDE system over a 2D domain. In [9], firstly, a signed distance function whose zero set is the observed surface is constructed to represent the 3D model, then the diffusion is applied to the representation, finally, the incomplete surface is extended to form a watertight model. The class of methods is usually used for the non-rigid objects.

In the methods above, the holes are filled or avoided using the neighborhood information of the hole in the reconstructed 3D model. In this paper, a multi-view dynamic reconstruction algorithm for rigid objects is proposed to fill the holes with the information of the space–time correlation in successive frames. In this algorithm, the space–time correlation in two successive frames is calculated by means of the feature motion parameters and iterative closest points (ICP), which are usually used for 3D face registration [10–12]. Owing to the space–time correlation, the temporal and spatial information are combined to reconstruct the shape of a rigid object, thus larger holes will be filled gradually.

97.2 The Framework of the Dynamic Reconstruction Algorithm

The dynamic reconstruction algorithm aims at filling or avoiding the holes with the space–time correlation. In this section, the detailed algorithm framework is presented (Fig. 97.1). The entire processing procedure is outlined as follows:

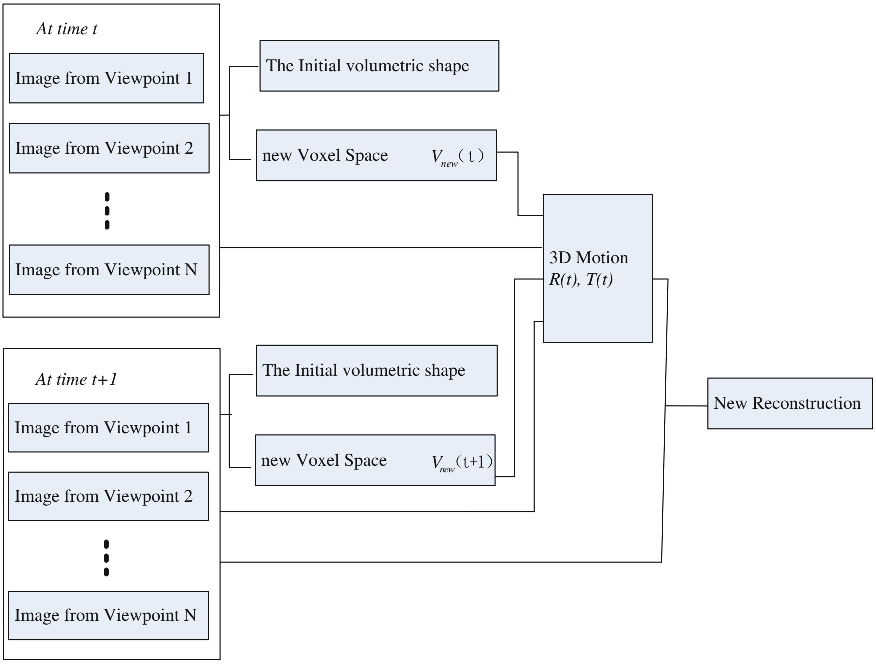


Fig. 97.1 The framework of the dynamic reconstruction

Step 1: *Initial reconstruction*. The initial volumetric shape with color and texture of an object is reconstructed with the improved generalized voxel coloring (GVC) [13] at different time in multi-view, and a new voxel space is obtained.

Step 2: *Space–time correlation calculation*. Calculate the 3D motion parameters of the rigid object by means of 3D feature motion and ICP [14], and the space–time correlation between the obtained successive voxel spaces is obtained.

Step 3: *The dynamic reconstruction*. The obtained voxel space of the previous frame is combined with the current frame images to reconstruct the object using the space–time correlation.

97.3 The Dynamic Reconstruction Algorithm

Given a rigid object O in the multi-view system. m calibrated digital cameras are put around O to capture its images.

97.3.1 The Initial Shape Reconstruction in Multi-View

The improved GVC is used to reconstruct the volumetric model of O. The silhouettes of the object O are extracted in m images from multi-view.

Assumed the initial voxel space is $V(0) = \{(x, y, z)|x_s \leq x \leq x_e, y_s \leq y \leq y_e, z_s \leq z \leq z_e\}$, and the size of a voxel is $l \times w \times h$ at time t. For each voxel $\text{vox}_i(\text{vox}_i \in V)$, the following steps are implemented.

Step 1: Calculate the Euclidean distance between the voxel and the optical centers of m viewpoints $\text{dist}(\text{vox}_i, \text{center}_j)$, where $j = 1, 2, \ldots, m$, and determine the visibility of the voxel vox_i.

Project vox_i to m viewpoints:

$$p_{i\_j} = \text{proj}_j(\text{vox}_i) \;\; j = 1, 2, \ldots, m \tag{97.1}$$

where $\text{proj}_j(x)$is the projection function of the voxel x in the viewpoint j, and p_{i_j} is the 2D projection pixel zone. If partial p_{i_j} belongs to the object silhouettes, assign some properties to these pixels inp_{i_j} according to $\text{dist}(\text{vox}_i, \text{center}_j)$, else vox_i is discarded because it does not belong to the object. When a pixel in p_{i_j} is also in the silhouette of the viewpoint j, the following properties are assigned:

(a) The pixel is projected by another voxel whether or not: is Used

$$\text{is Used} = \begin{cases} 0 & \text{The pixel is not projected} \\ 1 & \text{The pixel is projected by another voxel} \end{cases}$$

(b) The corresponding voxel: v (its projection in viewpoint j is p_j)
(c) The distance between the corresponding voxel and the viewpoint j: distance (the initial value is a larger value.)

When $\text{dist}(\text{vox}_i, \text{center}_j) < \text{distance}$, the labels of these pixels are changed (Eq. 97.2) and put the voxel vox_i in a temporary voxel space V_{temp}.

$$\begin{aligned} \text{is Used} &= 1; \\ v &= \text{vox}_i; \\ \text{distance} &= \text{dist}(\text{vox}_i, \text{center}_j) \end{aligned} \tag{97.2}$$

Step 2: Determine whether the voxel belongs to the shape of O according to the color-consistency in different cameras and the continuity of the color in the same camera.

The voxel vox_k $\left(\text{vox}_k \in V_{\text{temp}}\right)$ is re-projected to m viewpoints with Eq. 97.1. If some pixels in its projection zone p_{i_j} of the viewpoint j are labeled and the corresponding voxel v is vox_k, the voxel vox_k is visible in the viewpoint j. The color mean $\mu^c_{k\_j}$ and variance $\text{var}^c_{k\_j}$ of these pixels are calculated $(c = R, G, B)$.

The following items are examined to determine whether vox_k belongs to the shape of O.

(I) The color of vox_k's projection zone is continuous in the viewpoint j when $\text{var}^c_{k\_j} < T_{\text{continuous}}$;

(II) The color-consistency of vox_k is satisfied in its visible viewpoints when $\sigma_k < T_{\text{consistency}}$, where

$$\sigma_k = \frac{1}{3}\sqrt{(\sigma_k^R)^2 + (\sigma_k^G)^2 + (\sigma_k^B)^2} \tag{97.3}$$

and

$$(\sigma_k^c)^2 = \frac{1}{M_{\text{visible}} - 1}\left(\sum_{\substack{\text{vox}_k \text{ is visible} \\ \text{in viewpoint} j}} (\mu_{k\_j}^c)^2 - \frac{1}{M_{\text{visible}}}\left(\sum_{\substack{\text{vox}_k \text{ is visible} \\ \text{in viewpoint} j}} \mu_{k\_j}^c\right)^2\right) \tag{97.4}$$

where $c = R, G, B$.

M_{visible} is the number of viewpoints where vox_k is visible. $T_{\text{continuous}}$ and $T_{\text{consistency}}$ are the predefined thresholds.

Step 3: The shape $S(t)$ of O and a new voxel space $V_{\text{new}}(t)$ are obtained.

When vox_k is visible only in the viewpoint j, vox_k and its color $\mu_{k\_j}^c (c = R, G, B)$ are put in $S(t)$ and $V_{\text{new}}(t)$, if (I) is satisfied.

If vox_k is visible in several viewpoints, vox_k and its color $\mu_k^c (c = R, G, B)$ are put in $S(t)$ and $V_{\text{new}}(t)$ when (I) and (II) are concurrently satisfied.

$$\mu_k^c = \frac{1}{M_{\text{visible}}} \sum_{\substack{\text{vox}_k \text{ is visible} \\ \text{in viewpoint} j}} \mu_{k\_j}^c \tag{97.5}$$

If vox_k is invisible in m viewpoints, vox_k is an occluded voxel and is put in $V_{\text{new}}(t)$. It is obvious that $V_{\text{new}}(t)$ is a solid voxel space containing the shape of the object.

97.3.2 The Space–Time Correlation

At time $t + 1$, to utilize the information from 3D reconstruction result at time t, it is indispensable to get the space–time correlation between the obtained voxel spaces at time t and time $t + 1$. To get the space–time correlation, the 3D motion parameters of the obtained voxel space at time t, $R(t)$ and $T(t)$, are calculated. The $R(t)$ and $T(t)$ are obtained by combining the 3D feature motion and ICP. The feature motion is used to get a coarse correlation and the corresponding 3D motion parameters are $R_{\text{feature}}(t)$ and $T_{\text{feature}}(t)$. The coarse correlation is fine tuned with ICP, and the corresponding 3D motion parameters are $R_{\text{ICP}}(t)$ and $T_{\text{ICP}}(t)$. The $R(t)$ and $T(t)$ will be obtained by combining $R_{\text{feature}}(t)$, $T_{\text{feature}}(t)$ and $R_{\text{ICP}}(t)$, $T_{\text{ICP}}(t)$.

1. The 3D feature motion

Some corresponding 3D features at time t and time $t+1$can be used to calculate the 3D motion parameters of the object. In this paper some corner features are used. Assumed the corresponding 3D coordinates of corner at time t and time $t+1$ are $\{P_1^t, P_2^t, \ldots, P_n^t\}$ and $\{P_1^{t+1}, P_2^{t+1}, \ldots, P_n^{t+1}\}$, the 3D motion parameters satisfy:

$$R_{\text{feature}} P_i^t + T_{\text{feature}} = P_i^{t+1} \quad (i = 1, 2, \ldots, n) \tag{97.6}$$

2. FineTuning Based on ICP

ICP algorithm is a common method for rigid point set registration due to its simplicity and low computational complexity 2. It calculates correspondences of two point sets with the closest distance criterion and the least squares rigid transformation iteratively.

In this algorithm, fine tuning based on ICP is used to refine the space–time correlation between the successive voxel spaces further after the 3D feature motion correlation. The 3D motion parameters between the voxel space after the 3D feature motion and the obtained voxel space at time $t+1$ are calculated: $R_{\text{ICP}}(t)$, $T_{\text{ICP}}(t)$.

According to $R_{\text{feature}}(t)$, $T_{\text{feature}}(t)$ and $R_{\text{ICP}}(t)$, $T_{\text{ICP}}(t)$., the 3D motion parameters can be calculated as follows:

$$R(t) = R_{\text{ICP}}(t) \times R_{\text{feature}}(t) \tag{97.7}$$

$$T(t) = R_{\text{ICP}}(t) \times T_{\text{feature}}(t) + T_{\text{ICP}}(t) \tag{97.8}$$

97.3.3 Dynamic Reconstruction

$V_{\text{new}}(t)$ is transformed as the voxel space at time $t+1$, $V(t+1)$, through the motion $R(t)$and $T(t)$. In $V(t+1)$, the shape $S(t+1)$ is reconstructed and the new voxel space $V_{\text{new}}(t+1)$ is obtained with the method in Sect. 97.3.1. Some special operations are implemented for voxels belonged to $S(t)$ after the transformation: these voxels are put into $S(t+1)$ directly, when the projections belong to the silhouettes of O at time $t+1$.

97.4 Experiments

In this section, some 3D reconstruction experiments are presented with our method. In the experiments, 16 cameras are placed around the object, and the resolution is 648 × 490. Figure 97.2 presents the images from eight viewpoints at time $t1$, $t2$, $t3$, and $t4$.

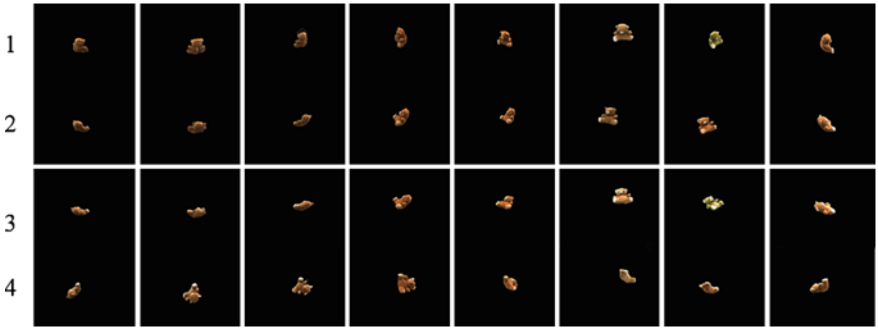
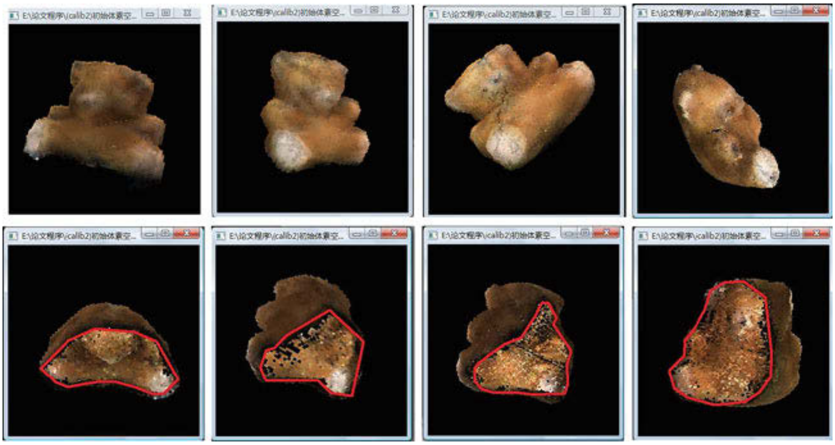


Fig. 97.2 The images from 16 viewpoints at time *t*1, *t*2, *t*3 and *t*4 from camera 1, 3, 5, 7, 9, 11, 13, 15



The initial reconstruction result at time $t1$

The initial reconstruction result at time $t2$

The initial reconstruction result at time $t3$

The initial reconstruction result at time $t4$

Fig. 97.3 The initial reconstruction (The holes, stemmed from the occlusion or view insufficiency, are labeled by *red* loop)

Firstly, the 3D shape of the object at a moment is reconstructed and the result is shown in Fig. 97.3. In the reconstructed 3D object, there are larger holes due to the insufficiency of viewpoints.

Secondly, in order to get the space–time correlation, the feature motion and ICP fine tuning are combined to calculate the 3D motion parameters between the voxel spaces from two successive frames. Figure 97.4 shows the 2D projections of $V(t2)$ in images at time *t*2. It is obvious that the projections cover all the silhouette zones, and $V(t2)$ can be used as the voxel space of the dynamic reconstruction at time *t*2.

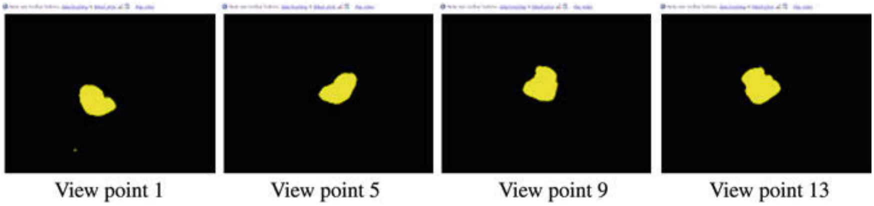

View point 1 View point 5 View point 9 View point 13

Fig. 97.4 The 2D projections of $V(t + 1)$ in images at time *t2*

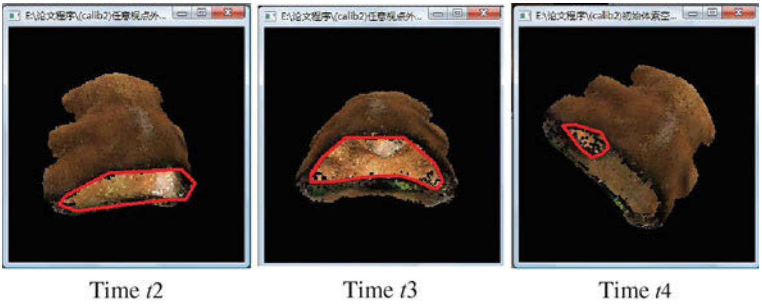

E:\论文程序\(calib2)任意视点外...
E:\论文程序\(calib2)任意视点外...
E:\论文程序\(calib2)初始体素空...
Time *t*2 Time *t*3 Time *t*4

Fig. 97.5 The results of dynamic reconstruction

Thirdly, the dynamic reconstruction results are shown in Fig. 97.5. Obviously, the larger holes in the initial reconstruction are reduced gradually in the dynamic reconstruction comparing with the initial reconstructions in Fig. 97.3.

97.5 Conclusions

In this chapter, a dynamic 3D reconstruction algorithm for a rigid object shape is presented to avoid the holes due to the insufficiency of viewpoints. In this algorithm, the 3D motion parameters are calculated to get the space–time correlation of an object at different time. According to the space–time correlation, the more complete 3D rigid object is reconstructed, and the large holes are reduced in the dynamic reconstruction gradually.

References

1. Cheung GKM, Baker S, Kanado T (2003) Visual hull alignment and refinement across time: a 3D reconstruction algorithm combining shape-from-Silhouette with stereo. In: Proceedings of the 2003 computer society conference on computer vision and pattern recognition, 2:II-375–382

2. Liang C, Wong K-YK (2010) 3D reconstruction using silhouettes from unordered viewpoints. Image Vis Comput 28:579–589
3. Branch J, Prieto F, Boulanger P (2006) Automatic hole-filling of triangular meshes using local radial basis function. In: Proceedings of the third international symposium on 3D data processing, visualization, and transmission, pp 727–734
4. Tong W, Tang C (2006) Multiresolution mesh reconstruction from noisy 3D point set. In: The 18th international conference on pattern recongnition, pp 5–8
5. Ikeuchi K, Miyazaki D (2008) Hole filling of 3D model by flipping signs of signed distance field in adaptive resolution. IEEE Trans Pattern Anal Mach Intell 30(4):686–699
6. Wang J, Oliveira MM (2007) Filling holes on locally smooth surfaces reconstructed from point clouds. Image Vis Comput 25(1):103–113
7. Wang J, Oliveira MM (2003) A Hole-filling strategy for reconstruction of smooth surfaces in range images. In: Proceedings of the XVI Brazilian symposium on computer graphics and image processing, pp 11–18
8. Park S, Guo X, Shin H, Qin H (2005) Shape and appearance repair for incomplete point surfaces. In: Proceedings of the tenth international conference on computer vision, 2:1260–1267
9. Davis J, Marschner SR, Garr M, Levoy M (2002) Filling holes in complex surfaces using volumetric diffusion. In: Proceedings of the first international symposium on 3D data processing visualization and transmission, pp 428–861
10. Xiaoli L, Feipeng D (2010) A rapid method for 3D face recognition based on rejection algorithm. Acta Aumatica Sinica 36(1):153–158
11. Lu X, Jain AK, Colbry D (2006) Matching 2.5D face scans to 3D models. IEEE Trans Pattern Anal Mach Intell 28(1):31–36
12. Chang KI, Bowyer KW, Flynn PJ (2005) An evaluation of multi-modal 2D+3D face biometrics. IEEE Trans Pattern Anal Mach Intell 27(4):619–624
13. Culbertsion WB, Malzbender T, Slabaugh G (2000) Generalized voxel coloring. Vision algorithm 99, pp 100–115
14. Myronenko A, Song X (2009) Point set registration: coherent point drift. IEEE Trans Pattern Anal Mach Intell 32:2266–2275

Chapter 98
The Environmental Footprint of Data Centers: The Influence of Server Renewal Rates on the Overall Footprint

Willem Vereecken, Ward Vanheddeghem, Didier Colle,
Mario Pickavet, Bart Dhoedt and Piet Demeester

Abstract The environmental footprint of ICT is rising. Data centers are key contributors to this footprint. In this Chapter we investigate the influence of the renewal rate of servers on the footprint of the data center. We take into account both the use phase power consumption as well as the contributions of the other life cycle stages. Based on this we construct an analytical model. From the results, we demonstrate that in a scenario where the data center needs to keep up with the increasing processing capacity of the servers, the footprint increases annually and keeping the servers in operation as long as possible is necessary. However, when the capacity remains constant, the footprint is decreasing and an optimal renewal rate is obtained.

Keywords Green IT · Carbon footprint · Data center · Life cycle assessment

W. Vereecken (✉) · W. Vanheddeghem · D. Colle · M. Pickavet ·
B. Dhoedt · P. Demeester
Internet Based Communication Networks and Systems,
Ghent University, IBBT Gaston Crommenlaan 8 bus 201,
9050 Ghent, Belgium
e-mail: Willem.Vereecken@intec.ugent.be

W. Vanheddeghem
e-mail: Ward.Vanheddeghem@intec.ugent.be

D. Colle
e-mail: Didier.Colle@intec.ugent.be

M. Pickavet
e-mail: Mario.Pickavet@intec.ugent.be

B. Dhoedt
e-mail: Bart.Dhoedt@intec.ugent.be

P. Demeester
e-mail: Piet.Demeester@intec.ugent.be

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_98,

98.1 Introduction

In the past years, it has become clear that increasing carbon emissions are a global challenge. In every sector of the economy, initiatives are being taken, and endorsed by governments to reduce carbon footprint. Also in ICT these challenges need to be tackled. Studies [1, 2] have shown the power consumption of ICT is growing even faster than the world's global power consumption, thus being responsible for an increasing fraction of this global power consumption. Data centers are currently responsible for about 1/6th of the ICT footprint. It has already been demonstrated that also their power consumption is increasing [3, 4].

On the other hand, it is estimated that ICT can play an important role in the reduction of the global carbon footprint. Through dematerialization of streams, ICT services enable people to massively reduce their carbon footprint while still fulfilling their needs. In the SMART 2020 report [2], it is estimated that ICT can reduce up to five times its own footprint. It is important to note that these claims will only be achieved when a certain adoption of these technologies is obtained.

Hence, it is essential that the ICT sector can keep its own carbon footprint under control, and even reduce it in the near future in order to be able to support the claims of the carbon footprint reduction capabilities of ICT. The sector itself realizes this and has taken several initiatives to tackle this issue. In the work that is being performed in these initiatives (e.g. [5, 6]), however, the focus is mainly on energy consumption of equipment. However, in order to estimate the full impact of a technology, we need to take into account the full impact of a product's life cycle (i.e. from material extraction until disposal of the product). As we will demonstrate in this Chapter, if the life cycle impact is not taken into account, wrong conclusions might be drawn and efforts to reduce the carbon footprint could lead to an increase of carbon footprint. Vice versa, decisions that seem to increase the carbon footprint could actually lead to overall reductions thanks to other life cycle stages that are more advantageous.

In this work, we will focus on the carbon footprint of data centers. In Sect. 98.2, we will investigate the power consumption of the ICT equipment in a data center. Next, we will relate this power consumption to full life cycle impact of the ICT equipment. In Sect. 98.3 we identify values of the key parameters in the model. Then, in Sect. 98.4, we compare the influence of the different factors on the overall datacenter power footprint. Finally, in Sect. 98.5 we analyze these results and draw the main conclusions.

98.2 Modeling the Footprint of the Data Center

We construct a model that describes the footprint of a data center. We analyze the server power consumption and the related footprint. For this, we consider two scenarios. In the 'constant number of servers' scenario, we assume that the number of servers in a data center remains constant, and every removed server is replaced by a new server. In the 'constant data center capacity' scenario, we assume the processing capacity of the data center remains constant and that, with increasing server capacity, a higher number of old servers is replaced by a lower number of new servers.

98.2.1 Server Power Consumption

We consider a data center. In this data center a number of servers are present. Every year, new servers are brought into the data center. In this model, we assume this happens at the beginning of the year. At the same time, old servers are removed from the data center. We assume that a server is used for n years before being removed. This means, in a given year y, there are servers present from year y, $y-1$, …, $y-n+1$. We denote the number of servers added in year y as N_y. When we denote the average power consumption of a server purchased in year y as P_y^s we get for the total server power consumption in the data center:

$$P_y^{\mathrm{dc}} = \sum_{(i=y)}^{(y-n+1)} N_i P_i^s. \qquad (98.1)$$

We assume an exponential growth rate for the server power consumption. We denote the growth factor as β. This implies:

$$P_{y+i}^s = P_y^s \beta^i. \qquad (98.2)$$

Constant number of servers. First, let us assume that every old server gets replaced by a new server. This means that N_y is constant. Since every year we replace N_y servers which remain in the data center for n years, the total number of servers in the data center is $n \times N_y$, which we denote as N. With (98.2), we get for the power consumption:

$$P_y^{\mathrm{dc}} = \frac{N}{n} P_y^s \sum_{i=y}^{y-n+1} \beta^i. \qquad (98.3)$$

Or, with the formula for geometric series:

$$P_y^{\mathrm{dc}} = \frac{N}{n} P_y^s \frac{1-(1/\beta)^n}{1-(1/\beta)}. \qquad (98.4)$$

Constant data center capacity. In the previous case we assumed the number of servers in the data center to remain constant. However, this does not take into account the fact that server capacity is increasing. From Moore's law, we know that the processing capacity of a server doubles every 18 months. This means, if we replace servers, that we can replace them by a smaller number of servers if we want the processing capacity of out data center to remain the same.

Let us denote the processing capacity of a server as C^s_y. Again, we assume an exponential growth for this capacity:

$$C^s_{y+i} = C^s_y \gamma^i. \tag{98.5}$$

The total capacity present in the data center is C^{tot}. Since this remains constant, and each year we replace a fraction $\frac{1}{n}$, we get:

$$\frac{C^{\text{tot}}}{n} = N_y C^s_y. \tag{98.6}$$

This needs to be valid every year. Hence:

$$N_{y+i} C^s_{y+i} = N_y C^s_y. \tag{98.7}$$

or, with (98.5):

$$N_{y+i} = N_y (1/\gamma)^i. \tag{98.8}$$

Similar to the calculation of (98.4), we get:

$$P^{\text{dc}}_y = N_y P^s_y \frac{1 - (\gamma/\beta)^n}{1 - (\gamma/\beta)}. \tag{98.9}$$

In this formula, we can no longer simplify by eliminating N_y.

98.2.2 Server Footprint

After determining the power consumption of the server, we need to calculate the carbon emissions associated to this power consumption. First we need to multiply the power consumption (in W) with a factor of 8.766 $\left(= \frac{365 \times 24}{1000}\right)$ to get the yearly electrical energy consumption (in kWh). Second, we need to account for the carbon emissions associated to this energy consumption. These carbon emissions are expressed as carbon emission intensity I, i.e. the mass of CO_2 (in g) emitted per used kWh. This emission intensity is dependent on the production technology (e.g. based on oil, gas, charcoal, etc.).

Next to the power consumption, we need to take into account the full life cycle of the server. Life cycle assessment is a field in which the environmental impact of a product or service is measured taking into account the material extraction, production, transportation, use, and disposal. Only the use phase impact is determined by the time the product is operational. Thus, we can model the other life cycle impacts as a single emission at the moment the server is purchased. We denote the non-use-phase impact of a server as L. Every year N_y servers with this footprint are purchased.

Based on the above, we get for the footprint (in kg CO_2) in year y:

$$F_y = N_y L + 8.766\, I\, P_y^{\mathrm{dc}}. \tag{98.10}$$

with P_y^{dc} as denoted by either (98.4) or (98.9).

98.2.3 Power Usage Effectiveness

Next to the power consumption of the servers, other equipment is present in a data center as well. This equipment also consumes power. It is on one hand that other ICT equipment such as switches, storage networks, etc. and on the other hand equipment used for cooling, uninterruptable power supplies, lighting etc. The latter overhead is considered proportional to the ICT equipment power consumption and is expressed by power usage effectiveness (PUE). By multiplying the ICT power consumption with the PUE one obtains the total data center power consumption. The PUE is considered to be approximately two [7].

In many studies, the PUE is also considered in the modeling of the data center. This in itself is valid. However, in this Chapter we want to incorporate the life cycle impact of the servers. If we would incorporate the PUE, this would require accounting the life cycle impact of the other equipment as well. As we consider these impacts out of scope for the effects we wish to model, we do not account for PUE.

98.3 Parameter Value Estimation

After constructing the model, we identify the value of the different parameters in order to be able to draw conclusions.

First, we need to understand how the server power consumption evolves. We used the server power consumption measurements submitted to SPEC using the SPEC power benchmark [8]. In these measurements two values are important. The idle server power consumption P_0 and the full load server power consumption P_{full}. With linear interpolation the average server consumption can be determined in function of the server CPU load λ:

$$P^s = (1 - \lambda)P_0 + \lambda P_{\mathrm{full}} \tag{98.11}$$

In Fig. 98.1 we have displayed P_0 and P_{full} for the data collected in different years. We see that the fraction $P_0 \backslash P_{\mathrm{full}}$ is decreasing and thus the server power consumption becomes more and more dependent on the load. This means the server power consumption will have a different growth rate depending on the assumed load. We assume an average load on the servers of $\lambda = 80\%$, which means that the server capacity in the data center is well used.

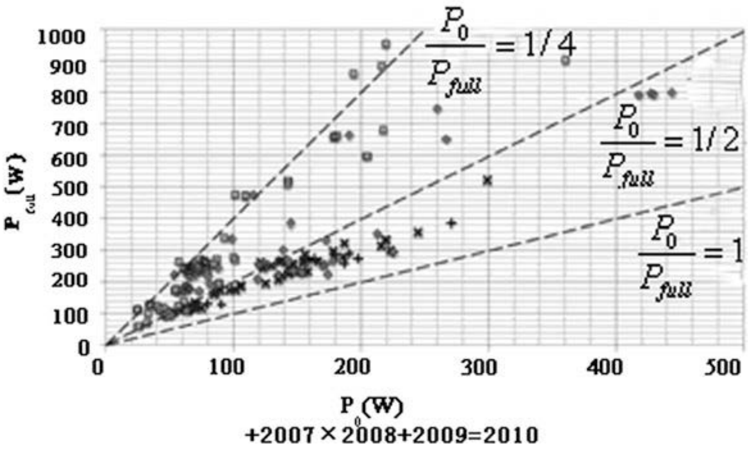


Fig. 98.1 Server power consumption P_0 and P_{full} for different years

Based on this data we performed a linear regression to determine the growth factor of the server power consumption. Note that in the benchmarking results in [8], next to the year of submission, the month is given as well. This allows for a more fine-grained analysis. As a result, we get for the server power consumption in 2007 a value of 187 W. The growth rate of the power consumption is 5.2% p.a. (i.e. $\beta = 1.052$).

The dataset we used to perform the linear regression has the advantage of being a consistent dataset with 217 data points. On the other hand, the servers submitted in this dataset will probably be better performing in terms of power consumption than the average server as the submission of the results imply a consideration for power consumption aspects during the design of the machines.

Second, we model the growth of the *capacity* of servers. Moore's law states that the number of transistors on a chip doubles every 18 months. In the past, this increase was enabled by ever increasing the clock frequency of the CPU. Currently, the trend is to provide multiple cores per server. In this work, we assume Moore's law also describes the increasing processing capabilities of servers. This implies $\gamma = 1.59$ $(= 2^{2/3})$.

The value for the *carbon emission intensity* is based on either production technology of the energy or either the weighted average of the technologies used in the energy mix for a certain location. In [9] we can find values for the several energy production technologies as well as the regional averages. In this Chapter, we assume the world average emission intensity of $I = 504$ gCO_2/kWh.

Finally, we need to determine the non-use-phase footprint of the server. In [10] an estimation is made expressed in Mega-Joules. Again using the world average emission intensity we get $L = 1903$ $kgCO_2$ per server. Note that the operating model of the considered server in [10] is significantly different from the assumptions made in this Chapter.

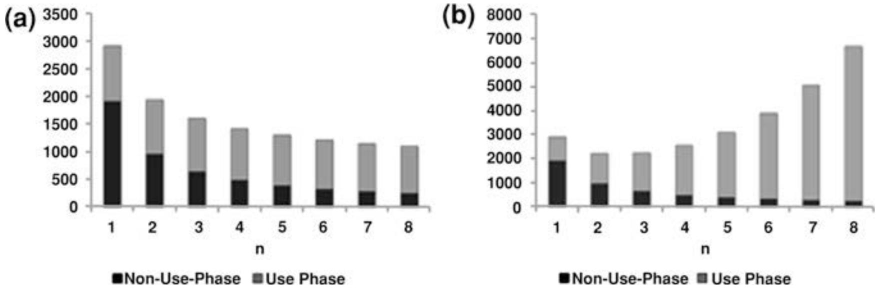


Fig. 98.2 Normalized data center footprint in function of the server renewal rate n in 2011. **a** Constant number of servers, **b** Constant server capacity

98.4 Results

We evaluate the constant server and the constant capacity scenario for a data center in 2011. For the constant server scenario, we normalize the footprint to the footprint of 1 server. This implies dividing (98.4) by *N*.

In the constant capacity scenario, we normalize to the capacity of one server in 2011. This capacity in itself remains undefined but it allows us to make quantitative comparisons. This assumption implies substituting N_y by $1/n$ in (98.9). Note that in these assumptions, for $\gamma = 1$, formulas (98.4) and (98.9) are equal.

In Fig. 98.2, we have displayed the footprint for the considered data center in function of the renewal rate *n*. In Fig. 98.2a we consider the scenario with a constant number of servers. The longer the renewal rate, the lower the footprint becomes. This is because the non-use-phase footprint decreases and a direct consequence of the lower number of replaced servers. Additionally, since the server power increases yearly, the number of servers that is replaced is lower, and the overall power consumption decreases.

In Fig. 98.2b, we have displayed the footprint for the considered data center in the constant capacity scenario. Now, there is an optimum for $n = 2$. Because the server capacity is increasing, every year we need to install less servers. Since the capacity is increasing faster than the power consumption ($/ > 1$), the large amount of old servers outweighs their lower power consumption on the long term. This means, in this scenario, it is best to handle a replacement period of two years in order for the carbon emissions to be minimal.

This analysis could lead to the conclusion that it is better to replace a larger number of servers. When we compare the values for $n = 8$, one notices in the constant server scenario the footprint is a lot lower than in the constant capacity scenario. This is a wrong conclusion, though, and originates in the normalization to the server capacity from 2011. In Fig. 98.3, we represent both cases on a longer term. We now normalize for the year 2007 in the same way as described before

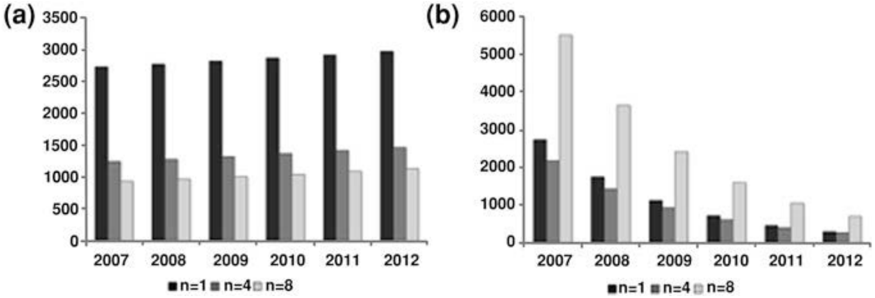


Fig. 98.3 Data center footprint evolution in function of the server renewal rate *n*. **a** Constant number of servers, **b** Constant server capacity

and we look at the evolution of the total footprint until 2012. In the constant number of servers scenario the footprint increases yearly, due to the increasing server power consumption. In the constant capacity scenario, the footprint decreases due to the reduced equipment requirement.

98.5 Conclusions

We evaluated the evolution of data center power consumption. We analyzed two scenarios. One in which we assume every old server needs to be replaced by new server and one where the processing capacity of the data center remains constant.

If during the replacements, the number of servers remain constant, we see that over the years the footprint of the data center increases. In this case it is essential to extend the lifetime of the servers as long as possible.

In the constant capacity case, there is a yearly decrease of the footprint. In this case, we also observe there is an optimal replacement period of about two years.

The driver for the capacity is the software running on the servers. The fact that data center power consumption is still increasing indicates that software is increasingly demanding capacity. As long as this situation persists, it is key to keep servers in operation as long as possible in order to reduce the full life cycle cost of the servers.

It is however important to strive for the scenario in which the software allows for the processing capacity to remain constant. In this case the power consumption of data centers will reduce. At this time however, we need to carefully evaluate the optimal renewal rate for data centers.

Acknowledgments The work described in this Chapter was carried out with the support of the TREND project (“Towards Real Energy-Efficient Network Design”), a Network of Excellence funded by the EU’s Seventh Framework Program.

References

1. Pickavet M, Vereecken W, Demeyer S, Audenaert P, Vermeulen B, Develder C, Colle D, Dhoedt B, Demeester P (2008) Worldwide energy needs for ICT: the rise of power-aware networking. Advanced networks and telecommunication systems, 2008, ANTS '08. 2nd International Symposium on, pp 1–3
2. Webb M (2008) SMART 2020: enabling the low carbon economy in the information age. The climate group
3. Koomey J (2007) Estimating total power consumption by servers in the U.S. and the world. Analytics Press, Oakland
4. U.S. Environmental Protection Agency (2008) Report to congress on server and data center energy efficiency, Public Law 109–431
5. Lefevre L, Orgerie A-C (2010) Designing and evaluating an energy efficient cloud. J Supercomput 51:352–373
6. Vereecken W, Deboosere L, Simoens P, Vermeulen B, Colle D, Develder C, Pickavet M, Dhoedt B, Demeester P (2010) Power efficiency of thin clients. Eur Trans Telecommun 26(6):479–490
7. The Green Grid (2007) Green grid metrics: designing datacenter power efficiency
8. SPECpower_ssj2008 Results. http://www.spec.org/power_ssj2008/results/
9. IEA (2010) CO_2 emissions from fuel combustion—highlights. http://www.iea.org/co2highlights/
10. Hannemann CR, Carey VP, Shah AJ, Patel C (2008) Lifetime exergy consumption of an enterprise server. IEEE international symposium on, Electronics and the environment, 2008. ISEE 2008, pp 1–5

Chapter 99
Traffic Spatiotemporal Data Model on Urban Road Network Under Adverse Weather Conditions

Xi-qiao Zhang, Long-hai Yang and Shi An

Abstract Based on analysis of the applicability of current international spatiotemporal data models and with considering the characteristics of urban road traffic network under adverse weather conditions, the event-based object-oriented spatiotemporal data model is brought forward. First, the intrinsic law of road traffic network is depicted and the traffic supply-demand influential factors under adverse weather conditions are analyzed. After that, the definitions of 'class' and 'objects' are given based on the above analysis, and unified modeling language (UML) are applied to expressing the traffic attributes and relations, the spatiotemporal data modeling on urban traffic network under adverse weather conditions was got. In the end, example analysis shows that this model provides a much more clear and flexible mode for forecasting, management, and decision-making in traffic supply-demand researches. This study also provides a solid foundation for studying urban traffic problems vividly and dynamically.

Keywords Traffic network · Spatiotemporal data model · Adverse weather

Yang Long-hai (1970), male, Anhui Province, Harbin Institute of Technology, associate professor, Ph.D. research area: traffic demand analysis and traffic simulation.

X. Zhang · L. Yang (✉) · S. An
School of Transportation Science and Engineering, Harbin Institute of Technology,
Harbin 150090, China
e-mail: yanglonghai@hit.edu.cn

X. Zhang
School of Management, Harbin Institute of Technology, Harbin 150001, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_99,

99.1 Introduction

In recent years, with the development of urban economy and the spread of urban scope, adverse weather, which influence traffic while it does not destroy traffic infrastructures, has increasing influencing force on urban traffic. For example, in December 2002, a heavy snow in Beijing results in entire paralysis of the urban traffic. Another case in point is rainstorm in Beijing on 10 July, 2004, which resulted in the chaos of urban traffic for six hours. Therefore, it is necessary to adopt up-to-date information technology to analyze traffic conditions under adverse whether condition so as to enhance information level of urban traffic management.

Depiction and modeling on urban road traffic network by abstract manner is the basis for analysis on road network. With the development of science and technology, geographic information system for traffic (GIS-T) is applied as an important method in analyzing and solving traffic problems, which have complex influencing factors, vast data and multiple results. Current data models on traffic network cover node-arc model, linear referencing systems model, navigable data model, spatiotemporal data model, etc. [1, 2]. In the above models, spatiotemporal data model can efficiently represent the dynamic characteristics of road networks, discover the laws of traffic problems, forecast the tendency of traffic problems and realize simulation on traffic process; therefore, it becomes a hot research topic for international researchers [3, 4].

Since Langran and Chrisman proposed temporal geographic information system in 1988, referred to as TGIS, TGIS became a hot topic of geographic information system. Time is omnipresent objective attribute in nature, all information have corresponding tense attributes. The playmaker TGIS is spatiotemporal database, which concept is based on spatiotemporal data model. Based on reference and analysis of previous spatiotemporal data models, adverse weather is taken as event and road traffic infrastructures is taken as object in this paper in order to realize the forecast and analysis on dynamic spatiotemporal characteristics of urban road traffic.

99.2 Modeling Analysis on Traffic Network Under Adverse Weather

99.2.1 Variation Types of Spatiotemporal Data

Spatiotemporal data models should take the process and application of various types into considerations since variation is one of basic characteristics of geographic entity and phenomena. Studies on variation types or basic variation laws of geographic entity conduce to our profound understanding spatiotemporal linguistic

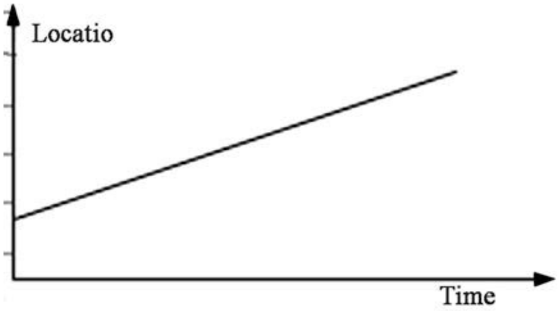


Fig. 99.1 Continuous

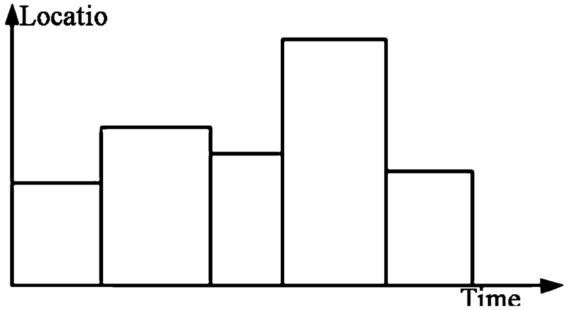


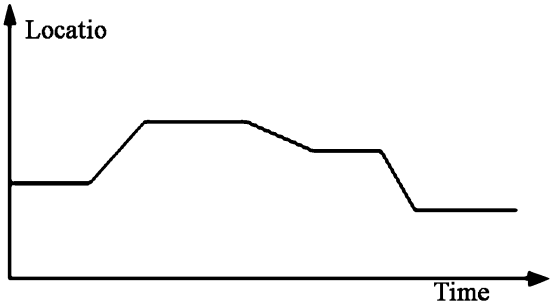
Fig. 99.2 Discrete

meaning of data model. According to variation rhythm, spatiotemporal data models can be classified as the following three types [4]:

(1) Continuous: in this type, spatiotemporal object can been deemed as forever moving substance, such as water or air follow, and such information as its attribute, figure, etc. changes continuously, as is shown in Fig. 99.1.
(2) Discrete: spatiotemporal object of this type always in quiescence state, but this state will be broken when certain geographic event takes place. Here, the spatial location and attribute of spatiotemporal object are likely to change, as is shown in Fig. 99.2.
(3) Stepwise: spatiotemporal object of this type belongs to quiescence state and movement state at times, for example, population, traffic vehicle, etc. Characteristics of spatiotemporal object of this type are represented only in variation of spatial location, and its attribute and shape retain unchanged, as is shown in Fig. 99.3.

Spatiotemporal data models discussed in this paper aim at tense problem of road traffic network under adverse weather conditions. According to variation characteristics of attributes of roads and traffic under adverse weather, tense problems of road traffic network management accord with above second type.

Fig. 99.3 Stepwise



99.2.2 Relationship Between Event and Time

In 1984, the concept of event is introduced by Copeland and Maier [5], they believe that it is event that leads to the change of object' characteristics and it has necessary relationship with new characteristics caused by change. We can believe that event takes place in a certain time or time span, and time attribute should be included when depicting the event. Variations in a thing' life cycle is a developmental process from one state to another and this process is driven by events.

'Event time' is defined as the time when spatiotemporal object change in real world. Time granularity is used to express the discrete degree of time memory. Time granularity is the unit of measurement for time record and it is the basic unit for time denotation and inference. Selection of time granularity should accord with concrete applications, and time granularities with various precisions will influence the expression and calculation of time data structure.

There is a corresponding relationship between event and time. The occurrence of one event corresponds to beginning and end time of the event. Let E denote the set of events, $E = \{e_1, e_2, e_3 \cdots e_n\}$, and T denote the set of times, $T = \{t_1, t_2, t_3 \cdots t_n\}$, t can be understood as time spot or time span, then $\forall e \in E, \exists t \in T$ and there is a only one t. Current realization approaches of tense in spatiotemporal data models can be classified as two kinds, one is taking time as time dimension to the models and another is considering time as one of attributes. The latter is adopted by in this paper, and the occurrence of adverse weather is taken as beginning time, then granularity selection is carried out to partition time span.

99.2.3 Object-Oriented Spatiotemporal Topology Relation

Object-oriented technology in spatiotemporal data modeling was first cared by Michael F. Worboys and Donna J. Peuquet [6]. Although there is no common recognition on object-oriented modeling concept, theoretical bases and

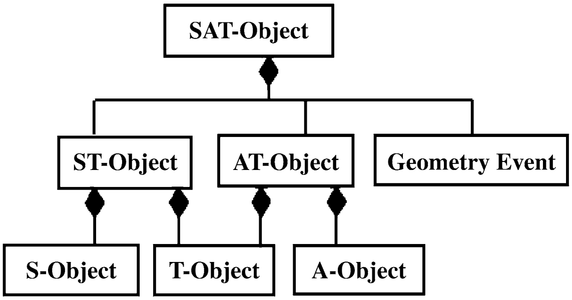


Fig. 99.4 Object-oriented spatial–temporal data model based on event

implementation technologies, the powerful expression ability of this technology is highlighted. The basic modeling unit is object, no matter how complex the entity is, it can be denoted by one object and relation among object can be set up by object marks [7]. Object-oriented modeling can directly express one-to-many relation. It can support not only record with various lengths but also object set and can depict object from the perspectives of geometrical information, special topic, semantic information and tense information. Object-oriented models make it easier for users incepting and understanding upper logic concepts, and hence avoid too fussy technological details so as to make models convenient to be set up and applied.

The kernel of object-oriented models is to organize geography space time according to basic ideas of object-oriented. The objects are conceptual entities with independent encapsulation and unique mark. Tense, space characteristics, attribute characteristics, related behavior operation and other object' relations are encapsulated in every spatiotemporal object. Based on object-oriented viewpoints, object-oriented construction manners are adopted in this paper. According to different geometry characteristics (spot, line, profile) of space objects in road traffic system, road entities can be designed respectively into different object classes. Time signs are earmarked on various data structure units so as to forming spatiotemporal object classes, which are bases for models design.

99.2.4 Object-Oriented Spatiotemporal Data Model

Tense GIS spatiotemporal object (SAT-Object) [8–10] (Fig. 99.4) is synthesized by spatial object with time characteristics (ST-Object), attributes objects with time characteristics (AT-object) and event object class. we get events happening at a certain time or a certain period though accurately expressing the form of spatiotemporal objects at one time or a certain period, at the same time it polymerize of the event object using this modeling method, and we can better express spatiotemporal semantics.

99.3 Data Models on Traffic Network in Adverse Weather

99.3.1 Data Structure of Spatiotemporal Objects

According to the previous modeling analysis, spatiotemporal data model in this paper will be set up based on events and object-oriented theory. Event time dimension is taken as a kind of attribute adding to every space object, and hence ordinary data structure can be denoted as <Obiect:{ID, Attr(*t*), Spatial(*t*), Temporal(Tv, Td), Actions} > [6], in which ID is the mark of traffic infrastructures and it can denote a spatiotemporal object uniquely; Attr(*t*) denotes non space attributes that change along with time; while Spatial(*t*) denotes space attributes that change along with time; Temporal (Tv, Td) denotes time attributes of objects and reflects objects' produce, state change and wither away process, where efficient time Tv and affair time are orthogonal; Actions denote operations on objects that include various calculative operations in defining time, space and attributes of objects, realizing interactions among homogeneous or inhomogeneity objects, so as to tightly interrelate objects' data and operations as depicted as Fig. 99.5.

99.3.2 Objects Composition of Road Traffic Network

Analyzes of road traffic under adverse weather are based on huge information analyzes work, the entities in the information can be abstracted as various spatiotemporal objects and further classified to three classes that are spot, line, and profile from the perspective of geometry, and road users and vehicles can be seen as one kind of behavior attributes of space entities, furthermore, when attach time marks to various data structure units, spatiotemporal objects class will come into being [11, 12]. By applying unified modeling language (UML), various objects and their relations in the event-based object-oriented spatiotemporal data models can be depicted by Fig. 99.6. Linearity reference benchmark is applied in space depiction. As the entity of profile class, traffic sub-zone is surrounded by boundary arc sections; as line entity, road section is composed by one or many arcs, which can be further depicted by nodes and reference spots. In this depiction, distance of arc sections, distance among reference spots and distance among nodes of road networks are taken as systematic parameters, moreover, co-factor matrix of pending parameters, which is irrespective of observation value, is applied in operational design on linearity reference benchmark and control. In certain space range, such as traffic network, various space entity objects can constitute a complex object, that is to say this object is composed by nodes (reference spots, intersections), line (road sections, road) and profiles (traffic sub-zones) [13, 14].

Fig. 99.5 Data structure of spatiotemporal objects

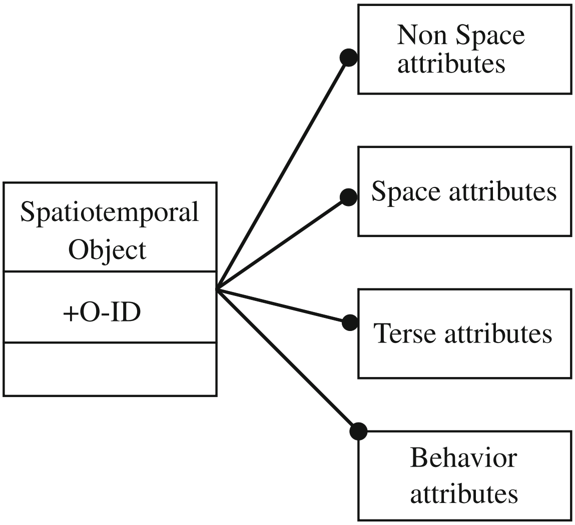
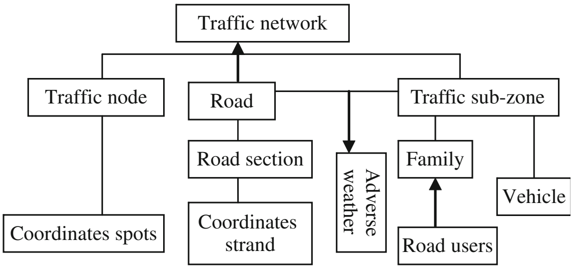


Fig. 99.6 Spatiotemp oral data models class of road traffic network under adverse weather



99.3.3 Example Depictions of Spatiotemporal Data Models

In the construction of ‘urban road traffic management system answering for adverse weather’, event-based object-oriented spatiotemporal data models are applied into the system design and obtained favorable results [15]. Take the snowfall weather for example tense is divided into three sections: time of snowfall, time of icy road, time of recovered road. The key of systematical design lies in labor on each time section. Table 99.1 shows some partial information of certain road section in each time section of snowfall weather.

The road capacity in Table 99.1 is the basic data for calculation of traffic supply ability, and it is obtained by fuzzy set according to apace information, such as gradient of road section, and traffic condition that change along with event time, such as friction coefficient or visibility.

Adverse weather is the precondition of ‘urban road traffic management system answering for adverse weather’ researches, which take road supply ability and traffic demand as research objects to study the dynamic equilibrium relation of

Table 99.1 Analyzes results of road section under snowfall weather

RoadID road name	Attr_1(*t*) road capacity	Attr_2(*t*) friction coefficient	Attr_3 (*t*) visibility	Temp_1 road gradient	Ts start	Te end	Other
3079	2700	0.55–0.6	1.0	0.05	–	–	Normal
	1500	0.25	0.6	0.05	12–17 9:20	12–17 13:00	Snowfall
	1450	0.17–0.22	0.8–0.9	0.05	12–17 13:00	12–18 9:00	Icy
	2170	0.3–0.35	0.9–1.0	0.05	12–18 9:00	12–18 15:00	Recover

traffic supply and demand under adverse weather so as to realizing efficient management and inquiry on spatiotemporal data, for example, forecasting traffic jam in a certain road section for a certain time, or carry through traffic demand planning and traffic system planning according to road supply forecasting, or road selection for road users.

99.4 Conclusions

Researches on variation laws of event-based spatiotemporal data provide theoretical foundation for design of spatiotemporal data models. A kind of event-based object-oriented spatiotemporal data model is brought forward in this paper. First, based on the event of adverse weather, this model disposes time in a discrete manner, and then determine time granularity and hierarchical structure of objects. Second, depiction is given to data structure of road objects and illumination is shadowed on various objects of road objects and their relation by applying UML language. Finally, based on analytical structure of spatiotemporal data models, design for urban road traffic management system answering for adverse weather is put forward, which can realize spatiotemporal analysis function with practical importance such as traffic condition inquiry, dynamic traffic management planning, road selection, etc.

Acknowledgments This work is supported by China Postdoctor Foundation (No. 20090450857), by The Ministry of Education of Humanities and Social Science Research Youth Foundation (No. 10YJC630391), by Heilongjiang Province Youth Science Foundation (No. QC2009C25), by Natural Scientific Research Innovation Foundation in Harbin Institute of Technology, by The Ministry of Education of Specialized Research Fund for the Doctoral Program (New Teachers Fund), by Heilongjiang Province Postdoctor Foundation.

References

1. Miller HJ, Shaw SL (2001) Geographic information systems for transportation: principles and applications. Oxford University Press, Oxford
2. Shi L-j, Xu G-h, He M, Song Y (2004) Research progress on GIS-T data models. J Beijing Univ Technol 30(3):318–322 (in Chinese)

3. Shaw S-L (2000) Moving toward spatiotemporal GIS for transportation applications. In: Proceedings of the twentieth annual ESRI user conference. vol 1420. Building Knoxville, pp 205–210
4. Donggen W, Tao C (2001) A spatio-temporal data model for activity-based transport demand modeling. Geogr Inf Sci 156(6):561–585
5. Yu H, Shaw S-L (2004) Representing and visualizing travel diary data: a spatio-temporal GIS approach. In: 2004 ESRI international user conference, San Diego
6. Zhang S-S (2003) Object-oriented spatiotemporal data model of urban transportation planning. Comput Appl 23(6):56–59 (in Chinese)
7. Cao Z-Y, Liu Y (2002) An object-oriented spatio-temporal data model. Acta Geodaetica et Cartographica Sinica 31(1):87–92 (in Chinese)
8. Yansheng L, Chuan Q, Bo T (2003) A new object-oriented spatio-temporal data model based on attribute-bit. J Huazhong Univ Sci Technol (Nat Sci) 31(3):52–54 (in Chinese)
9. Fonseca F, Davis C, Camara G (2003) Bridging ontologies and conceptual schemas in geographic information integration. Geoinformatica 7(4):355–378
10. Pelekis N, Theodoulidis B, Kopanakis I, Theodoridis Y (2004) Literature review of spatio-temporal database models. Knowl Eng Rev 19(3):235–274
11. Elias F, Kostas G, Nikos P et al (2005) Nearest neighbor search on moving object trajectories. Advances in spatial and temporal databases. Springer, Berlin, pp 328–345
12. Parent C, Stefano S, Esteban Z (2006) Conceptual modeling for traditional and spatio-temporal applications: the MADS approach. Springer, Berlin, pp 428–431
13. Worboysm F (2005) Event-oriented approaches to geographic phenomena. Int J Geogr Inf Sci 19(1):1–28
14. Reitsma F, Albrecht J (2005) Implementing a new data model for simulating processes. Int J Geogr Inf Sci 10(19):1073–1090
15. Mcintosh J, Yuan M (2005) Assessing similarity of geographic processes and events. Trans GIS 9(2):223–245

Chapter 100
The Research on Cultural Algorithm of Shortest Path in the Digital Map

Yan Liu, Zhaosheng Yang and Xiufeng Han

Abstract For the current problem of urban traffic congestion, the topology structure is proposed in this paper, based on GIS using cultural algorithm to solve the shortest path problem. The simulation results show that the actual optimal path of the improved algorithm is superior to the traditional optimal path algorithm in speed and accuracy.

Keywords Topology structure · Cultural algorithm · The shortest path

100.1 Introduction

With the development of China's economy, urban transportation is playing a significant role in promoting the exchange of resources. According to statistics, the annual losses due to traffic congestion is 5–8% in China. Thus, a reasonable path planning is necessary for traffic decongestion. Vehicle routing problem first

Y. Liu (✉)
Institute of Computer Science and Technology, Changchun Normal University,
and College of Traffic and Transportation, Jilin University,
Changchun 130022, Jilin, China
e-mail: liuy78@126.com

Z. Yang
State Key Lab of Automobile Dynamic Simulation, Jilin University,
Changchun 130022, Jilin, China
e-mail: yangzs@mail.jlu.edu.cn

X. Han
Changchun City Planning, Changchun 130011, Jilin, China
e-mail: 16554912@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_100,

concerned Dantzig and Ramser in 1959. Thereafter, domestic and foreign scholars researched on a wide range of vehicle routing problems with in-depth research, and found that in vehicle routing problem the main problem is the shortest path problem. The shortest path not only geographically refers to the shortest distance, but can also be applied to other parameters, such as time, cost, traffic, and so on. Accordingly, the shortest path problem has become the fastest, lowest cost problem.

Currently, the algorithm which solves the shortest path problem has about 17 species, three of whose results were better. They are: TQQ (graph growth with two queues): TQQ graph growth algorithm is based on the theory for the calculation of a single point to all other shortest distances between points; DKA (the Dijkstra's algorithm implemented with approximate buckets) and DKD (the Dijkstra's algorithm implemented with double buckets) is based on the Dijkstra algorithm for computing the shortest path between two points [1]. The shortest path of the vehicle routing problem between two points considers not only the shortest path, but also the weather, road quality, traffic, and many such uncertain factors. The cultural algorithm is a multi-population evolution based on the calculation model for the evolution of search mechanism and knowledge of storage which provides a framework for the combination. The writer stored the GIS data in the form of the adjacency matrix to build the topology structure of digital maps; on this basis, the cultural algorithm is used for solving the optimal value of the actual road path.

100.2 Construct Digital Map Topology Structure

GIS data (such as roads) to the shortest path calculation must first be at the relationship between nodes and arcs in the abstract, which is called building GIS network topology.

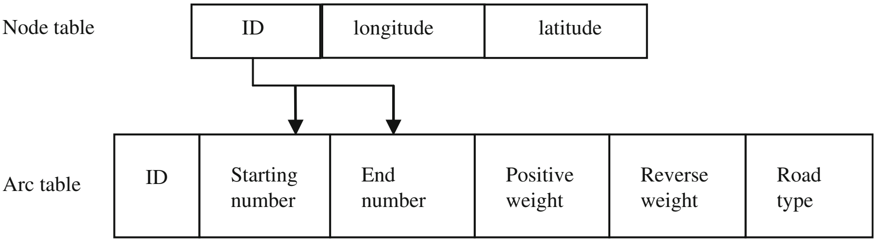
100.2.1 Abstract Digital Map

The network topology which is extracted from vector map abstract between the section of road and intersection layer nodes and arcs by the relationship can use the following method.

Object data table from the road extracts sections of code, the initial node identity, end node identification, length information, and generate arc topology table;

Object data from the node table, extracts the node to join node topology data identifies the table, and checks for duplication, if any, to remove duplicate endpoint information from node to extract the corresponding object data table identifying the longitude and the node latitude, generated node topology table (Table 100.1).

Table 100.1 Node table and arc table format



Node topology table include: the identification number of nodes (ID), unique identifier nodes; longitude, longitude coordinates of nodes; latitude, latitude coordinates of nodes.

Arc table: including the identification number of arc (ID), that is, section ID, which is the only positive integer that identifies the arc; number and the starting point of the arc end number, the identification number of the corresponding node; positive arc Reverse weight and type of road attribute table corresponding to the weight and type.

100.2.2 The Establishment of Digital Map Topology

The adjacency matrix stores the sparse road network, the disadvantage is large and low efficiency of the storage space occupied. Using the neighbor table does not have this problem, and can show large-scale transportation network. So it uses the network topology adjacency list structure; the two tables are used to represent the network topology in this article. One of the tables are used to store the arc-related data, called the arc table, and the other table is used to store and associate data node, called the node table.

All sections of the network up to the starting point for the order are arranged in an array; the starting points in the same section can be in any order. Adjacency matrix of the array section is similar to the compression storage, its content is adjacent to multiple tables with the characteristics of an arc that is expressed in two nodes. The attribute data of section (such as starting point, end point, weight, etc.) is stored in sections of the property sheet. Nodes equivalent to a set of end points of the degree of record index table, get this through out-degree nodes connected to it and the first section of the position. Groups in the nodes, which are the first node I elements and the corresponding node i are stored in out-degree of node i which is as a starting point in the section of the first section of the location [2].

100.2.3 Algorithm Complexity Analysis

Assume the road network to have n road sections and n nodes; it produces sections of segment table and nodes of any one section of the table to determine when the endpoint is a repeat of its nodes when nodes need to traverse the entire table up to m times, so that the time complexity degree O ($n \times m$). When the data structures for each node need to find the section that contains it to traverse the whole road section of the table, so the time complexity is O ($m \times n$), the total time to create topological complexity is O ($n \times m$).

100.3 The Cultural Algorithm for Solving the Shortest Path

100.3.1 Computational Framework for Cultural Algorithm

Cultural algorithms have two major evolutionary spaces [3]: the belief space which is in the evolutionary process from the experiences and knowledge acquired; and the population space which is a group of individuals from the specific space. Its computing infrastructure is shown in Fig. 100.1, which is the space of two specific protocols through the exchange of information.

The main idea of the algorithm is: the individual groups in the evolutionary process space, the formation of individual experience, through the accept () function will be delivered to the belief space of individual experiences, beliefs, individuals will receive the experience of space according to certain rules of behavior compared and the optimization of group experience and the experience under the belief space of existing and new situations of individual experience with function update () which updates the group experience. Belief space in the formation of the experience of space through the influence () function modifies the rules to allow the evolution of individual space for higher efficiency on the behavior of individual groups of space.

As can be seen from the above analysis: cultural algorithm to solve constrained optimization problems is the key to express and handle the constraints, to get from the individual experience of space groups, belief space to refine the experience and knowledge to solve problems and to guide and improve population space; namely, how to design population space and belief space.

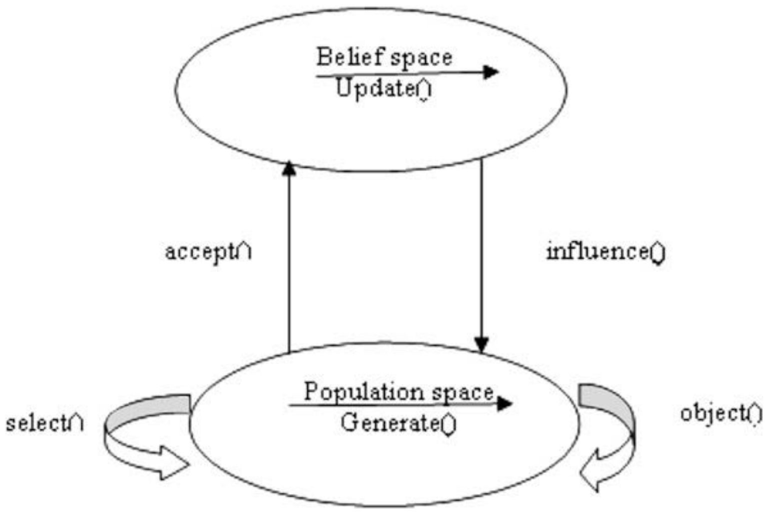


Fig. 100.1 Cultural algorithm framework

100.3.2 Algorithm Analysis and Design

100.3.2.1 Design Population Space

In this paper, the writer embeds ant colony algorithm into the cultural algorithm [4–6] framework. The writer presents a bionic intelligent algorithm to solve travel route planning.

Let m be the number of ant colony systems, the initial moment in the algorithm, the m ants randomly distributed to the n cities, U on behalf of the ant k point i of departure from the feasible point set (in a way finding process, traverse city has been removed from the collection), each ant randomly from the starting point to embark on the transition probability p_{ij} under proportional rule to choose to move to the cities. In t time ant i moves from city k to the city j on transfer rules:

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{[\tau_{ij}^{k}(t)]^{n}\times[\eta_{ij}(t)]^{\beta}}{\sum\limits_{i\subset U}^{n}[\tau_{ij}(t)]^{\alpha}\times[\eta_{ij}(t)]^{\beta}}, & \text{if j} \in U\\ 0, & \text{otherwise}\end{cases} \tag{100.1}$$

where $U = \{1,2, ..., n\}$—tabu$_k$ that ants choose the next step to allow all the cities, the list tabu$_k$ records the current ant k traversed the city, when all the cities n are added to the tabu$_k$ when, ant k has completed a cycle, then the path of ant k passed a candidate solution is the problem; τ_{ij} (t) is time t edge (i, j) on the pheromone; η_{ij} (t) is time t edge (i, j) the visibility; α, β are two parameters, representing the information heuristic factor and expectations of the heuristic factor. When all the ants are after the successful completion of a path finding, the amount of pheromone on the path according to Eq. 100.2 is to adjust

$$\tau_{ij}(t+n) = (1-\rho) \times \tau_{ij} + \Delta\tau_{ij}(t) \tag{100.2}$$

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}(t) \tag{100.3}$$

$$\Delta\tau_{ij}^{k}(t) = \begin{cases} \frac{Q}{L_k}, & \text{if the ants } k \text{ in this loop through the } (i,\ j) \\ 0, & \text{otherwise} \end{cases} \tag{100.4}$$

where: ρ—evaporation coefficient of pheromone, 1-ρ—residual factor pheromone; Q—pheromone strength, which to some extent affect the convergence speed; L_k —K-ants in this cycle by taking the total length of the path.

100.3.2.2 Design Belief Space

The core of belief space is updating and description for knowledge. Belief space in the evaluation of the receiving function from the individual population spatial sampling, sample extraction using knowledge update function implicit information carried by the individual to be summarized in the form of knowledge, describes the experience and memory and behavior problem solving [4, 5]. Function obtained by taking the optimal solution, a two-stage method can be a further optimized solution. As the road network is a more complex network, you can use 3-OPT algorithm to further optimize the path to shorten the length of the path accelerating the convergence speed of the ant colony algorithm. Set the path to any three points i, j, k, the current best path is: cs···cici + 1···cjcj + 1···ckck + 1···ct,if d(ci, ci + 1) + d(cj, cj + 1) > d(ci, cj) +d(ci + 1, cj + 1), the elements of the path (cj, ···, ci +1) is the reverse order; If d(cj, cj + 1) +d(ck, ck + 1) > d(cj, ck) + d(cj + 1, ck + 1), the elements of the path (ck, ···, cj + 1) is the reverse order.

100.3.2.3 Specific Steps of the Algorithm

Knowledge of the population space changes to these steps: [7]

Initialization parameters, so that time $t = 0$ and the number of cycles Nc = 0, set the maximum number of cycles Ncmax, so that for each road section (i, j) the initial pheromone τij (0) = C, C is a constant, $\Delta\tau ij$ = 0;

Will be placed in each ant starting point O;

According to state transition rules to select the next node, and a local pheromone update. Pheromone updates the scope of no more than [τmin, τmax];

Repeat (3) steps, until all the ants reach the terminal;

Record update the global optimal solution, and by accept () function is passed to the belief space;

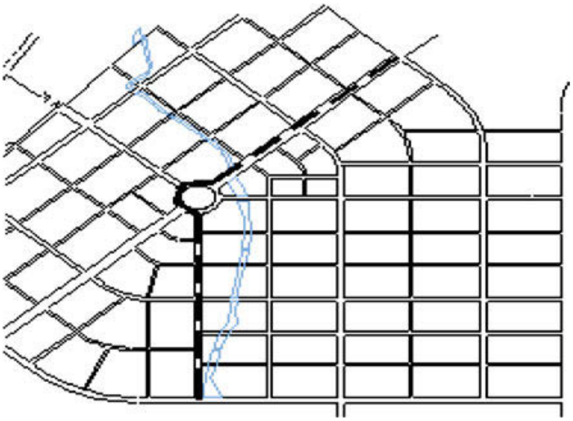


Fig. 100.2 An urban road traffic network with the shortest path

The global pheromone update;

If each ant’s path is a path algorithm it converges to the same end;

If the number of cycles Nc $\geq$ Ncmax, then end the algorithm; otherwise go to (2).

Knowledge of the belief space changes to these steps:

Optimal solution of population space will be passed to the belief space;

Optimal path for the exchange operate 3-OPT;

If the first t iterations, the path length is after the crossover operation L (t), if L (t) is less than L ($t-1$), then return after the first cross-t path, or return to the $t-1$ generation of the Optimal path, t++.

100.4 Test

In order to verify the validity of the algorithm, we use VC programming language [8] to implement the algorithm. Experimental data is the actual data in an urban road traffic network (Fig. 100.2).

100.5 Conclusion

This digital map from the shortest path to solve the problem introduced the establishment of digital map topology, and the cultural algorithm for solving the shortest path, and the actual path planning algorithm is applied. Experiments show that the algorithm can efficiently alleviate the premature convergence of traditional ant colony algorithm in the local optimal solution, slow convergence and other shortcomings, can be better applied to the actual GIS project development. As each city is different, the Algorithm is only suitable for certain cities.

References

1. Wang L, Duan J, Wang B (2005) Algorithms research and simulation of shortest path in GIS. Comput Simul 22(1):117–120
2. Xiong S, Zhao P, Li J (2005) The urban road network topology automatically generated in mapinfo. Control Technol 24(3):67–68
3. Reynolds RG, Shinin Z (2001) Knowledge-based function optimization using fuzzy cultural algorithms with evolutionary programming. IEEE Trans Syst Man Cybern Part B: Cybern 31(1):1–18
4. Guo Y, Wang H (2009) Summary of cultural algorithms. Comput Eng Appl 45(9):41–46
5. Qi Z, Liu M (2008) The research of cultural algorithm. Comput Technol Dev 18(5):126–130
6. Liu S, Zhu F, You X (2010) Research progress of ant colony algorithm for solving TSP problem. Comput Eng Des 31(14):3274–3276
7. Xue X, Nan Z, Zhao W (2011) Cultural algorithms based on improved ant colony algorithm for optimal routing problem. J Jiamusi Univ (Nat Sci Ed) 29(1):54–57
8. Microsoft Corporation (1998) Visual C++ 6.0 programmer's guide

Chapter 101
The Denoising of Enhanced Edge Image Based on Wavelet

Wenxian Xiao, Junhui Fu, Zhen Liu and Wenlong Wan

Abstract Some of the traditional noise suppression techniques are often at the expense of image edges and details. In order to remove image noise and preserve image edge and texture details well enough at the same time, the method of using edge detection was proposed to detect image edge and texture details. Making it fused with the image and decomposed with noisy image is by using the second generation wavelet to denoise the image high-frequency adaptively. The simulation results show that the denoising method is superior to the traditional wavelet threshold denoising method.

Keywords Wavelet transform · Edge enhancement · Image denoising

101.1 Introduction

With the increasing improvement in wavelet theory, it gets more and more attention with its own good time–frequency characteristics of denoising in the image areas, but in the practical application process, the transform of traditional wavelet has defects of large amount of calculation, large consumption of storage space and floating point calculation [1, 2]. To this end, in 1994, Sweldens proposed a Fourier transform not relying on new wavelet construction method—improving methods of wavelet transform which is based on lifting scheme and is called the second generation wavelet transform (second generation wavelet) [3, 4]. The improving method does not depend directly on the Fourier transform but completes the

W. Xiao (✉) · J. Fu · Z. Liu · W. Wan
Henan Institute of Science and Technology,
Xinxiang 453003, Henan, China
e-mail: Xwenx@yeah.net

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_101,

calculation in spatial area and situ calculation (no additional storage space), and has the characters of realizing the transformation from integer to integer easily and a smaller amount of calculating [5], being a new hot spot of research and application of wavelet.

Although the traditional wavelet denoising method to denoising techniques was introduced into the frequency domain, to a certain extent, and retained a large high-frequency coefficients in the process of denoising, those small edges and textures of the wavelet coefficients are reduced or even removed. Therefore it cannot maintain the edges and details very well.

In this paper, characteristics of the second generation wavelet transform to remove the noise of the images and at the same time retain the image edge and texture details of the method are given. To this end, taking into account before denoising method using edge detection to detect the edges of noisy images and texture first and making it fused in a certain percentage with noisy image, then decompose the images which have been fused in the second generation wavelet transform and denoise image high-frequency adaptively, finally threshold the wavelet coefficients after reconstruction and get the denoised image.

101.2 Principles of the Second Generation Wavelet Transform

Improving the method gives a simple and effective construction method of biorthogonal wavelet, using polynomial interpolation to obtain the basic high-frequency signal components, and then by constructing a scaling function to obtain the signal of the low frequency components. The basic idea of improving method is to decompose the existing wavelet filter into the basic building blocks, and complete wavelet transform step by step. In 1998, Daubechies and Sweldens proved that any wavelet can be used to enhance the program [6] to achieve. The improving method divides the process of wavelet transform into the following three stages [7–10].

Splitting. The input signal sj is splinted into two disjoint subsets $\mathrm{sj}-1$ and $\mathrm{dj}-1$, the commonly used algorithm is that the input signal is divided into two subsets according to parity,

$$\mathrm{Split}\ (\mathrm{sj}) = (\mathrm{sj1}, \mathrm{dj1}) \tag{101.1}$$

Forecasting. For the correlation among the data, using the predictive value of $\mathrm{sj}-1$ p $(\mathrm{sj}-1)$ to predict the odd sequence of $\mathrm{dj}-1$. That is, after making the dual role filter of palooka it as the predictive value of the odd signal. The actual value and the predictive value of odd signals are subtracted as the residual signal. If using a subset of the $\mathrm{dj}-1$ and the predictive value of p $(\mathrm{sj}-1)$ the difference to replace $\mathrm{dj}-1$, then this value of difference reflects the degree of approximation between the two. If the forecast is reasonable, then the information of the distinct data sets is much less than the original subset of $\mathrm{dj}-1$. The expression of the forecasting process is

$$\mathrm{dj}-1=\ \mathrm{dj}-1-p(\mathrm{sj}-1) \tag{101.2}$$

Updating. As they are broken down into subsets, some of the features of the original are lost; by updating the data of the subset the same features with the original data of collection remains. Generating a better sub-data set sj − 1 by Operator *U*, so that some features of the original data sets of sj remain. The expression of sj − 1 is

$$\mathrm{sj}-1=\ \mathrm{sj}-1+U(\mathrm{dj}-1). \tag{101.3}$$

Lifting scheme is a fully reversible process; the improving formula in the reconstructing of the data is the same as the formula and decomposition, only needing changing order and symbolic computation.

$$\mathrm{sj}-1=\ \mathrm{sj}-1-\ U(\mathrm{dj}-1), \tag{101.4}$$

$$\mathrm{dj}-1=\ \mathrm{dj}-1+\ P(\mathrm{sj}-1), \tag{101.5}$$

$$\mathrm{sj}\ =\ \mathrm{Merge}\ (\mathrm{sj}-1,\ \mathrm{dj}-1)), \tag{101.6}$$

To merge is to combine. That is, reconstructing a split subset dj−1 and sj−1 into the original signal sj.

101.3 Image Denoising Based on the Second Generation Wavelet

101.3.1 Denoising Model

Suppose *T* (*i*, *j*) for the noisy image which can be expressed as an *M* × *N* matrix:

$$T(i,j)\ =\ F(i,j)\ +\ e(i,j),$$

$$i=0,\ 1,\ \ldots,\ M-1,\quad j=0,\ 1,\ \ldots,\ N-1,$$

where *F* (*i*, *j*) of the original image, *e* (*i*, *j*) is stationary zero mean white noise. At present, the implementation of fractional calculus of computation methods of solving a variety of commonly used methods are mainly geometric approximation method and estimate correction method. Because geometric approximate potter figure approximation method in fitting frequency interval at both ends of the existing biggish error is easy to cause the frequency response distortion, so in discussions, fractional order nonlinear systems such as chaos pseudo exist complex phenomena that may appear chaotic, a growing number of engineering and technical personnel began to consider using a more reliable numerical study method journal of commonly used methods is one of the generalized

Adams–Bashforth–Moulton method. Next, to facilitate further analysis, this paper was the first to adopt such guesstimation correction methods.

101.3.2 Enhancing Edge Information

Denoising and the reservation of edge detail are the two existing problems of denoising. The purpose of denoising is to remove noise, especially those present in high-frequency noise, but the details of the image are also in the high frequency part of signal. Traditional denoising techniques such as denoising of the time-domain wiener filter are at the expense of reducing image edges and detailers, in order to denoise and at the same time to preserve image edge information well. This paper works on edge enhancement before denoising of images.

Derivative of the function reflects the significant changes in the image intensity level. The first derivative of the local maximum value and zero-crossing of the second derivative of image intensity changes are a great place. Therefore, these derivative values can be the border strength as the corresponding point of the method by setting the threshold and extract edge.

Soberts operator is a partial differential operator by the operator to find the edge; it deals with gray shades image processing and noise better .

Using soberts edge detection operator, to detect edges in noisy images $G\ (i, j)$, and $G\ (i,\ j)$ and the noisy image $T\ (i,\ j)$ for the wavelet transform, wavelet transform coefficients obtained their $W * g\ (i, j)$ and Wt (i, j), then $W * g\ (i, j)$ and $Wt\ (i,\ j)$ were fused, in order to avoid excessive edge enhancement, adding the adjustment factor k,

$$N(i,j) = T(i,j) + k * G(i,j), \tag{7}$$

where $N\ (i,\ j)$ for the fused wavelet coefficients, the $N\ (i,\ j)$ edge enhanced reconstruction of noisy image $N^*\ (i,\ j)$, using the second generation wavelet transform, the shareholding.

This paper will, by using this algorithm, point out of new nearly mention chromatography a hyperchaos class system corresponding to fractional chaotic system dynamics and synchronous control problem.

The classical differential equation theory believed that autonomous system to produce the minimum-order chaotic project should not be <3 times, when the introduction of the concept of fractional derivative, the autonomous system to produce chaos head into a minimum order number can be less than three, or even smaller. Previous studies showed that: power system has the integral of fractional order dynamical systems that are not characteristic, for example, fractional order Chua circuit in the order of 2.7 to the possible chaos, such an interesting phenomenon has attracted many Physics, Mathematics, and Engineering and Technical personnel of interest. Currently, the fractional order chaotic system control and anti-control theory in the ascendant has become a frontier field in the nonlinear science field.

101.3.3 Image Denoising

By using the second generation of two-dimensional (2D) image of wavelet decomposition n level ultimately there will be $(3n + 1)$ different frequency bands, which contain $3n$ high-frequency bands and a low frequency. After the noisy image was decomposed by the wave with multi-scale, the noise energy is mainly distributed in the order of high-frequency sub-band and wavelet coefficients in the low-scale, having a high proportion of noise energy; as the wavelet transform reduced, the growth series gets small, so the appropriate threshold is chosen in the high-frequency sub-band to filter out the noise [11–13].

As the traditional soft threshold denoising using a single threshold on wavelet coefficients are processed, the wavelet coefficients of different scales use the same threshold, but a single threshold function will not give a good separation at each level between single and noise. Therefore, in wavelet decomposition, the first and second floors by the level of high-frequency coefficients, vertical and diagonal high frequency coefficient of high frequency coefficients, and then use decamp function to calculate the default threshold value of each frequency, and process the various high thresholds.

The proposed method and the traditional wavelet threshold denoising from the mean square error and peak signal to noise ratio shown in Table 107.1 can be seen from T in order to achieve a better denoising result for the different high-frequency wavelet coefficients. The article calculated the corresponding default threshold by decamp [14], after a threshold of the reconstruction of wavelet coefficients, to obtain the denoised image.

Fang results of this study are to secure communication research in the field of persons with partial reference value, in future studies, we also will further research and realize the arbitrary differentiable relations of the generalized synchronous observer design scheme; whether using single scalar signal realizes synchronization observer design can surely be the next phase that needs to be further studied.

101.4 Analysis of Experiment and Simulation Results

The woman and the cameraman image on the Matlab 7.5 system images adds the Gaussian white noise with zero mean and standard deviation of 18, to the choice of the sym4 wavelet basis function, and obtains the second generation wavelet based on the improvement method (improve sym4 wavelet), the method proposed in this

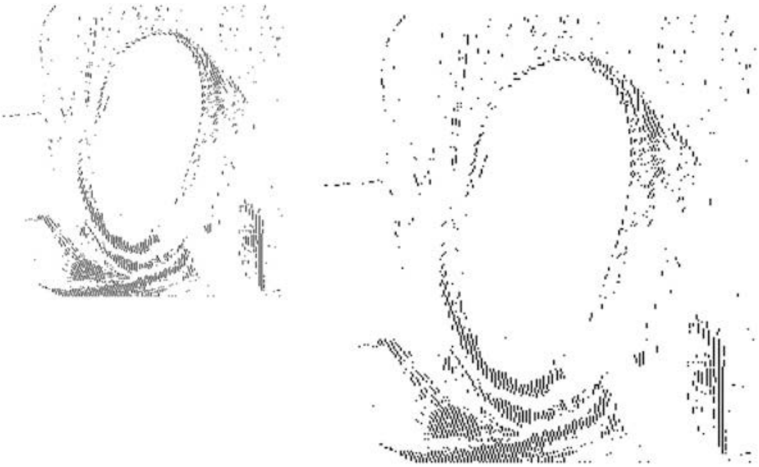


Fig. 101.1 Image edge

paper and traditional methods of wavelet thresholding for denoising experiments. The denoising results were compared and analyzed by peak signal to noise ratio (PSNR) and mean square [15, 16].

PSNR = 10 log 10 (2,552/MSE),

MSE = $(M \times N){-}1$

where, M and N are the images corresponding to the number of rows and columns; F $(i,\ j)$ and Y $(i,\ j)$ respectively as the original image and denoising. Experimental steps are as follows:

- The sorbets operator, call the MATLAB system edge () function to extract the edge of G (i, j), shown in Fig. 101.1.
- Select sym4 wavelet, obtained the second generation wavelet by the improvement on the 2D images with noise T $(i,\ j)$ and its edge of G $(i,\ j)$ wavelet decomposed wavelet coefficients Wt (i, j) and $W{*}g$ (i, j).
- Wavelet coefficients Wg (i, j) and Wt (i, j), acceding to the press (7) integration. And then reconstruct the image edge enhancement of noisy N * (i, j).
- For enhancing the edge of the noisy image N * $(i,\ j)$ for two layers that the proposed method for image denoising images from a smaller mean square error, peak signal to noise ratio is greater, the better the denoising effect is.
- Use the processed wavelet coefficients to get the denoised image by inverse transform, and evaluate the denoising results numerically.

From Fig. 101.2, the proposed denoising method to denoising image texture is clearer, and the image is subjectively better.

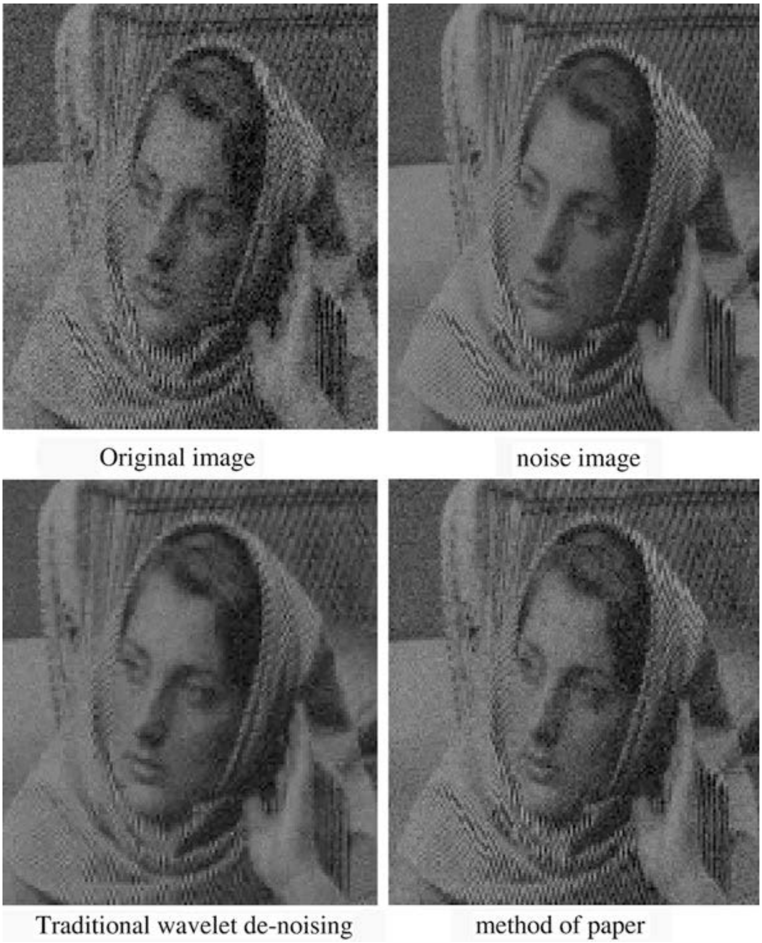


Fig. 101.2 Comparison of image denoising: **a** Original image, **b** noise image, **c** traditional wavelet de-noising, **d** method of paper

101.5 Conclusions

In order to better protect the image edge and texture, combining with the advantages of second generation wavelet transform, edge enhancement was made in the paper before denoising. Experimental results show that the method not only reduces the mean square error, but also improves the performance of the image PSNR, having a good suppression to the image noise, and better maintains the image edge and texture features.

References

1. Defeng Z (2009) MATLAB wavelet analysis [M]. China Machine Press, Beijing, pp 206–209
2. Sweldens W (1997) The lifting scheme: a construction of second generationwavelets [J]. SIAMJ Math Anal 29(2):511–546
3. Donoho DL (1995) De-noising by soft-thresholding [J]. IEEE Trans Inform Theory 41(3): 613–627
4. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation viawavelet shrinkage [J]. Biomerika 81(4):425–455
5. Daubexhies I, Sweldens W (1994) Factoring wavelet transforms into lifting steps [J]. J Fourier Ana l 4(3):245–267
6. Sweldens W (1997) The lifting scheme: a construction of second generation wavelets [J]. SIAM J Math Anal 29(2):511–546
7. Gezhe science, Javert (2007) Wavelet analysis theory and MATLAB R2007 to achieve [M]. Electronic Industry Press, Beijing, pp 375–380
8. Zhangjun B, Xia J et al, ROCKETS (1999) MATLAB-based system analysis and design—wavelet analysis [M]. Xidian University Press, Xi’an, pp 226–230
9. Mei L, Chen L, Yan G (2008) Based on adaptive soft threshold denoising and edge enhancement [J]. Electron Meas Technol 31(7):4–6
10. Donoho DL (1995) De-noisingby soft-thresholding [J]. IEEE Trans Inform Theory 41(3): 613–627
11. Chang SG, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression [J]. IEEE Trans Image Process 9(9):1532–1546
12. Wang AL, Ye M, Deng Q (2008) MATLAB R2007 image processing technology and application [M]. Beijing Electronic Industry Press, Beijing, pp 116–169
13. Nan D, Chiang, Fuze X (2007) Based on second generation wavelet transform image denoising [J]. Natural Science and Engineering, Yantai University 20(1):4
14. Liu J (2009) Wavelet analysis in fabric pilling objective evaluation of [J]. Wuhan Institute of Technology 22(4):7–10
15. Donoho DL, Johnstone IM (1995) Wavelet shrinkage asympot [J]. Pia J Royal Stat Soc 57(2):301–369
16. Yang Z (2009) On a Meyer-type wavelet note [J]. Wuhan University of Science 22(4):32–34

Chapter 102
Reconstruction Based on the Principle of Hierarchical Image Analysis and Robot Vision Research

Zhiyong Zhang, Xiaoning Li and Xiaofeng Li

Abstract This paper analyzes the principle of hierarchical reconstruction on the impact of the robot image recognitions given to the affine point from the image reconstruction of the corresponding instance of calculated measures, which to some extent is a measure of the corresponding affine point reconstruction needs of the most basic structural information, such as the measure of similar objects based on reconstruction, metric reconstruction of specular reflection based, reconstruction based on the measurement of moving objects, and based on hidden consumption point (line) of the metric reconstruction to improve image clarity.

Keywords Stratified Reconstruction · Robot · Vision · Image

102.1 Introduction

Stratified Reconstruction from affine point of principle is the corresponding reconstruction method for calculated measures, first of all calculated from the point of the corresponding projective reconstruction [1]; affine points according to a corresponding projective reconstruction projective transformation, projective transformation by this Real feature vector to determine the plane at infinity in the projective reconstruction of space coordinates, which are affine reconstruction;

Z. Zhang (✉) · X. Li
Changchun Teachers College,
The Computer Science and Technology Institute,
Changchun 130032, Jilin Province, China

X. Li
Jilin University, Changchun 130032, Jilin Province, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_102,

reconstruction calculated from the affine camera intrinsic parameter matrix, and ultimately get metric reconstruction.

Affine point correspondence, of the two images (*I*, *I'*) between the two points correspond to

$$m_x = \{m_1, m_2, \ldots, m_k\} \leftrightarrow m_x^{'} = \left\{m_1^{'}, m_2^{'}, \ldots, m_k^{'}\right\}$$

$$m_y = \{n_1, n_2, \ldots, n_s\} \leftrightarrow m_y^{'} = \left\{n_1^{'}, n_2^{'}, \ldots, n_s^{'}\right\}$$

Satisfy the following properties: points corresponding to the spatial point sets and the difference between sets of points in space of an affine transformation, which makes the existence of affine transformation, denoted claimed that two points correspond, for the corresponding affine point. For example, the vertices of two six-sided compositions of two images corresponding to two points is the affine point correspondence, as between any two rectangules there exists an affine transformation to transform one set of vertices to another vertex set of knives. This exists in the following series of questions in both the corresponding affine points, so they can apply the method given in this section to solve.

102.2 Similar Objects

Two similar objects are known [2]: a set of corresponding feature points of the image and calculating a measure of scene reconstruction. Because objects *X* and *Y* are similar, it certainly makes the existence of a similarity transformation S2.

The similarity transformation is the affine transformation, so two similar objects, *Y* and *X* corresponding to any set of image feature points corresponding to constitute an affine point.

102.3 Mirror Reflection

Given an object and its mirror reflection [3] in the image of the projection, calculate a measure of the scene reconstruction. *X*, *Y* are objects that mirror reflection, it certainly makes the existence of a reflection transformation U3.

Affine transformation is due to reflection, so the object and its mirror reflection *X Y* corresponding to any set of image feature points correspond to constitute an affine point.

102.4 Moving Objects

Moving objects known at two different times was under the camera image (within the parameters of each camera position is fixed and constant) [4] to calculate a measure of reconstruction of the scene. Thus moving objects at time 1 on the world coordinate system is expressed as *X*, and the images of two cameras, respectively; at time 2 the world coordinate system is expressed as *Y*, and the images of two cameras, respectively. Since the object motion between the two moments can be expressed in Euclidean motion, it makes the existence of Euclidean transformation E4. Therefore, the point correspondence, must correspond to the affine point.

102.5 Binocular Device

Assuming a rigid body motion binocular device for general E, the image is under the camera from the eyes of the metric calculation of reconstruction of the scene [5].The problem with the above sports scene reconstruction problem is equivalent. Because binocular device for general rigid motion E of the images obtained with the eyes stationary object device on the scene in a fixed position relative to the rigid body motion for the resulting image is the same.

102.6 Three-Point Implicit Consumption

Three directions of space known to eliminate the implicit point of the projection in the image [6], calculated measures Rereading. Hutchison 3 infinity, the projection of two images, respectively, and also, take the two points Fuk image corresponds to 4, which corresponsd to the space point is denoted by order

$$m_x = \{m_1, m_2, m_3, m_4, m_5\} \leftrightarrow m_x^{'} = \left\{m_1^{'}, m_2^{'}, m_3^{'}, m_4^{'}, m_5^{'}\right\}$$

$$m_y = \{m_1, m_2, m_3, m_6, m_7\} \leftrightarrow m_y^{'} = \left\{m_1^{'}, m_2^{'}, m_3^{'}, m_6^{'}, m_7^{'}\right\}$$

They correspond to the two space points identified with the projective transformation which is a three-dimensional affine transformation, because the infinite plane projective transformation will transform plane at infinity. So, is the corresponding affine point.

102.7 Quasi-Affine Reconstruction

Reconstruction of quasi-expensive radio is a special projective reconstruction first introduced by the Hartiey and in-depth study dealing plane at infinity in a projective reconstruction of space coordinates [7]; if the convex Projective Reconstruction packages are located on the same side of the plane it is said to prevail in this projective reconstruction Affine Reconstruction (Quasi-Affine Reconstruction).Hartiey said that any projective reconstruction can be elevated to a quasi-affine reconstruction. Quasi-affine re-introduced the following algorithm.

Given n images of the point correspondence, orders, and 5 is the n pieces of a picture projective reconstruction. The projective reconstruction of space images in the camera i-matrix is the point corresponding to the projective reconstruction. Affine Reconstruction Projective Reconstruction subject is necessary and a sufficient condition (5) of the camera matrix and the space points satisfying the following conditions: for all i, j, where the projective depths have the same sign. Thus, projective reconstruction by changing the sign of the camera matrix and the space points are quasi-affine reconstruction of the symbol.

102.8 Calculating the Corresponding Point From the Affine Reconstruction of Order

Is affine point corresponding to a quasi-affine reconstruction, projective reconstruction of space if it is known in the plane at infinity [8].

You can get an affine reconstruction.Therefore, to calculate the affine reconstruction of quasi-affine reconstruction only in the space identified in the plane at infinity.

Consider X, Y, quasi-affine reconstruction, which are the last coordinates of the normalized vector. By the correspondence, when $k \geq 5$, we can get space in the quasi-affine reconstruction of a (point) projection change, which makes the projective transformation; note that this is a homogeneous transformation. If following the establishment of strict equality (9) the transformation matrix is unique and the transformation matrix is denoted.

Example 1 The projective reconstruction space, the plane at infinity is a positive eigenvalue matrix eigenvector.

Analysis What the geometry with their X, Y the difference between a (point) Photography transformation, i.e., there exists a matrix such that (10)

By (1) and (10), we can infer

General (9), we have (11)

So (12)

From (12) can be seen the feature vector if and only if it is the feature vector. A 3D space midpoint of duality, is the 3D Euclidean space which transforms a plane, so the feature vector is that it is a fixed surface, while that from the European space to the quasi-affine reconstruction of a face in the transformation of space, so the plane in Euclidean space into a quasi-affine reconstruction of space in the plane. As is the affine transformation, so the plane at infinity is a fixed plane, that is, a feature vector. So, yes, a real eigenvector, which is quasi-affine space plane at infinity. Because of having the same feature vector, it is also a real feature vector, which finally proves that the positive eigenvalues must be the eigenvectors, anyway: suppose the value of the negative characteristics of the eigenvectors. Thus, there are (13).

Thus the straight line connecting the two points. From (13) and (9), there and on both sides in the plane at infinity, and quasi-affine reconstruction of contradictions. Therefore, the eigenvalues must be the eigenvectors of the proof.

When *A* is a general affine transformation, the matrix may have a positive real value of the four feature vectors. In order to obtain an affine reconstruction it must be identified. A general discussion before the discussion *A* is a similar transformation for the special case, in addition to plane (movement) change, but can be uniquely identified in the quasi-affine reconstruction of space in the plane at infinity.

Example 2 Suppose *A* is a similarity transformation *S*.

(1) If *S* is not a plane (movement) change, then the quasi-affine reconstruction space can be a uniquely determined plane at infinity.
(2) When the S-dimensional (movement) changes, then the quasi-affine reconstruction space can be a uniquely determined plane at infinity having two solutions.

Analysis The similarity transformation *S* can be expressed as

Where, *s* is similar to the scale factor and *U* is an orthogonal matrix it is not difficult to calculate.

Its physical characteristics such as values, which, when *U* is a rotation matrix, take a positive sign; when the *U* matrix is for the reflection taking a negative sign, it is easy to check on the plane at infinity which is the characteristic value of a feature vector, considering the transformation. Among them, the meaning is as described earlier.

(1) When *s* is not equal to 1, it will feature as the value of tangible and (in general, because (14) is a homogeneous second-class type, that is a constant multiple of the difference between the sense of equality). Easy out on the eigenvectors of eigenvalues in the quasi-affine reconstruction of space in the plane at infinity. Since the eigenvalues are the eigenvalues of the reciprocal of the conjugate eigenvector corresponding to the same and therefore the value of physical characteristics such as eigenvalues and eigenvectors corresponding to a plane at infinity in the quasi-affine reconstruction of space coordinates, note that the norms and values of the other three characteristics of the module are not the only characteristic values equal to it, so this case can be uniquely identified in the proposed reconstruction of space in the affine plane at infinity.

(2) When $s = 1$, S is an isometric transformation; when the U matrix is for the rotation S is a Euclidean transformation; when the U matrix is for the reflection, S is an anti-Euclidean transformation, often referred to as reflection transformation (such as specular reflection.) The following points S and reflection transformations of Euclidean transform to discuss two cases:

When S is the Euclidean transformation, it must be tangible, such as the characteristics of value, so that value is also characteristic. If S is flat (sport) transformation, i.e., the translation vector S is not orthogonal with the plane parallel to the axis of rotation, by Proposition 4, it can be extrapolated infinite plane in the corresponding experimental characteristics of a double feature vector only, by the only real feature vector, which is quasi-affine reconstruction of space in the plane at infinity; if S is flat (sport) transformation, i.e., the translation vector S perpendicular axis of rotation parallel to the plane, then the corresponding two real eigenvalue 1 has two linearly independent eigenvectors, so there must be a positive double eigenvalue, and the double value corresponding to the two linearly independent eigenvectors. Thus, this time unrelated to the sense of linear, infinite plane has two solutions.

When the S is reflective exchange, it will be visible as a characteristic value, so value is also characteristic. Therefore, only one positive eigenvalue, according to Proposition 1, can undoubtedly determine the quasi-affine reconstruction of space in the plane at infinity. Proved.

102.9 When a General Affine Exchange, according to the Camera Parameters are Constant and Changes in the two cases, given the Method Chosen

(1) Within the parameters of constant

If all the camera intrinsic parameters are equal, then the first j-1 and a point of view between the plane at infinity homography matrix model eigenvalue 3 are equal (often referred to as model constraints). The camera projection matrix [9].

J-1 shows a view with the plane at infinity between the single-matrix of all possible solutions should be as

Under normal circumstances, there is only one j K for every matrix that satisfies the model constraints. In this way, you can verify that all the characteristics of the matrix model are equal to the value determined.

(2) Changes in the parameters

Two intrinsic parameters are not equal, the model equivalent constraints no longer hold. At this point, the calculation of all points corresponding to the quasi-affine reconstruction:

If then there must be the same symbol. So, you can verify whether the same symbol is to be determined.

Note that when the camera intrinsic parameters are not equal, the above conditions are necessary to determine that seen.

(3) Metric Reconstruction

Suppose there are $n + 1$ view of the image. A view from the $n + 1$ image, according to the previous section, can get an affine reconstruction.

If the first one point of view known within the parameters of the camera matrix K, you can get a metric reconstruction.

Therefore, to calculate metric reconstruction, the first one only needs to calculate the point of view of the camera intrinsic parameters matrix K.

Remember the first i-view camera image is absolute conic (IAC) , by the affine reconstruction it can be one of the constraint equations IAC which is the first i-1 with the first view of the infinite homograph 1 and the camera viewpoint the IAC. This method can be used to solve the intrinsic parameters.

102.10 Summary

This paper analyzes that the hierarchical principle by the robot image recognition reconstruction can improve the clarity by about 5% [10].

References

1. Zhang G (2006) Computer control [N], Computer World
2. Zhi S, Fu Z (2009) Roots: a multi-scale mathematical morphology classification of similar objects invariant moments, Beijing Aeronautics and Astronautics (04):15–16
3. Zhuo Z, Zhao Y, Wu YL (2007) Too summer, phase clouds. granite surface bidirectional specular component and diffuse component of comparative study. Infrared and Millimeter Waves (11):078–080
4. Chen M, Liu Y (2010) A moving object based on motion vector extraction method. Electronic Design Engineering (06):11–13
5. Zheng ZY, Yao GZ, Jin G, Chengli H, Zhang H, Dai J (1983) Binocular stereo vision information processing II, Three-dimensional spatial filtering effect on the eyes. Psychol (12):031–034
6. Chi T (2006) Remote sensing technology: Strong support for the space calculation. Comput World (10):014–015
7. Zhu H (1998) Transformation method using vectors derived velocity and acceleration in cylindrical coordinates, the natural coordinates and spherical coordinates in the expression. Neijiang Teachers College (7):44–46

8. Sun F, Wu Fu-Chao, Hu Z-Y (2003) Determined by the projection plane parallel to the plane at infinity homography. J Software 14(5):935–946
9. Wu J, Wang T, cable Zhiyong, Zheng B (2009) DPSS projection matrix based on single-channel moving target detection method. Data Acquis and Process (4):024–028
10. Ye M (2000) Of standard definition TV and HDTV clarity. TV works

Chapter 103
An Improved Approach for Moving Object Detection Based on Markov Random Field

Buyu Xu, Hao Tang and Lei Zhou

Abstract Due to utilization of the relativity of every pixel of an image, the Markov random field (MRF) model is effective in solving the problem of detecting moving objects under a complex background. In this paper, the bits-segmentation of inter-frame difference images is used as the label field of MRF, and the compatibility function related to such labels and the hidden states is provided, so that an improved detection method for moving objects is proposed based on MRF. Compared with the traditional MRF method, the proposed approach can avoid the threshold selection process for obtaining the label field, which is a sensitive issue that may affect the detection negatively. The experiment results show that this approach is more effective and has a better adaptability than traditional methods.

Keywords Markov random filed · Inter-frame difference images · Moving object detecting · Belief propagation

B. Xu (✉) · H. Tang · L. Zhou
School of Computer and Information,
Hefei University of Technology, Hefei 230009, China
e-mail: bigfishery@163.com

H. Tang
e-mail: htang@hfut.edu.cn

L. Zhou
e-mail: zhouleizhl@hotmail.com

H. Tang
Engineering Research Center of Safety Critical Industry Measure
and Control Technology, Ministry of Education, Hefei 230009, China

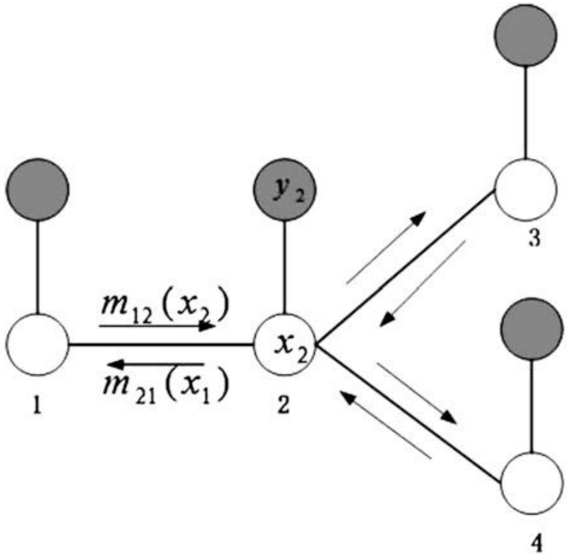
Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_103,

103.1 Introduction

Computer vision technology uses cameras to obtain images of the environment and uses the computer to process the visual information. With features of intuition and having a large amount of information, it is widely used in practice. Moving objects detection based on image sequence is a basic technology used in many computer vision applications. It has been widely used in security monitoring, intelligent transportation, visual navigation, etc. A widely used method for moving objects detection based on image sequence is the inter-frame difference method, which checks the changes of pixel's intensity between several adjacent frames within a short time. Inter-frame difference method usually extracts the moving regions through the use of two images pixel-based differential and thresholding [1, 2]. Inter-frame difference method has a strong self-adaptability for dynamic environment, it is fast and the computation is small. However, it cannot extract all the sports pixels for moving objects with uniform intensity and it may yield holes in such objects. The method based on Markov Random field considers the impact of neighboring pixels and uses the relativity of every pixel of an image. It transforms the problem, which determines whether a pixel belongs to the moving object or not, into a problem of solving a maximum posteriori probability estimation. The MRF model has a strong ability of anti-interference and is effective in solving the problem of detecting moving objects under a complex background. Yin [3] has taken account of the spatial and temporal dependency between pixels and established a three-dimensional MRF model. It is solved through belief propagation algorithm by asynchronous accelerated message updating and achieved good results for moving object detection [3]. How to find a good gray field model to better respond to the initial image features and how to get a good label field by segmentation have been a difficult problem in MRF segmentation theory. In [4], a Gaussian Markov model-based moving object segmentation algorithm was investigated, which described the inter-frame difference images of video sequence by Gaussian mixture distribution, and improved the standard MAP and used a quick method to calculate the posterior probability. In [5], the plan of gray field modeling based on weighted histogram was adopted to distribute trust degree to pixels of the coarse segmentation image, then established weighted histogram adopting the statistics of trust degrees, and finally built exact gray field model based on the weighted histogram.

In this Chapter, we use the 3D MRF model introduced in [3] and make some improvements on it, proposing a bits-segmentation method for acquisition of label field and creating the corresponding compatibility function. The method based on normal MRF model uses the binary image as the label field and an effective auto-threshold algorithm is needed. The method proposed in this paper uses the bits-segmentation to get a series of fixed threshold. Compared with the traditional MRF method, the proposed approach can avoid the threshold selection process for obtaining the label field, which is a sensitive issue that may affect the detection negatively. The experiment results show that this approach is more effective and has a better adaptability than traditional methods.

Fig. 103.1 MRF model



103.2 Markov Random Field Model

Pairwise Markov random fields can be represented as an undirected graph, the graph model has been shown in Fig. 103.1. In general, we assume that we observe some quantities about the node i, it is some labels which may represent the intensity of a pixel or some others, we note this observed data as y_i. And we want to infer some other quantities that cannot be observed directly, we note this hidden state as x_i. We further assume that there is some statistical dependency between x_i and y_i at each node i, which can be written as a joint compatibility function $\phi_i(x_i, y_i)$. The state of node i can also be influenced by its neighboring nodes, this dependency can be represented by a compatibility function $\psi_{ij}(x_i, x_j)$. Then we take the overall joint probability of hidden state x_i and observed data y_i to be:

$$P(\{x\}\{y\}) = \frac{1}{Z}\prod_{(ij)} \psi_{ij}(x_i, x_j) \prod_{i} \phi_i(x_i, y_i) \tag{103.1}$$

where Z is a normalization constant and the product over (ij) is over nearest neighbors on the graph.

For every node that belongs to $\{x\}$, we can find a state from the candidate states to maximize the joint probability $P(\{x\})$, and all the selected states will form the distribution of states as the optimum solution for $\{x\}$. The joint probability function contains all the observed data and hidden states, and for practical problems we usually have so many nodes that we can hardly get the results from this function directly. People have been focused on the problem of how to solve the MRF model efficiently for many years and have proposed some new efficient algorithms, such as Graph Cuts (GC) algorithm and Belief Propagation (BP) algorithm [6, 7]. The graph cuts algorithm converted the problem of solving the joint compatibility function into the problem of finding the optimal labels, it can efficiently find the optimal labels which correspond to the minimum energy. The belief propagation algorithm considered the statistical dependency between neighboring nodes, it is an efficient way that can be used to find the node's marginal distribution based on passing local messages.

In the BP algorithm, we introduced two concepts as message and belief. The message $m_{ij}(x_j)$ can intuitively be understood as "message" from a node i to node j about which state node j should be in. It will be a vector of the same dimensionality as x_j, with each component being proportional to how likely node i thinks it is that node j will be in the corresponding state. The belief at a node i is proportional to the product of the local message at that and all the messages coming into node:

$$b_i(x_i) = k\phi_i(x_i) \prod_{j \in N(i)} m_{ji}(x_i) \tag{103.2}$$

where k is a normalization constant (the beliefs must sum to 1) and $N(i)$ denotes the nodes neighboring i.

The messages are determined self-consistently by the message updating rules:

$$m_{ij}(x_j) \leftarrow \sum_{x_i} \phi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in N(i)\backslash j} m_{ki}(x_i) \tag{103.3}$$

Note that on the right-hand side, we take the product over all messages going into node i except for the one coming from node j. Compared with the joint probability function we can prove that $b_i(x_i)$ is equal to the marginal probability distribution $p_i(x_i)$ at node i.

103.3 Bits-Segmentaion for Label Field

Yin [3] constituted a 3D MRF model for moving object detecting problems and achieved good results. The 3D MRF model has considered the dependency between pixels both in spatial and temporal domain. In this model, every pixel will have a corresponding data node d_j and state node s_j. Data node d_j is acquired from thresholding on inter-frame difference images. State node s_j is a vector, each component represent (represents) a likelihood of how much the node belongs to the moving object. We perform asynchronous accelerated message updating to solve this model by belief propagation algorithm, and compute the minimum mean squared error estimate (MMSE) estimate as the best estimate for node s_j.

The normal detecting method based on MRF model uses binary motion detection result computed by inter-frame differencing as the label field, and then solve the model by some efficient algorithms, such as belief propagation algorithm,

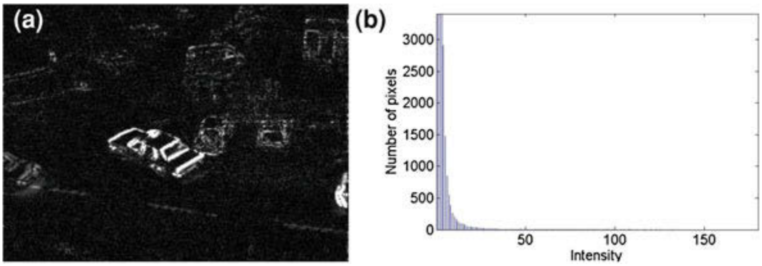


Fig. 103.2 **a** inter-frame difference image **b** histogram of Fig. 103.2a

to compute the optimal solution. We usually use a global threshold to separate the pixels of an image into two groups for the result of the binary motion detection. So choosing a proper threshold is very important. There are some existing auto-threshold methods, the method which tries from a rough value and then approaches to the best value iteratively was proposed in [9]. We can also find the threshold from the histogram, such as taking the maximum point of slope changes in the histogram as the threshold [10].

Figure 103.2b shows a typical inter-frame difference image's histogram, we can find that a large number of pixels were gathered in the lower end of the histogram. If the histogram has a distinct form of peaks and valleys, it can be easy to find a proper global threshold to classify the pixels. However, the typical inter-frame difference image's histogram shown in Fig. 103.2b has no such pattern, and experiments show that the above methods cannot get a very good thresholding effect in this case.

In the normal detecting method based on inter-frame difference, the result by thresholding the difference images was binary motion detection which can be used to indicate whether the pixel belonged to the moving object. In this case, the estimate state $\widehat{s}_j$ which represented the likelihood of how much the pixel belonged to the moving object was directly corresponding to the observation d_j. In the MRF model, the estimate state $\widehat{s}_j$ was computed from hidden state s_j with MMSE estimate, and the hidden state s_j was associated with observation d_j through the compatibility function $\phi_j(s_j, d_j)$. So we need a more inference process in the MRF model.

We can easily find from the inter-frame difference image, that the higher the pixel gray value is, the greater the likelihood of moving object will be and vice versa. However, the pixels with medium gray value are difficult to determine whether it belongs to moving object or the background. These pixels contain the information about possibilities of pixel's state. If we threshold these inter-frame difference images to classify the pixels into two groups with fixed label, the information contained in the pixels with medium gray value will be lost. We could use more thresholds to classify pixels into more groups in order to retain more such information about possibilities.

We usually have two candidate values for d_j, labeled as 0 and 1 from the binary motion detection, and the hidden state s_j usually has C candidates, $C > 2$. The more candidates we have for state s_j the more subtle the results will be to get through the inference process by belief propagation algorithm, but also the more the computational cost will be needed. In our experiments, we choose $C = 8$, this value can better balance the computational cost and detection results.

The compatibility function $\phi_j(s_j, d_j)$ describes the relationship between observation d_j and hidden state s_j. If d_j equals zero, no motion is detected at this pixel by inter-frame differencing, and a uniform distribution is used to represent "no-motion". However, when d_j equals one, motion is definitely detected by inter-frame differencing at this pixel, s_j has an impulse distribution at $s_j = C$ to show the confidence of existing motion regardless of what messages are passed to the node. The discrete form of compatibility function $\phi_j(s_j, d_j)$ was a mapping from d_j to s_j. Let s_j^p denote the pth state candidate at node $s_j(p \in [1, C])$, the message from observed data is a vector of C elements :

$$\phi(s_j^p, d_j) = \begin{cases} [\frac{1}{C} \cdots \frac{1}{C}]^T & d_j = 0 \\ [0 \cdots 01]^T & d_j = 1 \end{cases} \tag{103.4}$$

The computational cost of belief propagation was mainly in the message passing and belief updating, the mapping from d_j to s_j could be finished in the initialization phase, thus increasing candidates of d_j does not significantly increase the complexity of the normal MRF model.

Based on the above analysis, we proposed (propose) bits-segmentation method to get more candidates for d_j. For the 8-bits grayscale images, the selected thresholds were $2^i(i \in [1, 8])$, which is equivalent to classify the pixels by its highest non-zero bit:

$$\phi(s_j^p, d_j) = \begin{cases} s_j^p = \theta & p = d_j \\ s_j^p = \varepsilon & \text{otherwise} \end{cases} \tag{103.5}$$

where $d_j = \lfloor \log_2 I \rfloor + 1$, I is the pixel's intensity value, $\lfloor \cdot \rfloor$ means rounds "." to the nearest integers less than or equal to ".". As shown in formula 103.5, when d_j belongs to the pth candidate label, the possibility of s_j belong (belongs) to the pth candidate state is θ, the possibility to other candidates state is ε. Parameters satisfy the condition of $0 < \theta < 1$, $\varepsilon = (1 - \theta)/(C - 1)$ and $\theta > \; > \varepsilon$.

We can use more threshold (thresholds) to increase labels for d_j and the most simplified way to select these thresholds is to use uniform spacing. When C equals 8 the interval between uniform thresholds will be 32, causing nearly all the pixels to be classified into the first d_j group, which will be difficult to use for detecting. If we could increase the value of C the uniform threshold will be performed better, but it will also increase the computational cost for belief propagation.

The threshold selected by bits-segmentation is not uniform spacing, the label of d_j is determined by its highest non-zero bit. We will classify the pixels of low intensity with little threshold interval and the pixels of high intensity with large threshold interval. In this way we could as much as possible make every d_j contain a certain amount of useful information. $\psi_{ij}(s_i, s_j)$ defines a state transition function between two neighboring nodes via the Potts model:

$$\psi_{ij}(s_i^p, s_j^p) = \begin{cases} \theta & p = q \\ \varepsilon & \text{otherwise} \end{cases} \tag{103.6}$$

This function encourages neighboring nodes to have the same states. It also acts as a decay function to reduce the motion likelihood in the absence of current intensity differences.

There exists loops in the MRF model for image process, which will make the message pass into circles and difficult to converge. In this paper we perform asynchronous accelerated message updating to update the node's belief and schedule its message passing. For the 1D MRF model (chain), we update the node's belief (by formula 103.2) and pass the message (by formula 103.3) sequentially from one end of the chain to the other end, resulting in all the nodes of the chain to have its new beliefs; we call this process a 1D BP sweep. For the 2D MRF model (grid), we could perform four 1D BP sweeps (left to right, up to down, right to left, and down to up) individually and in parallel. Here, 1D BP means that BP on the spatial grid is executed simultaneously row by row, or column by column. The four 1D BP sweep beliefs (distributions for each pixel) are then fused together and the MMSE estimates will get from these fused beliefs.

103.4 Experiments

The normal MRF model uses binary motion detection result computed by interframe differencing as the label field. The detection results were very sensitive to the threshold, inappropriate threshold will cause a bad label field and lead to a fault detection. Our improvement mainly made more use of the original inter-frame difference image's information, using a fixed multi-threshold segmentation for the label field to obtain more stable results in the moving objects detecting problems.

When we luckily select the appropriate threshold we will get a good result from the MRF model. If the threshold is difficult to get or we get an inappropriate one, we may have fault results. Figure 103.3a shows the result from the normal MRF model, the threshold was computed by the method proposed in [9] and this is an inappropriate one. We can see that the detected object is not complete compared with Fig. 103.3b which is a result from the improved MRF model. Figure 103.3c, d was another comparison with these two methods, Fig. 103.3d is based on the improved method and gets a more complete object. From these two experiments on different scenes, we can see that the improved method could get a more stable result compared with the normal MRF model.

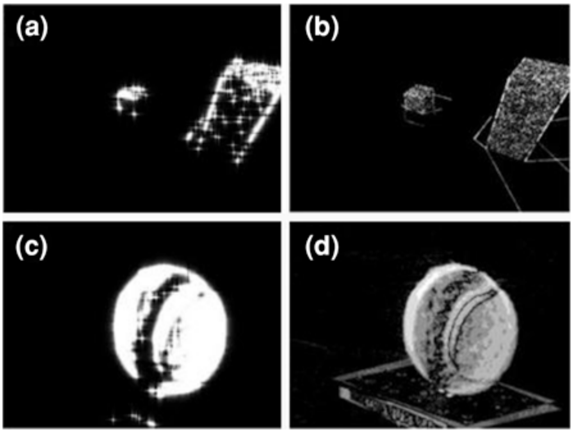


Fig. 103.3 Result of the normal MRF model with inappropriate threshold (**a**) and (**c**), compared with the result of the bit-segmentation MRF model (**b**) and (**d**)

103.5 Conclusion

The method proposed in this paper for acquisition of label field based on bits-segmentation has made some improvements for the normal MRF model used for moving objects detection. It avoid (avoids) the problem of having instability results caused by improper selection of threshold by auto threshold methods. Through experiments and analysis of inter-frame difference image's characteristics, the results show that this method can get more stable detection results. However, a large number of pixels at the low end of the histogram was caused by noise, our proposed method gives these pixels the same compatibility function, it will pass the influence of noise from observation to the estimates and may worsen the detection. In the future work, a more proper compatibility function will be explored.

References

1. Lipton A, Fujiyoshi H, Patil R (1998) Moving target classification and tracking from real-time video[C]. IEEE workshop on applications of computer vision, pp 8–14
2. Wang L-L, Li W, Gao X-R (2010) The research of moving object detection algorithm in video images. J Microcomput Inf 16(6):147–149
3. Yin Z, Collins R (2007) Belief propagation in a 3D spatio-temporal MRF for moving object detection[C]. In: IEEE conference on computer vision and pattern recognition, pp 1–8
4. Linghu Y-F, Guo X-M, Wang S-T (2007) Research on motion segmentation based on Markov random field models. J Comput Eng Appl 43(14):56–59
5. Long M, Wang L-P, Shen Z-K (2010) A slow-moving-object segmentation technology based on MRF under complex background. J Signal Process 26(6):911–916
6. Yedidia JS, Freeman WT, Weiss Y (2003) Understanding belief propagation and its generalizations[C]. In: Exploring artificial intelligence in the new millennium. Morgan Kaufmann Publishers, USA, pp 239–269

7. Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. J IEEE Trans Pattern Analysis Mach Intell 23(11):1222–1239
8. Tappen M, Freeman W (2003) Comparison of graph cuts with belief propagation for stereo using identical MRF parameters[C], In: ICCV, pp 900–907
9. Gonzalez C, Woods E (2003) Digital image processing[M], Electronic Industry Press, Beijing pp 485–486
10. Zhan J.-F, Qi F.-H, Wang H.-L (2000) Moving object segmentation technology based on spatial-temporal Markov random field[J]. J of China Inst Commun 21(11):63–68

Chapter 104
Study of Miniature Video Inspection Robot Design and Control

Xian-you Zhong, Chun-hua Zhao and Gang Wu

Abstract Utilization of nuclear power is of vital importance to the development of society. A miniature robot for video inspection is required to inspect in the narrow space of a nuclear power plant. At present there are no such robots that can solve the problem effectively in my country, and the foreign products are expensive. The design scheme which adapts to the condition in a nuclear plant was put forward based on the study of the developing condition of the pipeline robot and the underwater robot; the construction and principle are introduced in this article and the selection of sealing method and selection of motor are studied in detail. The robot can work in narrow space, such as in pipelines and sewers, and its prospect of application is wide in many fields.

Keywords Nuclear power · Pipeline robot · Static sealing · Dynamic sealing

104.1 Introduction

As the world's fastest growing economy and the second largest energy consumer, China is looking towards a more balanced mix of energy generating methods [1]. Developing nuclear power can meet the future energy needs without emitting

X. Zhong (✉) · C. Zhao · G. Wu
College of Mechanical and Material Engineering,
China Three Gorges University, Yichang 443002, China
e-mail: zhxy@ctgu.edu.cn

C. Zhao
e-mail: zhaolilic@ctgu.edu.cn

G. Wu
e-mail: wwwg2000@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_104,

carbon dioxide and other atmospheric pollutants. In order to ensure the safe operation of nuclear power plants, it is important to inspect the key areas of the plants, although there are a lot of operations with radioactives. Many pipes used in the modern industry and agriculture, petroleum, chemical, appear with flaws, corrosion, blockage after long time, and they need inspection.

104.2 General Scheme Design

104.2.1 System Composition and Working Principle

A miniature video inspection robot is controlled by computer with vision and image processing and editing functions. As shown in Fig. 104.1, it mainly consists of industrial PC A, cable and hoisting device B, crawling device C and cameras rotation device D. Operators can control the movement of the robot, which control industrial PC through the image obtained in real-time by CCD. Crawling device is driven by six wheels through synchronous belt transmission, with three wheels on the left, wheels is propelled by two motors separately. According to the requirements, operators control the robot to move through the cable video monitoring remotely, according to the needs of work, manipulator can be installed to catch the target. If the two stepper motors run at the same speed, mobile video robot can run straight forward or backward. If the two-stepper motors run at the same speed in the reverse direction, a mobile video robot can realize in situ rotation; if the two-stepper motors run at different speeds, mobile video robots can swerve. Motor and CCD cameras are sealed in the seal chamber of the robot, mobile video monitoring device of the joint parts such as connecting link head, using seal work crawling containers should maintain slightly superior water pressure tank of gas pressure, gas pressure sensors are set inside the seal cabin, real-time monitoring of sealing pressure, in order to ensure that the mobile video monitoring device in underwater can work reliably. Mobile robots can be evacuated through cable recovery device when finding leaks [2–6].

104.2.2 Camera Pitch Device

Due to the advantages of CCD camera devices of small volume, light weight, the inert small, low consumption, input light dynamic range wide, scanning no distortion, in this system (as shown in Fig. 104.2), the CCD adopted is of high sensitivity, low noise and high resolution; video detection camera angles in the camera lens is the angle of depression on the road. When camera lens height is constant, the image resolution will be taken down if the camera angles chosen are too large. If the angle of camera is too small, it is good to increase the image

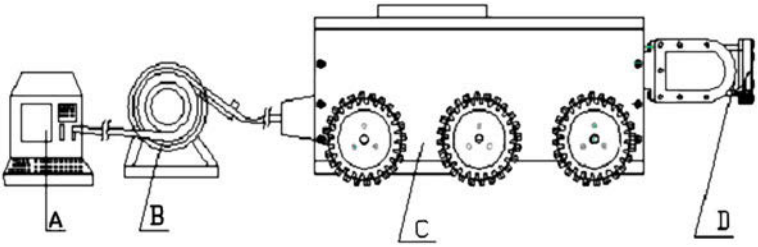


Fig. 104.1 Sketch of video inspection robot

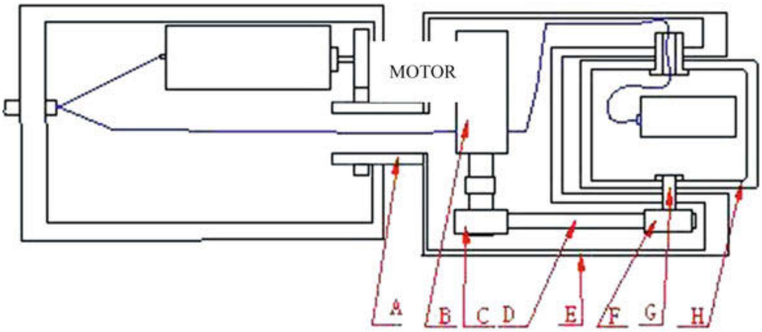


Fig. 104.2 Schematic of camera rotating mechanism's principle. *A* hollow shaft, *B* motor, *C* synchronization pulleys, *D* synchronous belt, *E* sealed cabin, *F* synchronization pulleys, *G* shaft, *H* sealed cabin

definition, but the camera scope is decreased. In the water, on the one hand, the light attenuation is very large, on the other hand, the dirt existing in the water make scattering of light quite serious. The above two points affect the quality of image. The first problem can be solved by strengthening illuminant, but for the second point, light scattering phenomenon increases with light strengthening, so it is necessary to design a camera device to expand the camera scope of the CCD to improve clarity of picture [7].

Motor driving rotation mechanism, embedded controller and sensors are sealed in crawling body, hollow shaft A fixed with sealed cabin E, motor drive hollow shaft A to rotate, and thus promote sealing cabin E, motor B drive shaft G to rotate through synchronous belt wheel C, shaft G fixed with sealed cabin H, so that camera device can rotate both in the vertical and horizontal directions, therefore camera device can be an omni-directional camera.

104.2.3 The Selection of Drive Mode

To meet the requirements of working in the narrow space of nuclear power plants, the volume of the robot must to be as small as possible. Usually drive mode has three ways, i.e. wheel drive, crawler drive, legged drive, caterpillar structure have good stable performance, the obstacle-surmounting performance and long life, but heavy crawler and a huge variety of driving wheel make its whole structure bulky; legged structure has good off-road capability and adaptability, but low efficiency; wheel drive structure is simple and easy to implement, moving with high efficiency. Comparing the advantages and disadvantages of these three ways, considering the requirements of working condition, wheeled driving mechanism is finally adopted. In order to increase the contact area between robot and pavement to moving more stable and efficient, six wheels drive is adopted [8].

104.2.4 Selection of Energy Supply Mode

Usually there are two ways to supply energy, i.e. have cable and no cable. For have cable, the main problem is that cable and surface friction will be very big when walking distance is very big, which seriously affects the robot's biggest walking distance. No cable supply of energy now has two kinds of schemes, one carrying battery, and the other carrying fuel generators. In addition to the huge volume that these two kinds of schemes have, they still have their respective weaknesses: the battery stores limited energy, and due to battery charging process quality, the influence of factors, the robot's walking distance is restricted. Because this robot walking distance required is not big, the cable is not long, if they carry battery or fuel generator, the weight and volume will surely increase. Using cable control operations, which can ensure control operation efficiency and reliability, when the equipment malfunctions, the robot can be withdrawn to a safety area through cable connections; thus the robot is controlled by cable.

104.2.5 The Selection of Sealing Arrangements

Sealing is key problem of developing underwater robots. Static sealing is easy to solve by O-ring and sealing gaskets. It is difficult to seal rotating axes (as shown in Fig. 104.3), because clearance existing between axes and chassis can produce leakage when shaft rotates. Due to the pressure not being high, and the wheel speed not high, O-ring sealing and ptfe combination sealing arrangements are adopted. As shown in Fig. 104.3, B is O-ring, and C is ptfe slip-ring, filler ptfe has good self-lubricity and wear-resisting performance.

104.2.6 Motor's Power Calculation and Motor Selection

Motor's power calculation and motor selection are one of the technical difficulties; the volume of the robot must be as small as possible. So in meeting the premise of

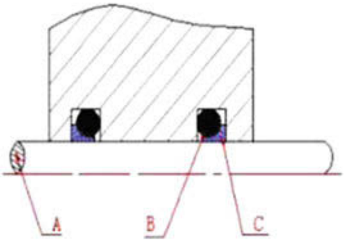


Fig. 104.3 O-rings-ptfe slip-ring sealing figure. *A* shaft, *B* O-ring, *C* ptfe slip-ring

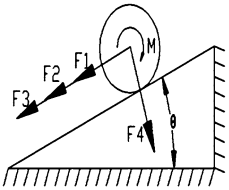


Fig. 104.4 Schematic of robot's climbing

power requirements, the motor chosen should be as small as possible. First, the power of robots required underwater is calculated as follows: resistance that robots suffered are as shown in Fig. 104.4 (without regard to water resistance)

Dynamic friction: $F1 = k1{*}F4 = k1{*}m1{*}\text{g}{*}\cos\theta$
Gravity weight: $F2 = m1{*}\text{g}{*}\sin\theta$
Cables and contact friction: $F3 = k2{*}m2{*}\text{g}{*}\cos\theta$
Uniform uphill by resistance: $F = F1 + F2 + F3$
Uniform uphill required motor minimum power: $P = F{*}V$
Cars have to overcome the minimum resistance moment: $M = (F1 + F2 + F3){*}\,\text{R}$
$k1$ for the robot and contact rolling friction coefficient
$m1$, $m2$ for quality of robot and cable respectively
$k2$ for cables and contact friction
R for wheel radius
V for speed of robot

The robot's quality is 15 kilos, The robot's load is 6 kilos, The cable's quality is 15 kilos, k1 is 0.2, k2 is 0.3, R is 90 mm. When the robot climbs 30° slope, the minimum resistance that the robot must overcome is 7.36 NM, power required is 1.64 W when robot running velocity is at 6 m per min. Considering underwater operations, motor select faulhaber company dc servo motor, power is 84.5 W, choose planets wheel reducer with reduction ratio 134, maximum output torque 10 NM, maximum output torque of the robot: 10*2 = 20 NM, which is bigger than 7.36 NM, power of the dc servo motor is 84.5 W, and maximum output power of the two motor is 169 W, which is bigger than 16.4 W, so the designs reach the requirements.

104.2.7 Control System Design

The control system uses the PC—single-chip microcomputer control system. PC is responsible for the image processing and sends control instruction, single-chip microcomputer is responsible for rock-bottom control and robots walk speed signal feedback, video images are transmitted through the optical fiber communication system by CCD camera, video capture card installed in PC, and use a visual processing program docking to received display and storage video images, which make the operator observe the robot operating status in real-time.

104.3 Conclusions

In order to adapt to underwater operations and to some environments where people cannot work conveniently, this paper introduced a kind of miniature video inspection robot. In order to expand the scope of CCD camera, the camera pitch device was designed. The robot can also be widely used in various gas pipe, oil pipeline, sewers, etc. It can not only be applied in underwater detection, but also in anhydrous environments, where it has a broad prospect of application.

References

1. Information on http://news.xinhuanet.com/english/2007-02/06/content_5703701.htm
2. Roh SG, Choi HR (2005) Differential-drive in-pipe robot for moving inside urban gas pipelines [J]. IEEE Trans on Robotics 21(1):1–17
3. Murayama R, Makiyama S, Mitutoshi K, Yasutoshi T (2004) Development of an ultrasonic inspection robot using an electromagnetic acoustic transducer for a lamb wave and an SH-plate wave. Ultrasonics 6:825–829
4. Chen HJ, Li JY, Zhang XH, Deng ZQ (2005) Application of visual serving to an X-ray based welding inspection robot. In: Proceedings of IEEE international conference on control and automation, pp 977–982
5. Zagler A, Pfeiffer F (2003) "MORITZ" a pipe crawler for tube junctions[C] Proceedings of IEEE international conference on robotics and automation, Sept 14–19, Taipei, China Taiwan. Taipei: IEEE, 2003:2 954–2 959
6. Zhang XL, Zheng HJ, Zhao LY (2001) A small pipe-inspection robot[J]. Robot 23(7):626–629
7. Suzumori K, Miyagawa T, Kimura M et al (1999) Micro inspection robot for 1-in pipes. IEEE/ASME Trans on Mechatron 4(3):286–292
8. Deng ZQ, Chen J, Jiang SY et al (2005) Traction robot driven by six independent wheels for inspection inside pipeline [J]. Chinese J Mech Eng 41(9):67–72

Chapter 105
The Optical Characteristics and Measurement, Model and Solution of LED

Li Ke, Liu Fengling and Xing Chaofeng

Abstract Each LED needs to carry out the optical properties of the test after completed package to ensure that a range of LED products is in line with the factory requirements. The optical structure on optimization of LED includes the optical structure of chips and packages redesigned and optimized, so that the light distribution applications can meet the requirements of illuminators. Previous studies are based on the method of geometrical optics to simulate and optimize proposed structure. This method can accurately control a moving direction of single light, whereas it was difficult to perform the light intensity distribution on the overall. Therefore, large area of chips and small package structure were not very precise and difficult to optimize. So, new methods are needed to design package structure. This thesis aims to study and discuss new ways to achieve the LED light intensity distribution of the overall implementation, focusing on establishing the LED optical characteristics measurement and mathematics solution for computer simulation.

Keywords LED · Optical properties · Electrical parameters · Mathematics mode

L. Ke (✉) · L. Fengling
School of Science, University of Jinan, Jinan 250022, China
e-mail: ss_lik@ujn.edu.cn

L. Fengling
e-mail: 220090921062@ujn.edu.cn

X. Chaofeng
Zhangqiu no.4 middle school, Zhangqiu 250200, China
e-mail: hzq-1230@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_105,

105.1 Optical Properties and Measurement of LED

Flux. Flux is the most important term to express performance of LED, which is the amount of light or light energy that accurately radiates from light source, also known as emission of light [1]. Flux is the amount of light what light source emits per unit time. The part of radiation energy that radiated power can be felt by human eyes. It equals some band radiation in per unit time multiplying its seeing rate depending on the relative. Since human eyes sees different rate depending on the relative as to different bands, so while the radiation power of different wavelengths of light are equal,the flux is not equal. The symbol of flux is Φ, its unit is lumen (lm) [2]. According to spectral radiant flux, the luminous flux formula is derived.

$$\Phi = K_m \Phi(\lambda) V(\lambda) \mathrm{d}\lambda. \tag{105.1}$$

In the formula, $V(\lambda)$—the relative spectral efficiency of light; Km—the maximum of the relative spectral effectiveness radiated, lm/W is the symbol. In 1977, CIPM determined its value as 683 lm/W Km (λm = 555 nm) [3].

Illumination. Light intensity is the extent to which the object is illuminated, which can be expressed by the luminous flux accepted per unit area (Fig. 105.1).

Light illumination is related to lighting source, the illuminated surface and the location of light source in space, whose size is proportioned to the light intensity of light source and the cosine of the angle of incidence, but is inversely proportioned to the square of the distance light source to the illuminated surface. Point illumination on the surface is incident on the panel that contains the point on the luminous flux dΦ divided by the bin area of the business ds. The symbol of light illumination is Lux,1Lux=1 lm/m2. Objects with uniform irradiation by light, when getting luminous flux in the area of 1 square meter is 1 lumen, its illumination is 1 Lux.

Viewing angle. Notation: θ1/2; symbol: degree(o); definition: θ/2 refers to the luminous intensity as half the axial direction and strength values of the angle between the light axis. The angle defined is 2 times the half-value angle, also known as perspective (or as half-power angle), as shown in Fig. 105.2.

Illustration. Figure 105.2 shows that each LED in an LED luminous intensity angular distribution. The coordinates of 0 degree is the vertical (normal), the relative maximum luminous intensity. The greater it leaves the normal direction of the angle, the smaller the relative intensity is. From the figure, half-value angle or angle value can be concluded.

Luminous intensity. Luminous intensity ($I_V = \mathrm{d}\Phi/\mathrm{d}\Omega$) is what a light source emits, luminous flux, in a given direction of the unit solid angle. In the formula dΩ is a point source in the party up the sheets of solid angle element, as shown in Figs. 105.3 and 105.4.

I_V is the luminous intensity, which expresses emitted luminous flux in a certain direction of light within the spatial distribution of physical quantities. I_V value of the detected LED is usually defined as the line (on the light-emitting diode is its cylindrical axis) direction of the light intensity. If the radiation intensity on the

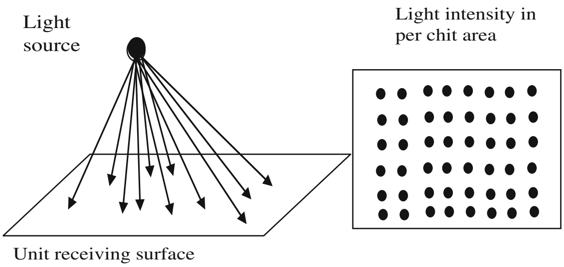


Fig. 105.1 Illumination

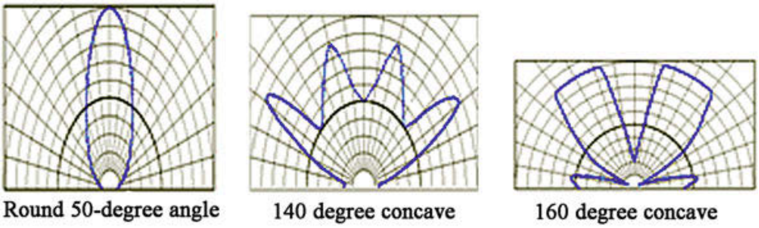
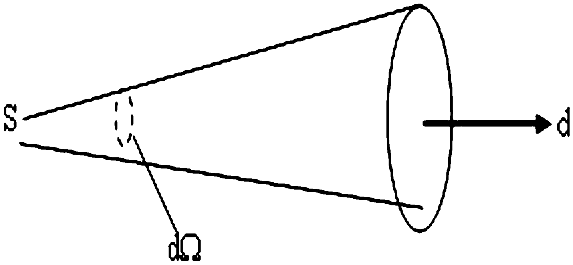


Fig. 105.2 Angular distribution of several LED light-emitting

Fig. 105.3 Luminous intensity and luminous



direction is (1/683)W/sr, then the light-emitting is 1candela (symbol for cd). As the luminous intensity of LED is in general smaller, so mills-candela is commonly used as unit for luminous intensity [4].

Measurement of Optical Properties. During the detection of LED optical and electrical properties, integrating sphere and the angle of light intensity distribution test system are commonly used.

Structure of integrating sphere. The basic structure of the integrating sphere is made of aluminum or plastic, an interval hollow ball. Multi-layer neutral diffusing material is coated evenly on the inner wall of the ball, such as magnesium oxide barium sulfate and PTFE. There are several holes opening on the ball, used as the incident light holes, installing detector, light source, etc. In order to prevent the

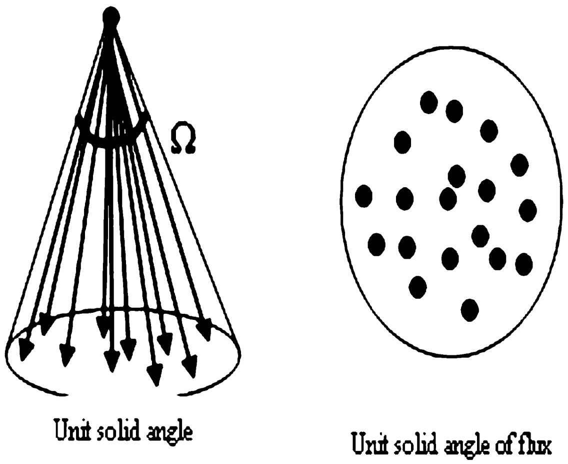


Fig. 105.4 Luminous intensity

incident light directly hitting the detector, the ball is also equipped with a blocking screen, as shown in Fig. 105.5.

Integrating sphere theory. The basic principle of integrating sphere is that the light incidences from the input hole, then is evenly reflected and diffused within this sphere, so the light got from the resulting output hole is the rather even diffused beam. And the incident angle of incident light, spatial distribution and seasonal will never impact the intensity and uniformity of the outputting beam. Yet, the light injects out until through the points within the integrating sphere. Therefore, integrating sphere can also be seen as a light intensity attenuator. The ratio of the output intensity and input intensity is the area of light output hole/ internal surface area of the integrating sphere.

The light intensity distribution of the test system of the angle. The light intensity distribution of the test system of the angle is for automatic test of LED light intensity angle distribution, the largest light intensity value of single LED, value of 0 light intensity, light intensity deviation angle, spread angle of light intensity, and LED light intensity at both ends of the load electrical parameters, providing distribution curve analysis function of $I - I_V$, $I - V_f$, $\theta - I_V$.

Figure 105.6 is the light intensity distribution of the test system of the angle made up of LED light intensity test platform, V detector, the host and so on [5]. Working theory: the rotating platform base (fixtures clipping LED) has an active door and covers the LED on the fixture,which can be regarded as a opaque black box (the outside light can be ignored), light emitted from LED eliminates glare through extinction effects of the barrel of extinction. When light enters the CCD detector probe, the probe converses optical signals to the host. After treatment, electrical signals are displayed from computer software.

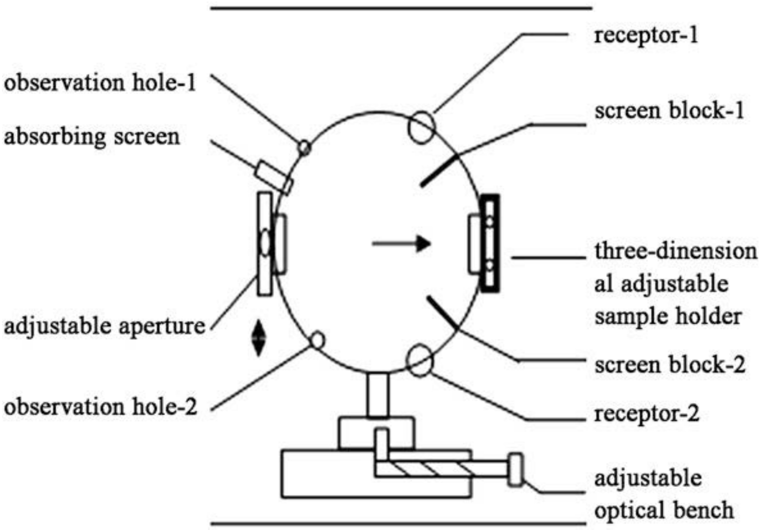
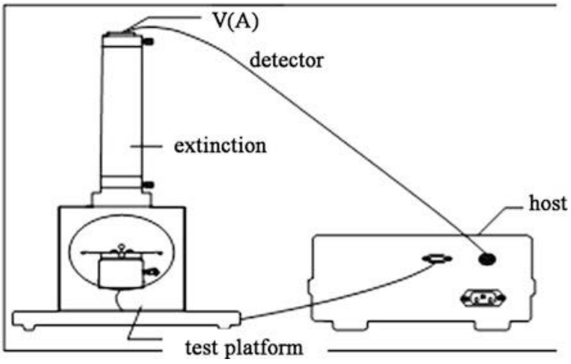


Fig. 105.5 Integrating sphere

Fig. 105.6 Test of spatial angle intensity distribution

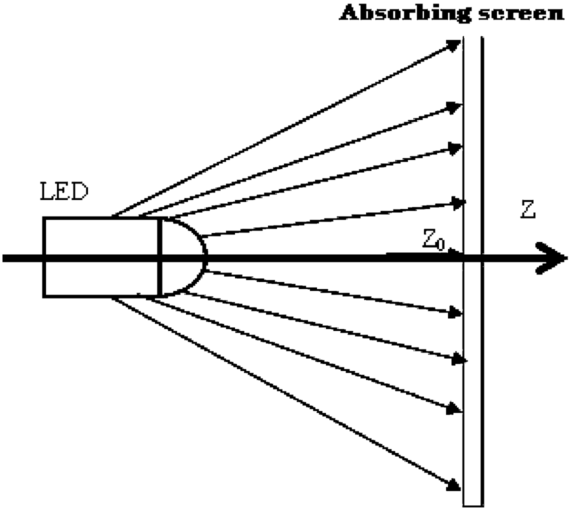


105.2 Measurement Model and Solution of LED

The characterization of the optical parameters of LED light sources and their measuring tool are discussed, the measuring tool (integrating sphere and angular distribution of the test system) is often used in factories and laboratories [6]. However, in the simulation,the main principle of this tool is for the mathematical model, for the establishment of photons emitted are needed to be described LED model, which will be output to the computer screen,getting the optical properties of simulated LED.

The Establishment of measurement model. In the simulation, graphics of LED's light intensity distribution, three-dimensional distribution of light intensity and

Fig. 105.7 Absorbing screen



space simulation distribution curve are mainly got. In practice, integrating sphere and angle distribution measurement system are used to achieve the absorption and the test of light from LED. In the simulation, a light absorption surface can be built to converge photon, in the following,such absorption surface in being established.

As shown in Figs. 105.7 and 105.8, the absorption surface can be defined as a plane. Its mathematical expression is:

$$Z = Z_0. \tag{105.2}$$

where, Z_0 is the point between absorbing surface and Z axis, the plane is parallel to the XOY plane.

Maintaining the integrity of the specifications. Flux: the size of flux can be expressed by the number of photons, so in the simulation it is OK as long as the random vector number of photons is defined. For example,the number of early shot photons can be defined 1,000, 10,000, 100,000 other, whose difference is that the photon statistics are not the same as the number and the length of running time of computer CPU is different. Light intensity distribution: if there is a point of intersection after the photons emit from the LED with the absorption surface, then the second intersection is recorded. The formula for calculating the point of intersection is:

$$\begin{cases} X = l\frac{U-Z_0}{n} + Z_0, \\ Y = m\frac{U-Z_0}{n} + Y_0, \\ Z = Z_0 \quad . \end{cases} \tag{105.3}$$

where, (l, m, n) is the outgoing direction vector of photon, (X_0, Y_0, Z_0) is the exit point coordinates of photo, U is the distance between the origin coordinates and the viewing screen.

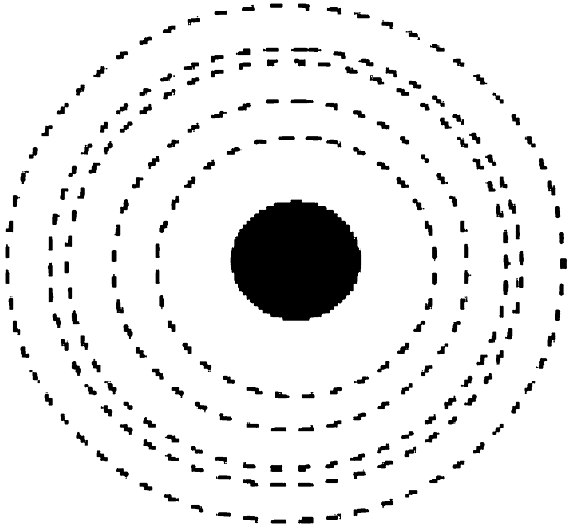


Fig. 105.8 Photon distribution on the screen

Displaying all intersection of points received by absorbing screen on the computer screen, then graphics of light intensity distribution is concluded.

The light distribution curve of light intensity is generally expressed by light distribution curve of LED light intensity. Simulation results are to statistic the point of view of its spatial distribution, resulting in photon quantities of unit space angle. The photon quantities of unit solid angle is the luminous intensity of space angle, to establish a 'solid angle-the photon number' artesian coordinate system that can be very intuitive to reflect light distribution curve of LED.

The light intensity distribution: In order to obtain the illuminate value of a point on the absorbing screen, to statistic photon quantities of the small area to this point as the center, to calculate light intensity values, which can be as the average photon quantities and illuminate values of this point.

Here, absorbing screen can be divided into $m \times n$ small areas to form a planar array. Each point value of each planar array expresses the statistical distribution of light intensity on the absorbing screen. According to point of each array illumination value, plane light intensity distribution to three-dimensional graphics are made, which can intuitively show the ratio between the size of light intensity of each point and the light intensity value of each point.

105.3 Summary

In short, understanding the optical characteristics and establishing measurement model and solution are very important steps for LED package. In addition, using the simulation method of Monte Carlo can also solve a more accurate optical structure model, and give specific imaged results of light distribution, the impact of modified light distribution can then be seen through correcting the parameters of light structure. Meanwhile,the combination of C++MFC with TRACEPRO software can be very simple and intuitive to perform the optical properties of LED model, and provide important reference basis for LED products designed for use to achieve the design requirements of industrialization.

References

1. Guo Q, Zhang L (2009) Design based on the LED light and electrical parameters of the test system[J]. OME Inf 26(6):29–32
2. Zhang J, Fang S, Hu Z (2008) LED package design of an optical system optimization[J]. Semicond Optoelectron 29(5):658–661
3. Tu D, Wu R, Hengliang Y (2008) The impact of intensity distribution to light structure of LED package. Optical Precis Eng 16(5):832–838
4. Hu H (2005) The design of LED lighting optical system and its array research about illumination distribution[J]. CNKI, 09(27)
5. Zhang J, Fang S, Hu Z (2008) LED package design[J] of an optical system optimization. Semiconductor Optoelectron 29(5):658–661
6. Chen Y (2009) The manufacturing technology and applicationof LED[M], 2nd edn. Electronic Industry Press, Beijing

Chapter 106
The Application of Remote Sensing Technology to Soil Salinization in Yanqi Basin

Xuejiao Qin, Chunfang Kong and Yunxia Ren

Abstract As soil salinization is one of the important factors influencing the agriculture production and ecological environment, it is of great significance to enhance the remote sensing monitoring for saline alkali soil. Landsat7 ETM+ remote sensing image was used to take the samples and make supervised classification by analyzing the spectral signature according to the different spectral patterns of ground objects. Meanwhile, the distribution of soil salinization in Yanqi basin is quantitatively assessed combined with field investigation and soil sampling as well as salt content measurement of the soil samples. The results show that mild saline soil accounts for 19.777%; moderate saline soil accounts for 5.378% and severe saline soil accounts for 0.209%.

Keywords Remote sensing · Yanqi basin · Landsat7 ETM+ · Soil salinization

X. Qin (✉) · C. Kong
College of Computer Science, China University of Geosciences,
Wuhan, China
e-mail: qxj227@126.com

C. Kong
e-mail: xkcf@163.com

Y. Ren
College of Earth Science and Tourism, Xinjiang Normal University,
Urumqi, China
e-mail: rhqjg@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_106,

106.1 Introduction

Soil salinization is a common form and phenomenon of dynamic land degradation in arid and semiarid areas, which seriously influences the ecological environment and the development of social economy. Xinjiang province located in the inland of northwest China, has a large distribution of saline areas. The area of saline soil in this territory accounts for as much as 19.25% of useable land while accounting for 30.85% of cultivated land at the same time [1]. It is of great significance to enhance the investigation and monitoring of soil salinization for the improvement and utilization of saline soil.

So far, the remote sensing technology is widely used in saline soil monitoring because it is macroscopic, dynamic, integrated, real-time, and less subject to ground conditions. Remote sensing of soil salinity on reflection is relatively sensitive, the higher the salt content, the stronger the spectral reflectance in the image picture. Usually, saline soil is shown as light-colored or gray in the image picture. Compared with the general land, salinization of soil has more intense spectral reflectance in the visible band and near infrared band [2].

106.2 General Situation of Study Area

Yanqi Basin is located in the hinterland of the Eurasian continent and has a typical temperate arid desert climate. As a mountain basin of South Tianshan in Xinjiang, Yanqi Basin is slightly diamond-shaped, the terrain slopes from northwest to southeast. The lowest is the largest inland freshwater lake Bosten Lake, the upper reaches of the lake is Kaidu River, and downstream is Kongque River. Bosten Lake is the water and salt influx center of the basin [3]. From the mountains surrounding the basin to the Bosten Lake, they are in sequence of Gobi, plain and lake. Besides, the dry climate, evaporation and groundwater depth which is mostly between 1 and 3 m, result in serious soil swamping and soil salinization in this area. Generally speaking, common types of saline soil include meadow saline soil, ordinary saline soil and swamp saline soil [4], all of these seriously affect the development and manufacturing of local agriculture.

106.3 Data Sources

Data used in this study include a Landsat7 ETM+ image imaging on Sep 26, 2009 and salt content data of soil samples received by field sampling and indoor determination.

106.4 Data Processing

Measured data. In this research 48 points of soil samples have been taken randomly through a field investigation in the Hejing county, which is located in the northwest of Yanqi Basin (Fig. 106.1). Then the salinity of soil samples was determined in the laboratory. According to the geographical environment of inland arid areas, compare with the soil salinity classification standard and divide soil salinization types of samples.

Remote sensing image preprocessing. Atmospheric correction, geometric correction, radiometric correction and so on were carried out before classification by using image processing software ENVI. Among them, control the absolute error of the geometric correction in less than 50 m and the relative error within the 15 m. In addition, fuse the resolution of the images, superpose different bands, and enhance the image (such as linear stretch, gray level transformation, etc.). After pretreatment, the image's spatial resolution is 15 meters and the image has 6 bands in total.

Selection of feature variables. A large number of remote sensing researches on soil salinization at home and abroad found that, visible band, near infrared band and shortwave infrared band are the key bands to identify soil salinity. Whether in the visible band or near infrared band, the image tone of saline soil is lighter than other soils in the remote sensing images. Furthermore, the higher the salt content, the stronger the spectral reflectance. In terms of sheer amount of information, a combination of TM 1, 3 and 5 bands contain the most information content of the data, but the accuracy of soil salinity information retrieval is not proportional to the remote sensing information content [5].

According to this research needs, the remote sensing image of study area was initially divided into 6 classes- vegetation (farming, forest, grassland), water, desert, desert, wetlands (including swamps), saline soil. A total of 20 sample points were selected for each of the typical objects after repeated selection, then obtain the spectral curves (Fig. 106.2). Through the spectral curve, what can be found is that saline soil can be distinguished rather well in band 2, band 3 and band 4. Also, it is easy to recognize other ground objects information in these three bands. Therefore, band 4, band 3 and band 2 were combined to get a 4–3–2 false color image.

Supervised classification. With conjoint analysis the field surveys and remote sensing image, first of all, the natural land, including saline soil and non-saline soil, can be separated from water, marshes, wetlands and vegetation zones. Second, according to different levels of soil salinity, different image features in the remote sensing image and different degrees of saline land can be distinguished. In fact, based on correlation, severe saline soil in the false color composite image map shows gray, rough texture; medium saline soil is light gray and has more smooth texture; slight saline soil in the image shows gray spots and smooth texture.

In this research, under the condition of ensuring 200 samples of each feature, the most relieved supervised classification method was used. The surface features

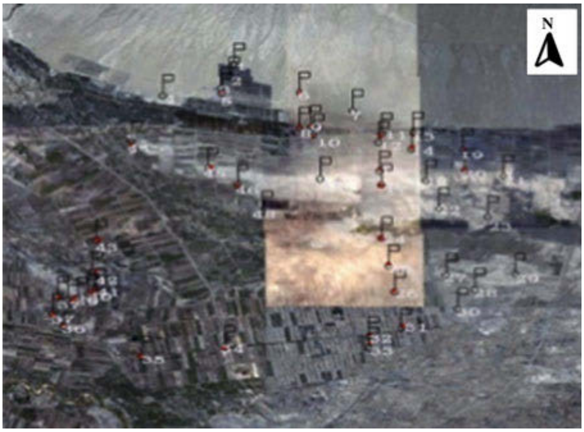


Fig. 106.1 Distribution of samples

were divided into eight categories. They are water, wetlands (including swamps), vegetation, Gobi, desert, severe saline soil, moderate saline soil and mild saline soil. Subsequently, the accuracy of classification results was evaluated (Table 106.1).

From the table of classification results confusion matrix, it is easy to conclude that the water is easy to distinguish from other surface features. Although there was a big mixture between vegetation (farming, forest, grassland) and wetlands (including marshes), it was due to the seasonal agricultural irrigation. Farmlands after irrigation and marshes are similar in the spectral features, but the study focuses on the delineation of saline soil, so the mixture had little effect on the research. In addition, either the mixture between saline soil and the Gobi or the mixture between different types of saline soil has been resolved fairly well.

Post classification. Whether using supervised classification or unsupervised classification, remote sensing image classification is processed separately for each pixel. So it is inevitable that noise appears in the classification surface features. In other words, heterogeneous surface features scattered distribute in large similar surface objects. In order to eliminate the impact of class noise, the class attribute of each pixel is compared with its neighboring pixels. If the pixel is inconsistent with the surrounding neighboring pixels, it should be adjusted to meet the consistency of the situation [6]. According to the features relatively continuity principle, 5 × 5 window was selected in the neighboring analysis for classification result graphs.

106.5 Analysis

Statistical results of this study show that severe saline soil occupies 0.209%, moderate saline soil occupies 5.378% and mild saline soil occupies 19.777% in the Yanqi Basin (Fig. 106.3 and Table 106.2).

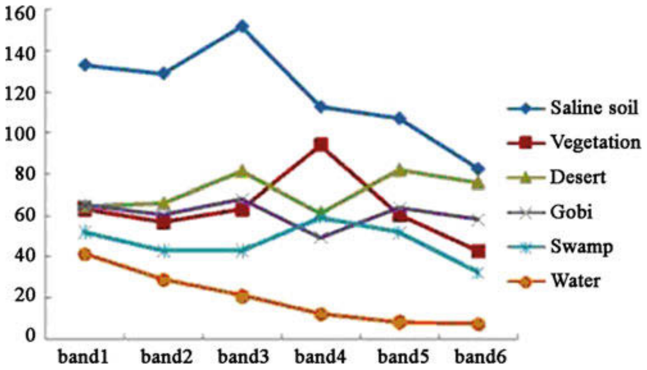


Fig. 106.2 Spectral curves of landmark

Table 106.1 Confusion matrix of classification results

	1	2	3	4	5	6	7	8	Total
1	195	0	0	0	0	0	0	0	195
2	5	174	2	0	0	0	0	4	185
3	0	1	198	0	0	0	0	0	199
4	0	0	0	200	0	0	0	0	200
5	0	0	0	0	170	8	0		178
6	0	0	0	0	30	192	0	0	222
7	0	25	0	0	0	0	200	0	225
8	0	0	0	0	0	0	0	196	196
Total	200	200	200	200	200	200	200	200	1600

1 Severe salinization, *2* moderate salinization, *3* mild salinization, *4* water, *5* wetlands (including marsh), *6* vegetation, *7* Gobi, *8* desert

Classification accuracy = 95.3125%, Kappa coefficient = 0.9464

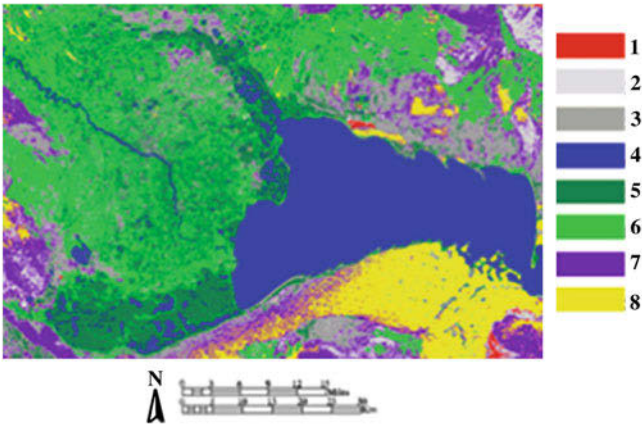


Fig. 106.3 Remote sensing image classification results. *1* Severe salinization, *2* moderate salinization, *3* mild salinization, *4* water, *5* wetlands (including marsh), *6* vegetation, *7* Gobi, *8* desert

Table 106.2 Soil salinization classification and area of Yanqi basin

Soil types	Area (km^2)	Proportion (%)
Severe saline soil	11.30670	0.209
Moderate saline soil	290.45025	5.378
Mild saline soil	1068.11145	19.777

106.6 Conclusions

Landsat7 ETM+ remote sensing image data is used in this research. First, spectral analysis is made for the ground objects. According to the spectral characteristic curve of different ground objects in different spectral bands, the 4 band, 3 band and 2 band were selected to obtain a false color composite image. Subsequently, the supervised classification was carried out and combined with field investigation and soil sample analysis,and cluster analysis at the same time. Thus, the degree and distribution of soil salinization in Yanqi basin is evaluated qualitatively and quantitatively. Kappa coefficient of the study was 94.64%. Therefore, the remote sensing data was analyzed and disposed to evaluate the soil salinization and the evaluation results were verified with the actual situation. The entire study area can be appraised from flat with this method, which consequently makes up the lack of information obtained from points in the evaluation of soil salinization. So the large quantity of time, manpower and material can be saved by using the remote sensing data to evaluate the salinization of soil. Hence, the method should be promoted widely in the soil salinization evaluation.

Acknowledgments This work has been supported from the Special Fund for Basic Scientific Research of Central Colleges (No: CUGL090232), Key Lab of Biogeology and Environmental Geology of Ministry of Education (No: BGEG1014), China University of Geosciences, Wuhan; and the Fund for Geological Survey of China Geological Survey (No: 1212010880404). The authors thank the anonymous reviewers for providing valuable comments on the manuscript.

References

1. Xu DQ (2004) Remote sensing technology for dynamic monitoring of soil salinity. Xinjiang Agriculture University, Urumqi
2. Rao BRM, Sankar TR, Dwivedi RS (1995) Spectral behaviour of salt-affected soils. Int J Remote Sens 16(12):2125–2136
3. Ma N, Yang L, Li JL (2008) Study on hyperspectral distinction of soil salinity: a case study of Yanqi, Xinjiang province. J Arid Land Resour Environ 22(2):114–117
4. Liu YF, Qi MG, Jin YC (2004) Characteristics analysis of soil salinization in Yanqi basin of Xinjiang Wei autonomous region. Bull Soil Water Conserv 24(1):49–52
5. Dwivedi RS (1992) Monitoring and the study of the effects of image scale on delineation of salt-affected soils in the Indio-Gangetic plains. Int J Remote Sens 13(8):1527–1536
6. Guan YX, Liu GH, Liu QS (2001) Remote sensing investigation of the Yellow River Delta saline. Int J Remote Sens 5(1):46–52

Chapter 107
A Watermarking Scheme Based on Video's Independent Features

Chunxing Wang and Xiaomei Zhuang

Abstract In this paper, a video watermarking scheme based on video independent static feature (VISF) and video independent dynamic feature (VISF) is proposed. In this scheme, we embed watermark into the dynamic component which is extracted from raw video based on ICA to guarantee the perceptual quality of the watermarked video. The simulations show that the proposed scheme has a good performance to resist Gaussian attack, salt and pepper attack, cutting attack, jpeg compression attack and MPEG-2 compression attack.

Keywords Video watermarking · VISF and VIDF · ICA

107.1 Introduction

Most studies on watermarking technology focus on the still image. In recent years, with the widespread application of digital video products in the network, effective copyright protection techniques are an urgent problem, either to prevent unauthorized copying or at least as an evidence of copyright infringement. The digital watermark used for protecting digital video is becoming a hot issue [1]; many algorithms of video watermarking have been proposed, and it has already been applied to Video-On-Demand (VOD), interactive multimedia

C. Wang (✉) · X. Zhuang
Shandong Normal University, Jinan,
Shandong Province, China
e-mail: cxwang@sdnu.edu.cn

X. Zhuang
e-mail: zhuangxiaomei1987222@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_107,

application system, DVD and so on [2]. Therefore, watermarking technology is becoming more and more attractive. It should meet requirements as follows: invisibility, blind detection, robustness to compression, synchronicity attack and collusion attack [3].

Video watermarking, in a sense, is similar to image watermarking in theory. So, many researchers applied results on image watermarking directly to video watermarking. But images have limited signal space, as video image sequences are different; they have an important feature that is time redundancy [4]. In video compression coding, by means of motion estimation and compensation, images information about "the front" and "the back" frames can be represented by motion parameters and residual images. The method enhances data compression ratio enormously. Consequently, video data are transmitted and stored in compressed form. Video data stream have to be decoded first when detecting compressed video watermarking.

In this paper, we propose a new video watermarking scheme based on video independent dynamic feature (VIDF). First, the raw video is divided into video independent static feature (VISF) component and VIDF component. We select VIDF to embed the watermark. Besides, the ICA is used before embedding. The simulations demonstrate that the proposed scheme can keep a good video fidelity and is robust to common attacks, such as Gaussian noises, salt and pepper noises, cutting, resizing, frame dropping, frame changing, as well as the MPEG-2 compression. This paper is organised as follows: Sect. 107.2 introduces Extracting the Independent Component, both the VIDF and the VISF; Sect. 107.3 describes the proposed watermarking scheme; the experimental results and conclusions are presented in Sect. 107.4 and 107.5 respectively.

107.2 Extracting the Independent Component

We regard an entire frame as a target and extract the motion and static vectors between every two successive frames [5–7]. If $f(i)$ is the current frame, $f(i-1)$ is the previous frame and $m(i)$ is the motion vector of previous frame, then their relationship can be presented as follows, Eq. 107.1:

$$f(i) = f(i-1) + m(i). \tag{107.1}$$

The VIDF $m_{m(i)}$ and the VISF $m_{f(i-1)}$ are extracted by means of ICA (Independent Component Analysis) [8, 9]. We can regard two successive frames as linear superposition of mutual static part $m_{f(i-1)}$ and relative motion component $m_{m(i)}$; that is Eq. 107.2 (Fig. 107.1):

$$\begin{bmatrix} f_{f(i)} \\ f_{f(i-1)} \end{bmatrix} = \mathbf{A} \begin{bmatrix} m_{f(i-1)} \\ m_{m(i)} \end{bmatrix}. \tag{107.2}$$

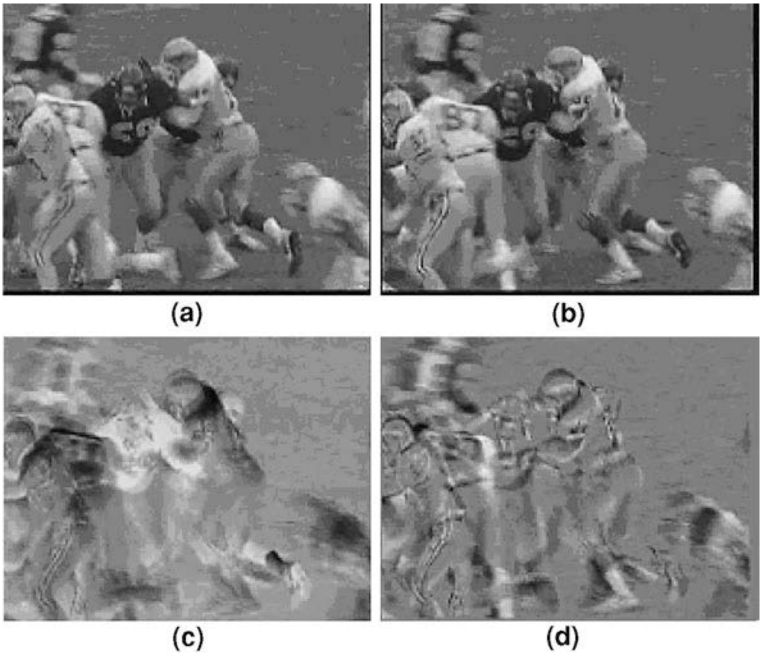


Fig. 107.1 VISF and VIDF are extracted from two successive frames in the "football" video. **a** The previous frame of fb.avi; **b** the back frame of fb.avi; **c** exacted VIDF; **d** exacted VISF

107.3 Watermarking Scheme

In our scheme, we embed watermark into VIDF component. Assume W is watermark, which is a binary image. We operate the embedding watermark in DWT domain (Fig. 107.2).

Referring to Eq. 107.2, we can use the matrix to obtain the motion component. The relative motion component is defined by $m_{m(i)}, m(i,j) \in M$, and the watermarking binary image is w, $w(k,l) \in W$. If the pixel value of video frames image is $p \times q$, then the pixel value of chosen W binary image is $\frac{p}{2} \times \frac{q}{2}$. Define that $a(k,l) \in LL_1$ and one vector set of MRR is $\overrightarrow{s}(k,l), \overrightarrow{s}(k,l) = (s_1(k,l), s_2(k,l), s_3(k,l))$. They meet the following relationship:

$$\begin{bmatrix} a(k,l) \\ s_1(k,l) \\ s_2(k,l) \\ s_3(k,l) \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \begin{bmatrix} g(2k,2l) \\ g(2k+1,2l) \\ g(2k,2l+1) \\ g(2k+1,2l+1) \end{bmatrix}. \tag{107.3}$$

Then calculate $\delta(k,l) = \left| \max_i \{s_i(k,l)\} - \min_j \{s_j(k,l)\} \right|$ mod 2 in the place (*k*, *l*).

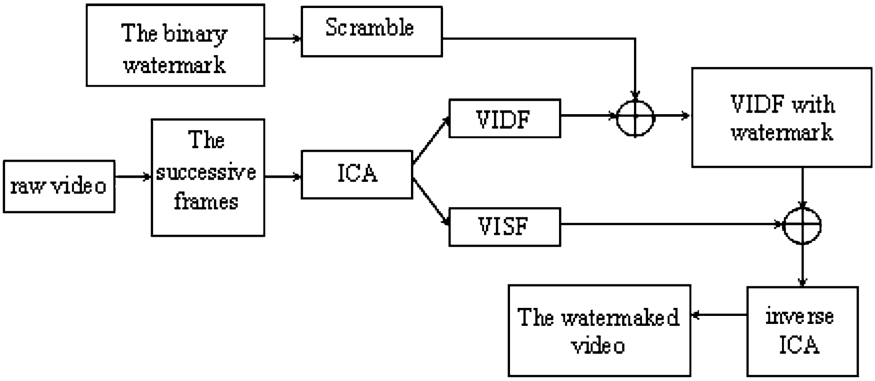


Fig. 107.2 Block diagram of embedding watermark into VIDF

When extracting the watermark from the to-be-detected video, we extract watermark from the watermarked video according to the following steps.

Obtaining δ (k, l) with the same method in the embedding process.

$$w(k,l) = \text{floor}\left(\frac{\delta(k,l)}{\Delta}\right) \bmod 2$$

107.4 Experimental Results

In simulations, we choose the two successive frames (frame 30 and 31 in the fb.avi 272*352) in the experiment. Transform the images from RGB into YCrCb and operate on the luminance component Y [10]. And we choose to use the FastICA algorithm. The value of Th is 0.7 (Fig. 107.3).

Divide the VIDF image into 8*8 blocks. Then process each block with DWT and embed the watermark of 68*110. The simulation results show that the PSNR between the original motion component image and the motion component of watermarked image is 39.3713 (Fig. 107.4).

In the restoring process, we combine the motion component signal with the static component signal to form the matrix of estimating signals. Then make use of the aliasing matrix to reconfig. the original signals. Restore the original grayscale images with the help of recovered sub-image and on the basis of that, restore the color images. The PNSR of the original grayscale images and the recovered grayscale images are respectively 41.1725 and 42.0443. The PNSR of the original color images is 41.2019 and the recovered is 42.0544.

The robust against selected attacks result is as follows:

The Gaussian attack (Fig. 107.5)

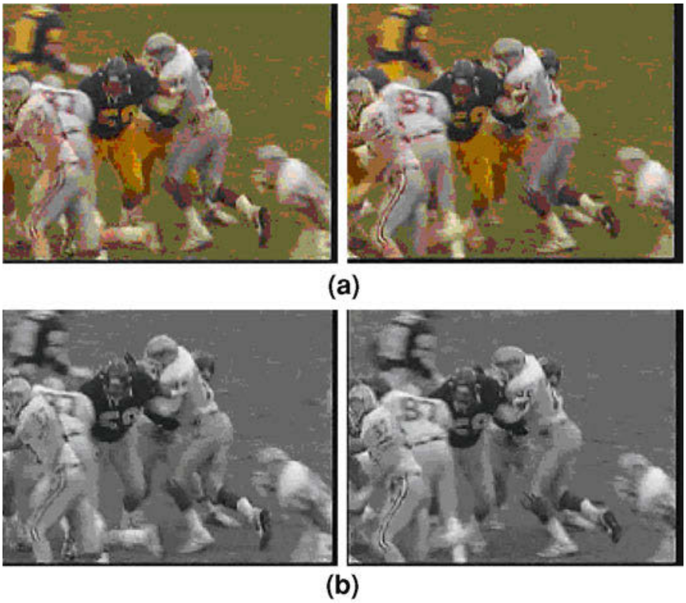
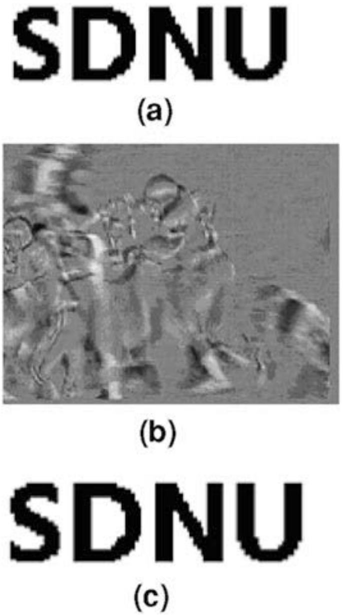


Fig. 107.3 The frames intercepted randomly. **a** The successive frames of the original video; **b** the grayscale images of the successive frames in the video

Fig. 107.4 The original motion component image, the motion component of watermarked image and the watermark, PSNR = 39.3713. **a** The original watermark; **b** VIDF with the watermark; **c** the extracted inverse scrambling watermark



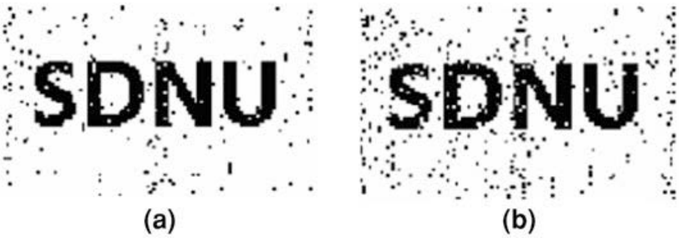


Fig. 107.5 The extracted watermark from Gaussian attack. Normalized correlation coefficients are respectively 0.9447 and 0.8810. (Here, *m* and var denote the mean and the variance of the Gaussian white noise respectively.) **a** $m = 0$, var $= 0.000007$; **b** $m = 0$, var $= 0.000005$

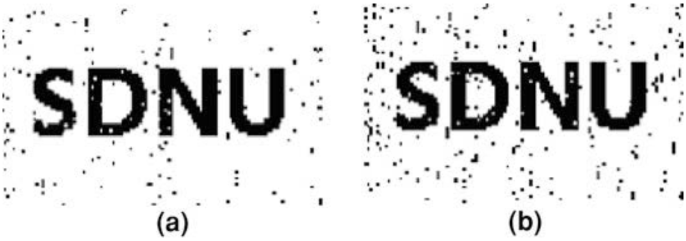


Fig. 107.6 The extracted watermark from the salt and pepper noise attack. The values of NC are respectively 0.9583, 0.9238. (Here, *d* denotes the noise density.) **a** $d = 0.00001$; **b** $d = 0.00011$

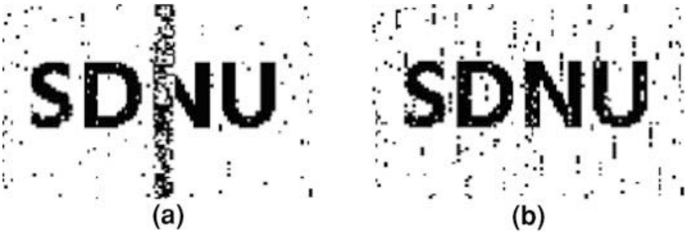


Fig. 107.7 The extracted watermark from cutting attack (NC = 0.9032, 0.9160). **a** Cut off some rows; **b** cut off some columns

The salt and pepper attack (Fig. 107.6)
The cutting attack (Fig. 107.7)
The jpeg compression attack (Fig. 107.8)
The MPEG-2 compression attack (Fig. 107.9)

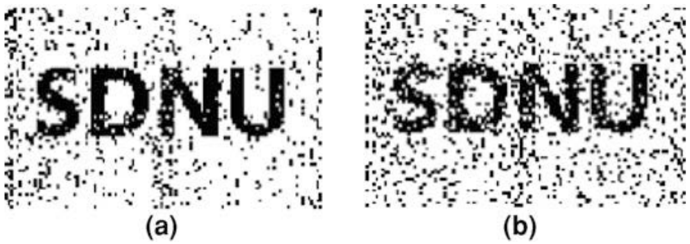


Fig. 107.8 The extracted watermark from compression attack (NC = 0.8000, 0.6909). **a** Compress 90; **b** compress 80

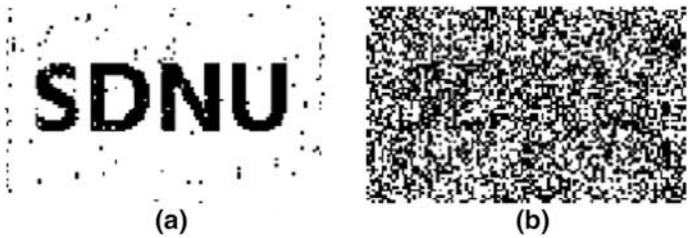


Fig. 107.9 The extracted watermark from MPEG-2 compression attack (NC = 0.9658, 0.1532). **a** Uncompressed; **b** compressed

107.5 Analysis and Discussions

The paper presents the scheme based on VIDF and VISF of successive video frames. The VIDF and VISF are obtained by use of the ICA technique. We embed watermark into the VIDF's high-frequency coefficients of the original video to have a good fidelity, and make use of DWT when embedding to get a good robustness. The results of simulation experiment prove that the scheme can resist some common video attacks such as Gaussian attack, salt and pepper attack, cutting attack, jpeg compression attack and MPEG-2 compression attack. In our ongoing work, we will work on the improvement of the property of real-time.

References

1. Zhang J, Jiegu Li,Zhang L (2001) Video watermark technique in motion vector. IEEE Comput Graph Image Process 963053:179–182
2. Szu H, Noel S, Yim S-B, Willey J, Landa J (2003) Protecting multimedia authenticity with ICA vaccination of digital bacteria watermarks. In: Proceedings of the international joint conference on neural networks, vol 1, pp 131–136

3. He Z, Liu J, Yang L (1999) Blind separation of images using edgeworth expansion based ICA algorithm. Chin J Electron 8(3):278–282
4. Sun J, Liu J, Zhang X (2003) A blind video watermarking scheme based on ICA. In: IEEE international conference on neural and networks & signal processing, vol 2, pp 1127–1130
5. Toch B, Lowe D, Saad D (2003) Watermarking of audio signals using independent component analysis. In: Proceedings of third international conference on web delivering of music, pp 71–74
6. Yu D, Satter F, Ma K-K (2002) Watermark detection and extraction using ICA method [J]. EURASIP J Appl Signal Process 1:92–104
7. Szu H, Noel S, Yim S-B, Willey J, Landa J (2003) Multimedia authenticity protection with ICA watermarking and digital bacteria vaccination. Neural Netw 16:907–914
8. Yu D, Satter F, Ma KK (2002) Watermark detection and extraction using ICA method. EURASIP J Appl Signal Process l:92–104
9. Bounkong S, Toch B, Saad D, Lowe D (2003) ICA for watermarking digital images. J Mach Learn Res 4:1471–1498
10. Piron L, Arnold M, Kutter M, Funk W, Boucqueau JM, Craven F (1999) OCTALIS benchmarking: comparison of four watermarking techniques. In: Proceedings of SPIE'99, vol 3657, pp 240–250

Chapter 108
Parallel Beamforming Technique Aimed at Increasing Frame-Rate of Medical Imaging

Lutao Wang, Gang Jin and Binjie Liu

Abstract Increasing frame rate is needed for high quality ultrasound imaging for research on tissue motion. The demand of increasing frame rate is even higher for 3D ultrasound imaging. In this chapter, we introduce a new technique which can increase frame rate without decreasing image quality. Wider transmit beams combined with four narrower parallel receive beams are formed to increase the frame rate by a factor of four. Through mainlobe width controlling and sidelobes level suppressing, the inherent gain loss for normal parallel beamfoming can be compensated in the maximal degree. Thus, better resolution and contrast can be obtained compared to normal window-based apodization beamformer. Because the computational cost is the same as delay and sum beamforming whose computation cost is linear with the number of array elements, this method has great advantages of possibility of real-time implementation.

Keywords Ultrasound imaging · Beam forming · Parallel acquisition · Contrast resolution

108.1 Introduction

When performing ultrasound imaging of tissue motion or moving heart, increasing the frame rates is needed. Because frame rate is inverse proportional to the number of transmit beams required to scan an image, the commonest way used to increase

L. Wang (✉) · G. Jin · B. Liu
Department of Automation, University of Electronic Science and Technology of China, Chengdu 611731, China
e-mail: wltuestc@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_108,

frame rates of ultrasound imaging without compromising the scan lines is to use parallel beamformers. According to this approach, a transmit beam combined with multiple receive beams which are simultaneously acquired from closely spaced regions are used to increase the scan lines for each transmit-beam. So the imaging frame rate is increased proportionally to the number of receive beams. But the mainlobe gain loss introduced by parallel beamforming will decrease the imaging quality [1].

Using narrower receive beams will increase lateral resolution. Synnevag [2] applied minimum variance (MV) beamforming to form receiving beams. The beam width was reduced to 1/4th and the imaging frame rate was increased by a factor of four and comparable imaging resolution was obtained. But the operation of matrix inversion which is used to find the optimize aperture weights in the MV beamforming is complexity proportional to the array size, O(L3) and this can hardly be implemented on real-time beamforming [3]. The robustness of MV beamformer also confined its application [4].

Our parallel Dolphy–Chebychev-rectangle (PDCR) beamformer presented in this article used the Minimum Beamwidth for Specified Sidelobe Level beamforming algorithm to find the transmit aperture weights. So the mainlobe width of transmit beams could be flexibly controlled and the mainlobe gain loss would be decreased. Rectangle apodization was applied on receiving aperture to get narrower receiving beams.

108.2 Parallel Receive Beamforming

After one beam is transmitted, there may be double time cost relative to the detection depth for conventional beamforming method to transmit another beam. That is to say the detection depth limits the frame rate. For conventional beamforming, only one scan line can be formed after each transmission and this confines the increase of frame rate. By simultaneously forming multiple receive beams steering closely spaced regions for each transmission beam, multiple scan lines can be obtained for single transmission. Thus the frame rate is increased proportionally to the number of parallel receive beams.

Figure 108.1 shows the corresponding one way Tx and Rx beampartterns using Parallel Receive Beamforming (PRB). Here, only one beam is transmitted, and four parallel beams are used on reception. The steering angles of receive beams is Rx1, Rx2, Rx3, Rx4 and central symmetry to the steering angle of the transmit beam.

Parallel forming several beams for each transmission can proportionally improve the frame rate to the number of parallel receiving beams, but array gain may be decreased if targets appear between scan lines. So the mainlobe of transmit beam should be wider to compensate the potential amplitude loss arising from mismatch between receive beams and transmit beam, and the sidelobes level should be lower to suppress noise and the off-axis interference signals.

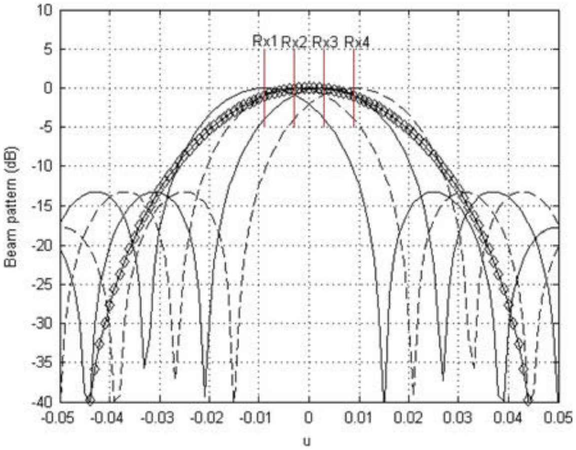


Fig. 108.1 Illustration of Parallel Receive Beamforming

108.3 Dolph–Chebychev Beamforming

The beam pattern of Dolph–Chebychev beamformer has the characters of minimum mainlobe width for a given sidelobe level [5]. For a symmetric, equally spaced, N elements array steered at broadside, it can be described by a polynomial of N−1. Setting this array polynomial equal to the Chebychev polynomial of N−1 degree and equating the array coefficients to the Chebychev polynomial coefficients, the beam pattern will correspond to a Chebychev polynomial of degree of N−1. Dolph–Chebychev beamformer weighting vector w_{DC} can be represented as:

$$w_{\text{DC}} = \left[V^H(\psi)\right]^{-1} e_1. \tag{108.1}$$

where $e_1 = [1 \quad 0 \quad \ldots \quad 0]^T$ and $V(\psi)$ is a $N \times N$ array manifold matrix be steered at ψ_{p} directions:

$$\psi_{\text{p}} = 2\ \cos^{-1}\left(\frac{1}{\text{x}_0}\cos\left(\frac{(2\text{p}-1)\pi}{(N-1)2}\right)\right), \quad x_0 = \cosh\left(\frac{1}{N-1}\cosh^{-1} R\right). \tag{108.2}$$

where R is the sidelobe level of beams.

108.4 Results and Discussion

Simulate a 18.5-mm linear array with 64 elements and equally spaced by half wavelength, the central frequency of the array elements was 3.5 MHz and the sample frequency was 100 MHz. The images were formed using 52 Tx and 52 Rx

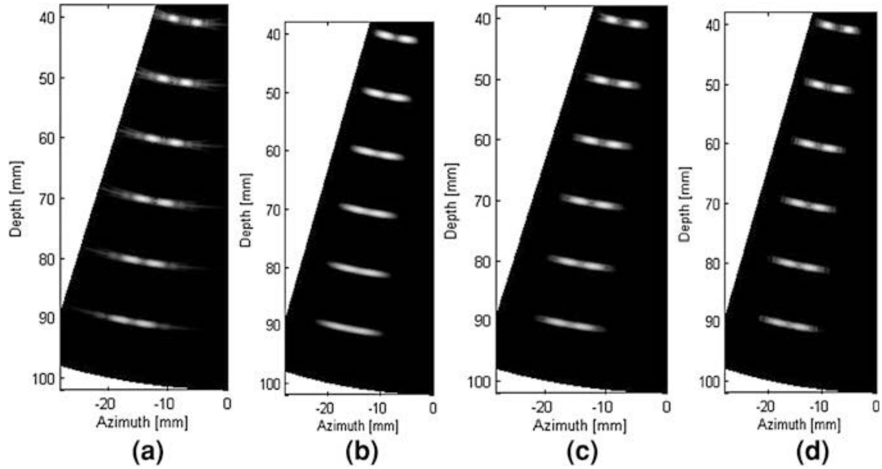


Fig. 108.2 Simulated beamformed responses of point targets using an 18.5 mm 64 elements, 3.5 MHz linear array. **a** Rectangle apodization. **b** Hanning apodization. **c** Dolphy–Chebychev (Tx)-Rectangle (Rx) apodization, nonparallel. **d** Dolphy–Chebychev (Tx)-Rectangle (Rx) apodization, four parallel receive beams

beams over a sector of 35°. For the PRB method, we used only 13 Tx beams and 4 Rx parallel beams for each Tx beam to form the image. The shift variance arises from misalignment of the transmission and receive beams using parallel beamforming compensated through interpolation on combinations of existing scan lines.

108.4.1 Simulated Point Targets

Figure 108.2 shows images of the point targets obtained using Rectangle, Hanning, Nonparallel Dolph–Chebychev-rectangle (NDCR) and Parallel Dolph–Chebychev-rectangle (PDCR) apodization over a 60 dB display dynamic range. All the simulated point targets were located at depths from 40 to 90 mm, deviation from the main axis about 10° and were separated by 10 mm in vertical and 3 mm in lateral.

As can be seen from Fig. 108.2, the beam response for Rectangle apodization has better lateral resolution, Hanning apodization can reduce sidelobe levels, but the lateral resolution is poor. NDCR apodization not only significantly reduces the sidelobe levels but also has the same lateral resolution compared to rectangle apodization. For PDCR method, comparing image quality can be obtained to NDCR, also the frame rate was increased by a factor of four.

Figure 108.3 shows the lateral variation of the beamformed responses at depths of 40 and 80 mm. As the results show, at both depths the proposed PDCR

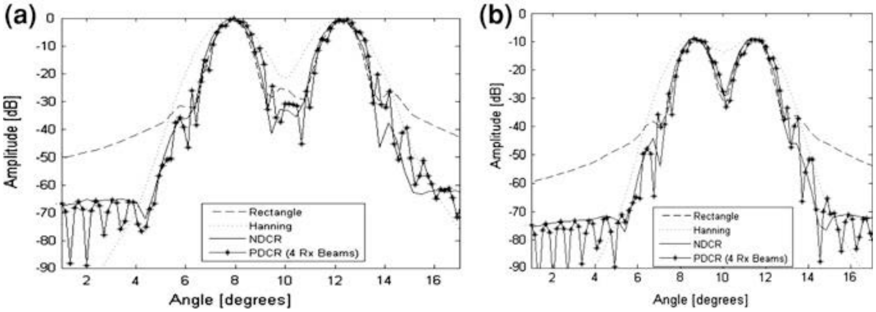


Fig. 108.3 Lateral variation of rectangle apodization, Hanning apodization, Dolph–Chebychev apodization, Dolph–Chebychev (4 Rx Beams) apodization beamformers using 64 elements, 3.5 MHz array at depths **a** 50 mm, **b** 80 mm

apodization presents narrower mainlobe width than Hanning apodization and lower sidelobe levels than Rectangle apodization.

108.4.2 Cystic Simulation

To investigate the imaging contrast for the beamforming methods, a cyst phantom in a speckle pattern was simulated. The resultant images are shown in Fig. 108.4. All images are displayed with 60 dB dynamic range.

Table 108.1 lists the relative CR and CNR for each beamforming method and confirmed the superiority of NDCR in terms of contrast ratio (CR), and the contrast to noise ratio (CNR).

As seen from Table 108.1, NDCR offers CR improvement of 6.72 dB and CNR about 29% in comparison with Rectangle apodization and the background standard deviation is approximately the same for the two beamformers. As the inherent gain loss for PRB beamformer, the mean intensity in the cyst region and background of PDCR beamformer are smaller than that of NDCR, and results in little decrease of CR and CNR. But compared with rectangle apodization, PDCR beamformer still presents a 5.78 dB increase of contrast resolution and about 19 enhancement of CNR.

108.5 Conclusion

For improving the quality of ultrasound medical imaging, it is desirable to increase solution and contrast simultaneously by means of providing narrower mainlobe width and suppressing sidelobe levels. Using PDCR, the wider transmission

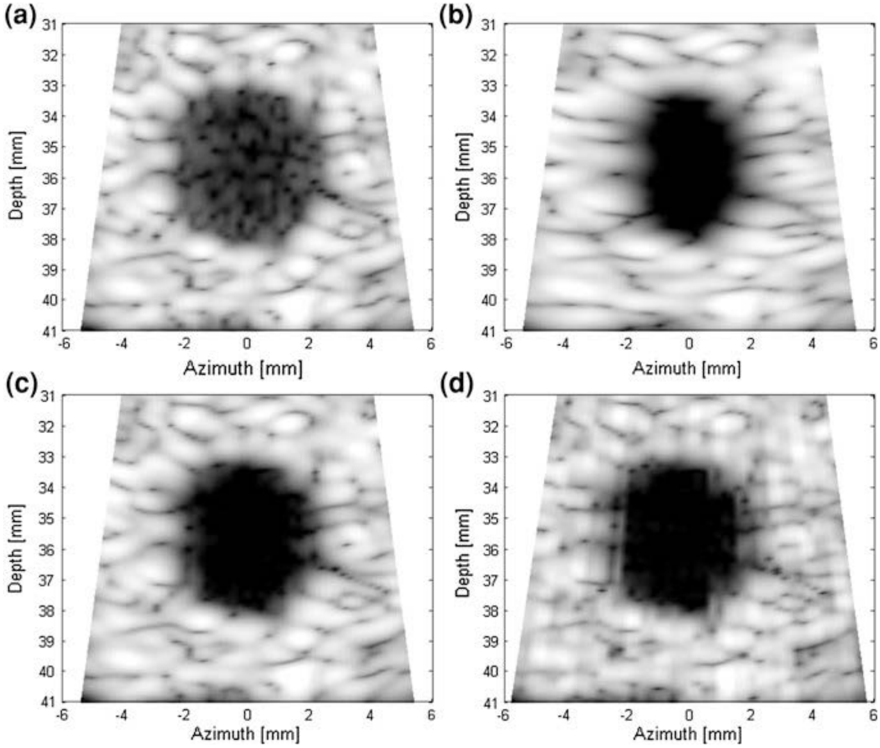


Fig. 108.4 Simulated cyst phantom using an 18.5 mm 64 elements, 3.5 MHz linear array. **a** Rectangle apodization. **b** Hanning apodization. **c** Dolphy–Chebychev (Tx)-Rectangle (Rx) apodization, nonparallel. **d** Dolphy–Chebychev (Tx)-Rectangle (Rx) apodization, four parallel receive beams

Table 108.1 Contrast parameters of the simulated cyst phantom for different beamformers

	Mean intensity in the cyst region (dB)	Mean intensity in the background (dB)	CR (dB)	CNR (dB)	Background standard deviation (dB)
Rectangle	−37.59	−14.29	23.30	3.12	7.46
Hanning	−42.78	−13.24	29.54	4.40	6.72
NDCR	−44.32	−14.29	30.02	4.04	7.43
PDCR(1:4)	−45.08	−16.00	29.08	3.71	7.84

mainlobe width compensates the gain loss induced by PRB and narrower receiving mainlobe width makes up the resolution loss because of wider transmit beam. Because of the gain loss of PRB cannot be fully compensated, the imaging quality of PDCR is a little decreased compared to NDCR. But much better contrast resolution and significant improvement of lateral resolution can be obtained than

Hanning and Rectangle appodization. Taking into account the two factors of detail resolution and contrast resolution comprehensively, we see that the image quality can be enhanced. As the computational cost is the same with conventional beamformer whose required computation cost is linear with the number of array elements, this proposed method is practical and may be implemented in real-time inexpensively.

References

1. Shattuck P, Weinshenker MD, Smith SW, von Ramm OT (1984) Explososcan: a parallel processing technique for high speed ultrasound imaging with linear phased array. J Acoust Soc Am 75(4):1273–1282
2. Synnevag JF, Austeng A, Holm S (2006) High frame-rate and high resolution medical imaging using adaptive beamforming. IEEE Ultrasonics Symposium, pp 2164–2167
3. Synnevag JF, Austeng A, Holm S (2005) Minimum variance adaptive beamforming applied to medical ultrasound imaging. In: Proceedings of IEEE Ultrasonics Symposium 2, pp 1199–1202
4. Holfort K, Gran F, Jensen JA (2007) Minimum variance beamforming for high frame-rate ultrasound imaging. In: Proceedings of IEEE Ultrasonics Symposium, pp 1541–1544
5. Van Trees HL (2002) Optimum array processing. Wiley, New York
6. Jensen JA (1996) Field: a program for simulating ultrasound systems. Med Biol Eng Comput 4:351–353

Part X
Internet Growth Modelling and Virtualized Networks

Chapter 109
SCPOPS: An Efficient and Effective Method for Service Discovery and Composition

Yuanyuan Jiang, Ying Zhang and Song Huang

Abstract In order to improve the efficiency of service discovery and composition, we first define a property ordered pairs called Property Ordered PairS (POPS) and some function of it in this paper. Then we propose a service discovery and a composition algorithm based on the POPS, the Service composition based on the POPS (SCPOPS) algorithm. We present the SCPOPS algorithm in detail and give an instance. Finally, we take some simulation experiment by the SCPOPS algorithm, the result shows it is efficient and effective for service discovery and composition.

Keywords SOA · POPS · Service composition · SCPOPS

109.1 Introduction

The recent years have witnessed an increasing adoption of Service-Oriented Architecture (SOA), which is an architectural approach for the implementation and delivery of loosely coupled distributed services [1]. Therefore, more and more enterprises are embracing SOA as a new paradigm to integrate and implement interoperable, robust and platform-independent distributed applications. Generally, services are considered as self-contained, self-describing, and modular applications that can be published, located, and invoked across the Web [2]. Most SOA implementations use Web Services. The early researches about service-orientation usually focused on the Web Service technology, while nowadays it is expanding to

Y. Jiang · Y. Zhang (✉) · S. Huang
PLA University of Science and Technology, Nanjing 210007, China
e-mail: zhywl66@163.com

Y. Jiang
e-mail: helen_nanjing@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_109,

wider scopes such as service-oriented systems, service-oriented computing, service-oriented applications, etc. However, with the rapid development of Internet technology and the increasing requirement of services, the quantity of services over the web and the complex service-oriented applications make it is unrealistic to fulfill the user needs with a single service. Thus, service discovery and composition are the two main tasks which have gained great momentum. Service composition is defined as the integration of a variety of existing services by a certain process logic to satisfy the users' requirements more effectively. The composition approach can be classified into manual composition, semi-auto composition, and automatic composition. As Web services are created independently by a variety of providers, service composition is a complex process [3].

There are many researches concerned with service discovery and composition has been studied by lots of researchers [4–15]. Bellwood [4] discovered services based on matchmaking of key words, but the accuracy of the services found are low. Lee [5] implements service composition by using data mining techniques for ubiquitous computing environments. Xu [6] presented a service discovery and composition by inverted indexing. Kuang [7] proposed a method for composition-oriented service discovery through generation of a tree. However, they do not pay attention to the relation between input and output concepts when building the inverted indexing. AI Planning techniques such as HTN [8, 9], Workflow techniques [10–13], Petri Net techniques [14, 15] are also widely used for service matchmaking, discovery and composition.

Based on their researches, we propose an efficient and effective service composition strategy based on Property Ordered PairS (POPS) in this paper. First of all, we define the POPS and the related functions. Then, we introduce our service composition algorithm based on POPS, called Service composition based on the POPS (SCPOPS). SCPOPS decreases the searching space of candidate services by matchmaking services with POPS retrieval records instead of the whole service repository, and focuses on the interface information both of outputs and inputs. Therefore the searching efficiency will be increased obviously.

The remainder of the paper is organized as follows: The POPS and its operations are defined in Sect. 109.2. The SCPOPS algorithm is presented in Sect. 109.3, while an instance for illustrating the algorithm is proposed in Sect. 109.4. Some experiments and conclusions are given in Sect. 109.5.

109.2 Preliminaries

109.2.1 Service Definition

A service can be formally defined as follows:

Definition 1 Service definition.

$$Service = \langle SN, SD, SF, SA \rangle$$

in which *SN* is the name of a service. *SD* is the functional descriptions of a service that normally use natural languages. *SF* is service functions including inputs, outputs, preconditions and results of a service. *SA* is the attributes of a service, such as the QoS attributes. In theory, all these four aspects should be considered during service composition. However, with the limited length of one paper, we predigest the service definition to a 3-tuple:

$$Service = \langle SN, I, O\rangle$$

in which *SN* is still the name of a service. *I* represents the input-set of a service while *O* represents the output-set of the service. Both *I* and *O* have semantic support by a domain ontology.

Definition 2 Service request definition.

$$SR = \langle I, O, T\rangle$$

in which *I* represents the input-set of a service request, and *O* represents the output-set of the service request. *T* is the time limit of request time.

109.2.2 POPS Definition

Definition 3 The property ordered pairs.

$$POPS(C) = \{\langle SN_1, I_1\rangle, \langle SN_2, I_2\rangle\ldots, \langle SN_n, I_n\rangle\}$$

in which SN_i, $1 \le i \le n$ represents the service name, $I_i, 1 \le i \le n$ represents the input concepts of SN_i. Therefore, $POPS(C)$ could represent all registered services that can provide the output concept in service repository. We build POPS retrieval records for service repository and update them periodically. If there is a new service registered successfully, its related information will be added to the records accordingly.

Definition 4 The *overlap* function of POPS.

$$overlap(POPS(C_1), POPS(C_2)) = \{\langle SN_1, I_1\rangle, \langle SN_2, I_2\rangle, \ldots, \langle SN_n, I_n\rangle\} \quad (109.1)$$

in which $\langle SN_i, I_i\rangle \in POPS(C_1) \wedge \langle SN_i, I_i\rangle \in POPS(C_2)$, $i = 1\ldots n$. The *overlap* function represents the services which could provide concepts of both C_1 and C_2. For instance, if there are $POPS(C_1) = \{\langle S_1, I_1\rangle, \langle S_2, I_2\rangle, \langle S_3, I_3\rangle\}$ and $POPS(C_2) = \{\langle S_1, I_1\rangle, \langle S_3, I_3\rangle\}$, then $overlap(POPS(C_1), POPS(C_2)) = \{\langle S_1, I_1\rangle,$ $left\langle S_3, I_3\rangle\}$. Actually, the return of *overlap* is still property ordered pairs.

Definition 5 The *getItem* function of POPS.

$$getItem(POPS(C), k) = \langle S_i, I_i\rangle \quad (109.2)$$

in which $\langle SN_i, I_i\rangle \in POPS(C)$. The *getItem* function is used to obtain an item of $POPS(C)$, and the ordinal ofthe item is k. For instance, if there is $POPS(C) = \{\langle S_1, I_1\rangle, \langle S_4, I_4\rangle\}$, then $getItem(POPS(C), 2) = \langle S_4, I_4\rangle$.

109.3 SCPOPS Algorithm

Currently, there are two ways for matchmaking service request and service in service repository. One is matchmaking inputs of service request and service at first. If one service needs inputs less than the service request provided, then checking its outputs could provide all the outputs needed by service request or not. If it could, it will be the matched service. Therefore, this is a Front-to-Back way which is easy to understand but has less efficiency. The other way is Back-to-Front, which matchmaking outputs first. If a service in the service repository provides all outputs described by service request, then matchmaking its inputs and service request. If the inputs are included by inputs of service request, the service is the matched service. Compare these two ways, the latter is goal-driven, thus it will avoid some meaningless searching of finding matched service inputs.

However, the two ways mentioned above both need to match the service request and all registered service in service repository one by one. They are time-consuming and not suitable because the size of service repository is becoming larger and larger. In order to solve this problem, we propose the SCPOPS algorithm as shown in Fig. 109.1.

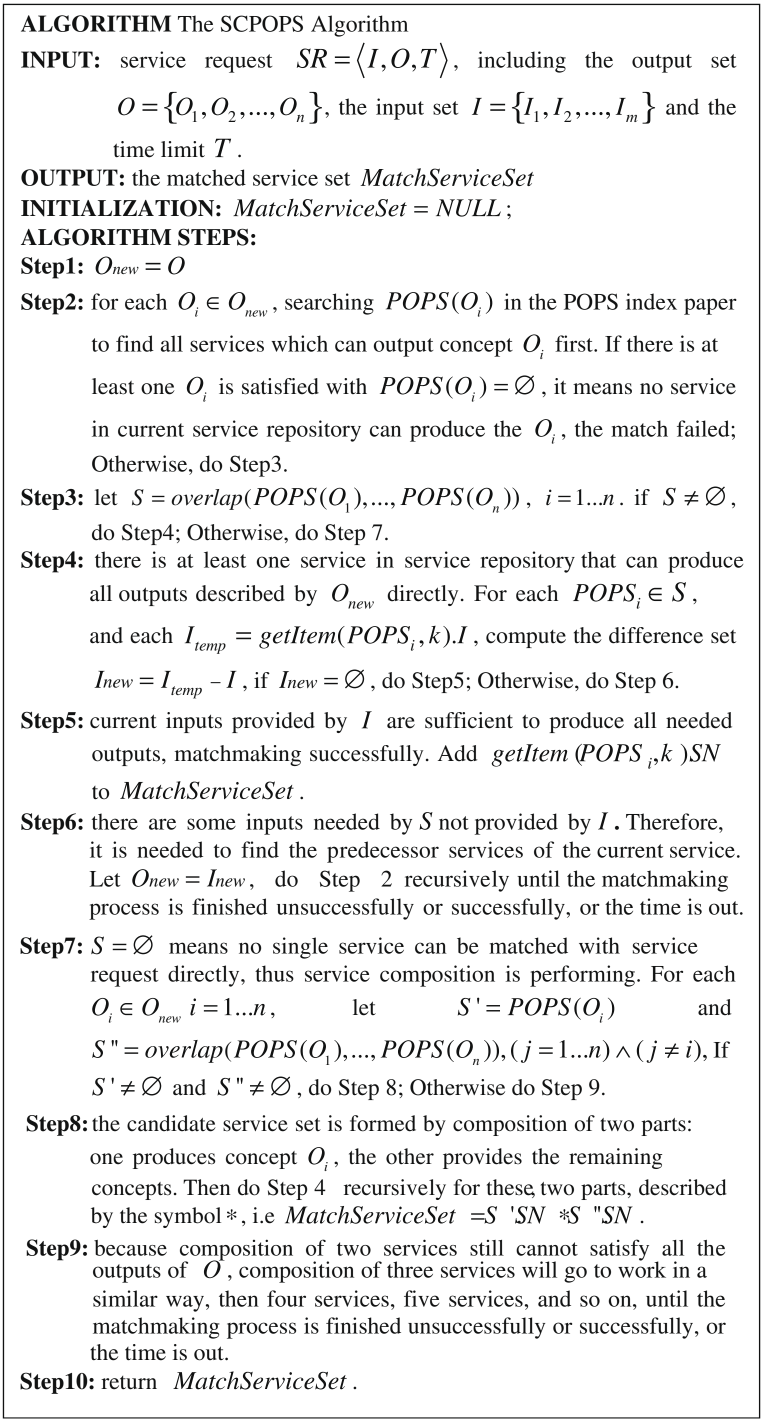
109.4 Example of SCPOPS Algorithm

In this section, we give an example for comprehending the SCPOPS algorithm better. There are six registered services in the service repository and the POPS retrieval records of the service repository as shown in Table 109.1.

Suppose the user provides input concepts $I = \{A, B, C\}$ and wants to get output concepts $O = \{E, F\}$. The algorithm executes as follows:

According to the output-set of service request O, $O_{new} = O = \{E, F\}$, do Step2.
Because $POPS(E) \neq \varnothing \wedge POPS(F) \neq \varnothing$, do Step3.
Because $S = overlap(POPS(E), POPS(F)) = \varnothing$, do Step7.
For each $O_i \in O_{new}$ $i = 1\ldots n$, let $S'_E = POPS(E) = \langle S3, \{A, C\}\rangle \neq \varnothing$ and $S''_E = overlap(POPS(F)) = \langle S4, \{B\}\rangle \neq \varnothing$, do Step 8.

The candidate service set is formed by composition of two parts: one produces concept E, the other provides concept F. First, for $O_{new} = \{E\}$, do Step 4. Because $I_{new} = getItem(POPS(E), 1).I - I = \{A, C\} - \{A, B, C\} = \varnothing$, do step5,



```
ALGORITHM The SCPOPS Algorithm
INPUT: service request $SR = \langle I, O, T \rangle$, including the output set
        $O = \{O_1, O_2, ..., O_n\}$, the input set $I = \{I_1, I_2, ..., I_m\}$ and the
        time limit $T$.
OUTPUT: the matched service set $MatchServiceSet$
INITIALIZATION: $MatchServiceSet = NULL$;
ALGORITHM STEPS:
Step1: $O_{new} = O$
Step2: for each $O_i \in O_{new}$, searching $POPS(O_i)$ in the POPS index paper
        to find all services which can output concept $O_i$ first. If there is at
        least one $O_i$ is satisfied with $POPS(O_i) = \varnothing$, it means no service
        in current service repository can produce the $O_i$, the match failed;
        Otherwise, do Step3.
Step3: let $S = overlap(POPS(O_1), ..., POPS(O_n))$, $i = 1...n$. if $S \neq \varnothing$,
        do Step4; Otherwise, do Step 7.
Step4: there is at least one service in service repository that can produce
        all outputs described by $O_{new}$ directly. For each $POPS_i \in S$,
        and each $I_{temp} = getItem(POPS_i, k).I$, compute the difference set
        $I_{new} = I_{temp} - I$, if $I_{new} = \varnothing$, do Step5; Otherwise, do Step 6.
Step5: current inputs provided by $I$ are sufficient to produce all needed
        outputs, matchmaking successfully. Add $getItem(POPS_i, k)SN$
        to $MatchServiceSet$.
Step6: there are some inputs needed by $S$ not provided by $I$. Therefore,
        it is needed to find the predecessor services of the current service.
        Let $O_{new} = I_{new}$, do Step 2 recursively until the matchmaking
        process is finished unsuccessfully or successfully, or the time is out.
Step7: $S = \varnothing$ means no single service can be matched with service
        request directly, thus service composition is performing. For each
        $O_i \in O_{new}$ $i = 1...n$, let $S' = POPS(O_i)$ and
        $S'' = overlap(POPS(O_1), ..., POPS(O_n)), (j = 1...n) \wedge (j \neq i)$, If
        $S' \neq \varnothing$ and $S'' \neq \varnothing$, do Step 8; Otherwise do Step 9.
Step8: the candidate service set is formed by composition of two parts:
        one produces concept $O_i$, the other provides the remaining
        concepts. Then do Step 4 recursively for these, two parts, described
        by the symbol $*$, i.e $MatchServiceSet = S'SN * S''SN$.
Step9: because composition of two services still cannot satisfy all the
        outputs of $O$, composition of three services will go to work in a
        similar way, then four services, five services, and so on, until the
        matchmaking process is finished unsuccessfully or successfully, or
        the time is out.
Step10: return $MatchServiceSet$.
```

Fig. 109.1 The SCPOPS algorithm

Table 109.1 The service repository and the POPS retrieval records

Service Repository	POPS Retrieval Records
$S1 : Input = \{A, B\}...Output = \{C\}$	$POPS(C) = \{\langle S1, \{A, B\}\rangle, \langle S2, \{D\}\rangle\}$
$S2 : Input = \{D\}......Output = \{C, G\}$	$POPS(E) = \{\langle S3, \{A, C\}\rangle\}$
$S3 : Input = \{A, C\}...Output = \{E\}$	$POPS(F) = \{\langle S4, \{B\}\rangle\}$
$S4 : Input = \{B\}......Output = \{F, H\}$	$POPS(G) = \{\langle S2, \{D\}\rangle, \langle S5, \{E\}\rangle\}$
$S5 : Input = \{E\}......Output = \{G\}$	$OPS(H) = \{\langle S4, \{B\}\rangle, \langle S6, \{A, G\}\rangle\}$
$S6 : Input = \{A, G\}...Output = \{H\}$	

$POPS(E).SN = getItem(POPS(E)).SN = S3$; Secondly, for $O_{new} = \{F\}$, do Step 4. Because $I_{new} = getItem(POPS(F), 1).I - I = \{B\} - \{A, B, C\} = \emptyset$, do step5, add $POPS(F).SN = getItem(POPS(F)).SN = S4$. Therefore, the result is $S3 * S4$.

109.5 Experiments and Conclusions

In order to evaluate the efficiency and effectiveness of SCPOPS, compared with the one by one Front-to-Back and Back-to-Front method, we take emulation experiments on an Intel Core2 Duo 2.33 GHz with 2 GB RAM running Eclipse. Figures 109.2 and 109.3 shows some of the results. In Fig. 109.2, 10 at 500 represents 10 service requests and the service repository has 500 services, and so on. From this picture, we can conclude that both the quantity of service requests and the size of service repository are the factors influencing the processing time of service composition. When the quantities of these two factors are not huge, the difference of responding time is small and the Front-to-Back method has the longest time cost. However, with the increase in service quantity, the SCPOPS tend to have obvious advantages in time consumption because it is using the POPS index table. For the same reason, as shown in Fig. 109.3, we try to discover and composite 15 service requests in service repository with 800 services. When the time limits is 60 s, the two algorithms can fulfill almost all service matchmakings except the Front-to-Back method which is not goal-driven; while the time limits up to 10 s, the SCPOPS has a bigger recall rate which means that it can match more services.

To sum up, although building and maintenance POPS index table will cost some time, the SCPOPS algorithm is still an efficient an effective method for automatic service composition. Because of the limited length of the paper, we do not discuss the semantic similarity in this paper, which is also an important part of service matchmaking. The future work will include researching the influence of QoS in service composition, as well as how to implement the composition of services based on fuzzy information, etc.

Fig. 109.2 The processing time of different cases

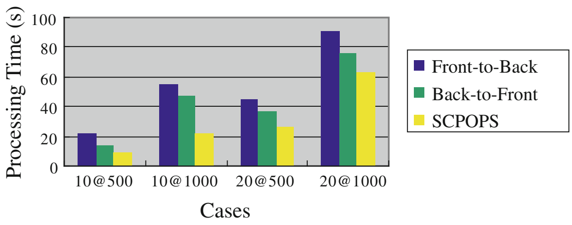
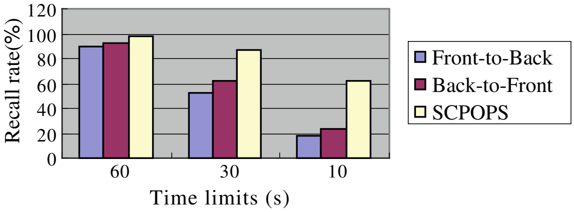


Fig. 109.3 The recall rate of different limits



References

1. Barker A, Walton CD, Robertson D (2009) Choreographing web services[J]. IEEE Trans Serv Comput 2(2):152–166
2. Rao J, Su X (2004) A survey of automated web service composition methods[C]. In: Proceedings of first international workshop on semantic web services and web process composition
3. Xiangwei L, Zhicai X, Li Y (2009) Independent global constraints-aware web service composition optimization based on genetic algorithm[C]. In: Proceedings of international conference on industrial and information systems
4. Bellwood T, Capell S, Clement L, Colgrave J, Dovey MJ, Feygin D, Hately A, Kochman R, Macias P, Novotny M, Paolucci M, von Riegen C, Rogers T, Sycara K, Wenzel P, Wu Z (2002) UDDI version 3.0[EB/OL]. http://uddi.org/pubs/uddi_v3.htm
5. Lee SY, Lee JY, Lee BI (2006) Service composition techniques using data mining for ubiquitous computing environments. Int J Comput Sci Netw Secur[J]. 6(9):110–117
6. Xu B, Li T, Gu Z et al (2006) SWSDS: quick web service discovery and composition in SEWSIP. In: Proceedings of the 8th IEEE international conference on ecommerce technology and the 3rd IEEE international conference on enterprise computing, e-commerce, and e-services
7. Kuang L, Li Y, Wu J et al (2007) Inverted indexing for composition-oriented service discovery[C]. In: Proceedings of IEEE international conference on web services (ICWS)
8. Qiu L, Shi Z, Lin F (2006) Context optimization of AI planning for services composition[C]. In: Proceedings of the IEEE international conference on e-business engineering, pp 610–617
9. Madhusudan T, Uttamsingh N (2006) A declarative approach to composing web services in dynamic environments[J]. Decis Support Syst 41(2):325–357
10. Bottaro A, Bourcier J, Escoer C et al (2007) Autonomic context-aware service composition[C]. In: Proceedings of 2nd IEEE international conference on pervasive services
11. Mingkhwan A, Fergus P, Abuelma'Atti O et al (2006) Dynamic service composition in home appliance networks. Multimed Tools Appl[J] 29(3):257–284
12. Pourreza H, Graham P (2006) On the fly service composition for local interaction environments[C]. In: Proceedings of IEEE international conference on pervasive computing and communications workshops, 393

13. Viroli M, Denti E, Ricci A (2007) Engineering a BPEL orchestration engine as a multi-agent system [J]. Sci Comput Program 66:226–245
14. Chi YL, Lee HM (2008) A formal modeling platform for composing web services[J]. Expert Syst Appl 34:1500–1507
15. Valero V, Cambronero ME, Díaz G et al (2009) A Petri net approach for the design and analysis of web services choreographies[J]. J Log Algebraic Program 78:359–380

Chapter 110
Logistic Regression Analysis of Influencing Factors on Postgraduate Entrance Exams

Dong Yong-quan

Abstract Using Logistic regression analysis methods, this paper analyzes some influencing factors on mathematics postgraduate entrance exams about four quantitative factors including specialized course normal academic records such as Mathematical analysis and Higher algebra, public course such as Politics and English, and three qualitative factors including whether the student is a leader, whether three goods student, and the number of scholarships.

Keywords Logistic regression · MLE · Statistics with R · Postgraduate entrance exams

110.1 Introduction

In recent years, with the reform of the college graduates employment system and the rapid expansion of higher education, the number of college graduates have increased rapidly, and employment issues are becoming increasingly prominent. Candidating for the master's degree is an option for many students to avoid the employment peak, and look for new and better employment opportunities. Of course, there are many students select ing the master's degree for the purpose of further improving their level of knowledge and better research. Many universities also work to enhance the success rate in graduates as the priority and as one of the indicators of educational assessment. Therefore, analyzing the main factors on

D. Yong-quan (✉)
Department of Mathematics and Information Science, Tangshan Teachers College,
Tangshan 063000, China
e-mail: xfx0502@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_110,

postgraduate entrance exams can guide the teaching and administrative departments to work out the right teaching strategies, and the students grasp the correct way of learning, and thus improve the acceptance rate. In this paper, we have a mathematical look at a total of 49 undergraduates (including 27 students passing first test post-graduate examinations) from an undergraduate graduating class of Department of Mathematics and Information Science of Tangshan Teachers College for logistic regression analysis of the influencing factors on postgraduate entrance exams.

Taking into account the pattern of mathematics postgraduate is $4 + x$, that is two basic courses of Mathematical analysis and Higher algebra, and two public courses of Politics and English are tested in first test, and specialized courses such as probability and statistics, real variable function, differential equation, etc. We first select the factors (covariates or explanatory variables) as follows:

Mathematics students need to learn mathematical analysis in almost four semester classes, so each student has four mathematical analysis results, which we take as mean of Mathematical analysis normal academic records x_1.

We have two Higher algebra results, which we take as the mean of Higher algebra normal academic records x_2.

Similarly, x_3 and $x_4$are respectively the normal academic records of Higher algebra, English and Politics.

In addition to the above four quantitative factors, some qualitative factors will also affect the performance of postgraduate students, such as

x_5 is a binary variable indicating whether student leader: yes, $x_{51} = 1$; no, $x_{50} = 0$.

x_6 is whether three goods student: yes, $x_{61} = 1$; no, $x_{60} = 0$.

x_7 is the number of scholarship: no, $x_{70} = 0$; once, $x_{71} = 1$; twice, $x_{72} = 2$; three times, $x_{73} = 3$.

Response variable$y$is a binary variable indicating whether postgraduate: yes, $y = 1$; no, $y = 0$, so ordinary linear regression model cannot be used to do regression analysis. Logistic regression will be used in this paper [1], and all statistical calculations completed by the statistical with R [2].

110.2 Logistic Regression Model

Logistic regression model is a kind of generalized linear model, being applicable to continuous data and discrete data, especially the latter, such as attribute data and count data. This is useful, especially in statistical analysis of biological, medical, economic and social data [3, 4].

Let the observational data be $(x_{i1}, x_{i2}, \ldots, x_{im})$, $i = 1, 2, \ldots, N$, where x_{ij} is the feature value$j$(referred to as covariates) of student i. y_i is the response variable of student i, indicating whether postgraduate: yes, $y_i = 1$; no, $y_i = 0$. $m$is the number of covariates, where $m = 7$. $N$is the sample size, where $N = 49$.

Table 110.1 Basic statistics of four quantitative data sets

Course	Mean	Std	Min	Max
Mathematical analysis	81.24	6.5270	62.95	90.65
Higher algebra	83.33	7.0019	65.85	94.00
English	78.73	3.9021	61.55	87.40
Politics	82.08	3.5176	73.65	89.05

Generalized linear model can handle such binary data, that is, introducing an appropriate strictly increasing function, called the link function G. Let

$$p(x) = P(y = 1|x). \tag{110.1}$$

be the probability of admitting to graduate school with the covariate being x, then

$$G(x) = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n. \tag{110.2}$$

is called the generalized linear models of link function G.

We use logit function as link function:

$$\log it(p) = \log \frac{p}{1-p}. \tag{110.3}$$

which has a clear statistical significance: the logarithm of the probability value of $y = 1$ to $y = 0$. The generalized linear models using Logit function as link function is known as the Logistic regression model.

$$\log it(P(y = 1|x)) = a_0 + \sum_{j=1}^{m} a_jx_j. \tag{110.4}$$

i.e.

$$P(y = 1|x) = \frac{1}{1 + e^{-a_0 - \sum_{j=1}^{m} a_jx_j}} \tag{110.5}$$

where $\xi = (a_0, a_1, \ldots, a_m)'$ is parameter vector being estimated by MLE method.

110.3 Empirical Analysis

There are eight data sets, including four quantitative data sets of normal academic records of Mathematical analysis, Higher algebra, English and Politics, with the basic statistics as shown in Table 110.1.

We will standardize the four types of quantitative data (raw data minus the average value), then do Logistic regression analysis, so the reference group is the average normal academic records.

Table 110.2 Statistical results of four qualitative data sets

Qualitative factors	Level	The proportion of all levels
Student leader	2(yes = 1, no = 0)	Yes: 12 (account for 24.5%) No: 37 (account for 75.5%)
Three good student	2(yes = 1, no = 0)	Yes: 21 (42.9%) No: 28 (57.1%)
Number of scholarships	4 (no, once, twice, three times)	No: 16 (32.7%) Once: 12 (24.5%) Twice: 13 (26.5%) Three times: 8 (16.3%)
Postgraduate	2 (yes = 1, no = 0)	Yes: 27 (55%); No: 22 (45%)

Table 110.3 Results of Logistic regression analysis

Variables	Parametric estimates	Standard error	z Values	p Values
Constant	−0.77911	0.88720	−0.878	0.3799
x_1	−0.06519	0.07993	−0.815	0.4148
x_2	−0.03661	0.07326	−0.500	0.6172
x_3	0.26644	0.15088	1.766	0.0774
x_4	−0.11161	0.16452	−0.678	0.4975
x_{51}	−0.11238	0.89466	−0.126	0.9000
x_{61}	1.48242	1.05972	1.399	0.1619
x_{71}	0.18511	1.28020	0.145	0.8850
x_{72}	1.85085	1.53878	1.203	0.2291
x_{73}	−0.03015	1.55596	−0.019	0.9845

In addition, there are four qualitative data, including three explanatory variables: whether student leader, whether three goods student and the number of scholarship, and a response variable: whether admitted to graduate school. The statistics summary are shown as follows: (Table 110.2).

Using statistics with R, the results of Logistic regression analysis of the selected variables are shown in Table 110.3.

110.3.1 *p* Values in Table 110.3 Indicate That:

Postgraduate entrance exams success is significantly associated with English normal academic records x_3 at 1–0.0774 = 92.26% level, three goods student: yes, x_{61} at 83.81% level, and the number of scholarships: twice, x_{72} at 77.18% level second. Other effect factors followed by the strong to weak are the normal academic records of Mathematical analysis, Politics and Higher algebra, the number of scholarship: once, student leader: yes, the number of scholarships: three times.

Of the four quantitative factors, English normal academic records have the largest influence on postgraduate entrance exams, followed by Mathematical analysis. This is true of the actual situation of our school, Tangshan Teachers College, which belongs to Hebei Province. In the two B class institutions, 90% of students are from rural, low levels of English learning, so the English results are the main factors to their postgraduate success. Students in our school need to make efforts to learn English well, to lay a solid foundation for future studies. Higher Algebra in specialized courses is my teaching strength, with the basis of better teaching.

In contrast, Mathematical analysis is on the weaker; on the one hand, teachers with high qualifications and high professional titles less, on the other hand, not enough emphasis is given on teaching the basic course, following the same old teaching model. These should be the major adjustments. Part of the teachers' wealth in teaching experience and expertise is being expanded to the teaching of mathematical analysis, changing the large classes for small class teaching, to meet the needs of postgraduate students.

110.3.2 Parametric Estimates in Table 110.3 Indicate That:

The effect of English normal academic records increasing one point than the English average academic records (because we use the standardized data, the reference groups is the average academic records, the following Mathematical analysis, Higher algebra and Political are the same) to the success ratio of postgraduate is $e^{0.26644} = 1.3053$ times. There is still much room for improvement in English results, which should be given high priority.

Similarly, the effect of Mathematical analysis is $e^{-0.06519} = 0.9369$ times, which indicates the normal teaching of Mathematical analysis slightly different from postgraduate requirements. Mathematical analysis not a national examination, facing a wide range, the teaching is difficult to achieve full coverage of key and difficult topics, so it is not easy to improve the overall performance.

Higher algebra is $e^{-0.03661} = 0.9640$ times, which indicates that the difference in Higher algebra score is not clear. Because of the weak correlation to postgraduate (1-0.6172 = 0.3828), the quantitativeness of Higher algebra is not very meaningful.

Politics is $e^{-0.11161} = 0.8943$ times, which indicates that students' energy is attacked hard in the exams to learn more political subjects; this causes the prosperity of all kinds of political reasons for postgraduate entrance exam remedial classes.

Student leader is $e^{-0.11238} = 0.8937$ times, which indicates a student leader having weak negative correlation to postgraduate entrance exams success. As student leaders spend a certain amount of study time on management affairs, it affects their postgraduate entrance exams success rate, but the effect is relatively small.

Three goods student is $e^{1.48242} = 4.4036$ times, the greater positive correlation to postgraduate entrance exams, which indicates that the rating of three goods student in our academic largely take academia as the reference.

The effect of a scholarship increasing one point than no scholarship to the success ratio of postgraduate entrance exams is $e^{0.18511} = 1.2034$ times; secondary scholarship is $e^{1.85085} = 6.3652$ times; three scholarship is $e^{-0.03015} = 0.9703$ times. Overall, the impact of scholarship on the entrance examination is a positive correlation, especially the secondary scholarship having the biggest positive correlation to postgraduate entrance exams success, which indicates usually better academic performanc and generally a greater success chance in postgraduation. But three scholarships have weak negative correlation 0.9703 to postgraduate success, indicating a very good student achievement in peacetime, postgraduate entrance exams expectations that are too high; the schools applied for better resources, for example Tsinghua University, Beijing University and Chinese Academy of Sciences, may postgraduate entrance exams defeat. Therefore, guiding students correctly , being a good staff to postgraduate students, are the university teachers' bounden duty in addition to teaching students professional knowledge.

110.4 Conclusion

In summary, analyzing the main factors on postgraduate entrance exams can guide the teaching and administrative departments to work out the right teaching strategies, and the students grasp the correct way of learning, thus improving postgraduate acceptance rate. Additionally, this paper selected a typical class with better postgraduate performance in Department of Mathematics and Applied Mathematics of our school as a research object to analyze; the results coincided with the actual situation, we also found the teaching problems and shortcomings. If possible, more samples will be selected, and more meaningful conclusions will be obtained. Furthermore, Logistic regression method in this paper can also be used in the analysis of factors of other postgraduate professionals.

References

1. Jichuan W, Zhigang G (2001) Logistic regression model—methods & applications. Higher Education Press, Beijing
2. Dalgaard P (2002) Introductory statistics with R. Springer, New York
3. Lingxiang P et al (2007) Multivariate logistic regression analysis of ectopic pregnancy. Prog Obstet Gynecol 16:236–237
4. Yongquan D (2008) Evaluation and treatment efficacy prediction of AIDS. China Health Stat 25:204–206

Chapter 111
Efficient Variant of FastICA Algorithm for Speech Features Extraction

Xiaoli Huang and Huanglin Zeng

Abstract We propose an efficient variant of FastICA (EFICA) algorithm which is asymptotically efficient. The simulation results show superior performance of the EFICA algorithm compared with the fast fixed-point ICA (FastICA) algorithm and the Power ICA algorithm.

Keywords Efficient variant of FastICA · Interference signals ratio · Cramér–Rao bound · Fisher information matrix

111.1 Introduction

One of the central tools for speech features extraction is independent component analysis (ICA) [1, 2]. FastICA [3] is one of the most popular algorithms for ICA. This chapter presents an efficient variant of FastICA (EFICA) algorithm that extracts speech features which are robust to non-stationary noise contaminating the speech signal. The proposed algorithm outperforms classical FastICA and other algorithms in the noisy scenario [4]. Experiments with the speech database demonstrate consistent robustness to noise of varying statistics, yielding significant improvements in speech recognition accuracy over identical models trained using EFICA features and evaluated at independent sources that have been mixed linearly.

X. Huang (✉) · H. Zeng
Sichuan University of Science and Engineering, Zigong, China
e-mail: hxlpiaoran@163.com

H. Zeng
e-mail: zhl@suse.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_111,

111.2 Independent Component Analysis

The most common method for generating spatially localized features is to apply ICA to produce basis vectors that are statistically independent. ICA is a statistical technique to extract non-Gaussian and statistically independent source signals given only the observed or measured data [5]. We model the observations vector (at some time instant t) of measured signals $X(t) = [x_1(t), \ldots, x_l(t), \ldots, x_m(t)]^T$ as a linear mixing via the unknown mixing matrix A and the unknown source signals matrix $S(t) = [s_1(t), \ldots, s_l(t), \ldots s_n(t)]^T$:

$$X(t) = AS(t) \tag{111.1}$$

where $S(t)$ is a statistically independent zero-mean real-random process, A is a $n \times n$ mixing matrix. In the subsequent derivation, we shall assume $m = n$. ICA makes the solutions to Eq.111.1 well-posed by forcing independence between the components of the basis. It is worth noting that independence does not imply orthogonality, indeed the latter is a considerably weaker constraint.

The new dataset $X(t)$ was processed by ICA; the independent components (ICs) have the following expression:

$$Y(t) = WX(t) = MS(t) \tag{111.2}$$

where $Y(t) = [y_1(t), \ldots, y_n(t)]^T$ is the output of separation systems, W is the demixing matrix computed according the extended, and $M = WA$ is the global matrix. After the complete separation of the source signal, we have:

$$M = WA = I \tag{111.3}$$

where I is an identity matrix. We have:

$$Y(t) = WX(t) = S(t) \tag{111.4}$$

111.3 Fast ICA

The FastICA algorithm is a computationally efficient and robust algorithm for ICA and blind source separation. It was introduced in [6] in two versions: A one-unit approach and a symmetric one.

By a unit we refer to a computational unit, eventually an artificial neuron, having a weight vector w that the neuron is able to update by a learning rule. The FastICA learning rule finds a direction, i.e. a unit vector w such that the projection $w^T x$ maximizes nongaussianity. Nongaussianity is here measured by the approximation of negentropy $J(w^T x)$ given in Eq. 111.5.

$$J(y) \propto [E\{G(s)\} - E\{G(v)\}] \tag{111.5}$$

The random variable s is assumed to be of zero mean and unit variance. v is a Gaussian variable of zero mean and unit variance, and $G(\cdot)$ is a non-quadratic function.

The one-unit algorithm of the preceding subsection estimates just one of the ICs, or one projection pursuit direction. To estimate several ICs, we need to run the one-unit FastICA algorithm using several units with weight vectors $W = (w_1, \ldots, w_l, \ldots, w_n)^T$.

To prevent different vectors from converging to the same maxima we must decorrelate the outputs $w_1^T x, \ldots, w_n^T x$ after every iteration. A simple way of achieving decorrelation is a deflation scheme based on a Gram-Schmidt-like decorrelation. This means that we estimate the ICs one by one. When we have estimated n ICs, or n vectors $w_1^T x, \ldots, w_n^T x$, we run the one-unit fixed-point algorithm for w_{n+1}, and after every iteration step subtract from w_{n+1} the projections $w_{p+1}^T w_j w_j, j = 1, \ldots, n$ of the previously estimated n vectors:

$$w_{n+1} = w_{n+1} - \sum_{j=1}^{n} w_{n+1}^T w_j w_j \tag{111.6}$$

then renormalize w_{n+1}:

$$w_{n+1} = \frac{w_{n+1}}{\sqrt{w_{n+1}^T w_{n+1}}} \tag{111.7}$$

In certain applications, however, it may be desired to use a symmetric decorrelation, in which no vectors are privileged over others. This can be accomplished by the classical method involving matrix square roots, let

$$W = \left(WW^T\right)^{\frac{-1}{2}} W \tag{111.8}$$

Under the assumption that each row s_i contains n independent realizations of non-Gaussian random variables ξ_i, if we express FastICA using the intermediate formula and write it in matrix form [6], we see that FastICA takes the following form:

$$W^+ = W + \mathrm{diag}(\alpha_i)\left[\mathrm{diag}(\beta_i) + E\left\{g(\xi)\xi^T\right\}\right]W \tag{111.9}$$

where W is the matrix $(w_1, \ldots, w_l, \ldots, w_n)^T$ of the vectors, $G(\cdot)$ is a non-quadratic function, $g(\cdot)$ and $g'(\cdot)$ denotes the first and the second derivatives of $G(\cdot)$ respectively. $\beta_i = -E\{g(\xi_i)\}$, $\alpha_i = -1/(\beta_i - E\{g'(\xi_i)\})$, and $E\{\cdot\}$ denote the expectation operator. Let G^{UN} be the gain matrix obtained by the symmetric variant of FastICA using a nonlinear function $g(\cdot)$. The main result shown in [7] was the following: Assume that the original signals ξ_i in the mixture have zero mean and unit variance that is sufficiently smooth. Then, the normalized gain matrix elements $\sqrt{n}G^{UN}$ have asymptotically Gaussian distributions $\Gamma(0, V_{ij}^{UN})$, with variances

$$V_{ij}^{UN} = \frac{\gamma_i + \gamma_j + \tau_j^2}{\left(\tau_i + \tau_j\right)^2} \tag{111.10}$$

where $\mu_i = E[g^2(\xi_i)]$,$\gamma_i = \mu_i - \beta_i^2$, and $\tau_i = |\beta_i - \alpha_i|$.

It was shown in [8] that the asymptotic interference signals ratio (ISR) matrix of FastICA has as elements

$$\mathrm{ISR}_{ij}^{EF} = \frac{1}{n}\frac{\gamma_i(\gamma_j + \tau_j^2)}{\tau_j^2\gamma_i + \tau_i^2(\gamma_j + \tau_j^2)} \tag{111.11}$$

where ISR is the ISR matrix, ISR_{ij}^{EF} is the i rows, and j columns element of it.

The variances in Eq. 111.11 are minimized if the function $g(\xi)$ equals the score function

$$\psi_i(\xi) = -\frac{d}{ds}\log p_i(\xi) = -\frac{p_i'(\xi)}{p_i(\xi)} \tag{111.12}$$

of the corresponding source distribution $p_i(\xi)$. The minimum variance can be shown to be close. It was shown that the Cramér–Rao bound (CRB) [8] is CRB $= \frac{1}{n}\frac{\kappa_i}{\kappa_i\kappa_j - 1}$ when $\kappa_i = E\left[\psi_i^2(\xi_i)\right]$.

CRB, which is the inverse of the fisher information matrix (FIM), can be used e.g., to show that an unbiased estimator is uniformly minimum variance unbiased estimator. CRB is also related to asymptotic optimality theory.

111.4 Efficient Variant of FastICA Algorithm

EFICA is essentially a modification of the popular FastICA algorithm [9], belonging to a wide family of ICA algorithms which exploit non-Gaussian of the sources distributions. The algorithm consists of three steps:

(1) Running the symmetric FastICA until convergence. The purpose of this step is to quickly and reliably get preliminary estimates of the original signals. In this step the optional nonlinearity in the original symmetric FastICA $g(\xi) = \tanh(\xi)$ is used due to its universality, but other possibilities seem to give promising results as well.

(2) Adaptive choice of different nonlinearities $g(\cdot)$ to estimate the score functions of the found sources, based on the outcome of step (1).

(3) A refinement or fine-tuning for each of the found source components by one-unit FastICA, using the nonlinearities found in step (2).

We use function $\hat{g}(\cdot)$ instead of common nonlinear function $g(\cdot)$:

$$\hat{g}(\xi) = [c_1\hat{g}(\xi_1), c_2\hat{g}(\xi_2), \ldots, c_n\hat{g}(\xi_n)] \tag{111.13}$$

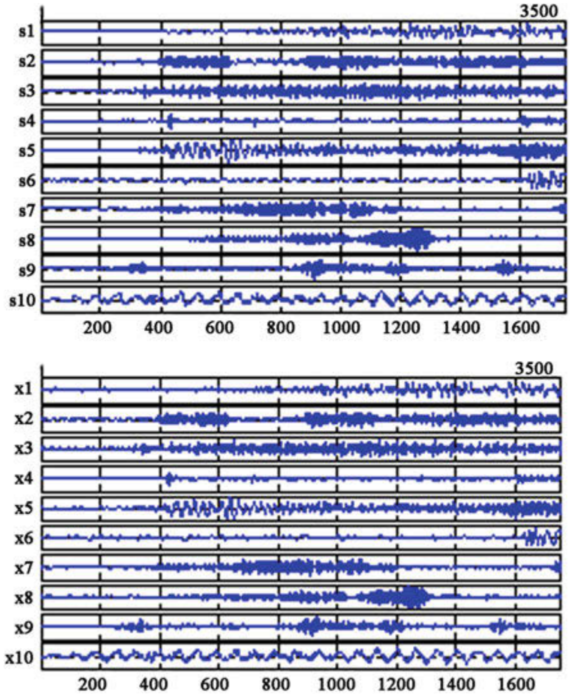


Fig. 111.1 The sources and the Mixed signals (**a**) sources (**b**) Mixed signals (Gaussian noise SNR = 15DB)

then

$$ISR_{ij}^{*} = \frac{1}{n}\frac{c_i\gamma_i(c_j\gamma_j + c_j^2\tau_j^2)}{c_j^2\tau_j^2 c_i\gamma_i + c_i^2\tau_i^2(c_j\gamma_j + c_j^2\tau_j^2)} \qquad (111.14)$$

In order to improve algorithm's optimal performance, c_{ij} is obtained as follows when $c_i = 1$:

$$c_{ij} = \min_{i=1, i\neq j} \mathrm{ISR}_{ij}^{*} = \frac{\tau_j\gamma_i}{\tau_i(\tau_j^2 + \gamma_j)} \qquad (111.15)$$

EFICA enhances FastICA by offering an elaborate data-adaptive choice of these nonlinearities, followed by a refinement step. It is shown that the asymptotic ISR matrix has as elements. As EFICA can choose nonlinear function and adjust weights adaptively, the theoretical ISR value can be very close to the actual one (CRB).

111.5 Simulation

In this section we examine performance of EFICA algorithm against the FastICA algorithm, Power ICA algorithm in terms of performance indices using the speech signals. To test the viability of the algorithm we performed computer experiments

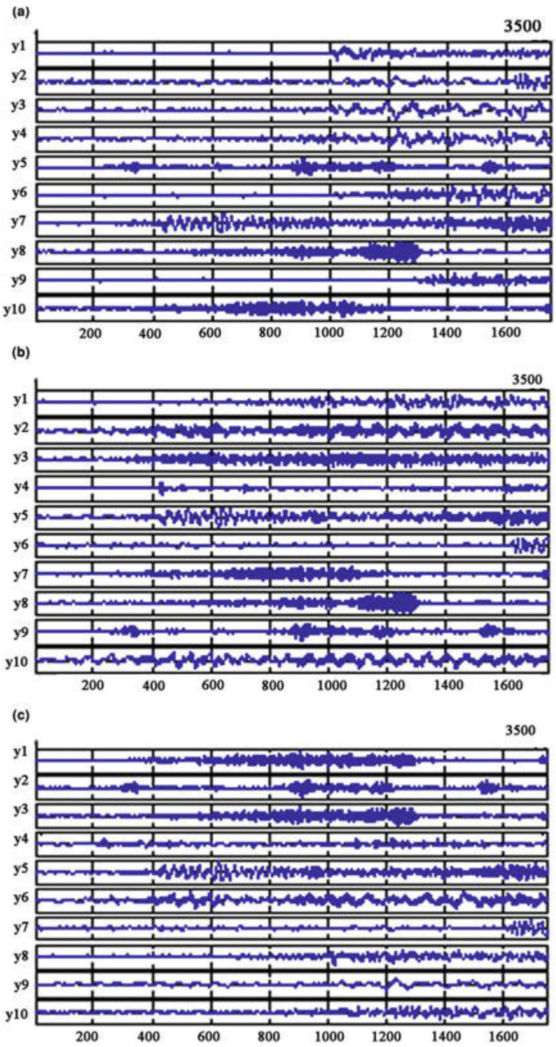


Fig. 111.2 The estimated sources (ICs) (**a**) Estimated sources (ICs) by fixed point ICA (**b**) Estimated sources (ICs) by power ICA (**c**) Estimated sources (ICs) by EFICA

on real sound data and synthetic data from a Laplace density. The mixed speech signals could look something like those in Fig. 111.1a. The first preprocessing step, which is common for both versions and for many other ICA algorithms, consists of removing the sample mean and decorrelating the data *X*. The subband decomposition and selection method is used to eliminate high-frequency noise effectively in Fig. 111.1b.

The mixed signals are separated through EFICA algorithm, Fixed-point ICA algorithm and PowerICA algorithm, respectively.

The separated signals are shown in Fig. 111.2.

Table 111.1 The predictive index and time-consuming of different algorithms

Algorithms	FastICA	Power ICA	EFICA
Time-consuming(s)	12.3	15.4	10.32
Predictive index	0.0824	0.0952	0.0548

In order to explain the proposed EFICA algorithm in more detail, the Amari distance between the true mixing matrix and the estimated one was monitored. We found that the method performed comparable to the other 2 widely used ICA methods, while enjoying the advantages mentioned above. Table 111.1 shows the performance of three algorithms, respectively, in the form of the Predictive index and time-consuming.

From Table 111.1 we can see that the EFICA algorithm is better and faster than other algorithms. At the same time it showed better versatility.

Acknowledgments This work was supported by the Foundation of Artificial Intelligence Key Lab of Sichuan Province Education Department (2009RY009) and the Foundation of Teaching Reform OF Sichuan University of Science & Engineering (JB-0918) .

References

1. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley Inter science, New York
2. Cichocki A, Amari S-I (2002) Adaptive signal and image processing: learning algorithms and applications. Wiley, New York
3. Hyv̈arinen A (1997) A fast fixed-point algorithm for independent component analysis. Neural Comput 9(7):1483–1492
4. Koldovský Z, Tichavský P (2006) Methods of fair comparison of performance of linear ICA techniques in presence of additive noise. In: Proceedings of IEEE internaional conference on acoustics, speech, and signal processing (ICASSP), vol V. Toulouse, France, May 2006, pp 873–876
5. Wada N, Yoshizawa S (2005) Robust speech feature extraction using RSF/DRA and burst noise skipping, ECTI transactions on Electrical Engineering, Electronics, and Communications 3(2), August 2005
6. Hyv̈arinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10(3):626–634
7. Tichavský P, Koldovský Z, Oja E (2005) Asymptotic performance analysis of the fixed-point algorithm (Fast-ICA) for independent component analysis, IEEE statistical signal processing workshop, Bordeaux, France, July 2005
8. Tichavský P, Koldovský Z, Oja E (2006) Performance analysis of the FastICA algorithm and Cramér-Rao bounds for linear independent component analysis. IEEE Trans Signal Process 54(4):1189–1203
9. Hyvärinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10(3):626–634

Chapter 112
Cryptanalysis of Two Event Signature Protocols for Peer-to-Peer Massively Multiplayer Online Games

Wei Yuan, Liang Hu, Hongtu Li and Jianfeng Chu

Abstract In 2008, Chan et al. presented an event signature (EASES) protocol for peer-to-peer massively multiplayer online games (P2P MMOGs). The authors declare that the EASES protocol is efficient and secure, and could achieve nonrepudiation, event commitment, save memory, bandwidth and reduce the complexity of the computations. However, we find that Chan et al.'s EASES protocol is not secure and gives the detailed steps to attack their protocol. Further, we point out that their dynamic EASES protocol is also not secure and can be cracked by the attacker. The attacking result shows that attacker can even replace any update event he wants to forge. Finally, we made a discussion about this problem and pointed the weakness existing in these three protocols.

Keywords Hash-chain · Cryptanalysis · Massive multiplayer online games · One-time signature · Peer-to-peer networks · MMOG · Network security

W. Yuan · L. Hu · H. Li · J. Chu (✉)
Department of Computer Science and Technology, Jilin University, Changchun, China
e-mail: chujf@jlu.edu.cn

W. Yuan
e-mail: yuanwei1@126.com

L. Hu
e-mail: hul@mails.jlu.edu.cn

H. Li
e-mail: li_hongtu@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_112,

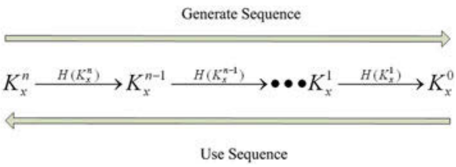
112.1 Introduction

Multiplayer online games are a rapidly growing segment of Internet applications in the recent years. By providing more entertainment and sociability than single-player games, it is fast becoming a major form of digital entertainment. In this kind of games, all players should connect with the server to send and receive event updates. An event update is cryptographic protocol by which a player generates an event message and sends it to the server for updating the game states. Traditional massively multiplayer online games (MMOGs) are conventional client–server models that do not scale with the number of simultaneous clients that need to be supported. To resolve conflicts in the simulation and act as a central repository for data, peer-to-peer (P2P) architecture is increasingly being considered as replacement for traditional client–server architecture in MMOGs. P2P MMOGs have many advantages over traditional client–server systems due to their network connectivity and basic network services in a self-organizing manner. Whenever a player wants to play the finger-guessing game, an event message is sent to the server and the server processes all the events and updates the game states to ensure a global ordering for game executions and fair plays. However, P2P MMOGs communicate on the Internet raise the security issues such as cheating a dishonest player can get valuable virtual items and even be sold for moneymaking. Recently, there are more and more efforts mounted to focus on event update protocols for online games in respect to the protection of sensitive communication and the provision of fair play.

In 2004, Dickey et al. [1] proposed a low latency and cheat-proof event ordering based on digital signatures and voting mechanism for P2P games. However, Corman et al. later show that Dickey et al.'s protocol is unable to prevent all cheats as claimed, and proposed an improvement called secure event agreement protocol [2], as digital signature requires a large amount of computations. To reduce heavyweight computations in every round of a game session, in 2008, Chan et al. [3] proposed an efficient and secure event signature (EASES) protocol using one-time signature with hash-chain key and claimed that their protocol has low computation and bandwidth costs, and is thus applicable to P2P-based MMOGs. Then they proposed a dynamic EASES protocol to avoid the pre-generation of hash-chain keys. Unfortunately, we find both the EASES protocol and the dynamic EASES protocol are not secure and attackers can easily forge a series of update event to replace the original one.

In this chapter, we briefly review the EASES protocol and the dynamic EASES protocol. Further, we introduce attacking methods to crack these protocols. Finally, we make a discussion on what our attack does.

Fig. 112.1 Construction of hash-chain keys


Generate Sequence
$K_x^n \xrightarrow{H(K_x^n)} K_x^{n-1} \xrightarrow{H(K_x^{n-1})} \bullet\bullet\bullet K_x^1 \xrightarrow{H(K_x^1)} K_x^0$
Use Sequence

112.2 Review of Chan et al.'s Event Signature Protocol for P2P MMOGs

The EASES protocol has four phases: the Initialization Phase, Signing Phase, Verification Phase and Re-initialization Phase.

112.2.1 Initialization Phase

In this phase, player P_i generates a series of one-time signature keys for a session and performs the following operations:

1. P_i chooses a master key MK_i to compute the nth one-time signature key $K_i^n = h(MK_i)$, where n represents the maximum number of rounds in a session.
2. P_i computes the other rth round one-time signature keys $K_i^{r-1} = H(K_i^r)$, where $r = (n-1), \ldots, 0$.
3. P_i signs the first one-time signature key by its private key to get the signature $\Delta_i = S_{\mathrm{sk}_i}(K_i^0)$. The hash-chain keys K_i^r will be used in the reverse order of their production during the subsequent rth rounds, where $r = 0, 1, 2, \ldots, n-1$. Figure 112.1 shows the production of hash-chain keys.

112.2.2 Signing Phase

In this phase, if P_i wants to submit event update messages to other online players in a game session with n rounds, he/she performs the following operations:

1. P_i computes the first round one-time signature key δ_i^1 by computing $\delta_i^1 = H(K_i^1 \parallel U_i^1), \Delta_i, K_i^0$. Then, P_x submits the first round message to other online players.
2. P_i computes the second round one-time signature key $\delta_i^2 = H(K_i^2 \parallel U_i^2), U_i^1, K_i^1$ and submits it to other online players.

3. P_i computes the rth round one-time signature key $\delta_i^r = H(K_i^r \parallel U_i^r), U_i^{r-1}, K_i^{r-1}$ and submits it to other online players in the subsequent rth round, where $r = 3, 4, \ldots, n$.

112.2.3 Verification Phase

In this phase, each online player P_j receives the event update message $\delta_i^1 = H(K_i^1 \parallel U_i^1), \Delta_i, K_i^0$ from the player P_i and performs the following operations:

1. In the first round, P_j first verifies $\Delta_i \overset{?}{=} D_{\mathrm{pk}_i}(K_i^0)$. If it holds, P_j confirms that the key K_x^0 is legitimate.
2. In the subsequent rth round, P_j verifies $K_i^{r-2} \overset{?}{=} H(K_i^{r-1})$ to check if the signature key K_i^{r-1} is legitimate, where $r = 2, 3, 4, \ldots, n$.
3. If above holds, P_j verifies $\delta_i^{r-1} \overset{?}{=} H(K_i^{r-1} \parallel U_i^{r-1})$ to check whether the update has been altered or not. If it passes verification, P_j convinces that no player has tampered with the update from P_i.

112.2.4 Re-Initialization Phase

If P_i wants to extend his/her game session for a few rounds, P_i regenerates a new master key and performs the following operations:

1. In the nth round, P_i chooses a new master key $MK_i^{'}$ and generates the new one-time signature keys $NewK_i^0, \ldots, NewK_i^n$. Then P_i has the new signature key $NewK_x^0$ with the key $K_i^n = H(MK_i^{'})$ to generate $\delta_i^n = H(K_i^n \parallel U_i^n \parallel NewK_i^0)$, U_i^{n-1}, K_i^{n-1}. P_i sends $\delta_i^n, U_i^{n-1}, K_i^{n-1}$ to other players as usual.
2. In the $(n + 1)$th round, P_i sends $\delta_i^{n+1} = H(NewK_i^1 \parallel U_i^{n+1}), U_i^n K_i^n NewK_i^0$ to other players.
3. In the $(n + 2)$th round, P_i sends $\delta_i^{n+2} = H(NewK_i^2 \parallel U_i^{n+2}), U_i^{n+1}, NewK_i^1, MK_i$ to other players.

Upon receiving new one-time signature keys from P_i, the other player, P_j, should perform the following verifiable operations:

1. In the (n + 1)th round, P_j verifies $\delta_i^n \overset{?}{=} H(K_i^n \parallel U_i^n \parallel NewK_i^0)$ to check if the new signature key $NewK_x^0$ is legitimate.
2. In the $(n + 2)$th round, in addition to the regular verifications, P_j must also verify $K_i^n \overset{?}{=} H(MK_i)$. If the above passes verification, P_j confirms the validity of

$NewK_x^0$. The series of new one-time signature keys $NewK_i^0, \ldots, NewK_i^n$ can be used after the $(n + 2)$th rounds.

112.3 How to Tamper the Update Message

In this part, we will show how the attack, P_k, tampers the update message when the normal user P_i communicates with P_j. We suppose the attacker can monitor, intercept and forge the message communicated between P_i and P_j. The detailed steps are described as follows.

1. When P_i starts to send the first message, $\delta_i^1 = H(K_i^1 \parallel U_i^1), \Delta_i, K_i^0$, to P_j and P_k intercepts it and records K_i^0.
2. When P_i sends the second message, $\delta_i^2 = H(K_i^2 \parallel U_i^2), U_i^1, K_i^1$, to P_j and P_k intercepts it and records K_i^1. Then P_k forges a new message $\delta_i^1 = H(K_i^1 \parallel U_i^{1*}), \Delta_i, K_i^0$ and sends it to P_j. P_j verifies $\Delta_i \overset{?}{=} D_{pk_i}(K_i^0)$ and records K_i^0 to his memory if the equation holds.
3. When P_i sends the third message, $\delta_i^3 = H(K_i^3 \parallel U_i^3), U_i^2, K_i^2$, to P_j and P_k intercepts it and records K_i^2. Then P_k forges a new message $\delta_i^2 = H(K_i^2 \parallel U_i^{2*})$, U_i^{1*}, K_i^1, and sends it to P_j. P_j verifies $K_i^0 \overset{?}{=} H(K_i^1)$ and $\delta_i^1 \overset{?}{=} H(K_i^1 \parallel U_i^{1*})$ and records U_i^{1*} and K_i^1 to his memory if the two equations hold.
4. When P_i sends the rth message, $\delta_i^r = H(K_i^r \parallel U_i^r), U_i^{r-1}, K_i^{r-1}$ $r = 3, \ldots, n$, to P_j and P_k intercept it and record K_i^2. Then P_k forges a new message $\delta_i^{r-1} = H(K_i^{r-1} \parallel U_i^{(r-1)*}), U_i^{(r-2)*}, K_i^{r-2}$, and sends it to P_j. P_j verifies $K_i^{r-3} \overset{?}{=} H(K_i^{r-2})$ and $\delta_i^{r-2} \overset{?}{=} H(K_i^{r-2} \parallel U_i^{(r-2)*})$ and records $U_i^{(r-2)*}$ and K_i^{r-2} to his memory.
5. Finally, P_j accepts and records all the forged message $U_i^{1*}, \ldots, U_i^{(n-2)*}$ instead of $U_i^1, \ldots, U_i^{n-2}$.

Above attack starts from the first message. Actually, the attacker can start his attack from arbitrary round. As the protocol is for the peer-to-peer online games, short time delay will not cause other players to find the existence of the attack.

112.4 Review of the Dynamic EASES

Dynamic EASES extends the “authentication via hashing” idea of the basic EASES to avoid the pre-generation of hash-chain keys. This helps to reduce memory usage. We achieve this by signing the messages without first preparing a series of hash-chain keys. The key point is to defer the revelation of committing key for two rounds. An adversary cannot intrude since he gets no knowledge about the committing key. As it is not necessary to initialize or re-initialize the hash-chain in the dynamic version, there are only two phases: signing and verification.

112.4.1 Signing Phase

When player$_i$ has to sign an update message U_i^r for player$_\mathrm{j}$ in round r, there are three cases to consider:

1. In the first round, player$_i$ picks an unpredictable random value K_i^1, computes $H(U_i^1|K_i^1)$ and sends out the signed hash value $S_{\mathrm{sk}}(H(U_i^1|K_i^1))$.
2. In the second round, player$_\mathrm{i}$ picks an unpredictable random value K_i^2, computes $H(U_i^2|K_i^2|K_i^1)$ and then sends out he signed hash value $S_{\mathrm{sk}}(H(U_i^2|K_i^2|K_i^1))$
3. In the subsequent rounds, player$_\mathrm{i}$ generates an unpredictable random value K_i^r, computes $H(U_i^r|K_i^r|K_i^{r-1})$ and sends $H(U_i^r|K_i^r|K_i^{r-1})$, K_i^{r-2} and U_i^{r-2}.

The player$_\mathrm{i}$'s out-going messages are summarized in the following equation:

$$\begin{cases} S_{\mathrm{sk}}(H(U_i^1|K_i^1)) & \text{in the first round} \\ S_{\mathrm{sk}}(H(U_i^2|K_i^2|K_i^2)) & \text{in the second round} \\ H(U_i^r|K_i^r|K_i^{r-1}), K_i^{r-2}, U_i^{r-2} & \text{in the subsequent } r\text{th round} \end{cases}$$

112.4.2 Verification Phase

Player$_\mathrm{j}$ performs the following actions to verify the authenticities of the received messages from $H(U_i^2|K_i^2|K_i^1)$ There are three cases to verify:

1. In the first round, player$_\mathrm{j}$ verifies the signature $S_{\mathrm{sk}}(H(U_i^1|K_i^1))$ by player$_\mathrm{i}$'s public key and then keeps $H(U_i^1|K_i^1)$ in the memory.
2. In the second round, player$_\mathrm{j}$ verifies the signature $S_{\mathrm{sk}}(H(U_i^2|K_i^2|K_i^1))$ by player$_\mathrm{i}$'s public key and then keeps $H(U_i^2|K_i^2|K_i^1)$ in the memory.
3. In the subsequent round, player$_\mathrm{j}$ can reveal the commitment and check the authenticity of the messages received in the previous two rounds. In other words, the authenticity of U_i^{r-2} is checked in the rth round. To verify K_i^{r-2} and U_i^{r-2} for the (r—2)th round, player$_\mathrm{i}$ simply computes $H(U_i^{r-2}|K_i^{r-2}|K_i^{r-1})$. Note that in the third round, only $H(U_i^1|K_i^1)$ needs to be computed.

The major difference between the dynamic and the basic EASES are:

1. Dynamic EASES needs two digital signature operations when initializing a new game session.
2. Hash-chain generation is no longer necessary in the dynamic version.

3. Dynamic EASES requires one more round than the basic EASES to reveal the plaintext of a message.

For example, in the basic EASES, the receiver reveals the commitment of δ_i^r and confirms the authenticity of K_i^r by deciding whether $K_i^{r-1} \overset{?}{=} H(K_i^r)$ after receiving U_i^r and K_i^r in the rth round. But in dynamic EASES, the player ensures that the message of the (r—2)th round is correct by verifying the message received in the rth round.

112.5 How to Intrude the Dynamic Eases

The dynamic EASES protocol is very similar with the EASES protocol, thus one can easily get the attacking method to this protocol. We should highlight that the first and the second messages cannot be tampered because the player's secret key sk is used to sign these messages. Thus attack can forge any message from the third one. Without losing the generality, we will describe how the attacker forges the rth message that player$_{\mathrm{i}}$ sent to player$_{\mathrm{j}}$. The detailed steps are described as follows.

In the rth round, player$_{\mathrm{i}}$ sends $H(U_i^r|K_i^r|K_i^{r-1})$, K_i^{r-2} and U_i^{r-2} to player$_{\mathrm{j}}$ and the player$_{\mathrm{j}}$ will check K_i^{r-2} and U_i^{r-2} with $H(U_i^{r-2}|K_i^{r-2}|K_i^{r-3})$, which he received in the (r—2)th round. The attacker can intercept $H(U_i^r|K_i^r|K_i^{r-1})$, K_i^{r-2} and U_i^{r-2}, forge a new message $H(U_i^{r*}|K_i^r|K_i^{r-1})$, K_i^{r-2} and U_i^{r-2}, and send it to the player$_{\mathrm{j}}$.

In the (r + 1)th round, the attacker does not intercept the message from player$_{\mathrm{i}}$.

In the (r + 2)th round, the attacker intercepts the message $H(U_i^{r+2}|K_i^{r+2}|K_i^{r+1})$, K_i^r and U_i^r, and sends $H(U_i^{r+2}|K_i^{r+2}|K_i^{r+1})$, K_i^r and U_i^{r*} to the player$_{\mathrm{j}}$. Since the hash value of U_i^{r*} has been modified in the rth round, player$_{\mathrm{j}}$ is sure to accept the forged message U_i^{r*}.

From this attack, we can see that attack can not only modify all messages from a selected round can but also modify the update event in some particular round. Hence, the dynamic protocol still does not achieve the original goal.

112.6 Discussions

The aim of Chan et al.'s protocol is to reduce the computational cost. They believe that public-key cryptosystem requires a large amount of computations. Then they propose the EASES and the dynamic EASES protocol. In these two protocols, only the first signature needs to be based on public-key cryptography, while others are based on the relationship between the hash-chain keys. As the above attack shows, they do not achieve their aim, because attackers can easily tamper the hash value based on the public message, like the associated key. Further, the update event can be forged and the receiver cannot find any questions. Thus, in Chan et al.'s protocol, the public key is necessary.

Acknowledgments The authors would like to thank the editors and anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China under Grant No. 60873235 and 60473099, the National Grand Fundamental Research 973 Program of China (Grant No. 2009CB320706), Scientific and Technological Developing Scheme of Jilin Province (20080318) and Program of New Century Excellent Talents in University (NCET-06-0300).

References

1. Dickey C, Zappala D, Lo V, Marr J (2004) Low latency and cheat-proof event ordering for peer-to-peer games. In: Proceedings of ACM international workshop on network and operating system support for digital audio and video, pp 134–139
2. Corman A, Douglas S, Schachte P, Teague V (2006) A secure event agreement (SEA) protocol for peer-to-peer games. In: The first international conference on availability, reliability and security
3. Chan M-C, Hu S-Y, Jiang J-R (2008) An efficient and secure event signature (EASES) protocol for peer-to-peer massively multiplayer online games. Comput Netw 52(9):1838–1854

Chapter 113
A Real-Time System Development Method Based on Aspect-Oriented

Wei Qiu and L. C. Zhang

Abstract Aspect-oriented modeling (AOM) raises the idea of separation of concerns to the level of software models. This approach applies aspect-orientation concepts to compose models that represent different concerns (business, security, persistence, etc.) into various base models. AOM techniques have been proposed for both static and behavioral models. The presented software design methodology can be used both in the case when software is developed alongside a hardware platform and in the case when such a platform is given from the start. In this paper I propose to take a look at the aspect reuse problem from an AOM point of view. I believe that the insights gained at the modeling level can provide deeper understanding of the fundamental problems of aspect reuse, and that the solutions I propose might be transferable to the programing language level as well.

Keywords Aspect-oriented modeling · Aspect-oriented programing · Crosscutting concerns · Specification and description language

113.1 Introduction

Aspect-oriented software development (AOSD) techniques aim to provide systematic means for the identification, separation, representation and composition of crosscutting concerns. Aspect-oriented ideas can be applied at any phase and at

W. Qiu (✉)
School of Computer Science, Jia Ying University,
Meizhou 514015, Guangdong, China
e-mail: qiuwei@jyu.edu.cn

L. C. Zhang
Faculty of Computer Science, Guangdong University of Technology,
Guangzhou 510090, Guangdong, China
e-mail: lchzhang@gdut.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_113,

any level of abstraction during software development. Aspect-oriented modeling (AOM) focuses on modularizing and composing crosscutting concerns within software models—models that can be used to describe or analyze properties of a system under development [1]. It has been shown that aspect-oriented programing techniques can be effectively used to increase code reuse. Even very general concerns such as distribution, concurrency, persistency and failures have been successfully implemented in an application-independent aspect, and then later composed with different applications. Reuse of many aspects within one application, however, has proven to be more challenging, since aspects can have complex dependencies and interactions. Currently, many researchers are working on aspect-oriented programing languages features that could make such reuse easier. In this paper I propose to take a look at the aspect reuse problem from an AOM point of view. I believe that the insights gained at the modeling level can provide deeper understanding of the fundamental problems of aspect reuse, and that the solutions I propose might be transferable to the programing language level as well. I demonstrate the validity of our approach by modeling the design of AspectOptima, a complex aspect framework that was created as a case study for studying aspect dependencies and interactions.

AOM in software engineering, separation of concerns refers to the ability to identify those parts of software artifacts that are relevant to a particular concept, goal, task or purpose. Concerns are the primary motivation for organizing and decomposing software into smaller, more manageable and comprehensible parts. Clarke and Baniassad [2] define an approach called Theme/UML. It introduces a theme module that can be used to represent a concern at the modeling level. Themes are declaratively complete units of modularization, in which any of the diagrams available in the UML can be used to model one view of the structure and behavior the concern requires to execute. In Theme/UML class diagrams and sequence diagrams are typically used to describe the structure and behavior of the concern being modeled. Just like in our approach, the binding to a base model is done by template parameter instantiation. In contrast to our approach, Theme/UML does not support model weaving. Similar to our approach, Whittle and Araujo [3] represent behavioral aspects with scenarios. Aspectual scenarios are modeled as interaction pattern specifications and are composed with specification scenarios. The weaving process is performed in two steps. First state machines are generated from the aspects and from the specification. Behavior is modeled using the specification and description language (SDL), a formalism related to state diagrams. In order to be able to reuse aspects, mappings have to be defined (equivalent to our instantiations) that link a reusable aspect to the application-specific context in which it is to be deployed. Reusability in [4] the authors propose framed aspects, an approach that uses AOP to modularize crosscutting and tangled concerns and frame technology [5] to allow aspect parameterization, configuration and customization.

113.2 Aspect-Oriented Software Development

AOSD supports the modularization of crosscutting concerns by providing the aspect abstraction that makes it possible to separate and compose them to produce the overall system. AspectJ [6] is an aspect-oriented extension to the Java programing language. Aspect is a modular unit of crosscutting implementation in AspectJ. Each aspect encapsulates functionality that crosscuts classes in a program. An aspect is defined by an aspect declaration, which has a similar form of class declaration in Java. Similar to a class, an aspect can be instantiated and can contain attributes and methods and it can be specialized in subaspects. An aspect is then combined with the classes it crosscuts according to specifications given within the aspect. Moreover, an aspect can introduce methods, attributes and interface implementation declarations into types by using the inter-type declaration construct. The essential mechanism provided for composing an aspect with other classes is called a join point. A join point is a well-defined point in the execution of a program, such as a call to a method, an access to an attribute, an object initialization and exception handler. Sets of join points may be represented by pointcuts. AspectJ provides various pointcut designators that may be combined through logical operators to build up complete descriptions of point cuts of interest. An aspect can specify advices that are used to define some code that should be executed when a point cut is reached. An advice is a method-like mechanism that consists of a piece of code to be executed before, after or around a point cut. An AspectJ program can be divided into two parts: a base code part which includes classes, interfaces and other language constructs for implementing the basic functionality of the program, and an aspect code part which includes aspects for modeling crosscutting concerns in the program. For further information about AspectJ, one can refer to [2].

In Figure 113.1 Principle of AOM has been proposed as a solution to cope with concerns that are difficult to capture with other development approaches, such as object-oriented development. AOM raises the idea of separation of concerns to the level of software models. This approach applies aspect-orientation concepts to compose models that represent different concerns (business, security, persistence, etc.) into various base models. Over the last years, many AOM techniques have been proposed for both static and behavioral models [7–10]. All these techniques provide a notion of model-based aspect and a model-weaving process. Figure 113.1 shows the principle of aspect-oriented modeling. Model-based aspects are typically made of point cuts and advices, where point cuts define where to affect the base model and the corresponding advices define what to do in the places identified by the point cuts. In the simplest form, point cuts and advices are expressed by model fragments based on the concrete syntax of the base model, but other sophisticated forms, such as predicates over model exist to select relevant model elements. The model-weaving process takes as input a base model and one or more aspects and produces a result model where elements that are expressed by advices are combined with elements from the base model each time point cuts

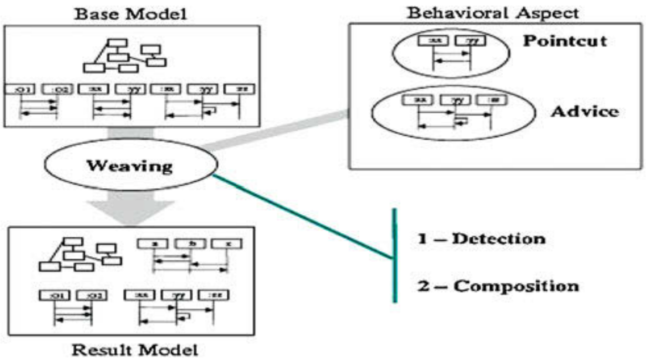


Fig. 113.1 Principle of aspect-oriented modeling

matches. In the following, we will manage variability by applying the principles of AOM at the meta level.

113.3 Aspect-Oriented Design Model

AODM is an UML extension that enhances the existing UML specification with aspect-oriented concepts that mimic the crosscutting characteristics of the AspectJ language.

The AODM defines:

(1) a special stereotype for standard UML classes (<<aspect>>) to capture the semantics of aspects;
(2) a new stereotype for standard UML operations (<<pointcut>>) to capture the semantics of AspectJ's pointcuts;
(3) a new stereotype for standard UML operations (named <<advice>>) that capture the semantic of AspectJ's advice;
(4) a new stereotype for UML collaboration templates (<<introduction>>) to describe inter-type declarations. Figure. 113.1 illustrates how an aspect-oriented implementation for the Observer design pattern [5] is modeled using the AODM. In the example, the Observer design pattern is used to make a color label (playing the Observer role) change its color whenever a button (playing the Subject role) is clicked. One abstract aspect (Subject Observer Protocol) implements the Observer pattern and a concrete aspect (Subject Observer Protocol Implementation) a particular instance of this pattern for the Button and ColorLabel classes. The AODM provides a visual notation for AspectJ's programs, where boxes that represent aspects are polluted with very detailed, implementation-specific information that is only useful for AspectJ. Note that, although the collaboration templates stereotyped with

<<introduction>> provide some means to modularize intertype declarations in a per-participant basis, the specification of join points (pointcuts) and advice are not properly modularized. Therefore, it lacks adequate support for dealing with heterogeneous aspects. In the AODM approach, all point cuts and advice are top-level elements and should be described in the aspect's Operations compartment. As a consequence, the notation does not provide means to express that both the point cut state Changes and the advice advice_id01 are related to the Subject participant. Moreover, the aspect's local operations are supposed to be mixed with pointcuts and advice. This design reflects the poor separation of concerns inside AspectJ's aspects [7] and leads to a poor separation of concerns inside AODM design-level aspects as well.

113.4 An Example System

The software design methodology described above has been tested in the development of a personal alarm device, used here to demonstrate different stages in the design process. Some details have been omitted for presentation purposes. The following functional specification of the device was given in the beginning of the design process. The personal alarm device is a battery-driven system worn by a person on his or her body, for example, by an elderly person at a care facility.

The device is capable of detecting the person's fall by analyzing acceleration. Once a fall has been detected, a fall alarm is sent wirelessly to an external receiver. The analysis requires that acceleration is sampled periodically every tperiod milliseconds. The device also includes an assistance call button that can trigger a separate kind of alarm sent in the same manner. An alarm must be sent within talarm milliseconds after a fall has been detected or after the button has been pressed. (1) Defining Extended System Specification. An extended system specification should include both functional and non-functional requirements. The functional requirements have to be expressed in terms of time-constrained reactions. Two such reactions can be identified by analyzing the original specification. The first reaction is sending an assistance alarm when the push button has been pressed. There is a timing requirement that the alarm is to be sent within talarm milliseconds. The second reaction is sending a fall alarm, which is triggered by fall detection. This is realized using a fall detection algorithm that requires sampling acceleration at regular intervals equal to tperiod milliseconds. The algorithm distinguishes two stages in fall detection: impact detection, with impact detected by acceleration exceeding a threshold value; and posture evaluation (see [7, 11] for a detailed description of the algorithm). Posture evaluation is performed tlag milliseconds after an impact has been detected, and is used to establish if the person is lying down, in which case a fall has been detected. The acceleration is sampled with the same periodicity both for impact detection and posture evaluation.

Hence the following timing requirements can be given for the second reaction: the acceleration sampling period tperiod; the lag between impact detection and posture evaluation tlag and the maximum period of time between fall detection and sending an alarm talarm. Both an assistance alarm and a fall alarm are sent using a radio transceiver and are received by external infrastructure which is outside the scope of the system. Therefore, the communication protocol (with its timing requirements) has to be part of the extended specification. Non-functional requirements for the system include a relatively small size (since the system has to be worn on the body, for example, at the hip), and a low-power consumption (as the device is to be powered by a battery [12–14]). (2) Formulation of System-Level Model Analyzing system specification, we can distinguish two events that the system should react to: an assistance call realized as an interrupt from a button; and the person's fall. The interrupt from a button can be modeled as an external input event. The person's fall, however, is something that is detected by the fall detection algorithm which is internal to the system and hence it is not an external event.

However, we can encode a periodic sampling of acceleration by the system as a reaction to a reset (an external input event) that starts up the system and triggers a reaction that includes sampling the acceleration (an external output "read" event) and posting a message with a delayed baseline that invokes another sampling after tperiod milliseconds, and so forth. The timing requirements on the first reaction consist of a relative deadline talarm milliseconds; the timing requirements on the second reaction are defined for each sampling that has a baseline equal to the baseline of the previous sampling plus tperiod milliseconds. Note that while the hardware for the button, the accelerometer and the radio transceiver are clearly a part of the system, the receiver of the alarm transmission is outside the developer's remit and should be viewed as an external service, and not as a system component. Thus the interface between the system and its environment is comprised on one hand, by reset interrupts and call button interrupts, and on the other hand, by the radio protocol used for communicating the alarms alongside the codes used to distinguish an assistance alarm from a fall alarm. (3a) Partitioning into Components. Let us now consider partitioning into components of our device. Analyzing the specification and the system-level model, we can see that the application will need the following independent resources: an acceleration sensor, a message sender (containing a radio transceiver) and a push button. Their independence warrants creating three separate components, each of them including both hardware and software parts (Fig. 113.2).

The next step is to define the interface of these components, bearing in mind that it should be complete but at the same time sufficiently abstract to accommodate various component implementations, which may possibly use different hardware to support the same functionality. The interface to the acceleration sensor should contain an input that can trigger sampling (sampleAcc), and an output that delivers the acceleration value once it has been acquired (consumeAcc) [15, 16]. Note that to preserve reactivity and component independence, we cannot allow the caller to block waiting for the sampling to complete. It is therefore necessary to implement callback functionality in the acceleration sensor to specify

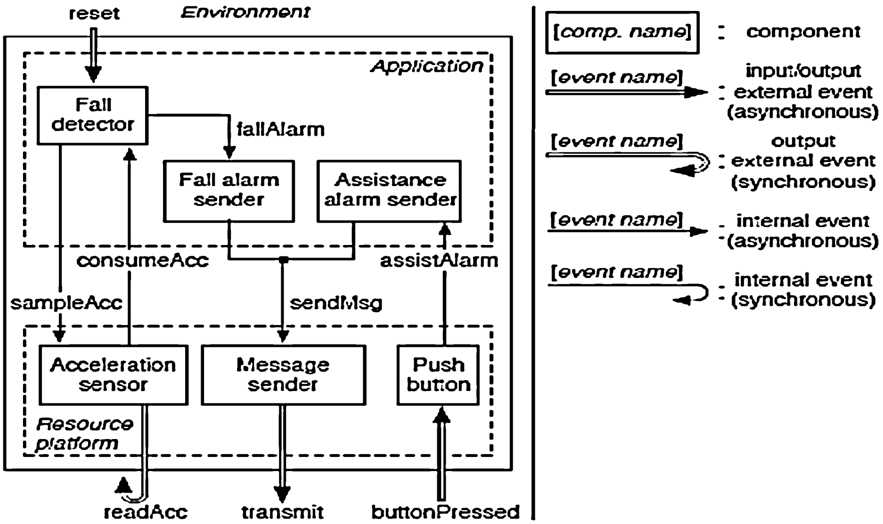


Fig. 113.2 AOM for personal alarm device

to which component the measured acceleration should be delivered. This can be done either when the acceleration sensor is instantiated (a static callback), or by passing a pointer to a function each time sampling is triggered (a dynamic callback) [17]. Similarly, to achieve the desired level of generality, the interface of the message sender should only contain one input—sending a message (sendMsg), and one output—delivery of a received message, but the latter is superfluous for our application. Note that the message sender represents a clear example of a shared resource—it can be used by any of the independent tasks of (a) fall detection, and (b) handling an assistance call. As such, it will have to include either message queuing or some kind of arbitration to synchronize access to the resource transparently to the components that may want to use it simultaneously. The interface of the last resource component—the button—is very simple, as it only needs one output to deliver the button event and the target component can easily be set statically. These three components naturally form a platform with clearly defined functionality and interface between it and any possible application. It now remains to partition the rest of the system—the application—into components. Here two independent activities can be identified: fall detection and assistance call handling, resulting in two separate components. At the same time, it is appropriate to de-couple the fall detection algorithm from how the system should react to a detected fall. For our application, this involves creating a message and forwarding it to the message sender, which can be done by a separate component—a fall alarm sender. If assistance call detection in the application is similarly de-coupled from the reaction to it, we will have two very similar components—a fall alarm sender and an assistance call sender. A possible implementation is to create them as two instances of the same component, a general alarm sender, with some parameter set to different values at initialization. Alternatively, they can be viewed as two different components. Timing requirements can be part of component specification as time constraints on the reactions. In this case, however, we skip this step and define the timing requirements directly at the object level. (3b) Search for Ready-Made Components. In our example, the personal alarm device is developed from scratch and there are no components that can be reused in the design. However, let us consider what components could be used in the future in similar applications. The first candidate for future use is, of course, the platform, consisting of an acceleration sensor, a message sender and a push button (all components combining hardware and software). This is most natural because a platform is always defined as a collection of hardware and software resources that can be used by a range of possible applications. At the same time, it is not inconceivable that such as components an acceleration sensor, a message sender or an alarm sender can be used separately in other designs. (3c) Hierarchical refinement of Component Structure. In the case of the example system, there is no room for hierarchical refinement of component structure due to the system's simplicity.

113.5 Conclusion

The presented modeling framework allows for a unified, consistent modeling of both hardware and software. Integration of these models is beneficial for development of embedded systems as they often exhibit a great degree of interdependency between hardware and software, and the specification often describes the system as a whole rather than only its software part. At the same time, inclusion of timing requirements in a functional specification in the form of time-constrained reactions allows us to specify, reason about, and verify real-time properties of embedded systems. Moreover, our modeling framework enables the developer to offer platform independent correctness/quality of service guarantees for hard/soft real-time systems, provided that the software can be scheduled on a given hardware platform so that all reaction deadlines are met. By combining this modeling framework with component-based design techniques and by expressing system functionality using reactive objects, our approach draws from the strengths of component-based design as well as from event-based, reactive, concurrent, object-oriented programing models. It facilitates software re-use and maintenance as well as separate development of parts of the system. This approach is realized in the concrete software design methodology presented above.

References

1. Graf S, Ober I, Ober I (2006) A real-time profile for UML. Int J Softw Tools Technol Transf 8(2):113–127
2. Klein J, Hélouet L, Jézéquel JM (2006) Semantic-based weaving of scenarios. In: AOSD 2006, ACM Press, pp 27–38

3. Klein J, Fleurey F, Jézéquel JM (2007) Weaving multiple aspects in sequence diagrams. Trans Aspect Oriented Softw Dev 4620:166–199
4. Bölükbasi G (2007) Aspectual decomposition of transactions. Master's thesis, School of Computer Science, McGill University, Montreal, Canada
5. Lindgren P, Nordlander J, Svensson L, Eriksson J (2005) Time for timber. Available: http://pure.ltu.se/ws/fbspretrieve/299960
6. Cunha CA, Sobral JL, Monteiro MP (2006) Reusable aspect-oriented implementations of concurrency control patterns and mechanisms. In: AOSD 2006, ACM Press, pp 134–145
7. Atkinson C, Bunse C, Gross H-G, Peper C (2005) Component-based software development for embedded systems: an overview of current research trends, ser. Lecture Notes in Computer Science. Springer-Verlag, Berlin
8. Clarke S, Baniassad E (2005) Aspect-oriented analysis and design. Addison-Wesley Professional, Boston
9. Reddy R, Ghosh S, France RB, Straw G, Bieman JM, Song E, Georg G (2006) Directives for composing aspect-oriented design class models. Trans Aspect-Oriented Softw Dev 3880: 75–105
10. Whittle J, Araujo J (2004) Scenario modelling with aspects. IEE Proc Softw 151:157–171
11. Nordlander J, Jones MP, Carlsson M, Jonsson J (2005) Programming with time-constrained reactions. Available: http://pure.ltu.se/ws/fbspretrieve/441200
12. Cottenier T, van den Berg A, Elrad T (2007) Stateful aspects: the case for aspect-oriented modeling. In: 10th Aspect-oriented modeling workshop, ACM Press
13. France R, Ray I, Georg G, Ghosh S (2004) Aspect-oriented approach to early design modelling. In: IEE proceedings software, August 2004, pp 173–185
14. Kangas M, Wiklander J, Vikman I, Nyberg L, Lindgren P, Ïamsä T (2007) Sensorband fall detector prototype: Validation through data collection and analysis. In: The 2nd international symposium on medical information and communication technology (ISMICT'07)
15. Kangas M, Vikman I, Wiklander J, Lindgren P, Nyberg L, Ïamsä T (2009) Sensitivity and specificity of fall detection in people aged 40 years and over. Gait and Posture 29:571–574
16. Kienzle J, Gélineau S (2006) AO challenge: implementing the ACID properties for transactional objects. In: Aspect-oriented software development—AOSD 2006, ACM Press, pp 202–213
17. Zhao Y, Liu J, Lee E (2007) A programming model for time-synchronized distributed real-time systems. In: 13th IEEE real-time and embedded technology and applications symposium (RTAS), pp 259–268

Chapter 114
The Research and Design for Registering Detection of Rotary Screen Printing Machine Based on Mean Shift and Harris Operator

Junfeng Jing, Guanyan Li and Yang Li

Abstract In order to realize the automatic registering system of rotary screen printing based on smart camera, this paper proposes the color printing images segmentation technology based on Mean Shift algorithm and the block matching algorithm based on Harris corner detection. The extended form of the Mean Shift algorithm is used in the segmentation of printing images, and the experiments show that the algorithm is good for the segmentation of printing images. Each chromaticity region is extracted from the segmented standard image,and the feature points on each region are detected by Harris operator. With these feature points as centers, select the standard matching blocks, and find the best matching blocks in the dealt image caught in real time to calculate the register error. The algorithms are proved to be feasible by simulations and experiments, which lays foundation for the closed-loop control of online register detection.

Keywords Rotary screen printing machine · Register detection · Mean shift · Harris · Block matching

114.1 Introduction

The accuracy of the mis-shift pattern is a key factor which affects the quality of fabric printing [1]. In order to guarantee the accuracy of printing patterns, and to ensure that there is no printing quality problems, such as the mis-shift pattern, all

J. Jing (✉) · G. Li · Y. Li
College of Electronics and Information, Xi'an Polytechnic University,
710048 Xi'an, China
e-mail: jjfeng2011@126.com

J. Jing
College of Mechanical and Electrical Engineering, Xi'an University
of Electronic Science and Technology, 710126 Xi'an, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_114,

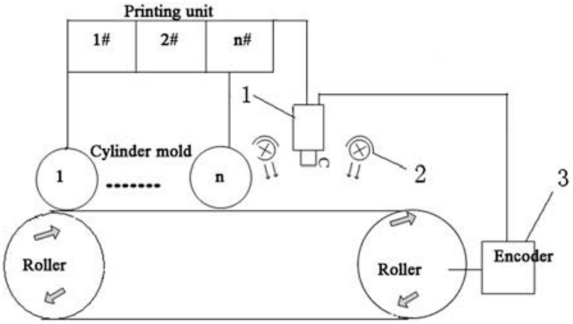


Fig. 114.1 The diagram of printing image detection system

printing cylinders of rotary screen printing machine must keep the same pace precisely with the printing conduction band [2, 3]. In the actual production process, as the rotary screen printing machines may affected easily by the wear of transmission parts, loose gear of the mis-shift pattern and fabric deformation and other factors, the relative position between the cylinders and the bands may changed easily, if not adjusted timely, the mis-shift pattern may appeared. For this question, it is still determined and justed artificial, and the truely intelligent closed loop precision control of the mis-shift pattern not yet been realized.

This research, which based on proposing error detection scheme on the basis of the mis-shift pattern that based on the smart camera, against the patterns for color printing and dyeing characteristics of the texture noise, image segmentation by the use of Mean Shift, and with the Harris operator to extracts the features of all the regions chromatic that has been extracted, and then make the feature point as the center, choosing the right template in the standard image, and with which making the block matching on the target image when real time acqusition, and then calculates the error of the mis-shift pattern [4, 5]. To make the preparation for the mis-shift pattern precision control system of intelligent closed loop.

114.2 Error Detection Programs of the Mis-Shift Pattern Based on Smart Camera

When rotary screen printing started, it will adjusts the position and speed about each cylinder of the system, when adjusted the speed of the band and the relative position between the cylinders and the bands well, if there is no the mis-shift pattern appeared, the location of the logo on the fabric printed by the cylinders is determined.

Figure 114.1 is a rotary screen printing of the mis-shift pattern detection system diagram, 1 stands for smart camera, 2 stands for light, 3 stands for intelligent camera shutter control encoder, the dye of the fabric carried by the condition band is the direct detection object of the smart camera, the system installed the CCD camera on the fixed point of the condition band, when the conduction band carried the fabric through the entire several times of therepeat sizes (the mis-shift pattern error is the accumulated error, there is no need to collect each repeat sizes), the smart camera receives the trigger pulse of the encoder and then collects the image. And then the smart camera calls the arithmetic which used for image processing inside, making the chromatic segmentation, the extraction and the matching of each chromatography, calculating the error of the mis-shift pattern and compared with the threshold seted before, according to the result, adjusting the relative rotary screen or the conduction band.

114.3 Mean Shift in Image Segmentation

114.3.1 Introduction of Mean Shift Algorithm

Mean Shift algorithm is an effective feature space iterative clustering algorithm for statistical, which was original proposed by Fukunaga and others in 1975, this research will use Mean Shift algorithm that extended on the fabric image segmentation, through combining the colour and spacial feature clustering [6]. And then solve the segmentation problem of fabric image well.

Extended Mean Shift vector as follows:

that is:

$$M_h(x) = \frac{\sum_{i=1}^{n} G(\frac{x_i - x}{h}) w(x_i) x_i}{\sum_{i=1}^{n} G(\frac{x_i - x}{h}) w(x_i)} - x \tag{114.1}$$

order

$$M_h(x) = \frac{\sum_{i=1}^{n} G(\frac{x_i - x}{h}) w(x_i) x_i}{\sum_{i=1}^{n} G(\frac{x_i - x}{h}) w(x_i)} \tag{114.2}$$

$$M_h(x) = \frac{\sum_{i=1}^{n} G(\frac{x_i - x}{h}) w(x_i)(x_i - x)}{\sum_{i=1}^{n} G(\frac{x_i - x}{h}) w(x_i)} \tag{114.3}$$

Where, $G(x)$ is a unit of core function, Unit uniform core function and Epanechnikov kernel function are as follows:

$$\mathrm{F(x)} = \begin{cases} 1 & \text{if } \|x\| < 1 \\ 0 & \text{if } \|x\| \geq 1 \end{cases} \tag{114.4}$$

$$K_E(x) = \begin{cases} C_E(1 - \|x\|^2) & \text{if } \|x\| \leq 1 \\ 0 & \text{if } \|x\| > 1 \end{cases} \tag{114.5}$$

C_E is a constant to ensure $\int_{R^d} |K_E(x)| dx = 1$ achieved.

From 1 to 3, h is a window width parameter which is >0, $w(x_i) \geq 0$ is the weight factor that assigned to the sampling points x_i. The reason why the

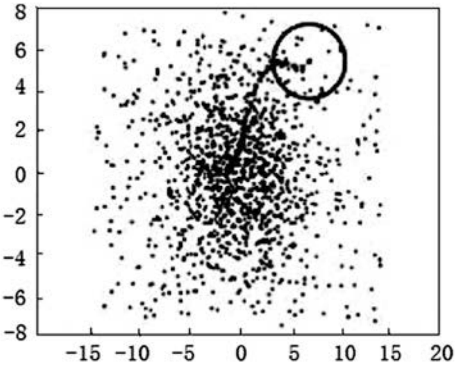


Fig. 114.2 The diagram of Mean Shift algorithm searching mode

introduction of the weight coefficient w (x_i) of the sampling points x_i is because as long as the sampling points into Sh, the nearer to x, the greater impact on the statistical characteristics.

The Mean Shift vector and density gradient estimated relationship are as follows:

$$M_{h,\,g}(x)=\frac{1}{2}h^2c\frac{\overset{\wedge}{\nabla f}_{h,k(x)}}{\hat{f}_{h,G(x)}} \tag{114.6}$$

where, c is bigger than zero, g (x) is the profile function of the core function $G(X)$, at the point of x, Mean Shift vector calculated by a core function $G(X)$ and the density gradient estimate standardized by a core function $K(X)$ are proportional, Therefore, this vector always points to the direction of the largest density. Here standardization was calculated by the probability density estimate which was calculated by the core $G(X)$. Mean Shift vector $M_{h\cdot g}(x)$ moved to the direction that the density increased maximum. At the same time, the step may changed, when the density is small, the step is long, and when the density is big, that is closer to the peak of the probability density (mode), the step is short. Under certain condition, Mean Shift algorithm will converge to a point x that near the peak. Mean Shift algorithm searching process, such as Fig. 114.2.

114.3.2 Mean Shift Image Filtering

When the color printing image determined the color space(here select the color space of uniform color model Luv) that is the color information (L, U, V) and the space information (X, Y), then you can get each pixel on the five-dimensional feature space, that is (L, U, V, X, Y). And then Mean Shift algorithm clustered through combining the color and spatial characteristics. With $x_i(i = 1,\ldots n)$ respected the n-

pixel of the image, started with one of these pixels $x_j(j = 1,\ldots n)$, got the Mean Shift vector $Mh.\text{g}(x)$.

In order to understand the role of the core function $G(X)$ in the Mean Shift iterative process clearly, using $\{y_j\}$, $j = 1, 2,\ldots$ indicates a series of center as $G(x)$ moved, then you can get

$$y_{j+1} = \frac{\sum_{i=1}^{n} G(\frac{x_i - y_i}{h}) w(x_i) x_i}{\sum_{i=1}^{n} G(\frac{x_i - y_i}{h}) w(x_i)} \quad i = 1, 2, \ldots n \tag{114.7}$$

In which, at the beginning $y_1 = x_{j,}$ y_{j+1} is the weighted average calculated by the core function $G(X)$ in y_j, also the next continuous point calculated in y_j. Density estimation corresponding in these continuous center point formed a sequence, as follows:

$$\left\{ \hat{f}_{h,k(j)} \right\} = \hat{f}_{h,k(y_j)} \quad j = 1, 2, \ldots \tag{114.8}$$

In which, $\hat{f}_{h,k(y_j)}$ is a series of probability density estimation calculated by K, because the Mean Shift algorithm is convergent, so that the sequence $\{y_j\}$, $j = 1, 2,\ldots$ is convergent, from the starting point y_1, the sequence $\{y_j\}$, $j = 1, 2,\ldots$ The corresponding sequence of density estimates (114.8) also becomes larger and larger, namely: the density becomes larger and larger, until converge to the maximum density.

Specific filtering steps are as follows:

Order $\{x_i\}$, $i = 1, 2,\ldots, n$ and $\{z_i\}$, $i = 1, 2,\ldots, n$ indicates the original and filtered pixel in five-dimensional respectively. To each pixel:

1. Initialization: $j = 1$, $y_{i,1} = x_i$;
2. According to the formula (114.7), calculate $y_{i,j+1}$ in the core whose center is $y_{i,j}$ until it convergent, denoted by $y = y_{i,c}$;
3. Order $z_i = (x_i^s, y_{i,c}^r)$, s and r of which denote the space and the chroma components of the vector.

In step 2, we use the Epanechnikov kernel function. The filtered data in the space position x_i may use the color field component of the convergence point y_c. The spatial components makes use of x_i, and color field component with y_c, and also the color value convergent is assigned to the point that treated.

114.3.3 Mean Shift Segmentation

In this research, image means texture image, which should classify the variegated caused by texture on the weft on the same region. Therefore, in the process of cluster in the merging regional, with some artificial constraints, then set the minimum area size limit M.

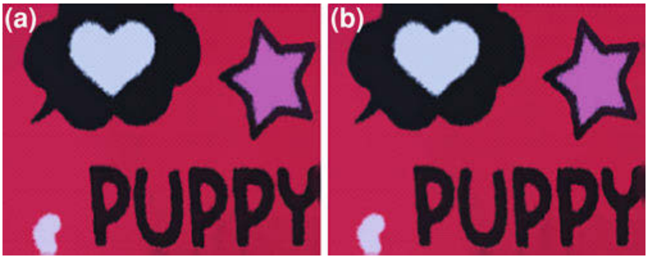


Fig. 114.3 The segmentation results of the printing image by Mean Shift **a** The original printing image **b** The segmentation results by Mean Shift

This research chose 24-bit color fabric image, whose size is 228 × 176, resolution is 96 dpi. The separate effects which is controlled by three parameters, that are spatial bandwidth h_s, color field bandwidth h_r and the minimum regional M, setting parameter $[h_s,\ h_r,\ M] = [8.5,\ 8,\ 1{,}000]$, the segmentation results as Fig. 114.3. At sharing time 3.57 s, you can see from the segmentation result that the segmentation based on Mean Shift algorithm can overcome the effect of the outside world, In particular, it can filter out the texture noise of each trapping area, This can directly make the chromatic extraction on the segmentation results in Fig. 114.3b.

114.4 Error Detection Method of the Mis-Shift Pattern Based on Block Matching

From the intuitive sense, corner is a point that contain enough information and can extract from current frame and the next frame. Harris operator is a stability operator, which is a kind of feature extraction that based on signal and proposed by Harris and Stephens, which is charactered by simple calculation and reasonably and uniform corner that extracted. The process is expressed as:

$$M = \mathrm{Gauss}(\tilde{s}) \otimes I \tag{114.9}$$

$I = \begin{bmatrix} I^2x & IxIy \\ IxIy & I^2y \end{bmatrix}$ is the dual gradient Matrix, Ix and Iy are the first-order partial derivatives of the image in the direction of the x, y, $\mathrm{Gauss}(\tilde{s})$ is Gaussian model,

$$R = \det(M) - k(\mathrm{trace}(M))^2,$$
$$k = 0.04\text{–}0.06$$

det is the determinant of the matrix, trace is the matrix trace, k is a free parameters of the corner detection. Each pixel of the matrix and the corner point corresponding of the original image match one by one, only when the corner is greater

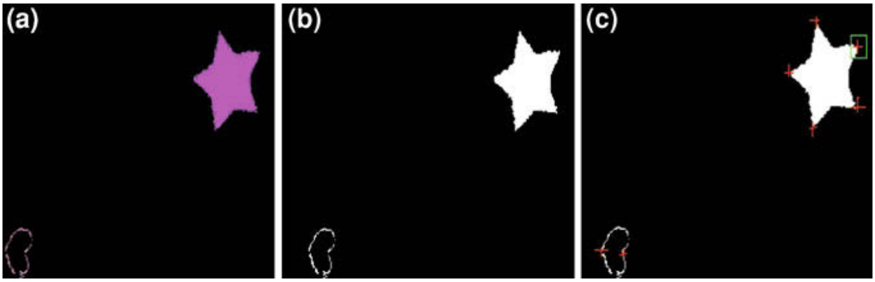


Fig. 114.4 The corner points detection on the parted different color region (**a**) chromatic *pink* (**b**) binary (**c**) extract corner point

than a certain threshold is considered the corner, As the threshold is too small, which will increased the number of feature points and the amount of subsequent calculation of matching, if too big, it will increased the likelihood of leakage election, so you can select the appropriate threshold by your experience in practice.

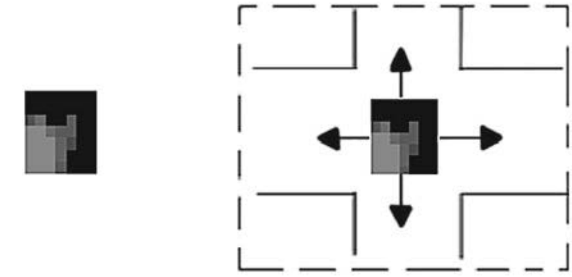
In order to achieve fast matching, you should filter the feature points. Screening principles as following: First, the distance between two feature points not less than the size of the cut, second, select a square area as the center is the feature points and the length is Bs, matching on the standard image first, if you find two or more than two matching regions, then you should exclude the feature point. Making the chromatic pink as example to indicate the extraction process of the corner Harris. Figure 114.4a is a chromatic pink that extracted from the segmentation results from the original image, Fig. 114.4b is the binary. After corner detection, marked the corners that after selected in the Fig. 114.4c.

Image matching technology is a process of finding the most similar region between the target image and template image. The principle of block matching break up the image into two-dimensional and sub-blocks of a certain size, and then find the best match of these sub-blocks through its adjacent frames by BMA matching criterion, the relative position (dx, dy) between which and the current block is the motion vector. This chapter refers to the above ideas, for the standard image, each feature point P_i $(x_{i,}\ y_i)$ selected of the feature points P, which as the center, chose the small area which size is Bs × Bs as the feature block, In the target image, make the corresponding point O_i $(x_{i,}\ y_i)$ of P_i as the center, chose a cross space whose length is Ss(>Bs) and then find the best matching block, Fig. 114.5 makes the example of the corner which marked in the upper right of Fig. 114.4c, expressing the matching process in the target image of the selected image.

In this chapter, matching criteria is normalized correlation function, namely the seeking of cross correlation:

$$R(i,j) = \frac{\mathrm{Rst}}{\sqrt{\mathrm{Rss}\cdot\mathrm{Rtt}}}$$

Fig. 114.5 The diagram of feature blocks searching and matching



feature block

searching region

Table 114.1 Coordinate standard points and matching point

Corner	The corner coordinate of standard points	The matching points coordinate of target image
1	(5, 83)	(6, 85)
2	(15, 72)	(15, 73)
3	(36, 80)	(36, 82)
4	(33, 93)	(34, 94)
5	(16, 94)	(16, 96)
6	(69, 9)	(69, 12)
7	(69, 19)	(69, 21)

$$i,j \in \left(P(i,j) - \frac{\mathrm{Ss}}{2}P(i,j) + \frac{\mathrm{Ss}}{2}\right)$$

of which:

$$\mathrm{Rst} = \sum_{x}^{\mathrm{Bs}}\sum_{y}^{\mathrm{Bs}}(T(x,y) - \overline{T}) \times (S(x+i,y+j) - \overline{S}(i,j)) \tag{114.10}$$

$$\mathrm{Rtt} = \sqrt{\sum\nolimits_{x}^{\mathrm{Bs}}\sum\nolimits_{y}^{\mathrm{Bs}}(T(x,y) - \overline{T})^2} \tag{114.11}$$

$$\mathrm{Rss} = \sqrt{\sum\nolimits_{x}^{\mathrm{Bs}}\sum\nolimits_{y}^{\mathrm{Bs}}(S(x+i,y+j) - \overline{S}(i,j))^2} \tag{114.12}$$

The steps (1) calculate the Rtt of matching block; (2) calculate Rst and Rss in the searching region respectively; (3) compare the correlation coefficient, then compare the maximum absolute value with the value of a threshold function, if max(|*R* (*i, j*)|) > *Threshold*, then it is the most suitable.

114.4.1 Result

When the image matching finished, such as Table 114.1, you can get the the average of the seven points dx = 0.142857 mm, dy = 0.928572 mm. That is the pink colour area of the image turn 0.928572 mm in the right, turn down to 0.142857 mm. Because we have selected the feature points before, and put the searching zone to a "+" area, so the matching arithmetic enhance our efficiency, the time-consuming is just 0.41 s.

Acknowledgments The authors would like to thank for the support of Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 11JK0910).

References

1. Cheng Y (1995) Mean-shift mode seeking and clustering[J]. IEEE Trans Pattern Anal Machine Intell 17(8):790–799
2. Fukunaga K, Hostetler LD (1975) The estimation of the gradient of adensity function, with applications in pattern recognition[J]. IEEE Trans Pattern Anal Machine Intell 21:32–40
3. Xu M (2007) The study of a fabric image segmentation algorithm[D]. Zhejiang University, Zhejiang
4. Zhiqiang W, Zhixing C (2007) The convergence analysis of mean shift algorithm[J]. J Softw 182:205–212
5. Harris C, Satephens MJ (1988) A combined corner and edge detector[J]. Image Vis Comput 6(1):121–128
6. Lu H (2010) Research of block-matching motion estimation algorithm[D]. Jilin University, Jilin

Chapter 115
Simulation of Road Circular Curve Design Specifications in Cold Regions

Xi-qiao Zhang, Long-hai Yang and Shi An

Abstract This chapter focuses on the phenomenon that road adhesion coefficient decreases in the winter of cold regions, analyzes the outside brake forces of vehicles when turning on level road surface, and establishes the model of circular curve without setting superelevation. Then it measures data to determine the structure of the vehicle, speed, adhesion coefficient, sight distance, braking time and other parameters, decides theoretical security design specifications values of circular curve by simulation methods and finally gives the recommended radius values of circular curve without setting superelevation in consideration of driving comfort, which provides a frame of reference for horizontal alignment of the road in cold regions.

Keywords Cold regions · Road · Circular curve design · Line shape simulation

This work is supported by Ministry of Transportation West Fund Project (No. 200631882090).
Yang Long-hai (1970-), male, Anhui Province, Harbin Institute of Technology, associate professor, Ph.D. research area: traffic demand analysis and traffic simulation.

X. Zhang (✉) · L. Yang · S. An
School of Transportation Science and Engineering,
Harbin Institute of Technology, Harbin 150090, China
e-mail: yanglonghai@hit.edu.cn

X. Zhang
School of Management, Harbin Institute of Technology,
Harbin 150001, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_115,

115.1 Introduction

In cold regions, long winter, low temperature and heavy snow make the ice and snow hard to be cleaned up; therefore, the road condition is bad and the adhesion coefficient of road surface is usually very low, which can easily lead to sideslip and failure of rotation ability of front wheel in time of braking. This phenomenon has a relation with the selection of highway route design indexes.

Circular curve is a commonly used linear element, the design of circular curve is primarily to determine its radius value. When vehicles brake on the circular road, it will be subjected to lateral forces such as a centrifugal force. When the lateral counter-force reaches the lateral adhesion, the car will slip along the direction of the lateral force, which will cause significant impact on driving safety. In the smaller circular curve radius sections, in order to reduce or offset the effect of centrifugal force, it can be used by a large superelevation. If too high superelevation is used the vehicle will risk slipping along the largest road slope of synthesis. In particular, for a low-speed vehicle or a stopped vehicle, the centrifugal force is close to or equal to zero. At this time the lateral adhesion of the car should be equal to the force of gravity acted on cross-slope. It shows that the maximum is correlated with the adhesion coefficient. Therefore, it is important to establish the model of the relationship among adhesion coefficient, the circular curve radius and the large superelevation, in order to reveal the variation among them.

115.2 Model of Vehicle Braking in Circular Curve

Circular curve can be differentiated as crown slope curve and superelevation curve. The load-carrying of vehicles in different circular curves are different, and accordingly, the dynamic models of vehicle brakings are different. Assuming the wheels are rigid, then the angle of steering wheel is in proportion to the angle of the front-wheel. The angle between velocity reversal of middle point in the front wheel and vertical axis can be denoted as $\delta = 0.5(\delta_L + \delta_R)$, thereinto, δ_L and δ_R represent the left and right front wheels respectively. The angle between velocity reversal of vehicle centroid and vertical axis is denoted as γ. F_{X1} and F_{X2} denote the longitudinal force exerted on front and behind wheels respectively. F_{Y1} and F_{Y2} denote the transverse force exerted on front and behind wheels respectively. F_C denotes the centrifugal force. V denotes the velocity of vehicle centroid. m denotes the total mass of vehicle. l denotes wheel base. l_1 and l_2 denote the distance from the centroid to the front axle and rear axle respectively. hg denotes the height of the centroid. B denotes the tread (the front tread and the rear tread are the same). The coordinate system is set up and the forces exerted on the vehicle are shown in Fig. 115.1.

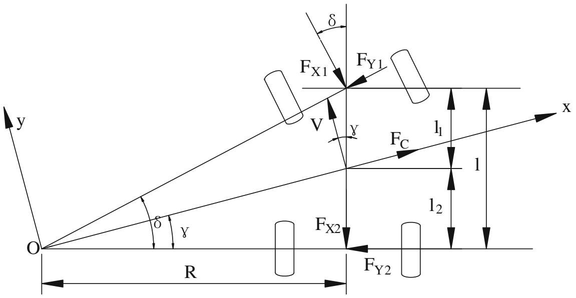


Fig. 115.1 The forces exerted on vehicle in time of turning in horizontal road surface

115.2.1 Modeling Assumptions

A vehicle dynamics model can represent the motion characteristics of vehicles, and hence lays the basis for describing the vehicle braking status. Due to the complexity of vehicle motion, where solving the MDOF problem is involved, thus in order to simplify the mathematical model while sufficing to represent the motion characteristics of vehicles, in this chapter the whole vehicle movement along three axes, longitudinal displacement, transverse displacement, and yaw angle displacement, and the rotary motion of four wheels, have been selected to establish the braking model for the whole vehicle. The following assumptions have been made: (1) the vehicle body as a whole is considered to be completely rigid; (2) the influence of the road roughness on the accuracy of the result of theoretical analysis is neglected; (3) the load transfer between the coaxial left and right wheels caused by the side tilt of the vehicle body, is not considered; (4) there is no vertical movement for the vehicle; (5) the pitch and side tilt movements of the vehicle body are not considered when it is moving; (6) tires have the same mechanical properties; (7) the air resistance and wheel rolling resistance are neglected.

115.2.2 Vehicle Circular Curve Braking Model Without Setting the Superelevation

When the radius of the circular curve is greater than a certain value, it can be done without setting the superelevation, and permits setting the same crown slope as a straight section. Assume the slope angle of the crown slope is β, and consider the vehicle is moving along the bidirectional crown slope. Due to the fact that it is most adverse when the vehicle is moving near the curve outside, and according to the force analysis when the vehicle is braking near the curve outside, as shown in Figs. 115.1 and 115.2, the dynamic model for vehicle braking is obtained as

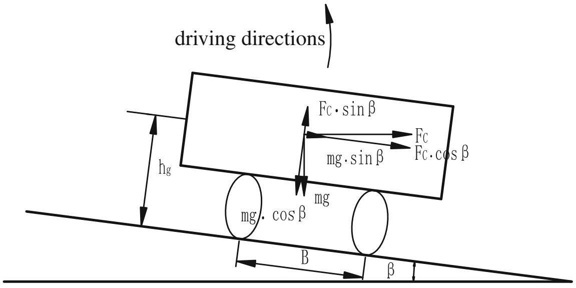


Fig. 115.2 The force of braking in the outside of corners when vehicle without setting superelevation

$$\begin{cases} mV' = \sin\gamma \cdot F_{Y2} - \cos(\delta-\gamma)\cdot F_{X1} - \sin(\delta-\gamma)\cdot F_{Y1} - \cos\gamma\cdot F_{X2} \\ \cos\beta\cdot F_C + \sin\beta\cdot mg = \cos(\delta-\gamma)\cdot F_{Y1} - \sin(\delta-\gamma)\cdot F_{X1} + \sin\gamma\cdot F_{X2} + \cos\gamma\cdot F_{Y2} \end{cases} \tag{115.1}$$

Front axle normal reaction force is:

$$F_{Z1} = \frac{(\cos\beta\cdot mg - \sin\beta\cdot F_C)l_2}{l} - \frac{mh_g}{l}V' \tag{115.2}$$

Rear axle normal reaction force is:

$$F_{Z2} = \frac{(\cos\beta\cdot mg - \sin\beta\cdot F_C)l_1}{l} + \frac{mh_g}{l}V' \tag{115.3}$$

The centrifugal force of vehicle when turning is:

$$F_C = m\cdot\frac{V^2}{R} \tag{115.4}$$

The braking in time of vehicle swerving will induce longitudinal and transverse forces on it as shown in Fig. 115.1. The relations among longitudinal force, longitudinal adhesion coefficient and normal reactive force are depicted as

$$F_{X1} = \varphi_X\cdot F_{Z1} \quad F_{X2} = \varphi_X\cdot F_{Z2} \tag{115.5}$$

The relations among transverse force, transverse adhesion coefficient and normal reactive force are depicted as

$$F_{Y1} = \varphi_Y\cdot F_{Z1} \quad F_{Y2} = \varphi_Y\cdot F_{Z2} \tag{115.6}$$

φ_X and φ_Y denote longitudinal and transverse coefficient respectively.

Formula 115.1 is a model of circular curve without setting superelevation when turning and breaking. Put formula 115.5, 115.6 into formula 115.1:

$$\begin{cases} mV' = \sin\gamma \cdot \varphi_Y F_{Z2} - \cos(\delta - \gamma) \cdot \varphi_X F_{Z1} - \sin(\delta - \gamma) \cdot \varphi_Y F_{Z1} - \cos\gamma \cdot \varphi_X F_{Z2} \\ \cos\beta \cdot F_C + \sin\beta \cdot mg = \cos(\delta - \gamma) \cdot \varphi_Y F_{Z1} - \sin(\delta - \gamma) \cdot \varphi_X F_{Z1} + \sin\gamma \cdot \varphi_X F_{Z2} \\ + \cos\gamma \cdot \varphi_Y F_{Z2} \end{cases} \tag{115.7}$$

To simplified, define $\cos\beta \cdot mg - \sin\beta \cdot F_C = A$, put formula 115.2, 115.3 into the formula 115.7

$$\begin{aligned} lmV' =& \varphi_Y \sin\gamma(l_1 A + mh_g V') - \varphi_Y \sin(\delta - \gamma)(l_2 A - mh_g V') \\ & - \varphi_X \cos(\delta - \gamma)(l_2 A - mh_g V') - \varphi_X \cos\gamma(l_1 A + mh_g V') \end{aligned} \tag{115.8}$$

$$\begin{aligned} (\cos\beta \cdot F_C + \sin\beta \cdot mg)l = & \varphi_Y \cos(\delta - \gamma)(Al_2 - mh_g V') + \varphi_X \sin\gamma(Al_1 + mh_g V') \\ & - \varphi_X \sin(\delta - \gamma)(Al_2 - mh_g V') + \varphi_Y \cos\gamma(Al_1 + mh_g V') \end{aligned} \tag{115.9}$$

Straighten formula 115.8 as

$$V' = \frac{[\varphi_Y l_1 \sin\gamma - \varphi_Y l_2 \sin(\delta - \gamma) - \varphi_X l_2 \cos(\delta - \gamma) - \varphi_X l_1 \cos\gamma]A}{m\{l - h_g[\varphi_Y \sin\gamma + \varphi_Y \sin(\delta - \gamma) + \varphi_X \cos(\delta - \gamma) - \varphi_X \cos\gamma]\}} \tag{115.10}$$

Straighten formula 115.9 as

$$V' = \frac{(\cos\beta \cdot F_C + \sin\beta \cdot mg)l - [\varphi_Y l_2 \cos(\delta - \gamma) - \varphi_X l_2 \sin(\delta - \gamma) + \varphi_X l_1 \sin\gamma + \varphi_Y l_1 \cos\gamma]A}{[\varphi_X \sin(\delta - \gamma) + \varphi_X \sin\gamma + \varphi_Y \cos\gamma - \varphi_Y \cos(\delta - \gamma)]mh_g} \tag{115.11}$$

$$\varphi_Y l_1 \sin\gamma - \varphi_Y l_2 \sin(\delta - \gamma) - \varphi_X l_2 \cos(\delta - \gamma) - \varphi_X l_1 \cos\gamma = B$$

$$l - h_g[\varphi_Y \sin\gamma + \varphi_Y \sin(\delta - \gamma) + \varphi_X \cos(\delta - \gamma) - \varphi_X \cos\gamma] = C$$

$$\varphi_Y l_2 \cos(\delta - \gamma) - \varphi_X l_2 \sin(\delta - \gamma) + \varphi_X l_1 \sin\gamma + \varphi_Y l_1 \cos\gamma = D$$

$$[\varphi_X \sin(\delta - \gamma) + \varphi_X \sin\gamma + \varphi_Y \cos\gamma - \varphi_Y \cos(\delta - \gamma)]h_g = E$$

That formula 115.10 is simplified as:

$$V' = \frac{BA}{mC} \tag{115.12}$$

Simplify formula 115.11 as

$$V' = \frac{(\cos\beta \cdot F_C + \sin\beta \cdot mg)l - DA}{mE} \tag{115.13}$$

Simultaneous formula 115.8 and 115.9

$$C(\cos\beta \cdot F_C + \sin\beta \cdot mg)l = (BE + CD)A \tag{115.14}$$

Put $\cos\beta \cdot mg - \sin\beta \cdot F_C = A$ and formula 115.4 into formula 115.14, through simultaneous equations, the braking model in circular curve without setting superelevation can be denoted as:

$$R = \frac{[Cl\cos\beta + (BE + CD)\sin\beta]V^2}{[(BE + CD)\cos\beta - Cl\sin\beta]g} \tag{115.15}$$

Thereinto, B, C, D and E denote relevant parameters that have relation with adhesion coefficient, steering angle and wheel base.

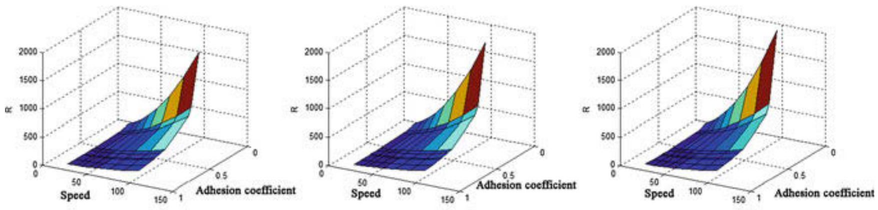
115.3 Simulation Analysis on Vehicle Circular Curve Braking Model

Simulation technology has become an indispensable tool in the study of complex systems analysis and design. The relationship between road adhesion coefficient and road alignment indicators is complex. The relational model is a multivariate one, and it is difficult to quantify and intuitive analysis their correlation. The simulation technology can be used to solve the mathematical model and simulate data into two-dimensional or three-dimensional graphics which can be analyzed and processed on the graphics. In this chapter, MATLAB simulation technique has been used to simulate the circular curve braking model, in order to lay the foundations for providing the recommended values for the design of circular curve.

115.3.1 Determining Simulation Parameters

According to the established vehicle circular curve braking model, parameters such as vehicle structural parameters, vehicle speed, and adhesion coefficient, should be determined.

1. *Determining vehicle structural parameters.* For different types of vehicles, structure dimensions, vehicle speed, gross vehicle mass (GVM), may be distinct, and hence requires different road geometric design indexes. In this chapter, the most adverse vehicle type, truck, is selected to implement the simulation study on the model of road geometric design indexes.
2. *Determining the vehicle speed.* The relationship between vehicle speed and the design speed is: when the design speed is in the range of 120–80 km/h, the vehicle speed is 85% of the design speed; when the design speed is between 80 and 40 km/h, the vehicle speed is 90% of the design speed; when the design speed is below 40 km/h, the vehicle speed equals the design speed [1].
3. *Determining the adhesion coefficient.* The transverse adhesion coefficient φ_Y is restricted by vehicle driving status, when there is no tangential force, but only transverse force retains, φ_Y takes its extreme value φ; when there is no



road crown slope 1.5% road crown slope 2.0% road crown slope 2.5%

Fig. 115.3 Adhesion coefficient and simulation models without setting radius of superelevation circular curve

transverse force, but only tangential force, φ_X takes its extreme value φ; when both forces exist, the longitudinal adhesion coefficient φ_X and transverse adhesion coefficient φ_Y are related to the adhesion coefficient φ as $\varphi = \sqrt{\varphi_X^2 + \varphi_Y^2}$. It is commonly used $\varphi_X = 0.8{-}0.7\varphi$ and $\varphi_Y = 0.6{-}0.7\varphi$ both experimentally and empirically, and among various adhesion coefficient values for different road conditions and vehicle speeds, the minimum value could be chosen as the adhesion coefficient under current road conditions [2].

4. *Determine other parameters.* From a security point of view, choose a lower value of vehicle eye height as the eye height for the driver, viz., the stopping sight distances should be chosen as $d1 = 1.2$ m, $d2 = 0$; when the transient sight distance is used, $d1 = d2 = 1.2$ m should be adopted. One vehicle braking process can be divided into several time periods, including the reaction time 1.36 s for the driver, the coordinating time for the braking system 0.04 s, and the growing deceleration time 0.2 s [3].

115.3.2 Simulation Results

According to the vehicle braking model without setting superelevation circular curve, that is formula 115.7 and truck types, this chapter carried out a simulation when the road crown slope was 1.5, 2.0 and 2.5% and calculated adhesion coefficient and depicted simulation models without setting radius of superelevation circular curve as shown in Fig. 115.3.

115.3.3 Simulation Results Analysis

Simulation studies suggest that adhesion coefficient is closely related with radius values of circular curve without and with setting superelevation, its variation is complex, and there will be a big difference in the calculated values of road alignment indicators under different attachment conditions [4, 5]. Therefore, in

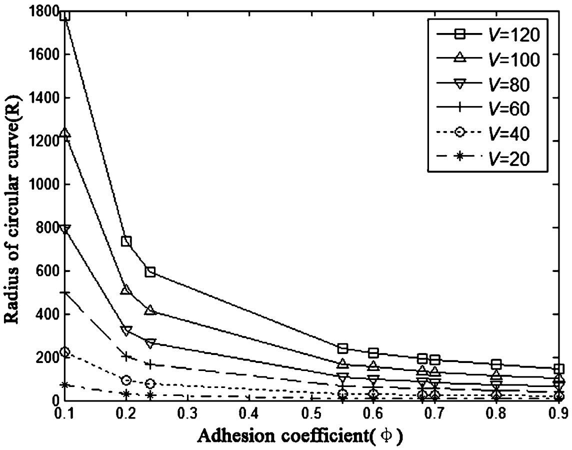


Fig. 115.4 Variation of circular curve radius with adhesion coefficient

order to meet the road alignment requirements of traffic safety, the impact of adhesion coefficient in cold regions should be fully considered when determining the road alignment indicators.

If the 2.0% camber slope and the 8% superelevation are taken into account in simulation results of circular curve radius, we may discover the difference, as shown in Fig. 115.4. Figure 115.4a shows variation of circular curve radius with adhesion coefficient when setting the camber slope as 2.0% without any superelevation, while Fig. 115.4b shows variation of circular curve radius with adhesion coefficient when setting the superelevation as 8%. The following are the conclusions according to the simulation:

1. With and without setting superelevation, the trend of circular curve radius with adhesion coefficient are almost the same, and the circular curve radius of no superelevation would be bigger than the other one when adhesion coefficient does not change.
2. The relationship of circular curve radius and adhesion coefficient is nonlinear. Lower adhesion coefficient, bigger circular curve radius, especially when the adhesion coefficient is lower than 0.5.

115.4 Recommended Values of Circular Curve Radius

The values of circular curve radius should satisfy the requirement of road safety and travelling comfort at the same time, which is the principle for designing circular curve radius.

Experiments showed that the coefficient value of transverse force affect travelling comfort. Therefore, the model in this chapter was built based on adhesion

Table 115.1 Recommended values without setting radius of superelevation circular curve (*m*)

Design speed (km/h)		120	100	80	60	40	20
Wet pavement and snow pavement	$i = 1.5\%$	1455	835	455	225	90	50
	$i = 2.0\%$	1555	880	475	230	90	50
	$i = 2.5\%$	1670	930	500	240	95	50
Icy pavement	$i = 1.5\%$	1650	1150	750	450	200	65
	$i = 2.0\%$	1800	1300	800	500	250	70
	$i = 2.5\%$	2000	1400	900	600	300	80

coefficient that could satisfy the requirement of travelling safety, i.e. the circular curve radius without setting superelevation (R_{12}) can satisfy the travelling safety requirement. Select the bigger one of integral R_{11} and R_{12}, i.e. $R_1 = \max\{R_{11}, R_{12}\}$, as the recommended values of circular curve radius without setting superelevation. By calculating and comparing, we can obtain the recommended values without setting the radius of superelevation circular curve under circumstance of three kinds of road conditions such as wet pavement and snow icy road conditions, as shown in Table 115.1.

115.5 Conclusion

There are vast lands of endless skies, complex and varied climate in China. The temperature, humidity, rainfall and snowfall have a significant difference throughout the year in most areas. According to the phenomenon of winter road adhesion coefficient reduction in cold areas, this chapter establishes a model of circular curve brake and gets the theoretical safe value of circular curve design index through measured data and simulation method. This chapter presents suggested values of circular curve design specifications as to a comfortable degree and this can provide a reference for road alignment design.

References

1. Yang TG (2005) Theory for solving highway traffic sight distance. Liaoning Transp Sci Technol 6:42–44
2. Pei YL (2005) Road survey and design, vol 46. Harbin Institute of Technology Press, Harbin, pp 123–125
3. Gao YL (2004) Automobile application engineering. People’s Press of China, China, p 95
4. Design rule of highway route (JTG D20-2006), People’s Communications Press, Beijing
5. Kai H, Middleton D (2006) Enhancing highway geometric design: development of interactive virtual reality visualization system with open-source technologies. Transp res rec (7):134–142

Chapter 116
Research on Threat Assessment Based on Dempster–Shafer Evidence Theory

Wen Jiang, Deqiang Han, Xin Fan and Dejie Duanmu

Abstract Based on D−S evidence theory, a new method of multi-target threat evaluating and sequencing is presented. Firstly, the concerned problems are defined and formalized using fuzzy sets and fuzzy number. Secondly, the BPA is obtained by using the similarity measure between generalized fuzzy numbers. Lastly, D−S evidence theory is applied to combine the BPA to derive a creditable result of multi-target threat evaluating and sequencing. The proposed method changes the uncertain problem of multi-target threat evaluating and sequencing to a certain one. The results coincide well with real-time air situation, and it is very valuable for scientific decision-making in air defense operation.

Keywords D−S evidence theory · Fuzzy set · Threat evaluating · Similarity measures · Basic probability assignment

116.1 Introduction

The modern air attack environment has become more complex because of the use of high technology in military affairs. Multi-target threat evaluating and sequencing is one of the key taches within the processes of command and control in aerial defense battle. It is a process to evaluate the threat degree of the air attacking target combining various characteristic information of the air attacking

W. Jiang (✉) · X. Fan · D. Duanmu
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
e-mail: jiangwen@nwpu.edu.cn

D. Han
Institute of Integrated Automation, Xi'an Jiao Tong University, Xi'an, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_116,

target delivered by sensors. Threat evaluating and sequencing of the air target is a comprehensive evaluation or decision for threat degree influence to multi-targets and multi-factors. Therefore, threat assessment substantially belongs to a type of multi-attribute evaluation or decision problem.

D−S evidence theory is a powerful and flexible mathematical tool for handling uncertain, imprecise, and incomplete information [1, 2]. In the process of building multi-target threat evaluating and sequencing system, it is complex to measure the targets' threat degree because many factors, such as arrive boundary time, navigate shortcut, target rate, target type, and interferential power, can affect it. Their influence on targets' threat degree assessment is different. Taking these into consideration, the concerned factors are defined and formalized using fuzzy sets and fuzzy number. Then D−S evidence theory is applied to combine all the issues, regarded as evidences, in order to derive a compellent result of multi-target threat evaluating and sequencing. In this paper, a model to assess air target threat is constructed based on D−S evidence theory, which provides an effective and simple method for the targets' threat degree assessment and the targets' sequencing in large-scale air attack.

116.2 Model of Multi-Target Threat Evaluating and Sequencing

The purpose of this study is to design and develop a new model which can evaluate and sequence the threat degree of air attack target. Detailed descriptions are presented in the following sections.

116.2.1 Main Factors Affecting Threat Degrees

The threat degrees of air attack targets are determined by the tactics geometrical conditions and air attack targets' capability [3, 4]. Parameters describing tactics geometrical conditions are target distance, target altitude, target speed, azimuth angle, and targets' attack intention, etc. Parameters describing targets' air-to-ground capability are target types, interfere capability, and so on. If we calculate all parameters mentioned above, the algorithm is too complicated to be easily carried out. Therefore, target distance, target altitude, target speed, target types, and target interference ability are investigated in this paper.

It is different for the above five factors to affect the threat degrees of air attack targets. Target distance, target altitude, and target speed are quantitative index and their random value can be obtained through radar scanning. Target types and target interference ability are qualitative index and should be fuzzed, so we use triangular fuzzy numbers to express them. The above factors' influence on target threat can be described specifically as follows:

1. The closer the target is, or the smaller the target distance is, the greater the target's threat is to the ground-to-air missile system.
2. The lower the target altitude is, the greater the targets' threat is to the ground-to-air missile system.
3. The quicker the target speed is, the greater the target's threat is to the ground-to-air missile system.
4. Target type's influence on threat is tactic ballistic missile > air-to-ground missile > cruise missile > stealth aircraft > fighter > attack helicopter > early warning aircraft > reconnaissance aircraft > jammer aircraft. Their qualitative threaten can be described by quantitative triangular fuzzy numbers, as follows: (0.8, 0.9, 1.0), (0.7, 0.8, 0.9), (0.6, 0.7, 0.8), (0.5, 0.6, 0.7), (0.4, 0.5, 0.6), (0.3, 0.4, 0.5), (0.2, 0.3, 0.4), (0.1, 0.2, 0.3), (0.0, 0.1, 0.2).
5. The stronger the target interference ability is, the greater the target's threat is to the ground-to-air missile system. The interference ability of target is divided into very weak, weak, moderate, strong, and very strong, which can be represented, respectively, as follows: (0.0, 0.1, 0.2), (0.2, 0.3, 0.4), (0.4, 0.5, 0.6), (0.6, 0.7, 0.8), (0.8, 0.9, 1.0).

116.2.2 Membership Function of Threat Degrees

In 1965, the notion of fuzzy sets was first introduced by Zadeh [5], providing a natural way to deal with problems in which the source of imprecision is the absence of sharply defined criteria of class membership. A brief introduction of Fuzzy sets is given as follows.

Definition 3.1 A fuzzy set A is defined on a universe X may be given as:

$$A = \{\langle x, \mu_A(x)\rangle | x \in X\}$$

where $\mu_A : X \to [0, 1]$ is the membership function A. The membership value $\mu_A(x)$ describes the degree of belongingness of $x \in X$ in A.

Firstly the threat degrees of target are divided into very low, low, medium, high, and very high. Then let X be the universe of discourse, which contains five linguistic variables describing the degree of threat, $X = \{$very low, low, medium, high, very high$\}$, assuming that only two adjacent linguistic variables have the overlap of meanings. And let A be a fuzzy set of the universe of discourse X subjectively defined as follow:

where f_{VeryLow}, $\mathrm{f}_{\mathrm{Low}}$, f_{Medium}, f_{High} and f_{VeryHigh} are the membership functions of the fuzzy sets, which are shown in Fig. 116.1.

$$\begin{array}{rcll}
f_{\text{VeryLow}}(x) & = & -4x+1 & 0 \le x \le 0.25 \\
f_{\text{Low}}(x) & = & \begin{cases} 4x & 0 \le x \le 0.25 \\ -4x+2 & 0.25 \le x \le 0.5 \end{cases} & \\
f_{\text{Medium}}(x) & = & \begin{cases} 4x-1 & 0.25 \le x \le 0.5 \\ -4x+3 & 0.5 \le x \le 0.75 \end{cases} & \\
f_{\text{High}}(x) & = & \begin{cases} 4x-2 & 0.5 \le x \le 0.75 \\ -4x+4 & 0.75 \le x \le 1 \end{cases} & \\
f_{\text{VeryHigh}}(x) & = & 4x-3 & 0.75 \le x \le 1
\end{array}$$

116.2.3 A Method for Determining the Degree of Similarity Between Two Generalized Fuzzy Numbers [6]

Let $\widetilde{A}$ and $\widetilde{B}$ be two generalized trapezoidal fuzzy numbers, where $\widetilde{A} = (a_1, a_2, a_3, a_4; w_{\widetilde{A}})$ and $\widetilde{B} = (b_1, b_2, b_3, b_4; w_{\widetilde{B}})$, $0 \le a_1 \le a_2 \le a_3 \le a_4 \le 1$ and $0 \le b_1 \le b_2 \le b_3 \le b_4 \le 1$.

The measure of similarity is defined as follows:

$$S(\widetilde{A}, \widetilde{B}) = \left[1 - \frac{\sum_{i=1}^{4} |a_i - b_i|}{4}\right] \times \frac{\min(P(\widetilde{A}), P(\widetilde{B}))}{\max(P(\widetilde{A}), P(\widetilde{B}))} \times \frac{\min(A(\widetilde{A}), A(\widetilde{B})) + \min(w_{\widetilde{A}}, w_{\widetilde{B}})}{\max(A(\widetilde{A}), A(\widetilde{B})) + \max(w_{\widetilde{A}}, w_{\widetilde{B}})} \tag{116.1}$$

$P(\widetilde{A})$, and $P(\widetilde{B})$ are the perimeters of the two generalized trapezoidal fuzzy numbers which are calculated as follows:

$$P(\widetilde{A}) = \sqrt{(a_1 - a_2)^2 + w_{\widetilde{A}}^2} + \sqrt{(a_3 - a_4)^2 + w_{\widetilde{A}}^2} + (a_3 - a_2) + (a_4 - a_1) \tag{116.2}$$

$$P(\widetilde{B}) = \sqrt{(b_1 - b_2)^2 + w_{\widetilde{B}}^2} + \sqrt{(b_3 - b_4)^2 + w_{\widetilde{B}}^2} + (b_3 - b_2) + (b_4 - b_1) \tag{116.3}$$

On the other hand we have $A(\widetilde{A})$ and $B(\widetilde{B})$ which are the areas of the two fuzzy numbers, and they are calculated as follows:

$$A(\widetilde{A}) = 0.5 w_{\widetilde{A}}(a_3 - a_2 + a_4 - a_1) \tag{116.4}$$

$$B(\widetilde{A}) = 0.5 w_{\widetilde{B}}(b_3 - b_2 + b_4 - b_1) \tag{116.5}$$

116.2.4 Determining BPA

Consider the frame of discernment with five hypotheses $\Theta = \{$Very Low (VL), Low (L), Medium (M), High (H), Very High (VH)$\}$, which describe the threat degree (see Fig. 116.1). Based on the value of the five factors, the evidences are defined in D−S evidence theory.

For the first three factors, including target distance, target altitude, and target speed, their values are accurate and different types, so we normalize their value as follows:

$$r_{ij} = (x_{ij} - M_i^-)/(M_i^+ - M_i^-), \quad i = 1, 2, \ldots, 5; \quad j = 1, 2, \ldots, n \tag{116.6}$$

x_{ij} is the value of factor i of target j.

Target distance and Target altitude are inversely proportional to the targets' threat, so M_i^+ and M_i^- are defined as follows:

$$M_i^+ = \min_j\{x_{ij}\}, \quad M_i^- = \max_j\{x_{ij}\}$$

Target speed is directly proportional to the targets' threat, so M_i^+ and M_i^- are defined as follows:

$$M_i^+ = \max_j\{x_{ij}\},\ M_i^- = \min_j\{x_{ij}\}$$

A new method to obtain BPA is proposed based on the similarity measure [6] between generalized fuzzy numbers.

We randomly choose a normalized datum of target distance, for example $r_{ij} = 0.6$, which corresponds to a generalized trapezoidal fuzzy number (0.6, 0.6, 0.6, 0.6; 1.0). In Fig. 116.2, the relation between (0.6, 0.6, 0.6, 0.6; 1.0) and the membership functions of threat degrees are distinctly shown. The algorithm of our new method can be listed step by step as follows.

Step 1: we apply Eqs. 116.1–116.5 to obtain the similarity between the generalized trapezoidal fuzzy number (0.6, 0.6, 0.6, 0.6; 1.0) and the linguistic terms of threat degrees A, where $A \in \Theta = \{\text{VL, L, M, H, VH}\}$. The similarities are shown as follows:

$$\begin{aligned} S(\text{VL}) = 0.1803;\ S(\text{L}) = 0.2030;\ S(\text{M}) = 0.2577;\ \text{S(H)} \\ = 0.2498;\ S(\text{VH}) = 0.2582 \end{aligned}$$

We can easily conclude that if the similarity is more, the probability, which threat degree is A, is higher.

Then, we define

$$S(\Theta) = 1 - \max\{S(\text{VL}), S(\text{L}), S(\text{M}), S(\text{H}), S(\text{VH})\} = 1\ -\ 0.2582 = 0.7418$$

Step 2: Normalize the similarity measure to obtain the BPA function. The BPA can be obtained as follows:

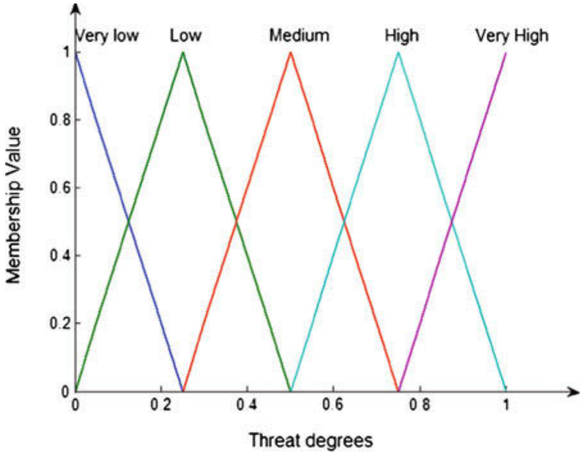


Fig. 116.1 Membership function of threat degrees

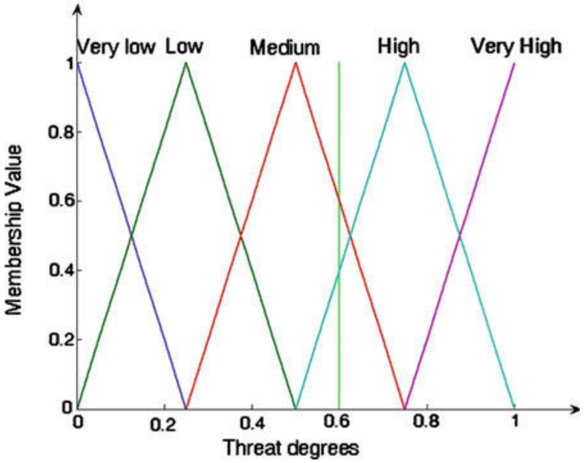


Fig. 116.2 The instance and the membership function of threat degrees

$$m(\mathrm{VL}) = 0.0953 \quad m(\mathrm{L}) = 0.1074 \quad m(\mathrm{M}) = 0.1363$$

$$m(\mathrm{H}) = 0.1321 \quad m(\mathrm{VH}) = 0.1366 \quad m(\Theta) = 0.3923$$

For the last two factors, namely target type and target interference ability, their values are fuzzy, which are represented by triangular fuzzy numbers.

We randomly choose a kind of target type, for example tactic ballistic missile, represented as (0.8, 0.9, 1.0). In Fig. 116.3, the relation between the instance (0.8, 0.9, 1.0) and the membership functions of threat degrees are distinctly shown.

In the same way, we apply Eqs. 116.1–116.5 to calculate the similarity between the instance (0.8, 0.9, 1.0), which corresponds to a generalized trapezoidal fuzzy

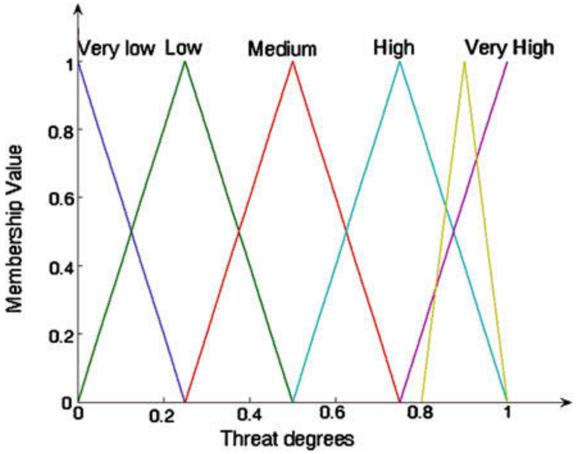


Fig. 116.3 The instance and the membership function of threat degrees

number (0.8, 0.9, 0.9, 1.0; 1.0) and the linguistic terms of threat degrees A, where $A \in \Theta = \{\text{VL, L, M, H, VH}\}$. The similarities are shown as follows:

$$\begin{aligned} S(\text{VL}) &= 0.1540;\ S(\text{L}) = 0.2657;\ S(\text{M}) = 0.4555;\ S(\text{H}) \\ &= 0.6453;\ S(\text{VH}) = 0.8882 \end{aligned}$$

Then, $S(\Theta)$ is defined as follows:

$$S(\Theta) = 1 - \max\{S(\text{VL}), S(\text{L}), S(\text{M}), S(\text{H}), S(\text{VH})\} = 1 - 0.8882 = 0.1118$$

Similarly, the obtained similarity S is normalized and the BPA can be obtained as follows:

$$m(\text{VL}) = 0.0611 \quad m(\text{L}) = 0.1054 \quad m(\text{M}) = 0.1807$$

$$m(\text{H}) = 0.2560 \quad m(\text{VH}) = 0.3524 \quad m(\Theta) = 0.0444$$

116.2.5 Combining the Evidences

Once the BPA is accomplished, we can use Dempster's rule of combination to combine all the evidences. However, counter-intuitive results may be obtained by classical Dempster combination rule when collected evidences highly conflict with each other [7]. Many methods have been proposed to solve this problem.

In this paper, the method suggested by Murphy [8] to achieve certainty with averages is adopted, which provides more accurate record of contributing beliefs. However, averaging lacks convergence. In order to handle that, averaging is integrated into the combining rule. If there are n pieces of evidences, the masses assigned to the same set should be averaged, which avoids overdependence on a

Table 116.1 The parameter of threat objectives

Target	Factor				
	Distance (km)	Altitude (m)	Speed (m/s)	Types	Interference ability
A_1	80	2,500	480	Tactic ballistic missile	Weak
A_2	150	1,500	150	Cruise missile	Moderate
A_3	100	1,200	230	Fighter	Very strong
A_4	120	1,000	180	Attack helicopter	Strong
A_5	180	2,000	520	Early warning aircraft	Very weak

Table 116.2 The BPA of target A_1

	m(VL)	m(L)	m(M)	m(H)	m(VH)	$m(\Theta)$
m_1	0.0163	0.0523	0.1046	0.1569	0.2448	0.4251
m_2	0.2448	0.1569	0.1046	0.0523	0.0163	0.4251
m_3	0.0411	0.0691	0.1173	0.1551	0.2126	0.4049
m_4	0.0611	0.1054	0.1807	0.2560	0.3524	0.0444
m_5	0.2219	0.2374	0.2110	0.1451	0.1055	0.0791

single piece of conflicting evidence. Then, use Dempster's rule of combination to combine the evidence with itself $n - 1$ times.

116.3 Numerical Example

Assume five targets to carry out an air attack to the protected important position. The type, distance, altitude, speed, and interference ability of each target are measured and shown in Table 116.1.

Based on the proposed method above, the BPA of A_1 can be calculated, and shown in Table 116.2.

Firstly, we use Murphy's rule to derive the averaged evidence as follows:

$$m(\mathrm{VL}) = 0.1170 \quad m(\mathrm{L}) = 0.1242 \quad m(\mathrm{M}) = 0.1436$$

$$m(\mathrm{H}) = 0.1531 \quad m(\mathrm{VH}) = 0.1863 \quad m(\Theta) = 0.2757$$

Secondly, the combination rule of DS theory shall be applied to combine the averaged evidence with itself four times. The result is shown below:

$$m(\mathrm{VL}) = 0.1255 \quad m(\mathrm{L}) = 0.1399 \quad m(\mathrm{M}) = 0.1842$$

$$m(\mathrm{H}) = 0.2091 \quad m(\mathrm{VH}) = 0.3153 \quad m(\Theta) = 0.0258$$

In the same way, the BPAs of five targets can be obtained and combined by Murphy's rule and Dempster's rule. The final combination result is shown in Table 116.3.

Table 116.3 The combination result

Target	m(VL)	m(L)	m(M)	m(H)	m(VH)	$m(\Theta)$
A_1	0.1255	0.1399	0.1842	0.2091	0.3153	0.0258
A_2	0.2191	0.1895	0.2304	0.1762	0.1580	0.0268
A_3	0.0942	0.1191	0.2018	0.2305	0.3323	0.0221
A_4	0.1562	0.1623	0.2175	0.2013	0.2339	0.0289
A_5	0.3565	0.2361	0.1784	0.1076	0.0970	0.0245

With all the evidences combined together, the decision maker can evaluate the threat degree of the five targets. Consequently, the final threat degree sequencing of the five targets is $A_3 > A_1 > A_4 > A_2 > A_5$. It can be seen that our proposed method has good results in multi-target threat evaluating and sequencing problem.

116.4 Conclusion

In the modern air attack environment, multi-target threat evaluating and sequencing is one of the key taches within the processes of command and control in aerial defense battle. In this paper, a new model to assess target threat degree is proposed. Based on fuzzy sets and D−S evidence theory, the proposed method in our model is flexible and practical to cope with qualitative and quantitative factors of the targets. The model has proven its potential in helping with evaluating the threat degree. By example verification the method can give more authentic threat evaluation value and can conduct threat line up. It is anticipated that the present study can provide some reference to the establishment of ground-to-air multi-target defense model and the research of command decision-making.

Acknowledgments This work was supported in part by a grant from National Natural Science Foundation of China, Grant No. 60904099 and Aviation Science Foundation, Grant No. 20095153022.

References

1. Dempster A (1967) Upper and lower probabilities induced by multivalued mapping. Ann Math Stat 38:325–339
2. Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
3. Fan CY, Han XM, Wang XF (2003) Target threat evaluating and sequencing method based on the maximum degree of membership. Syst Eng Electron 25(1):47–48
4. Zhang RC, Liu ZL, Wang S (2005) Evaluating and sequencing of the air target threat using fuzzy MADM theory. Mod Def Technol 33(1):15–18
5. Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353
6. Hejazi SR, Doostparast A, Hosseini SM (2011) An improved fuzzy risk analysis based on a new similarity measures of generalized fuzzy numbers. Expert Syst Appl 38(8):9179–9185

7. Zadeh L (1986) A simple view of the Dempster–Shafer theory of evidence and its implication for the rule of combination. J AI Mag 7(1):85–90
8. Murphy CK (2000) Combining belief functions when evidence conflicts. Decis Support Syst 29(1):1–9

Chapter 117
Target Positioning of Picking Robot Fusing Laser Ranging and Vision

Jian Song

Abstract A target positioning method merging laser ranging and vision for eggplant picking robot is proposed for the sake of overcoming such disadvantages as high complexity and computation of algorithm in the binocular vision positioning method. The laser ranging sensor is used to measure the distance of the fruit target. Information of the two-dimensional images of the target is acquired by the color video camera. Three-layer BP neural network is established with the image centroid coordinates and the laser ranging sensor measuring distance as the input quantity and with the picking space of points coordinates as the output quantity. The modified BP algorithm is adopted to train the weight of the neural network. Valid network weights are obtained after 126 times of cycling. It is determined through experiments that in the measuring distance range from 250 to 650 mm, the mean error value of the root-mean-square (RMS) of the eggplant space positioning value is 11.2 mm and the average time used is 0.34 s. The positioning method merging laser ranging and vision is of good intelligence and wide adaptability, able to meet the requirements for target positioning of the picking robot.

Keywords Laser ranging · Neural network · Target positioning

117.1 Introduction

Harvesting or picking is the most effort-requiring and time-consuming link in vegetable production operation, which, according to statistics, approximately accounts for from 50 to 70% of all the working out [1]. With the aging of the

J. Song (✉)
College of Machinery, Weifang University, Weifang 261061, China
e-mail: sjian11@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_117,

population and the decrease of the farming labor force, it is of great significance to develop mechanized harvesting technology and to research on vegetable picking robots [2]. At present, color camera is the frequently- used target detection tool for picking robot. A color camera can only acquire the two-dimensional positioning information and the maturity information of the target [3].

In recent years, some scholars abroad have begun to apply the multi-sensor fusion technology to the fruit target positioning system, in which information is collected simultaneously by multi-sensors of diverse types or in different positions and is automatically and comprehensively analyzed through computer technique in order to obtain the target position information [4]. Mitsuji Monta of Japan developed one vision system for identifying tomato by fusing information from color video camera, laser ranging device and infrared sensors, which has gained the test result with an identification rate of 65–70% [5]. Naoshi Kondo of Japan installed three color video cameras and four illumination devices with polarizing filter on the vision system to detect the information about strawberry stems [6]. Murali of University of Florida, America, developed an orange identification vision system composed of a color video camera and four ultrasonic sensors [7].

In this paper, on the basis of the foreign scholars' experience, with the eggplant in the growing environment as the object of study, BP neural network method is adopted to study the target location of the picking robot by fusing laser and visual information, with a view to enhancing the intelligence and real timing of the vision system.

117.2 System Structure of the Picking Robot

According to the requirements of the picking task and the modularity design idea, an open-type system structure for the picking robot is adopted (as shown in Fig. 117.1) in order that the picking robot has good expansibility, commonality and capability flexible operation. The system includes PC machine, PHI-LD90-50 type high-performance laser ranging sensor, 2-DOF automatic PAN/TILT, DMC multi-axis motion controller, Yaskawa AC servo drive system, ZINO DH-CG320 machine vision system, National WV-CP470 color video camera, and 4-DOF articulated robot body.

117.3 Laser Ranging System

117.3.1 Laser Ranging Sensor

Laser ranging can be divided into two types: continuous wave (CW) phase position ranging and pulse ranging. Laser pulse ranging method can achieve adequate measuring precision and measuring speed for the measurement of long distance target [8]. PHI-LD90-50 type high-performance Laser ranging sensor is selected

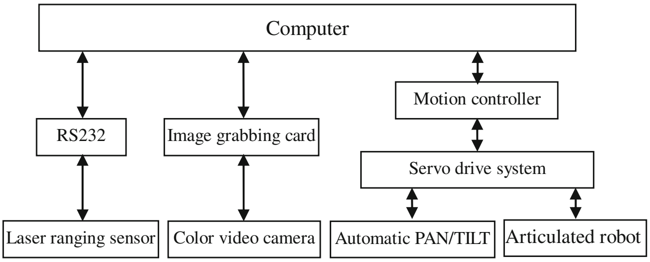


Fig. 117.1 Block diagram of picking robot

and used, in consideration of the fact that picking robot needs to work in various kinds of environment conditions like moist and overcast and rainy ones, with a comparison of sensors of different kinds of performances.

This sensor, produced by Dalian Peng Hui Xin Da Technology Co., Ltd of US PHI Industry Group, employs the pulse ranging method. Because of the fusion of "high penetration technology", it can work under extremely hostile environment and measure quickly and accurately the target distance. The measuring result can be transmitted to the circumjacent devices with RS232/422 protocol interface via the RS232/422 interface of the sensor for the purpose of such uses as detection, measurement and control. Meanwhile, the control of the laser ranging sensor may also be fulfilled through the computer or other connected devices.

117.3.2 Automatic Pan/Tilt

The function of the automatic Pan/Tilt is to expand the ranging sensor's vision. To this end, 2-DOF automatic Pan/Tilt with 0–360° in the horizontal direction and −90 to +90° in the vertical direction is chosen. Stepping motors are adopted for both the horizontal electromotor and the vertical electromotor, controlled by DMC multi-axis motion controller, by which the laser ranging sensor fixed on the Pan/Tilt is driven. It conducts real-time measurement on the ambient environment with the picking robot as the reference point.

117.4 Visual Identification of the Eggplant Fruit

117.4.1 Image Segmentation

Due to the complexity of the eggplant growing environment, it is quite difficult to acquire preferable segmentation effect by the fixed-threshold-based segmentation method, while it is hard for such automatic threshold segmentation methods as

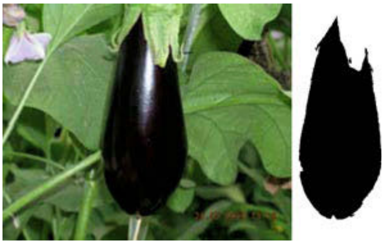


Fig. 117.2 Result of image segmentation

OSTU, iteration method, Minimum error probability method to adapt to the picking robot's requirements with a relatively slow segmentation speed and an unsatisfactory timeliness. On account of the above considerations, the threshold segmentation algorithm based on brightness is put forward which can realize the ideal segmentation of the eggplant fruit target. In line with the color feature analysis of the eggplant fruits and their surroundings, G-B grayscale images are the most favorable for segmentation of the fruit target. In view of the two points above, brightness-based threshold segmentation algorithm is adopted to segment the G-B grayscale images.

The main principle is:

$$T_h = G_{av} + (G_{\max} - G_{av}) * f \tag{117.1}$$

where: T_h is the segmentation threshold value; G_{av} is the mean gray value; $G_{\max}$ is the maximum gray value; f is the weight factor

Through segmentation experiments on 30 eggplant images, f values range from -0.5 to 0.5. In general, when $f = 0.1$, it can get preferable segmentation effect. The image segmentation effect is shown in Fig. 117.2.

117.4.2 Feature Extraction of the Fruit Target

Feature extraction refers to the features that some kind of object possesses in the image. For fruit and vegetable picking operation, the feature parameters of the picking target such as center of gravity, area, circumscribed rectangle and cut-out point are finally provided by the robot vision system. Binary image is acquired through image segmentation, and then the contour is marked in the dimensional array after edge extraction and contour tracing for the image. In this way, the fruit target features can be extracted conveniently. Figure 117.3 is the rendering of feature extraction.

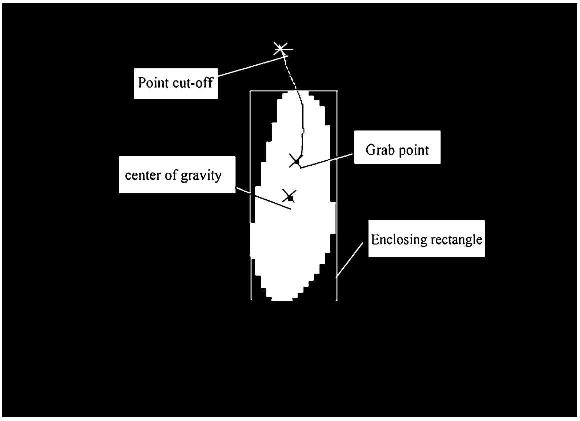


Fig. 117.3 Result of feature extraction

117.5 Target Positioning Algorithm Based on BP Neural Network

117.5.1 Selections of the Input Signal and the Output Signal

The x-coordinate and the y-coordinate of the center of gravity of the image acquired by video camera and the measurement distance of the laser ranging sensor (X_l, Y_l, S) are conducted as the input quantity, and the special position coordinates (X, Y, Z) of the grasped position as the output quantity. By using the nonlinear mapping function of the neural network, 3-layer BP neural network is adopted, and then the three-dimensional coordinates of the position can be obtained to realize the binocular vision positioning of the picking robot.

117.5.2 Establishment of BP Network

In the design of BP neural network, the number of hidden nodes has significant influence on the network performance. Furthermore, there has not yet been definite theoretical direction for choosing hidden nodes number. If the hidden nodes are too little, the training frequency will be increased and the information retrieval accuracy will be reduced. On the other side, if they are too many, the volume of network metric training will be increased and the network reliability will be lowered so that its ability to process the non-sample data will be affected. The following empirical formula is adopted in this project:

$$N_2 = \sqrt{N_1 + N_3} + a \tag{117.2}$$

where :N_1 and N_3 are respectively input and output nodes numbers, and α is a constant between 1 and 10.

After many experiments in the research, it is assumed that $N_2 = 6$, and BP neural network with a 3-5-3 structure is composed to realize the eggplant fruit positioning.

The neuron input and output function is a dual bend function:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{117.3}$$

The square error sum of the neuron output of the output layer and the teacher signal is defined as:

$$E = \frac{1}{2}\sum_{k=1}^{k}(d_k - o_k)^2 \tag{117.4}$$

where:o_k indicates the output of neuron k of the output layer while d_k indicates the teacher signal to the output layer.

117.5.3 Speediness of Learning Algorithm

Momentum term is introduced in order to realize the speediness of learning algorithm. "Momentum" is borrowed from the inertia concept in mechanics, which considers not only the current value but the past influence. When amending the network weights at the moment $t = n$, it adds the amendment quantity of the previous moment$(t = n - 1)$, i.e.,

$$\Delta w_{ji}^{n} = -\eta \frac{\partial E^{n-1}}{\partial w_{ji}^{n-1}} + \alpha \Delta w_{ji}^{n-1} \tag{117.5}$$

where: η is the learning rate; E^{n-1} is the square error sum of the output layer till the moment $n - 1$; α is the momentum term coefficient, $0 < \alpha < 1$.

The amending direction of the current value is different from the direction of the previous weight, i.e., when the first and the second terms on the right of formula (117.5) are opposite in sign, the absolute value of the summation amendment quantity becomes less to prevent overregulation; on the contrary, when the amendment quantity is in the same direction, the amendment quantity becomes a bigger value after addition so as to realize the speediness of the learning algorithm. It assumes $\eta = 0.06$, $\alpha = 0.8$ in this paper.

117.6 Experimental Results and Analysis

117.6.1 Experimental Methods

With the laser ranging sensor and the CCD video camera fixed on the manipulator base, the manipulator grips four eggplants respectively and motions to 25 desired positions, thus 4 × 25 = 100 groups of data can be acquired.

Using the recognition algorithm for eggplant fruit mentioned above, it can get the centroid coordinates (X_l, Y_l) of the eggplant image collected by the camera and the laser ranging sensor measuring distance S, which serve as the input quantity of BP neural network, and the actual coordinates of the robot serve as the output quantity, so that the fast learning algorithm is applied for training. After 126 times of iteration training, the overall error reduced to 0.001, and convergent valid network weight is obtained.

117.6.2 Experimental Result

Experiments are conducted on the ten eggplants within a range from 250 to 650 mm in different positions by using the trained network. The RMS errors e, $e_{\max}$, $e_{\min}$, R_e, and T_e of the measurement coordinate values obtained through the neural network and the recorded actual coordinate values of the manipulator motions serve as the indexes for evaluating the measurement precision:

$$e = \sqrt{\frac{(X_c - X_s)^2 + (Y_c - Y_s)^2 + (Z_c - Z_s)^2}{3}} \tag{117.6}$$

In the formula, X_c, Y_c and Z_c are respectively the X-coordinate, Y-coordinate and Z-coordinate calculated by using neural network measurement; X_s, Y_s and $Z_s$are respectively the actual X-coordinate, Y-coordinate and Z-coordinate of the spatial point obtained by actual measurement.

$$e = \sqrt{\frac{(X_c - X_s)^2 + (Y_c - Y_s)^2 + (Z_c - Z_s)^2}{3}} \tag{117.7}$$

where: X_c, Y_c and Z_c are respectively the X-coordinate, Y-coordinate and Z-coordinate calculated by using neural network measurement; X_s, Y_s and Z_s are respectively the actual X-coordinate, Y-coordinate and Z-coordinate of the spatial point obtained by actual measurement.

$$e_{\max} = \max_{i=1,2,\cdots,I} e(i) \tag{117.8}$$

Table 117.1 Test result

Item name	$R_e(mm)$	$e_{\max}(mm)$	$e_{\min}(mm)$	$T_e(s)$
Calculated value	11.2	24.6	2.9	0.34

$$e_{\min} = \min_{i=1,2,\cdots,I} e(i) \tag{117.9}$$

$$\mathrm{Re} = \sqrt{\frac{\sum_{i=1}^{I} e^2(i)}{I}} \tag{117.10}$$

$$Te = \frac{\sum_{i=1}^{I} T_i}{I} \tag{117.11}$$

where: $e(i)$ is the error of No.i point, I is the number of the test points. Test results are shown in Table 117.1.

117.6.3 Experimental Analysis

In the lab conditions, the mean error value of the root-mean-square (RMS) of the eggplant space positioning value is 11.2 mm and the average time used is 0.34 s. The main reasons leading to the errors are: (1) The measurement errors of the laser ranging sensor; (2) The errors resulted from large distortion of the video camera; (3) The errors of centroid coordinates caused by the fact that it is impossible to eliminate radically though the segmented image is processed by filtering, consequently the accuracy of the teacher signal is affected. (4) The reliability of the neural network weights are restricted by the quantity and the diversity of the teacher signal samples.

The positioning method merging laser ranging and visual information belongs to adaptive algorithm, which possesses strong approximation ability and generalization capability, good intelligence and wide adaptability, being able to basically meet the requirements for target positioning of the picking robot.

117.7 Conclusion

A target positioning method fusing laser ranging and visual information for eggplant picking robot is proposed for the sake of reducing the complexity and computation of the picking target positioning algorithm. Because BP neural network possesses strong approximation ability and generalization capability, it guarantees the precision of the picking robot visual positioning and shortens the measurement time. It is determined through experiments that the mean error value of the RMS of the eggplant space positioning value is 11.2 mm and the average time used is 0.34 s.

Compared with the traditional binocular vision method, the intelligence and real timing are improved. Although this method can meet the requirements for target positioning of the picking robot, there is still some room for further improvement in its measurement precision and real timing. In the future research, it calls for optimization of the systematic algorithm design, accurate standardization of the camera parameters and increase of the quality of the teacher samples, thereby to develop the fruit target location algorithm that can be applied to the actual task of the picking robot.

Acknowledgments This work is supported by Shandong Provincial Natural Science Foundation, China (No.Y2008G32).

References

1. Song J, Zhang TZ, Xu LM (2006) Research actuality and prospect of picking robot for fruits and vegetables. Trans Chin Soc Agric Mach 37(5):158–162
2. Tang XY, Zhang TZ (2005) Robotics for fruit and vegetable harvesting: a review. Robot 27:90–96
3. Kondo N, Monta M, Fujiura T (1996) Fruit harvesting robot in Japan. Adv Space Res 18(2):181–184
4. Hayashi S, Ganno K, Ishii Y (2002) Robotic harvesting system for eggplants. JARQ 36(6):163–168
5. Monta M, Namba K, Kondo N (2004) Three dimensional sensing system using laser scanner. ASAE, ASAE/CSAE meeting, Beijing, IEEE, pp 437–443
6. Kondo N, Monta M, Fujiura T (1996) Fruit harvesting robot in Japan. Adv Space Res 18(1/2): 181–184
7. Regunathan M, Lee W (2005) Citrus fruit identification and size determination using machine vision and ultrasonic sensors. ASAE, ASAE meeting, Hokkaido, IEEE, pp 639–644
8. Zhang JW, Zhang SM (2008) Design of ship anti-collision and position system based on laser rangefinder sensor. Appl Laser 28(6):497–501

Chapter 118
Research and Design of the Reconfigurable ForCES TML

Ligang Dong, Ling Cai, Lianfang Zhu and Weiming Wang

Abstract The researches of routers include both performance improvement and function flexibility. Forwarding and Control Element Separation (ForCES) is a kind of network device architecture that will greatly improve the function flexibility. Transport Mapping Layer (TML), as an important part of the ForCES router, is responsible for transmitting the ForCES messages. The paper designs and implements a reconfigurable TML architecture, which has been supported by the experiments. The reconfigurable TML can dynamically apply different TML implementation and different transport protocols to transmit ForCES messages properly on the different media. This feature brings much flexibility and reconfigurability to the ForCES router.

Keywords ForCES · TML · Open interfaces · Reconfigurability

L. Dong (✉) · L. Cai · L. Zhu · W. Wang
Institute of Network and Communication Engineering, Zhejiang Gongshang University, No.18, Xuezheng Str., Xiasha University Town,
Hangzhou 310018, People's Republic of China
e-mail: donglg@mail.zjgsu.edu.cn

L. Cai
e-mail: cailing@pop.zjgsu.edu.cn

L. Zhu
e-mail: zhulianfang@pop.zjgsu.edu.cn

W. Wang
e-mail: wmwang@mail.zjgsu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_118,

118.1 Introduction

Forwarding and Control Element Separation (ForCES) [1] is based on open programable idea and is considered to be used widely. According to the ForCES requirements (RFC 3654) [2] and ForCES framework (RFC 3746) [3], a ForCES network element (NE) structure is depicted in Fig. 118.1.

In Fig. 118.1, control element (CE) is responsible for signaling, controlling protocol processing and network management . Forwarding element (FE) uses the underlying hardware to provide packet processing and forwarding. Fp is the interface between CE and FE. The ForCES protocol runs on Fp reference point, which is single-hop or multi-hop network.

Transport mapping layer (TML) transmits ForCES protocol messages, and provides a transparent and secure transmission channel for ForCES protocol layer (PL). The TML addresses how the protocol messages are mapped to different transport media (such as TCP/IP, ATM and ethernet). The research issues of TML are about how to achieve transport level reliability, congestion control, multicast, and ordering.

Reconfigurability means the hardware and software can be dynamically reconfigured according to variable data flow or control flow in a system. In this paper, reconfigurable TML is an entity which can support the configuration of various transport protocols according to different transport media, network types and communication requirements to meet its QoS.

Till now, the researches of TMLs are summarized as follows:

TCP/DCCP-based TML

Kohler et al. proposed that TCP and DCCP (Datagram congestion control protocol) [4] can be chosen as TML transport protocols. TCP runs on the control channel, and data messages can be transmitted by DCCP, in which the security mechanism can choose TLS or IPSec.

SCTP-based TML

Salim and Ogawa proposed SCTP-based TML in 2007 [5], particularly described how the SCTP-TML satisfied the ForCES requirements, and listed the TML structure and corresponding interface descriptions.

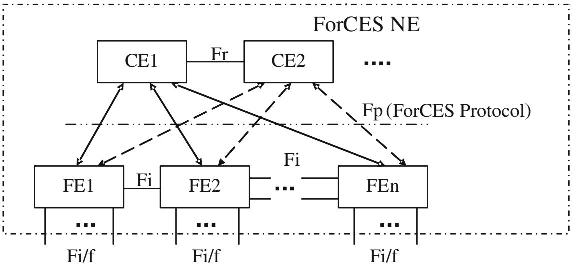
TIPC-based TML

Maloy et al. proposed TIPC (transparent inter-process communication) as TML transport protocol [6]. They described the premises and advantages of TIPC-based TML, and studied how the TIPC TML satisfied the ForCES requirements. Furthermore, they proposed the ForCES application services and LFBs about how to map to TIPC functional addressing. Also, they pointed out that TIPC could not provide security mechanisms and could apply to secure networks.

TCP/UDP-based TML

Hidell et al. proposed that TML control channel can choose TCP as transport protocol, while redirect messages can be transmitted by UDP [7]. This TML can choose window and ACK to ensure UDP reliability. We began to research ForCES

Fig. 118.1 ForCES architecture



since 2002, and were the first ForCES research team in China. And we choose TCP to transmit control messages and UDP to transmit redirect messages [8, 9]. Besides, we test and analyze the performance of the TML [10]. The results show that when we use a proper scheduling algorithm, the TML can avoid DoS (denial of service) attack from redirect messages.

The above researches analyze the application of different transport protocols in the ForCES router. However, no one has designed a structure which is flexible enough to support different transport protocol and different TML implementation reconfiguration for various transport media and environment. The contribution of this paper is the design and implementation of reconfigurable TML.

118.2 Design of Reconfigurable TML

Design of reconfigurable TML means the design of two kinds of interfaces: TML service interface and TML channel interface.

The TML service interface is the interface between PL and TML. By standardizing this interface, the same PL can be used for different TML implementation. We dynamically change TML without updating the PL. It is called TML overall reconfigurability.

The TML channel interface is between the common used part (called TML core) and different transport protocol parts. By standardizing this interface, the same TML core can be used for different transport protocols. We can dynamically change the transport protocol without updating the TML core. It is called TML channel reconfigurability.

The structure of TML is shown in Fig. 118.2. Figure 118.3 is a TML example when SCTP is used as the transport protocol. Its detail will be introduced later.

With this kind of modular TML structure, PL, TML core, TML transport protocol can be developed and deployed separately. It will greatly improve the flexibility and reconfigurability of ForCES-based network devices.

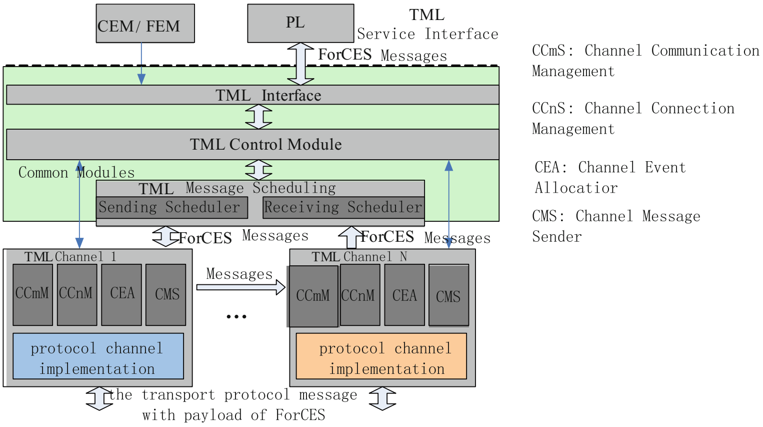

CEM/ FEM
PL
TML
Service Interface
ForCES Messages
TML Interface
TML Control Module
Common Modules
TML Message Scheduling
Sending Scheduler
Receiving Scheduler
ForCES Messages
ForCES Messages
TML Channel 1
CCmM
CCnM
CEA
CMS
protocol channel implementation
Messages
...
TML Channel N
CCmM
CCnM
CEA
CMS
protocol channel implementation
the transport protocol message with payload of ForCES
CCmS: Channel Communication Management
CCnS: Channel Connection Management
CEA: Channel Event Allocatior
CMS: Channel Message Sender

Fig. 118.2 Modular TML structure

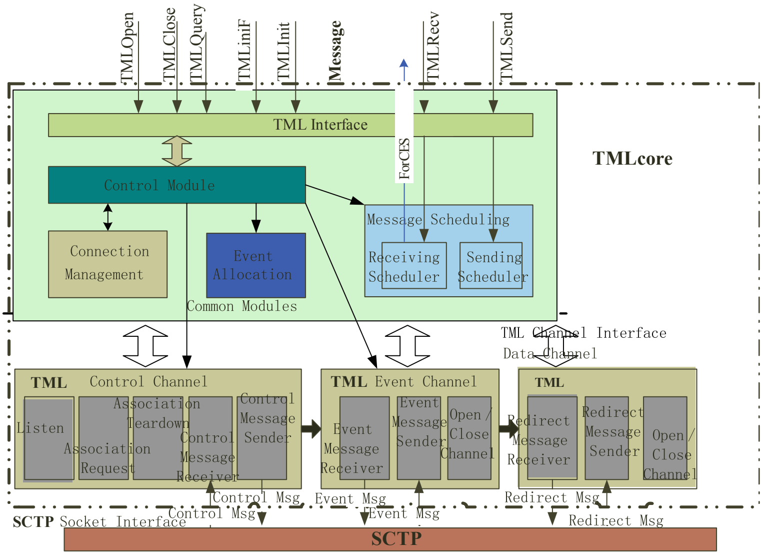
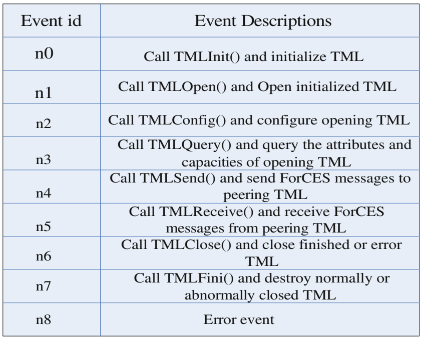


Fig. 118.3 SCTP-TML

118.2.1 TML Service Interface

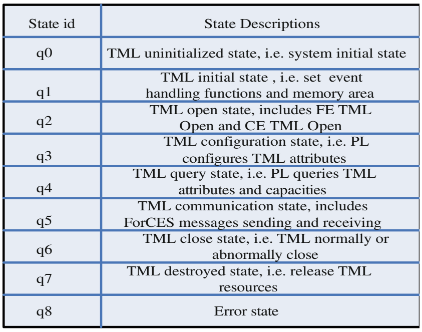
The ForCES protocol in the Fp reference point (see Fig. 118.1) is related to two layers: the PL and the TML. The PL and the TML are responsible for processing and transmitting the ForCES protocol messages, respectively.

Fig. 118.4 TML event descriptions[1]



Event id	Event Descriptions
n0	Call TMLInit() and initialize TML
n1	Call TMLOpen() and Open initialized TML
n2	Call TMLConfig() and configure opening TML
n3	Call TMLQuery() and query the attributes and capacities of opening TML
n4	Call TMLSend() and send ForCES messages to peering TML
n5	Call TMLReceive() and receive ForCES messages from peering TML
n6	Call TMLClose() and close finished or error TML
n7	Call TMLFini() and destroy normally or abnormally closed TML
n8	Error event

Fig. 118.5 TML state descriptions



State id	State Descriptions
q0	TML uninitialized state, i.e. system initial state
q1	TML initial state , i.e. set event handling functions and memory area
q2	TML open state, includes FE TML Open and CE TML Open
q3	TML configuration state, i.e. PL configures TML attributes
q4	TML query state, i.e. PL queries TML attributes and capacities
q5	TML communication state, includes ForCES messages sending and receiving
q6	TML close state, i.e. TML normally or abnormally close
q7	TML destroyed state, i.e. release TML resources
q8	Error state

TML need achieve mapping protocol messages to different transport media (such as IPand ATM). The TML have different implementations based on different underlying media and transmission capabilities. However, we would not expect that the PL changes along with different TMLs. In other words, one ForCES PL implementations must be portable across all kinds of TMLs.

To hide TML implementation details for PL, a set of TML service primitives have to be defined. We have submitted a draft of service primitives to IETF. Readers can refer to [11] for detail. When designing service primitives, we can view TML as a device, so that the PL can choose different TMLs, and dynamically load/unload TMLs. By borrowing idea from types of conventional network device operational primitives, we define six service primitives: TMLOpen, TMLClose, TMLConfig, TMLQuery, TMLSend and TMLReceive.

In order to describe better the state transitions of service primitives, TMLInit and TMLFini are also defined. Moreover, we view message sending and receiving states as message communication state. So TML event descriptions and state descriptions of service primitives are shown in Figs. 118.4 and 118.5. The state transition relationships in TML workflow are shown in Fig. 118.6.

However, since each TML is standardized, interoperability is guaranteed as long as both endpoints support the same TML.

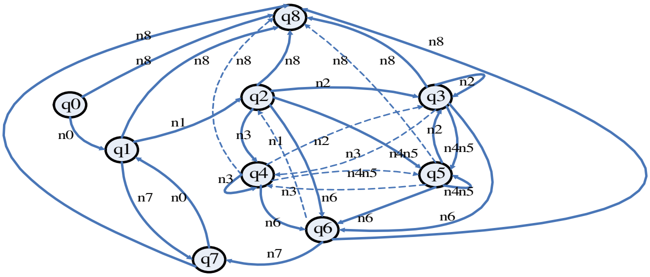


Fig. 118.6 State transition diagram in TML workflow

118.2.2 TML Channel Interface

As shown in Figs. 118.2 and 118.3, a TML consists of common modules, which are parts of the TML core, and TML channels.

There are the following common modules:

TML interface: provides a unified interface in accord with TML service primitives for PL.

Control module: provides interface implementations defined by TML Interface, including initialize, open, configure, query and destroy.

Message scheduling: contains two priority-based message schedulers, and each scheduler contains a thread. When TML initializes message channels, TML needs to register message queues and corresponding priorities to sending scheduler and receiving scheduler. When running, TML schedules messages run according to priority.

Event allocation: I/O event de-multiplexer based on select uses hash table to store registered event records, and its implementation is a single-thread. Moreover, it provides the interface to register and delete events for channel modules.

Connection management: stores, manages and maintains connection information and multicast list to provide services for other modules.

TML structure designed by this paper can support multiple transport protocols, such as TCP/UDP, SCTP and DCCP. The TML core need not change even if TML chooses different transport protocols.

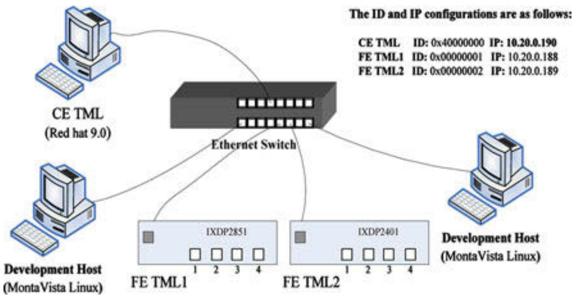
Channel modules are different for different transport protocols. Every channel provides the standard TML channel interface to TML core, so that the TML core can uniformly open and close channels, and send and receive event messages. In order to satisfy the TML requirements, ForCES messages are divided into control message, event message and redirect message. Consequently, a TML channel includes

Control channel, which is used for receiving and sending control messages.

Event channel, which is used for receiving and sending event messages.

Data channel, which is used for receiving and sending redirect messages containing data packets.

Fig. 118.7 TML functional test platform


The ID and IP configurations are as follows:
CE TML ID: 0x40000000 IP: 10.20.0.190
FE TML1 ID: 0x00000001 IP: 10.20.0.188
FE TML2 ID: 0x00000002 IP: 10.20.0.189
CE TML
(Red hat 9.0)
Ethernet Switch
Development Host
(MontaVista Linux)
IXDP2851
1 2 3 4
FE TML1
IXDP2401
1 2 3 4
FE TML2
Development Host
(MontaVista Linux)

Here we take SCTP-TML as an example. Figure 118.3 shows the structure and relationship among modules in SCTP-TML. SCTP-TML contains a TML core and a SCTP channel module.

118.3 Test and Analysis

The above descriptions introduce module design and interface design of reconfigurable TML. This section tests the TML functions and analyzes its performances based on above design.

118.3.1 Functional Test

TCP and UDP are the two most widely used transport protocols, so functional test and performance test consider TCP/UDP-TML control messages. The functional test platform is shown in Fig. 118.7.

The functional test platform consists of a PC with Red hat 8.0, Network Processor (IXDP2851, IXDP 2401), a development host and a ethernet switch. CE test programs run on PC, FE test programs run on NP and the development host installs Montavista Linux. Here we use ethereal capture packets, and the test results verify our design.

Furthermore, IETF ForCES working group carried out ForCES interoperability. We adopt reconfigurable TML to make the interoperability test with foreign research institutes. When testing, we can choose SCTP to be TML transport protocol by modifying the configuration file. The interoperability test contains seven scenarios: pre-association setup, including setting IDs and port numbers of CE and FE; TML priority channels connection; association setup-association complete; CE query; heartbeat monitoring; simple config command and association teardown. Further information is described in Ref. [12].

By the interoperability test with foreign research institutes, SCTP-TML achieves the successful and effective communication of various messages between

Fig. 118.8 TML performance test platform

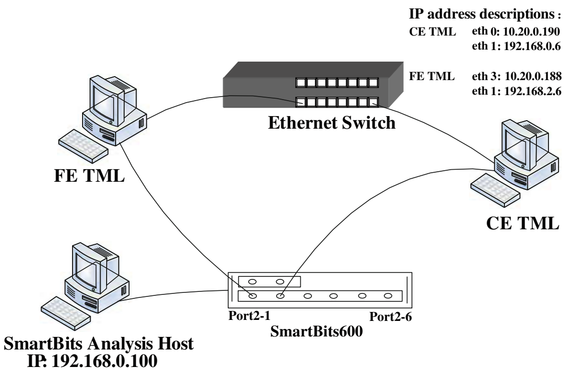
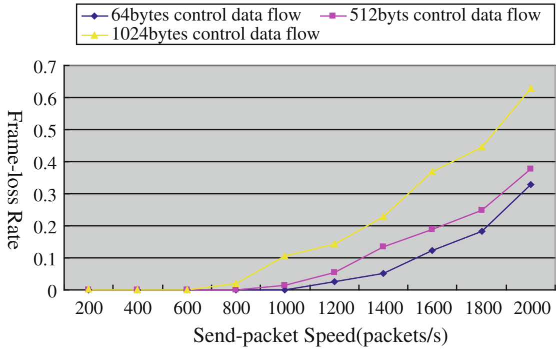


Fig. 118.9 Control message flow of TCP/UDP-TML



CE and FE. Combined with the above functional test of TCP/UDP-TML, the results fully show the correctness and stability of the reconfigurable TML.

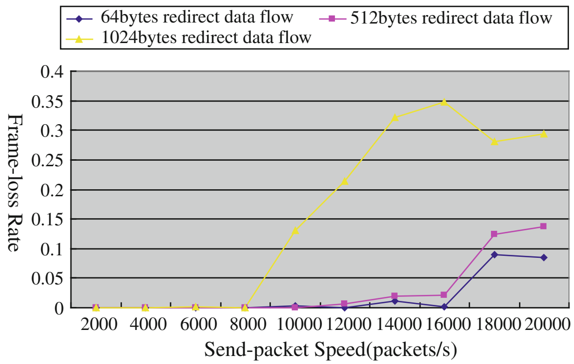
118.3.2 Performance Test

The performance test platform is shown in Fig. 118.8.

We use smartbits to generate packets for controlling the transmission of ForCES protocol message between CE and FE. As shown in Fig. 118.1, FE forwards packets from some Fi/f to some other Fi/f, but in some cases, FE also have to redirect some packets to CE. These packets are usually packets for routing protocols or network management protocols. These packets will be encapsulated as ForCES protocol messages (called redirect messages) and redirected to CE.

The test results are shown in Figs. 118.9 and 118.10. From the test results, we know loss rate is higher for larger length messages. Also, the results show the efficiency of UDP forwarding packets is clearly higher than TCP. So our choice of TCP control channel and UDP redirect channel is reasonable.

Fig. 118.10 Redirect message flow of TCP/UDP-TML



118.4 Summary

This paper proposes the design of reconfigurable TML based on our earlier ForCES researches, and verifies the rationality of this TML by testing.

However, this TML only achieves endpoint authentication based on IP address, and does not provide message encryption and decryption functions. So the next step should use IPSec to achieve TML security requirement.

Acknowledgments Supported by the National High Technology Development 863 Program of China under Grant No. 2008AA01A323, 2009AA01A334, the Natural Science Foundation of China under Grant No. 60903214, 60970126, Zhejiang Provincial NSF China No. Y1090452, Y1080078 and Zhejiang Sci & Tech Project No. 2009C31066, 2009C11050.

References

1. Forwarding and Control Element Separation (ForCES). http://www.ietf.org/html.charters/forces-charter.html
2. Khosravi H, Anderson T (2003) Requirements for separation of IP control and forwarding. http://www.ietf.org/rfc/rfc3654.txt. Accessed Nov 2003
3. Yang L, Dantu R, Anderson T, Gopal R (2004) Forwarding and control element separation (ForCES) framework. http://www.ietf.org/rfc/rfc3746.txt. Accessed Apr 2004
4. Kohler E, Handley M, Floyd S, Padhye J (2005) Datagram Congestion Control Protocol (DCCP). http://draft-ietf-dccp-spec-13.txt. Accessed Dec 2005
5. Salim HJ, Ogawa K (2010) SCTP based TML (Transport Mapping Layer) for ForCES protocol. http://tools.ietf.org/html/rfc5811.txt. Accessed Mar 2010
6. Maloy J, et al. (2004) TIPC: transparent inter process communication protocol, a layer 2 TML for the ForCES protocol. http://tools.ietf.org/html/draft-maloy-tipc-01.txt. Accessed Oct 2004
7. Hidell M, Sjödin P, Hagsand O (2005) Control and forwarding plane interaction in distributed routers. Proceedings of the 4th IFIP-TC6 networking conference: Networking 2005, Waterloo, May 2005

8. Wang W, Dong L, Zhuge B (2006) ForTER—an open programmable router based on forwarding and control element separation. In: Proceedings of DCABES 2006, vol II. Oct 2006, pp 1069–1077
9. Wang W, Dong L, Zhuge B (2007) TCP and UDP based ForCES protocol TML over IP networks. http://tools.ietf.org/id/draft-wang-forces-iptml-02.txt. Accessed Mar 2007
10. Zhuge B, Yang S, Wang W (2009/03) Performance analysis and optimization of TML in ForCES router. J Inf Eng Univ 10(1):125–128
11. Wang WM, Hadi Salim J, Alex A (2007) ForCES transport mapping layer (TML) service primitives. http://tools.ietf.org/html/draft-ietf-forces-tmlsp-01.txt. Accessed Feb 2007
12. Halenplidis E, Ogawa K, Wang X, Li C (2007) ForCES interoperability draft. http://tools.ietf.org/html/draft-ietf-forces-interoperability-04. Accessed Sept 2007

Chapter 119
The Study of a Mobility Supporting Coordination Mechanism for Supporting Mobile Devices

Ma Ying, Zhang Laomo and Wang Guodong

Abstract With the rapid growth of the mobile devices, a process with the device has high probability to surf on different wireless networks. Therefore the address of a host will be changed according to the located subnets and the convention software is difficult to run normally in this environment. In order to let the data delivering among processes that can support the mobility of hosts, we propose a new coordination mechanism for portable devices. We propose three methods to different environment: centralized mechanism, distributed mechanism, and hybrid mechanism, to satisfy communication not multifarious, multifarious communication, and communication habit unsure among procedures. Let users be able to select a most appropriate method to implement their own system with different need and different environment.

Keywords Coordination mechanism · Mobile devices · Wireless network

M. Ying (✉) · Z. Laomo (✉)
School of Software, Henan Institute of Engineering,
Zhengzhou 450053, Henan, China
e-mail: cyechina@163.com

Z. Laomo
e-mail: myfirst@163.com

W. Guodong
Graduate University of the Chinese Academy of Sciences,
Beijing 100049, China
e-mail: wgdaaa@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_119,

119.1 Introduction

The tremendous growth of the mobile users' population coupled with the portable devices is in contrast to the conventional programing languages, without the consideration of the mobility, to develop the related software. The mobility can be divided into: (1) terminal (portable device) mobility, (2) procedure mobility, and (3) service mobility. Therefore, providing the transplant mobility service, i.e., the users can use the service with considering the mobility, is critical.

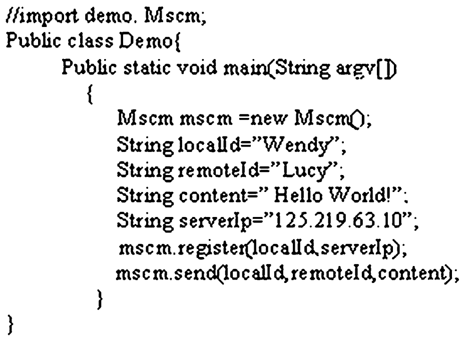
Therefore, when a local process communicates with other remote processes, no matter when and where, move everywhere or change a used portable device, they all hope can look like never moved, and used a same portable device, let the service continuously carrying on [1, 2]. Either the main stream nowadays or look good in the foreword, both of IPv4 and IPv6 have the characteristic of locality. That is IP is fixed and unchangeable in a region. So, no matter what types of movement had mentioned above, as long as moving, local procedure can not get real-time IP of remote procedure can not communicate consistently and it causes receiving lost content.

When local procedure and remote procedure communicate with each other, it can be divided into communication multifarious, communication not multifarious and general communication according to the communication model. And the so-called general communication is the communication sometimes multifarious and sometimes not in the uncertain condition and according to realistic situation it can be divided into two kinds (1) IP permanence: that is local procedure and the remote procedure uses a host and an IP carrying on communication continuously. On the normal on-line network, the communication model will not have errors but it does not have the characteristic of mobility either. (2) IP changed: local procedure and remote procedure use a same set of host to move everywhere, it does not have to consider the change of moving proceeding and it can still carry on communication. It has the characteristics of terminating machine mobility, procedure mobility and service mobility. Unfortunately, the communication of current procedure can't effectively support mobility [3].

Java agent development framework (JADE) [4, 5] is a platform which in charge of managing agent. It provides many toolboxes for programmers. It is not only used to debug during scheme development, but also use these tools to monitor or manage the built agent platform after scheme developing is completed.

For the communication between procedures can effectively support mobility and easily used by programmers, we custom a new class let the local procedure will not be influenced by itself or remote mobility on the usage. They are used to solve the problems to keep receiving the content correctly when IP or the host is changed, and let communication content will not be lost by users moving.

The rest of this paper is organized as follows. Section 119.2 introduces the mechanism. Section 119.3 presents the implementation and prototype. Finally, conclusions are addressed in Sect. 119.4.



```
//import demo. Mscm;
Public class Demo{
        Public static void main(String argv[])
           {
               Mscm mscm =new Mscm();
               String localId="Wendy";
               String remoteId="Lucy";
               String content=" Hello World!";
               String serverIp="125.219.63.10";
               mscm.register(localId,serverIp);
               mscm.send(localId,remoteId,content);
           }
}
```

Fig. 119.1 Registration and message sending

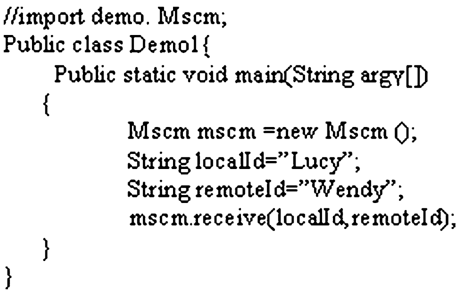
119.2 Mechanism

Our mechanism is implemented to a Java class, Mscm. This class combined with a JADE agent platform, is divided into a server and several procedures. To use this class to develop the function that processes can communicate with others without considering the mobility needs to select a host to run the Jade platform (the server of our mechanism). The goal of this JADE agent platform is real time to manage and update the location (IP addresses) of the hosts [6]. Accordingly, the programmers can use the methods of Mscm, including register(), send(), and receive(), by any program written in Java. The three methods are presented as follows:

- Mscm.register (String host_ID, String server_IP): In order to let two unknown procedures communicate with each other, each procedure must be assigned an Identification (ID). The method must be written before send() and its input is two strings. The first represents local procedure ID, which is registered from server; the second represents server IP address.
- Mscm.send (String sender_ID, receiver_ID, String Content): It is used to send content to a shared bugger space between local procedure and remote procedure. The method needs to send three strings: the first represents the ID of the sender; the second represents the ID of the corresponding receiver; the other string represents the content that the sender sends to this receiver. (Notably, the programmer needs to do is to assign a unique name of the procedure but assigns an IP address)
- Mscm.receive(String sender_ID, String received_Content): It is used to receive the content that is sent by a remote procedure. The method needs to send two Strings: the first String represents local procedure ID; and the second String represents remote procedure ID, representing to read which local procedure and remote procedure is in the pooling buffer content.

Figures 119.1 and 119.2 demonstrate an example to write the program. In Fig. 119.1 a mscm object is added first and four String variables are declared, then we used localID, remoteID, content, and severIP to be representation, and

Fig. 119.2 Message receiving



```
//import demo. Mscm;
Public class Demo1{
      Public static void main(String argy[])
    {
               Mscm mscm =new Mscm ();
               String localId="Lucy";
               String remoteId="Wendy";
               mscm.receive(localId,remoteId);
    }
}
```

users can define by themselves. However, it needs to register local procedure ID and select sever location, we used “Wendy” as the ID of local procedure and “125.219.63.10” as the sever IP. After finishing the above steps and calling send() method to communicate with remote procedure in any location of the following program. Here is an example that the process “Wendy” communicates with remote process “Lucy” with the content is “Hello World!”.

In Fig. 119.2, a new mscm object is added at first and two String variables are defined, we use local ID and remote ID to represent them, users can define by themselves; Calling receive() method in any location of the following procedure can get the needed content. On the side showing to grab the content of “Lucy” and “Wendy” these two procedures shared in the Buffer.

For managing the communicating content between local procedure and remote procedure effectively, in the system we propose three methods to adopt different the environments: (1) centralized (2) distributed (3) hybrid. Centralized mechanism satisfies that the communications among procedures frequently. The hybrid mechanism suits the general communication environment among procedures. The so-called general environment means the communication is sometimes multifarious and sometimes not.

Figure 119.3 explains that after procedure A started in host A and procedure B started in host B, procedure A and B can carry on communication through wired or wireless network.

When local procedure A communicates with remote procedure B, each host will store the content (history file) for the backup function. But when process A or process B leaves the storage history file, host and log from another host continues communication, the history file will not bring them but stay in the old host. So users can not get the history file when they want to look up in the new host.

I. *Centralized mechanism*. Among procedures need to reach receiving content not loss and get the latest history file anytime, Fig. 119.4 is the operation process of using centralized mechanism:

   (1) When starting procedure A and B can immediately deliver content to each other.

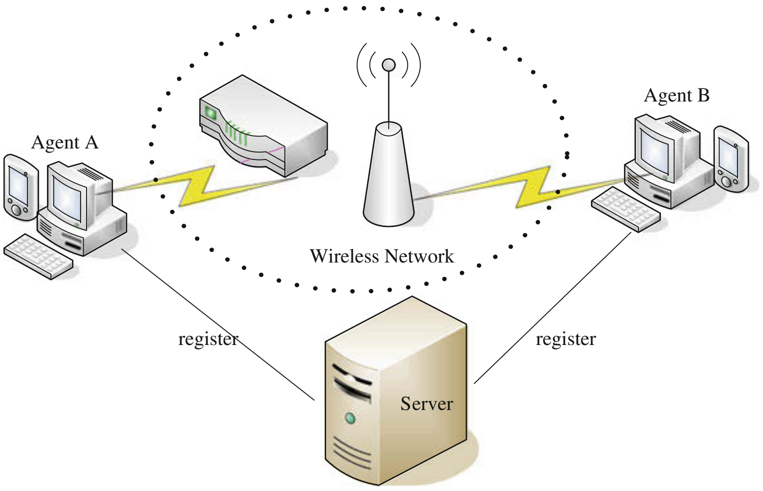


Fig. 119.3 Framework

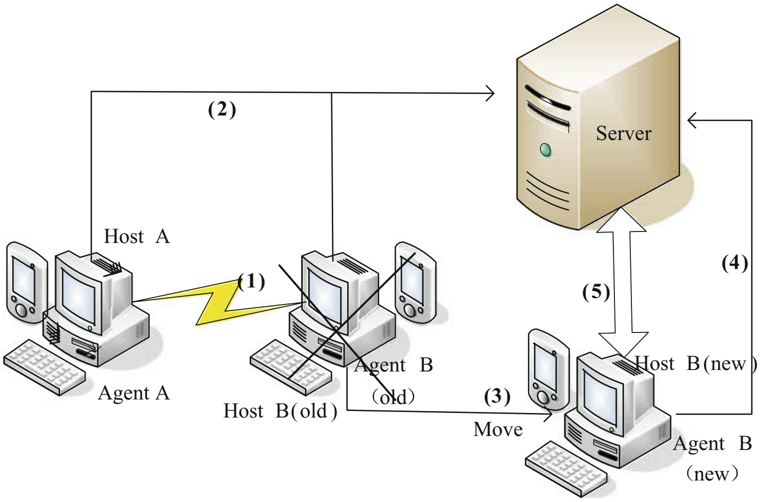


Fig. 119.4 Procedure of the hybrid mechanism

(2) Local procedure A delivers a content each time it updates history file in server after confirming remote procedure B receiving content; in contrast, procedure A receives a message content each time will also update the server.
(3) When users leave the current host B and shut down procedure B, log in again after users move to host B.

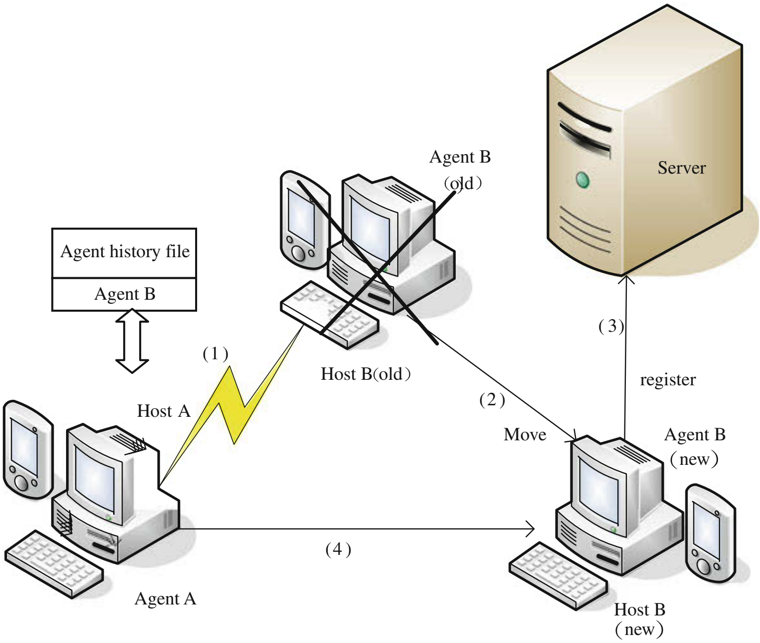


Fig. 119.5 Procedure of the distributed mechanism

(4) Procedure B can carry on communication with procedure A after registering first in the server.

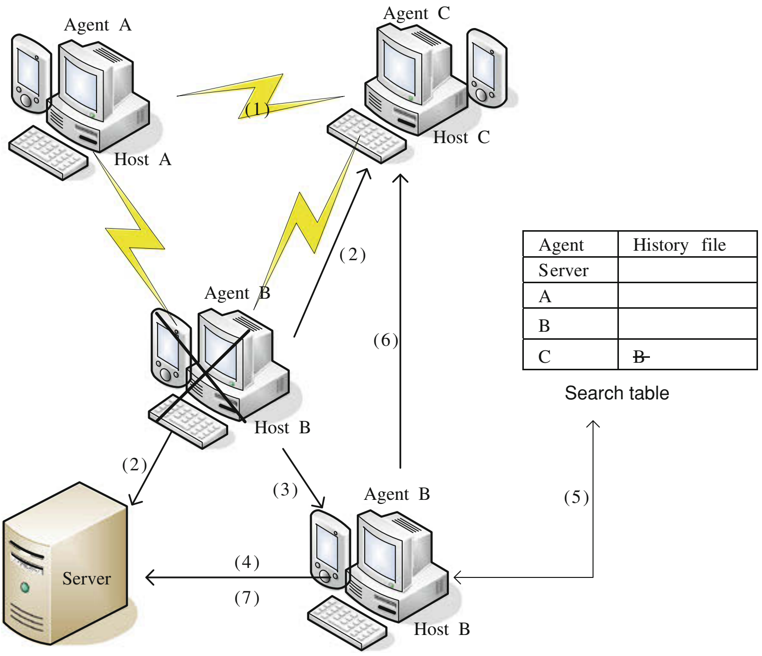
(5) Users ask the history file prior to procedure B from server and bring them to the current host B. The method suits the procedures communicating not multifarious, because procedure decreases storage action through server. On the contrary, the burden of server is not too big. And the system structure builds are most simple and cost effective.

II. *Distributed mechanism*. Among procedures need to reach receiving content not loss and get the latest history file at anytime. Fig. 119.5 is the operation process of using distributed mechanism:

(1) When starting procedure A and B can immediately deliver content to each other.

(2) Users leave the current host B and finish procedure B, which will deliver its history file to other remote procedure in current system and storage in host. Figure 119.5 is an example, for the other remote procedure in the current system, procedure A, the history file will be delivered to storage to procedure A of locating host A.

(3) Users login again in host B after a period of time and go first to re-register in server, then they continue carrying on communication with procedure A.



Agent	History file
Server	
A	
B	
C	~~B~~

Fig. 119.6 Procedure of re-login/data resumption mechanism

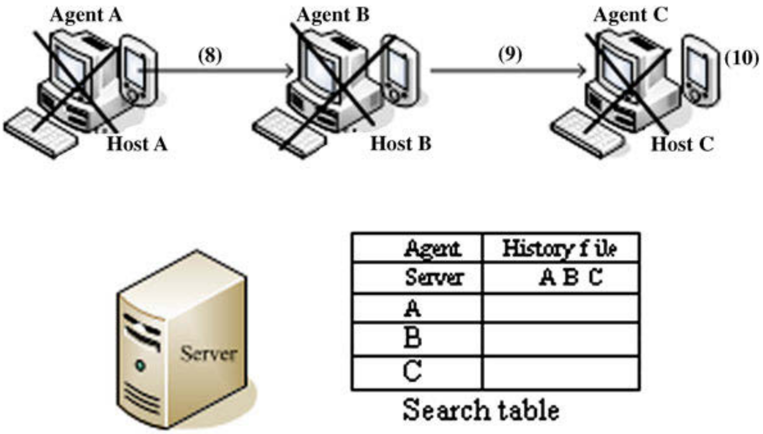
(4) Procedure A will build a table in its locating host A to record history file of owning those procedures. B is recorded in the table of procedure A. After procedure B re-login, the history file of procedure B will be delivered to procedure B.

The method suits communication multifarious among procedures because there is a procedure continuously in the system to exchange message content rapidly. If storage through server in this time, that will increase the burden of server and seriously affect the system execution capability. So we store through the host of remote procedure to raise the whole execution efficiency of system.

III. *Hybrid mechanism*. Among procedures that need to reach receiving content not loss and get the latest history file at anytime, Figs. 119.6 and 119.7 are the operation process of using hybrid mechanism:

(1) When starting procedure A procedure B and C can immediately deliver content to each other.

(2) Users leave the current host B and shut down procedure B will deliver history file to procedure C to storage in host C; and informing server which will update the history file of procedure C in the look-up table as B, the host that indicates the procedure C of locating the host that owns the history file of procedure B.



Agent	History f ile
Server	A B C
A	
B	
C	

Search table

Fig. 119.7 Procedure of continuously shutting down procedure

(3) Users re-login the host B after a period of time.

(4) Procedure B continues to communicate with procedure A and procedure C after registering from the server.

(5) Procedure B is inquiring the server and when the server receives the registration will respond that ID two Strings which represent "the procedure owns your history file" and "to deliver to do backup for the procedure of locating host".

(6) Procedure B knows that procedure C of locating host owns a own history file, asking an own history file directly to host C.

(7) Procedure B notifies a server to get history record back already; the server will empty the history file of procedure C in look-up table.

(8) When procedure A shuts down, will deliver its own history record to storage in the procedure B of locating host B, which will have the history file of procedure A and procedure B.

(9) When procedure B also shuts down will deliver its own procedure history file to procedure C (because in the system procedure C only remains) of locating host C to storage, which will own the history file of the procedure.

(10) When procedure C also shuts down, for there is no other procedures in the system, it will deliver to server all history file that it owns to storage, and the look- up table of the server, the history record of the procedure "Server" will be updated as A, B, C. That indicates the server owns the history file of procedure A procedure B and procedure C.

The method suits general communication among procedures for sometimes communication is multifarious and sometimes not among procedures, if the adopted centralized or distributed mechanism is unsuitable, so adopting this method to reach whole runtime capability of stabilization on system.

119.3 Implementation and System Model

We implement a Mscm class to make users communicate with remote procedure conveniently by calling Mscm class. And combining with JADE platform to implement a multi-agent system by the agent on terminal machine equipment of each procedure locating that can control the location of local and remote procedure at any time. When local procedure moves from a network region to another network region, the agent of local procedure will tell JADE platform the latest location, and notify the other remote procedure by JADE platform.

And local procedure can also periodically search the latest IP location of other remote procedures, to ensure JADE platform notifying the loss of latest location. And at the procedure carrying on communication aspect, local procedure has already known the latest location of remote procedure which want to communicate with through JADE platform, therefore can carry on a communication with that procedure directly. We adopt user datagram protocol (UDP) here. It can faster deliver the content to remote procedure, but UDP is an unreliable transport protocol so we do a check mechanism to judge if the content delivered to remote procedure correctly; and do the content file for a backup, adopting transmission control protocol (TCP) for TCP is a reliable transport protocol, although its efficiency is not as fast as UPD, to ensure the accuracy of the content file that we deliver, we use TCP as a transport mechanism to backup the content file.

119.4 Conclusion

When local procedure and remote procedure carry on communication, for the problem of mobility, it will cause receiving content loss. The key problem generated when the procedure is moving : the IP changes. We custom a new class let the local procedure of the users locating will be free from the influence of the movement and communicate with remote procedure conveniently to reach content without loss, accuracy receiving and delivering. Besides, the preservation method of communication content among procedures, we propose three methods to different environment: centralized mechanism, distributed mechanism, and hybrid mechanism, to satisfy communication not multifarious, multifarious communication and communication habit, unsure among procedures. Let users be able to select a most appropriate method to implement their own system in different need and different environment. We have already completed this system structure currently, we will enhance our system in the future and continuously join speech transmission function, etc.

References

1. Caire G (2005) JADE tutorial-jade programming for beginners, Version: JADE3.3, Mar 2005
2. Carrillo-Ramos A, Gensel J, Villanova-Oliver M, Martin H (2005) PUMAS: a framework based on ubiquitous agents for accessing web information systems through mobile devices. In: 20th annual ACM symposium on applied computing–handheld computing, vol 2, pp 1003–1008, Mar 2005
3. Chen E, Sabaz D, Gruzer WA (2004) Wireless distributed systems with JADE. In: IEEE international conference on systems, man and cybernetics, vol 1, pp 989–993, Oct 2004
4. JADE-Java Agent Development framework. http://jade.cselt.it
5. LEAP-lightweight extensible agent platform. http://leap.crm-paris.com/
6. Caire G (2005) LEAP user's guild, Version LEAP 3.3, March

Chapter 120
Goodness-of-fit Testing for Lifetime Distribution Based on Support Vector Machines

Xin-yao Zou and Xiao-ling Lin

Abstract This chapter describes a goodness-of-fit testing method for lifetime distribution that can be used to determine how well a sample of life data fits an assumed distribution when dealing with small sample failure data. This method transforms the examination of distribution function to machine learning problem based on support vector machine. We describe the statistical detail of the algorithms of least squares support vector machine and discuss how to make goodness-of-fit testing using it. And the application of this methodology on goodness-of-fit testing of Weibull distribution is presented.

Keywords Goodness-of-fit testing · Least Squares Support Vector Machines · Weibull distribution

120.1 Introduction

For reliability assessment, pinpointing the underlying failure distribution of microelectronic devices is an important task, the results of which can enable sound reliability and maintenance decisions to be made. Generally, when dealing with failure data, we make a hypothetic distribution according to histogram firstly, then

X. Zou (✉)
Department of Electron and Information Engineering,
Guangdong AIB Polytechnic College, Guangzhou 510507, China
e-mail: madelinexy@163.com

X. Lin
Science and Technology on Reliability Physics and Application of Electronic
Component Laboratory, China Electronic Produce Reliability and Environmental
Testing Research Institute, Guangzhou 510610 China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_120,

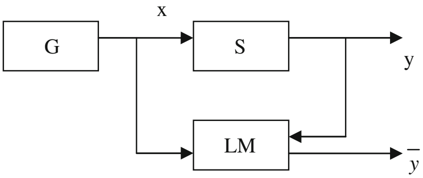


Fig. 120.1 A model of learning from example

to infer how well a sample of data fits the assumed distribution by using statistic methods. This procedure has been classified as "goodness-of-fit" (GoF) test, which provides the mathematical foundation for selecting the distribution model that best fits the data.

There has been extensive research on GoF test to judge the fit of a distribution to life data [1–4]. These GoF tests range from graphical plotting techniques (GPT), to tests which exploit characterization results for the specified underlying model. GPT is mostly used due to its simplicity and intuition, but the results of which has little accuracy because of human factor. Chi-squared test was proposed by Karl Pearson in 1990. This statistic is not suitable for many practical cases because it requires a large sample size. Kolmogorov–Smirnov (K–S) test may be used for any sample size, however this test has the following limitation: hypothesized distribution must be continuous and all parameters of the hypothesized distribution are available. Later, in the 1980s and 1990s, many researches have reported that the Cramer–vonMises (C-M) and Anderson–Darling (A-D) statistics are more powerful for GoF tests than the K–S statistic in most cases [5–7]. All those GoF tests are based on classical statistics developed for large sample. In fact, we usually do not have enough field failure data. This Chapter proposes a method based on statistical learning theory (SLT) [8], which is a new statistical theory framework established from finite samples, it provides a powerful theory fundament to solve machine learning problems with small samples. Statistical learning theory was introduced in the late 1960s. Before 1990s it was a purely theoretical analysis of the problem of function estimation from a given dataset. In the middle of the 1990s, support vector machines (SVM) based on the developed theory were proposed, which made SLT not only a tool for the theoretical analysis, but also a tool for creating practical algorithms for estimating multidimensional functions [9].

The idea of this method is to make a regression model according to the life dataset by using SVM. It transforms the prediction of distribution function to machine learning problem, shown in Fig. 120.1. The problem of learning is that of choosing the function that predicts the supervisor's response in the best possible way, and the model accuracy obtained by SVM regression can be used to GoF testing.

In the rest of this Chapter, we give a brief algorithm review of LS-SVM regression and describe how to make goodness-of-fit testing using it in Sect. 120.2. In Sect.120.3, we illustrate the feasibility of applying LS-SVM regression in lifetime distribution examination for Weibull distribution and the concluding remarks are given in Sect. 120.4.

120.2 Regression Using SVM and LS-SVM

120.2.1 Algorithms Review of LS-SVM Regression

Comparing with the traditional GoF statistics that judge the fitness according to an assumed significance level, what we care mostly is the generalization ability of learning machine for the goodness-of-fit testing by using machine learning. Bad generalization ability means wide difference between the sample distribution and total distribution. The generalization ability of SVM is based on the factors described in structural risk minimization (SRM), which has greater generalization ability and is superior to the empirical risk minimization (ERM) as developed for large sample and adopted in neural networks, least squares method in the problem of regression estimation and the maximum likelihood method in the problem of density estimation. For SVM, one needs to minimize (120.1) in order to find the regression estimation:

$$R(\alpha) \le R_{\mathrm{emp}}(\alpha) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{\mathrm{emp}}(\alpha)}{B\varepsilon}}\right) \tag{120.1}$$

where the first term is an estimate of the empirical risk (the number of errors on the training set) and the second is the confidence interval for this estimation. There may be overfitting if one just minimizes the empirical risk. In this case, even if one could minimize the empirical risk down to zero, the amount of errors on the test set could be big. Therefore to avoid overfitting and generalize well, SVM minimizes both of empirical risk and confidence interval.

SVM regression transforms this minimization problem into solving a convex optimization problem, more specifically a quadratic programming (QP) problem. This is obtained by employing the Vapnik's ε-insensitive loss function, formulating the optimization problem and exploiting the Mercer condition in order to relate the nonlinear feature space mapping to the chosen kernel function. However, this constrained optimization programing leads to higher computational burden. This disadvantage has been overcome by least squares support vector machines (LS-SVM), which work with equality instead of inequality constraints and a sum squared error (SSE) cost function [10, 11]. This reformulation greatly simplifies the problem in such a way that the solution is characterized by a linear system, which can be efficiently solved by iterative methods such as conjugate gradient [12]. So in the simulation experiment, we use LS-SVM regression instead of SVM to make the GoF testing for the lifetime distribution of MOS capacitors.

For LS-SVM, the regression estimation problem is formulated as the optimization problem:

$$\min_{w,b,e} J(w, e_i) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2$$

subject to the equality constraints

$$y_i = w^T \varphi(x_i) + b + e_i, i = 1, \ldots, N. \tag{120.2}$$

With the application of the Mercer's theorem on the kernel matrix Ω (as $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j), i, j = 1, \ldots, N$), it is not required to compute explicitly the nonlinear mapping $\varphi(\cdot)$ as this is done implicitly through the use of positive definite kernel functions $K(x_i, x_j)$. Usually, several choices for $K(x_i, x_j)$are possible.

(1) Linear kernel:$K(x_i, x_j) = x_i^T x_j$,

(2) Polynomial kernel: $K(x_i, x_j) = (x_i^T x_j / c + 1)^d$(polynomial of degree d, with c a tuning parameter);

(3) Gaussian Radial Basis Function (RBF) kernel: $K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma^2)$ (σ is a tuning parameter)

The solution of this optimization problem is obtained by Lagrangian function (120.3)

$$L_{LS-SVM} = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 - \sum_{i=1}^{N} \alpha_i (w^T \varphi(x_i) + b + e_i - y_i) \tag{120.3}$$

where $\alpha_i \in R$ are the Lagrange multipliers, the conditions for optimality are given by :

$$\begin{cases} \frac{\partial L_{LS-SVM}}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \\ \frac{\partial L_{LS-SVM}}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i = 0 \\ \frac{\partial L_{LS-SVM}}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma_i e_i, i = 1, \ldots, N \\ \frac{\partial L_{LS-SVM}}{\partial \alpha_i} = 0 \rightarrow y_i = w^T \varphi(x_i) + b + e_i \end{cases} \tag{120.4}$$

After elimination of w and e_i, the following linear system is obtained:

$$\begin{bmatrix} 0 & 1^T \\ 1 & \Omega + \mathrm{I}/\gamma \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{120.5}$$

where $y = [y_1, \ldots, y_N]^T$, $\alpha = [\alpha_1, \ldots \alpha_N]^T$

The LS-SVM regression formulation is then constructed as follows:

$$y(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b \tag{120.6}$$

where α, b are the solutions to (120.5).

As we can see from Eq. 120.6, we can obtain the most suitable regression function that predicts the supervisor's response if only we get α, b and kernel function K(x, xi). This means that the key of LS-SVM regression is to choose the kernel function and kernel parameters [13–15].

120.2.2 Selection of Kernel Function and Parameters

Choosing different kernel function means using different learning machine to train the data. Although there are several choices, we select RBF kernel as the kernel function for LS-SVM regression. Because in LS-SVM, the estimation of the support values is optimal in the case of a Gaussian distribution of the error variables due to the equality constraints [16]. For RBF kernel, there are two parameters σ^2 and γ to be determined, σ^2 is a tuning parameter and γ is a regularization parameter, which guarantees high generalization ability of the regression model. These parameters usually are chosen by using the k-fold cross-validation technique, which can prevent the overfitting problem. In k-fold cross-validation, we first divide the training set into k subsets of equal size. Sequentially one subset is tested using the model trained on the remaining subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the mean square error (MSE) of actual output value and predictor value [17, 18]. At the mean time, we recommend a coarse "grid-search" on $\sigma^2$and γ during using cross-validation. Basically pairs of (σ^2, γ) are tried and the one with the best cross-validation accuracy is picked. After identifying a "better" region on the grid, a finer grid search on that region can be conducted.

120.3 Application

120.3.1 Experiment

In this experiment, we use the LSSVM software package [9] based on Matlab to perform the model training and testing. A complete sample is generated from the two-parameter Weibull distribution with $\beta = 2.5$ and $\eta = 30$. The sample size is 20 and the failure times are as follows:

2.0541 s, 2.4248 s, 2.6247 s, 2.7663 s, 2.8565 s, 2.9653 s, 3.0253 s, 3.1046 s, 3.1570 s, 3.2149 s, 3.2809 s, 3.3499 s, 3.3911 s, 3.4404 s, 3.5086 s, 3.5410 s, 3.6109 s, 3.6584 s, 3.7495 s, 3.9396 s.

The parameters σ^2 and γ of RBF kernel are estimated by 10-fold cross-validation technique using the following steps [19]:

1. Set aside 2/3 of the data for the training/validation set and the remaining 1/3 for testing. Table 120.1 and Table 120.2 show the training dataset and testing dataset.

Table 120.1 Results of training/validation data under RBF kernel function

True value	Prediction value	True value	Prediction value
2.0541	2.0831	3.2809	3.2905
2.4248	2.4104	3.3499	3.3445
2.6247	2.6059	3.3911	3.3966
2.8565	2.8597	3.5086	3.4983
2.9653	2.9537	3.6109	3.6034
3.1046	3.1075	3.6584	3.6616
3.1570	3.1731	3.7495	3.7293
3.2149	3.2337		
MSE		1.9308e-004	

Table 120.2 Results of testing data under the LS-SVM regression model

True value	Prediction value	True value	Prediction value
2.7663	2.7474	3.5410	3.5498
3.0253	3.0350	3.9396	3.8229
3.4404	3.4476		
MSE		0.0028	

2. Starting from i = 0, perform 10-fold cross-validation on the training/validation data for each (σ^2,γ) combination from the initial candidate tuning sets $\sum_0 = \{0.5, 5, 10, 15, 25, 50, 100, 250, 500\}$and $\Gamma_0 = \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000, 5000, 10000\}$.
3. Choose optimal (σ^2, γ) from the tuning sets $\sum_i$ and Γ_i by looking at the best cross-validation performance for each (σ^2, γ) combination.
4. If i = imax (usually imax = 3), go to step 5; else i: = i + 1, construct a locally refined grid $\sum_i \times \Gamma_i$ around the optimal parameters (σ^2, γ) and go to step 3.
5. Construct the regression LS-SVM using the total data set for the optimal choice of the tuned parameters (σ^2,γ).
6. Assess the test set accuracy by means of mean square error (MSE).

After the above procedure, the two optimal parameters of RBF kernel function in the regression LS-SVM are $\sigma^2 = 34$ and $\gamma = 197$.

To illustrate the generalization of this regression model, we predict the output of training dataset and another test dataset with this regression model. Table 120.1 and Table 120.2 show the actual output, predictor output and MSE under training dataset and testing dataset, respectively.

120.3.2 Discussion

The model accuracy of LS-SVM regression is characterized by mean square error (Tables 120.1 and 120.2), which can be used to judge how well a sample of data fits an assumed distribution instead of calculating p value and significance level.

When we select a certain kernel function we should train the failure dataset with different parameters until we get optimal MSE and fit curve. Otherwise, we should choose other kernel function and parameters to train the data. From the MSE value of training set and testing set, we can infer that LSSVM regression can be used to goodness-of-fit testing of Weibull distribution.

This method is suitable for any sample size due to the solid statistical learning theory foundation. Even for small failure data, LS-SVM can find a suitable function that describes the inner relationship between the input and output by learning the training dataset and predict the future information. By choosing optimal kernel function and parameters, we can select the distribution model that best fits the training data, and the good generalization ability guarantee a random sample comes from some specific distribution. At the mean time, it has higher accuracy than GPT without human factor. And it is convenient to use because all of the calculations are conducted automatically with LS-SVM software.

120.4 Conclusion

This Chapter proposed an effective method based on statistical learning theory for goodness-of-fit testing of lifetime distribution. The key of this method is to select optimal kernel function and parameters in order to construct LS-SVM regression model to train the failure dataset. It provides new idea to deal with the reliability evaluation of microelectronic devices when we do not have enough failure data due to the small sample theory foundation. In our tests, we examine the feasibility of applying LS-SVM regression for goodness-of-fit testing of Weibull distribution. In addition, this method can be applied to other distributions as well.

References

1. Cirrone GAP et al (2004) A goodness-of-fit statistical toolkit. IEEE Trans Nucl Sci 51(5):2056–2063
2. Jeff Y, Ang T, Rastings NAL (1994) Model accuracy and goodness of fit for the Weibull distribution with suspended items. Microelectron Reliab 34(7):1177–1184
3. Balakrishnan N, Ng HKT, Kannan N (2004) Goodness-of-fit based on spacings for progressively type-II censored data from a general location-scale distribution. IEEE Trans Reliab 53(3):349–356
4. Littell JF (1982) Statistical models and methods for lifetime data. Wiley, NJ
5. Porter JE, Coleman JW, Moore AH (1992) Modified KS, AD and C-vM tests for the pareto distribution with unknown location & scale parameter. IEEE Trans Reliab 41(1):112–117
6. Gwinn DA (1993) Modified anderson-darling and cramervon mises goodness-of-fit tests for the normal distribution. NASA Tech Reprot AD-A262554
7. Park WJ, Seoh M (1994) More goodness-of-fit tests for the power-law process. IEEE Trans Reliab 43(2):275–278
8. Vapnik VN (1998) Statistical learning theory. Wiley Interscience Publisher, New York

9. Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20(3):273–297
10. Suykens JK, Gestel TV, Brabanter JD et al (2002) Least squares support vector machines. World Sci Press, Singapore
11. Wang H, Hu D (2005) Comparison of SVM and LS-SVM for regression. In: IEEE Proceedings of the international conference of neural networks and Brain, 279–283 pp
12. Vapnik V, Golowich S, Smola A (1997) Support vector method for function approximation, regression estimation and signal processing. Advance in neural information processing systems. MIT Press, Cambridge, pp 281–287
13. Scholkopf B, Burges CJC, Smola AJ (1999) Advances in kernel methods-support vector learning. MIT Press, Cambridge
14. Smola A, Scholkopf B, Muller KR (1998) The connection between regularization operators and support vector kernels. Neural Netw 11(4):637–649
15. Suykens JAK, Vanderwalle J, Moor BD (2001) Optimal control by least squares support vector machines. Neural Netw 14(1):23–35
16. MacKay DJC (1995) Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks. Netw Comput Neural Syst 6(3):469–505
17. LS-SVM Software [online] http://www.esat.kuleuven.ac.be/sista/lssvmlab/
18. Hu H.sh (2000) The reliability and breakdown mechanism of thin silicon oxide. Ph.D. thesis, metallurgy institute, chinese academy of sciences
19. Gestel TV, Suykens JAK, Baesens B et al (2004) Benchmarking least squares support vector machine classifiers. Mach Learn 54(1):5–32

Chapter 121
Research on Outsourcing or Self-Support Decision-Making of Products Recall Reverse Logistics

Guo Teng-da, Ou Chao-min and Yang Xi

Abstract To solve the problems of outsourcing or self-support decision-making in the process of products recall, an ANP network graph of products recall reverse logistics decision-making based on Analytic Network Process theory is established; cost of reverse logistics, ability to deal with reverse logistics, risk and organization of reverse logistics could be considered while making outsourcing or self-support decisions about products recall reverse logistics. The outsourcing or self-support decision-making processes which take HP PC recall incident as an example are simulated and sensitively analyzed through Super Decisions software; then weights of factors that influence products recall decision-making are observed. At the same time a quantify thought to solve the decision-making problems about reverse logistics is provided, and it can be useful for enterprises operation. Finally, a further discussion is put forward.

Keywords Products recall · Outsourcing · Self-support · Analytic network process

G. Teng-da (✉) · O. Chao-min
Computer Science Editorial, Springer-Verlag, Tiergartenstr. 17,
69121 Heidelberg, Germany
e-mail: tengdaguo@126.com

G. Teng-da · O. Chao-min
School of Information System and Management, National University
of Defence Technology, 410073 Changsha, China

Y. Xi
CNIAB, 100016 Beijing, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_121,

121.1 Introduction

With the complexity of manufacturing techniques, the high frequency of manufacturing technology and the individuation of customers demand, enterprises cannot evade defects absolutely in the processes of designing, producing, stocking, transporting, assembling and marketing. When some products could threaten a great number of customers' security in some certain scope, enterprises need to recall these products centrally; then, products recall logistics come into being. Products recall logistics usually have the non-core processes, such as checking, sorting, information imputing, assembling and transporting. These processes usually have large workload, high precision request and strict time limit, which make great challenge to enterprises. Thus, selecting excellent third-party reverse logistics enterprises or creating high level reverse logistics operation capability is a foundation job in products recall logistics [1].

Factors which influence the strategy of reverse logistics outsourcing and self-support are a lot. Nowadays, research on logistics outsourcing and self-support always focus on forward logistics. LIU Pingping studies logistics outsourcing risk, and considers outsourcing risk as an important basis when decisions are made [2]. Zhou Haixia analyzes outsourcing decision-making for manufacturing from the view of transaction cost economics, and concludes that complete outsourcing and self-support are all seldom in manufacturing, the mode of mix-outsourcing are selected more [3]. Yang Jin claims that switching cost in outsourcing decision-making are very important, then outsourcing decision-making tables are established [4]. The research above are usually from the point of risk and cost when considering factors about logistics outsourcing decision-making, and not consider reverse logistics outsourcing decision-making as an important field. Reverse logistics are more complex than forward logistics, and the establishment of information system is different from forward logistics, academe does not have the accomplished fruit about whether reverse logistics network needs to be integrated with forward logistics yet; thus, some theories about forward logistics decision-making cannot instruct reverse logistics. It is a strategic decision-making whether selecting third-party logistics or not, only the factors influencing reverse logistics are analyzed such as reverse logistics cost, reverse logistics disposal capability. Risk and reverse logistics organization in detail, could appropriate reverse logistics operation mode to be selected.

Products recall is a typical form of reverse logistics; the uncertain and unpredictable characters of reverse logistics are embodied in reverse logistics recall entirely. Basing on Analytic Network Process theory, this paper analyzes factors influencing products recall process, takes HP PC recall event as a case study, quantifies factors weight, simulates the decision-making processes, calculates its sensitivity and hopes to provide reference on products recall practice.

121.2 Analysis of Factors Influencing Outsourcing or Self-support Decision-making in Products Recall Processes

121.2.1 Theory of Analytic Network Process

Analytic Network Process (ANP) is a system analysis theory provided by Professor T. L. Satty of Pittsburgh University. It can be traced back to Analytic Hierarchy Process (AHP). ANP takes the opinion that: factors in network are interactional (A → B means A is influenced by B), in the meantime, feedback relationship exists in elements of factors, these influences and feedback must be considered in decision-making. The situation above is not concluded in AHP theory; AHP is an especial situation of ANP [5–7].

ANP describes the influence and feedback relationship by table form qualitatively; each side of arrowhead denotes that factors are interactive. At the same time, ANP describes the influence and feedback extent among factors quantitatively through matrix format, it compares factors indirect priority degree, establishes relationship judgment matrix, sorts variables got from latent root calculation method, after normalized and weighted function, weighted super matrix are gained. Then, make the weighted super matrix self-multiply several times, till it goes to stabilization which is to say that the sort variables are not changed, limit matrix are calculated. The row vector of limit matrix is limit mix weight; selection plan can be determined by these processes [8]. Comparatively essentiality index can be obtained by expert judgment and questionnaire.

121.2.2 Analysis of Factors Influencing Products Recall Reverse Logistics Outsourcing and Self-support

Based on above analysis, ANP network is established and shown in Fig. 121.1.

The aim of this chapter is to make the decisions of whether outsourcing or self-support in the processes of products recall. The problem is divided into each element group; element groups are made up of elements. The goal element group is reverse logistics decision-making element. It is influenced by four criteria element groups, which are reverse logistics cost, reverse logistics disposal capability, risk and reverse logistics organization.

1. *Reverse logistics cost subelements analysis.* Outsourcing Products recall reverse logistics can lead the happening of transaction cost and switching cost [4]. If enterprises want to outsource the whole reverse logistics activities in the processes of products recall, the cost of enterprises in this situation include trade cost, transaction cost and fees paid to third-party reverse logistics. These cost can be included in operation cost, the existence of trade cost and transaction cost always

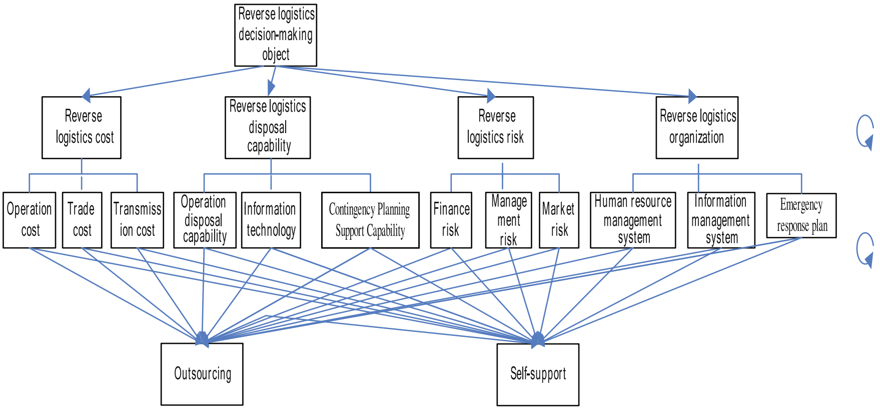

Reverse logistics decision-making object
Reverse logistics cost
Reverse logistics disposal capability
Reverse logistics risk
Reverse logistics organization
Operation cost
Trade cost
Transmission cost
Operation disposal capability
Information technology
Contingency Planning Support Capability
Finance risk
Management risk
Market risk
Human resource management system
Information management system
Emergency response plan
Outsourcing
Self-support

Fig. 121.1 ANP network of products recall logistics decision-making

cannot reduce cost in the early periods of logistics outsourcing [9]. If enterprises only want to outsource partial reverse logistics activities, the operation cost includes disposal cost when self-support and fees paid to third-party reverse logistics. If enterprises self-support the whole reverse logistics activities in the processes of products recall, all the cost in these processes could be included in operation cost. Enterprises need to take budgeting value of cost into consideration when making outsourcing or self-support decisions [3, 4, 9].

2. *Reverse logistics disposal capability subelements analysis*. Disposal capability focuses on activities to be decided whether outsourcing or not, such as information collection, sorting, transportation, assembling and so on. Information flow is the lead of logistics; the information technology in the processes of products recall includes transportation management system, stock management system, defect analysis system, quality ascending system, customer management system, etc. The perfect degree of information technology is an important basis when decision-making. Products recall has emergent characters; the factors lead to recall activities are always unfamiliar to enterprise. Thus, the capability of contingency planning support must be considered.

3. *Risk subelement analysis*. The uncertain and unpredictable characters of recall processes make the management risk happen when managing human resource, supervising and controlling copartners, mastering technology and solving conflict of economy and environment. Financial risk existing in reverse logistics is more obvious than that in forward logistics. Shadow cost are a lot in products recall reverse logistics; only these cost are taken into consideration, could financial risk be reduced. The integration of reverse logistics recall centers and transportation centers is not got success in practice, some bidirectional logistics managers claims that disposal processes are always overextended themselves when integration. Thus, separate reverse logistics recall centers are necessary. The R&D of

Table 121.1 Limited matrix

Cluster node labels		1 Goal		2 Criteria		
		Reverse logistics decision-making goal	Reverse logistics disposal capability	Reverse logistics cost	Reverse logistics organization	Reverse logistics risk
1 Goal	Reverse logistics decision-making goal	0.000000	0.000000	0.000000	0.000000	0.000000
2 Criteria	Reverse logistics disposal capability	0.287067	0.287067	0.287067	0.287067	0.287067
	Reverse logistics cost	0.113951	0.113951	0.113951	0.113951	0.113951
	Reverse logistics organization	0.428190	0.428190	0.428190	0.428190	0.428190
	Reverse logistics risk	0.170792	0.170792	0.170792	0.170792	0.170792

reverse logistics information system and the establishment of recall center occupy a lot of capital. Also, payoff period is long, which is easy to lead to the failure of marketing operation. In this situation, inter logistics innovation capability may be fallen after outsourcing. These factors constitute enterprises marketing risk.

4. *Reverse logistics organization subelement analysis*. To harmonize reverse logistics activities, a high-efficient and authoritative organization system must be set up. It can control the operation of logistics, and can dispose all kinds of problems and emergency duly and effectively [10]. The rationality of reverse logistics organization is important for outsourcing or self-support decision-making. Before making decisions, enterprise must clear the copartners' human resource management system and information management system, and at the same time emergency response plan must always must be considered.

The alternatives group includes two elements: outsourcing and self-support. The interaction of each element is described by arrowhead; annular arrowhead means that there are interaction and feedback system among elements. In Table 121.1, cost, reverse logistics disposal capability, risk, organization and their subelements are not separate. For example, among criteria group, the rationality of organization could affect the amount of cost and degree of risk simultaneously. The level of reverse logistics disposal capability could affect the amount of cost, and so on. Among subelements, the degree of information technology could affect process disposal capability and contingency planning support capability. The feedback function among these elements is included in ANP network and calculation subsequently.

121.3 Case Study: Recall of Defective PC in HP

A total of 900,000 PC recall incident makes HP become the first company to eat "recall crab" in electronic industry [11]. From discovering defects to putting out recall public announcement, HP Company is under a lot of pressure. Recall processes existing in HP Company are shown in Fig. 121.2.

These recall processes include sorting, checking, subdividing, discharging, information inputting, information sorting and other activities, which relate to non-core logistics activities in electronic manufacturing processes. And these activities above have an amount of workload and a high request for time and precision. HP Company takes these reverse logistics into consider whether outsourcing or self-support.

This chapter takes UPS as an outsourcing parameter reference, simulates HP Company decision-making processes according to HP data collected and provides a quantified view for decision-making.

121.3.1 Simulation

In ANP network shown in Table 121.1, comparing pair-wise elements which have feedback relationships, using professor judgment technology, and 16 relationship judgment matrix are obtained. As the calculation of ANP limit matrix is complex, especially in the situation of several elements having the feedback relationships, Super Decisions (SD) software are used to simulate.

Limited to the length of paper, relationship judgment matrixes are not laid in detail. Adjusting the value using SD software and making them according to consistency test, then stability about the results are calculated, limited matrix is obtained as in Table 121.1.

According to limited matrix, the degree of criteria group influenced by goal group is: reverse logistics organization > reverse logistics disposal capability > reverse logistics risk > reverse logistics cost, which means that the reverse logistics organization and disposal capability are important factors while decisions are made.

After clearing the weight index of each decision-making factors, the results are shown in Fig. 121.3, which means that the method of outsourcing is better than self-support. (outsourcing graphic is the above one).

121.3.2 Sensitivity Calculation

Sensitivity calculation is to analyze the influence of factors about final decision-making. Selecting reverse logistics disposal capability and combined elements

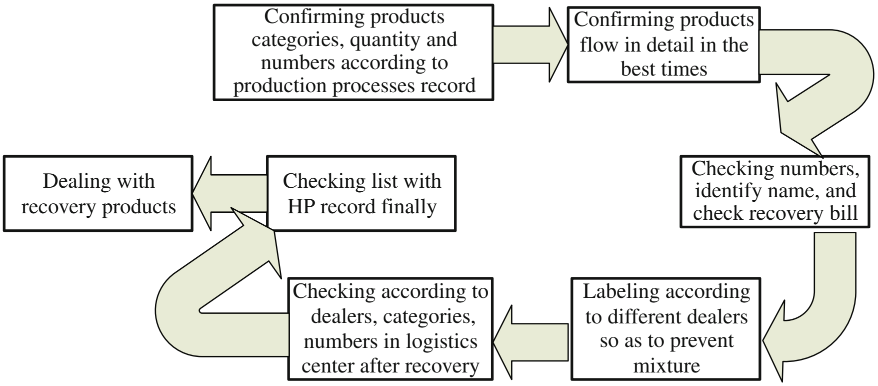
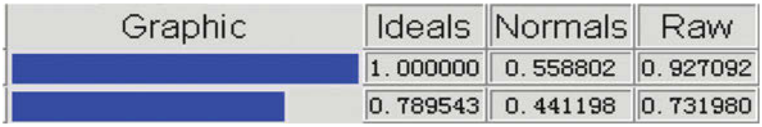

Confirming products categories, quantity and numbers according to production processes record
Confirming products flow in detail in the best times
Checking numbers, identify name, and check recovery bill
Labeling according to different dealers so as to prevent mixture
Checking according to dealers, categories, numbers in logistics center after recovery
Checking list with HP record finally
Dealing with recovery products

Fig. 121.2 Recall processes in HP



Graphic	Ideals	Normals	Raw
	1.000000	0.558802	0.927092
	0.789543	0.441198	0.731980

Fig. 121.3 Priority degree

which are composed by cost and reverse logistics capability as experimental objects, observing the change of decision-making when elements are changed, and then products recall logistics outsourcing or self-support are analyzed.

1. *Reverse logistics disposal capability is selected as separate experiment variable*. The weight is from 0 to 1, and 0.1 is considered as its step. The changes of priority of outsourcing and self-support when the weight of reverse logistics is changed are observed. In Fig. 121.4, vertical axis is described as priority; horizontal axis represents the step of experiment. When the weight of reverse logistics changes, the priority curves of outsourcing and self-support are changed accordingly. The difference of outsourcing and self-support curve is getting greater when disposal capability weight is increased, and the priority value of outsourcing is always more than self-support. This means that the third-party logistics selected have more priority than the company itself.

2. *Reverse logistics disposal capability and risk are selected as combined experimental variables*. The weight is from 0 to 1, 0.1 is considered as its step. When the weight changes, the priority of decision-making changes and shown in Fig. 121.5. In Fig. 121.5a, in the broken line position, these two variables combined weights are: (reverse logistics disposal capability 0.0001, risk 0.97), it is a position where self-support priority curve gets its top, and outsourcing priority curve gets its bottom. The self-support curve is over outsourcing curve.

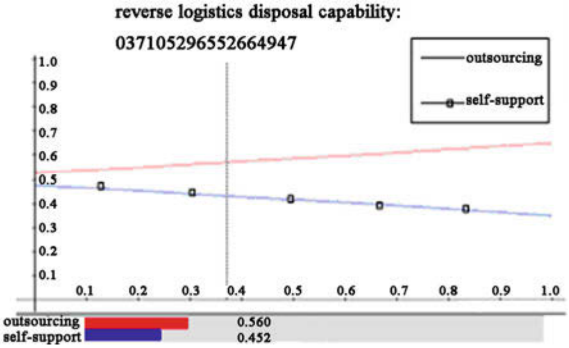


Fig. 121.4 Sensitivity analysis selecting reverse logistics disposal capability as separate experimental variable

We can see that, in this time, the difference of weight between reverse logistics disposal capability and risk is the most. The weight of disposal capability falls behind the weight of risk, which validates the result of experiment, and which is also to say that the third-party logistics selected has advantages in reverse logistics disposal capability. In the position of broken line in Fig. 121.5b, the two variables combined weights are: (reverse logistics disposal capability 0.80, risk 0.02). In this time, the priority curve of outsourcing is near its wave crest, and the priority curve of self-support is near its wave trough. Outsourcing priority curve is over self-support priority curve. In this time, the weight of reverse logistics is low. Thus, if the third-party logistics enterprise is selected, in the first period, there are no obvious effects on risk elusion, which is to say that operating logistics by itself has more advantages in risk control.

121.4 Conclusions

Outsourcing or self-support reverse logistics is influenced by reverse logistics cost, reverse logistics disposal capability, risk and reverse logistics organization. The quantitative analysis of factors influence reverse logistics can make enterprise clear the state and level of logistics system at present. After clearing what is more important to reverse logistics decision-making, enterprise can make decisions according to this.

This chapter provides a method about decision-making in manufacturing. ANP can be used in many intelligent decision-making fields, such as third-party reverse logistics suppliers' selection, reverse logistics system evaluation and so on. Qualified simulation is helpful for decision-making, different enterprises and different industry may set the parameters which relate to decision-making itself. However, ANP method has its limitations: first, managers need to be familiar with

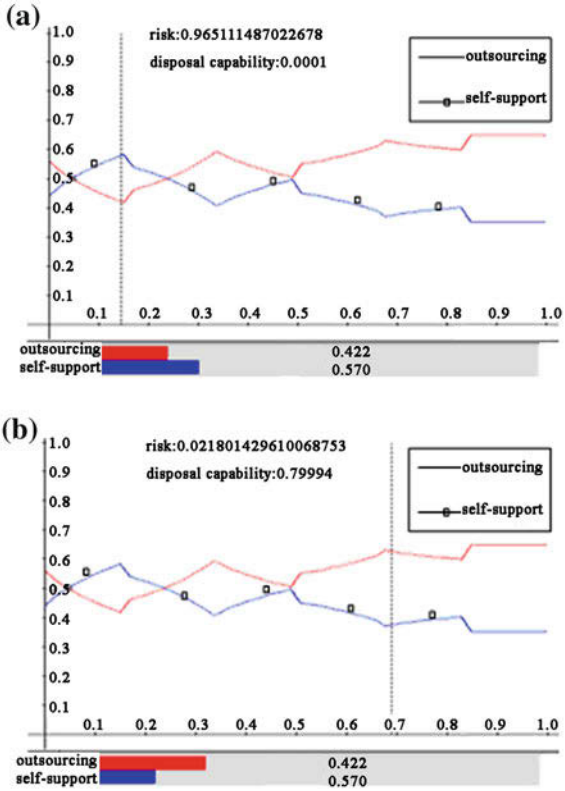


Fig. 121.5 Sensitivity analysis considering two combined elements as experimental variables

"what–if" judgment method, and the results of model highly depend on the weight assignment; second, it would be a complex job when factors in ANP network are a lot.

Acknowledgments We all thank China National Funds of Social Science for its help. (10GJ451-086).

References

1. Tang ZK (2005) Products recall and reverse logistics. Mod Manag 1:26–29
2. Liu PP, Pei BC (2009) Reverse logistics outsourcing risk and theory foundation. Commer Times 8:14–15
3. Zhou HX (2008) Transmission cost and logistics outsourcing. Foreign Invest China 8:139
4. Yang J, Wang Y (2004) Logistics decision-making considering transmission cost. Commer Econ Manag 150:23–25
5. Wang LF (2001) The theory and algorithm of analytic network process. Syst Eng Theory Pract 21:44–50

6. Satty TL (2001) Decision making with dependence and feedback: the analytic network process, 2nd edn. RWS Publications, Pittsburgh
7. Satty TL (2004) Decision making—the analytic hierarchy and network processes (AHP/ANP). J Syst Sci Syst Eng 13:1–35
8. Laura M, Joseph S (1998) Strategic analysis of logistics and supply chain management systems using the analytical network process. Transpn Res-E (Logist Transpn Rev) 34: 201–215
9. Shuai B, Sun CY (2006) Critical effect of decision making for logistics outsourcing of manufacturing enterprises. J Southwest Jiaotong Univ 41:296–299
10. Chen QS (2005) Modern logistics system. China Water Power Press, Beijing
11. Zhang G (2006) Supply chain revolution in HP. Logist Infor 1:61–63

Part XI
Network Components and Application

Chapter 122
Determining Similarity Between Concepts in Corpus

Li Chen, Zi-lin Song, Zhuang Miao and Cheng-jian Wang

Abstract The research on concept similarity plays a very important role in information retrieval, artificial intelligence and so on. In this paper, we focus on the method to measure the similarity between concepts using a corpus. We propose an approach to measure concept similarity using the association rule mining in a corpus. With the improvement of the most influential algorithm Apriori, we can measure the similarity between concepts in a corpus fast and precisely.

Keywords Concept similarity · Association rule mining · Association similarity · Apriori algorithm

122.1 Introduction

The study of concept similarity has long been an integral part of information retrieval, artificial intelligence and cognitive science. Whether in academia or industry, the measurement of similarity between concepts has been widely used, such as word sense disambiguation [1], detection and correction of word spelling errors (malapropisms) [2], images retrieval [3], documents retrieval [4], automatic hypertext links [5], etc. In particular the concept similarity has been used in

L. Chen (✉) · Z. Song · Z. Miao
Institute of Command Automation, PLA University of Science and Technology,
Nanjing 210007, China
e-mail: ivan_chenli@tom.com

C. Wang
Institute of Meteorology, PLA University of Science and Technology,
Nanjing 211101, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_122,

Semantic Web- related applications such as automatic annotation of Web pages [6], community mining [7], and keyword extraction [8].

The majority of the semantic similarity metrics employed today uses hand crafted resources, e.g., ontology and most commonly WordNet. The use and maintenance of ontology is a costly task. Also the resources are not ubiquitous and provide no information for words or concepts not included in the ontology. Continuously updating the ontology is a time-consuming and tedious task, demanding human labor and often expert knowledge.

Recently there has been much research interest in developing corpus-based approaches for estimating semantic similarity for concepts. Most approaches share a common assumption: similar concepts have similar distributional behavior in a corpus.

In this paper we mainly focus on the method to measure the similarity between concepts using a corpus. We propose an approach to measure concept similarity using the association rule mining in a corpus. With the improvement of the most influential algorithm Apriori, we can measure the similarity between concepts in a corpus fast and precisely.

The remainder of this paper is organized as follows: In Sect. 122.2 the knowledge of association rules in data mining is introduced. Then we propose an approach to measure association similarity in Sect. 122.3. The example of our approach is given in Sect. 122.4. Section 122.5 gives some related work. The last section presents the conclusions and the future work.

122.2 Association Rule Mining

Association rule mining [9, 10] is an important aspect of data mining. Association rule mining search for interesting relationships among items in a given data set. Let $J = \{i_1, i_2, \ldots, i_m\}$ be a set of all items (in this paper a concept is an item). Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq J$. Each transaction is associated with an identifier, called *TID*. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset J$, $B \subset J$, and $A \cap B = \varnothing$ The rule $A \Rightarrow B$ holds in the transaction set D with support s, where s is the percentage of transactions in D that contain $A \cup B$ (i.e., both A and B). This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence c in the transaction set D if c is the percentage of transactions in D contain A that also contain B. This is taken to be the conditional probability, $P(B|A)$. That is,

$$\text{support}(A \Rightarrow B) = P(A \cup B) \qquad (122.1)$$

$$\text{confidence}(A \Rightarrow B) = P(B|A) \qquad (122.2)$$

Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong.

A set of items is referred to as an item set. An item set that contains k items is a $k -$ itemset. The occurrence frequency of an item set is the number of transactions that contain the item set. This is also known, simply, as the frequency, support count, or count of the itemsst. An itemset satisfies minimum support if the occurrence frequency of the item set is greater than or equal to the product of min_sup and the total number of transactions in D. The number of transactions required for the itemset to satisfy minimum support is therefore referred to as the minimum support count. If an item set satisfies minimum support, then it is a frequent itemset. The set of frequent $k -$ itemset is commonly denoted by L_k.

Association rule mining is a two-step process:

1. Find all frequent itemsets: By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count.
2. Generation strong association rules from the frequent item sets: By definition, these rules must satisfy minimum support and minimum confidence.

Apriori is an influential algorithm for mining frequent itemsets. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties. Apriori employs an iterative approach known as a level-wise search, where $k -$ itemsets are used to explore$(k + 1) -$ itemsets. First, the frequent $1 -$ itemsets is found. This set is denotedL_1. L_1 is used to find L_2, the frequent $2 -$ itemsets, which is used to find L_3, and so on, until no more frequent $k -$ itemsets can be found. The finding of each L_k requires one full scan of database. In order to use the Apriori property, all nonempty subsets of a frequent item set must also be frequent. This property is based on the following observation. By definition, if an item set I does not satisfy the minimum support threshold, min_sup, then I is not frequent, that is, $P(I) < \text{min\_sup}$. If an item A is added to the item set I, then the resulting item set (i.e. $I \cup A$) cannot occur more frequently than I. Therefore, $I \cup A$ is not frequent either, that is $P(I \cup A) < \text{min\_sup}$.

122.3 Association Similarity

The Apriori algorithm is only used to find the association rules; we need another method to measure the similarity between concepts.

Let $FC = \{fc_1, fc_2, \ldots, fc_n\}$ be the set of feature concepts, $C = \{I_a, I_b\}$ be the set of testing concepts (the measurement of similarity cannot but process between two concepts), the vector of testing concept can be denoted as:

$$V_{I_a} = \{\text{con}(fc_1 \Rightarrow I_a), \text{con}(fc_2 \Rightarrow I_a), \ldots, \text{con}(fc_n \Rightarrow I_a)\} \tag{122.3}$$

$$V_{I_b} = \{\text{con}(fc_1 \Rightarrow I_b), \text{con}(fc_2 \Rightarrow I_b), \ldots, \text{con}(fc_n \Rightarrow I_b)\} \tag{122.4}$$

Table 122.1 The given corpus

TID	Concept sets	TID	Concept sets
01	$I_1, I_7, I_8, I_9, I_{14}$	06	I_1, I_4, I_7, I_{12}
02	I_2, I_3, I_9, I_{11}	07	$I_1, I_{10}, I_{13}, I_{14}$
03	$I_1, I_5, I_7, I_8, I_{12}$	08	I_2, I_{10}, I_{14}
04	$I_1, I_3, I_5, I_6, I_{13}$	09	I_5, I_6, I_7, I_{12}
05	$I_1, I_3, I_4, I_9, I_{14}$	10	I_1, I_2, I_5, I_7, I_8

Table 122.2 Candidate1 $-$ *concept sets*

Concept sets	Support count	Concept sets	Support count
I_1	7	I_8	3
I_2	3	I_9	3
I_3	3	I_{10}	2
I_4	1	I_{11}	1
I_5	4	I_{12}	3
I_6	2	I_{13}	2
I_7	5	I_{14}	4

Let V_{I_a} and V_{I_b} be the vector of testing concept I_a and I_b, the association similarity between I_a and I_b can be denoted as:

$$\{\text{sim}(I_a, I_b) = \frac{\sum_{i=1}^{n} V_{I_a i} \times V_{I_b i}}{\sqrt{\left(\sum_{i=1}^{n} V_{I_a i}^2\right)\left(\sum_{i=1}^{n} V_{I_b i}^2\right)}} \qquad (122.5)$$

122.4 The Example of Algorithm

In order to interpret the algorithm more clearly, we give a concrete example. Let $I = \{I_1, I_2, \ldots, I_{14}\}$ be the concept set which contains 14 concepts. There are ten contexts in a given corpus (note that the corpus in reality is much larger). Each context contains one or more concepts. The support count threshold $\min\sigma = 3$ (for simplicity, we use support count threshold here, corresponding support thresholdmin_sup $= 30\%$). We need to measure the similarity between the concept I_7 and the concept I_8.

(1) The given corpus; Table 122.1
(2) Scan the corpus and generate candidate1 $-$ concept sets; Table 122.2
(3) According to min_sup,. generate frequent 1 $-$ concept sets; Table 122.3
(4) Remove the testing concept I_7, I_8; Table 122.4
(5) Connect and generate the candidate2 $-$ concept sets; Table 122.5
(6) Generate the vector of testing concept:

Table 122.3 Frequent $1 - concept\,sets$

Concept sets	Support count	Concept sets	Support count
I_1	7	I_8	3
I_2	3	I_9	3
I_3	3	I_{12}	3
I_5	4	I_{14}	4
I_7	5		

Table 122.4 Processed frequent $1 - concept\,sets$

Concept sets	Support count	Concept sets	Support count
I_1	7	I_9	3
I_2	3	I_{12}	3
I_3	3	I_{14}	4
I_5	4		

Table 122.5 Candidate2 $- concept\,sets$

Concept sets	Support count	Concept sets	Support count
(I_7, I_1)	4	(I_8, I_1)	3
(I_7, I_2)	1	(I_8, I_2)	1
(I_7, I_3)	0	(I_8, I_3)	0
(I_7, I_5)	3	(I_8, I_5)	2
(I_7, I_9)	1	(I_8, I_9)	1
(I_7, I_{12})	3	(I_8, I_{12})	1
(I_7, I_{14})	1	(I_8, I_{14})	1

$$\text{confidence}(I_k \Rightarrow I_7) = \frac{\sigma(I_k \cup I_7)}{\sigma(I_k)} \times 100\%$$

$$\text{confidence}(I_k \Rightarrow I_8) = \frac{\sigma(I_k \cup I_8)}{\sigma(I_k)} \times 100\%$$

$$V_{I_7} = \{0.57, 0.33, 0, 0.75, 0.33, 1, 0.25\}$$

$$V_{I_8} = \{0.43, 0.33, 0, 0.5, 0.33, 0.33, 0.25\}.$$

(7) Finally calculate the association similarity between I_7 and I_8

$$\text{sim}(I_7, I_8) = \frac{\sum_{i=1}^{n} V_{7i} \times V_{8i}}{\sqrt{\left(\sum_{i=1}^{n} V_{7i}^2\right)\left(\sum_{i=1}^{n} V_{8i}^2\right)}} = 0.922.$$

122.5 Related Work

The approaches for semantic similarity based on ontology can be grouped into edge-based measures which consider the length of the paths that link the words, as well as the positions of words in the taxonomic structure[11], Information Content measures which find the difference of the contextual information between words as a function of their occurrence probability with respect to a corpus[12]. Hybrid methods combine synsets with word neighborhoods and other features[13] .

On the other side, corpus-based measures are based on a statistical analysis of a large text corpus. They have the advantage of being self-independent; they do not need any external knowledge resource, which can overcome the coverage problem in ontology-based measures. In this category, we can find co-occurrence-based measures [14] [15] and context-based measures [16][17].

Turney [15] presents a simple unsupervised learning algorithm for recognizing synonyms, based on statistical data acquired by querying a Web search engine. The algorithm, called PMI-IR, uses Pointwise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words. However, co-occurrence refers to the more general phenomenon of concepts that are likely to be used in the same context instead of the similarity of two concepts.

Sahami [16] measured similarity between two queries using snippets returned for those queries by a search engine. For each query, they collected snippets from a search engine and represented each snippet as a TF-IDF- weighted term vector. Similarity between two queries was then defined as the inner product between the corresponding vectors. They did not compare their similarity measure with ontology-based similarity measures.

122.6 Conclusions and Future Work

In this paper we first give the definition of the concept similarity. Then we propose an approach to measure association similarity using improved Apriori algorithm which is usually used to mining association rules in large database. According to our approach, we can measure the similarity between concepts fast and precisely.

Our future work is about the improvement of similarity measurement including the better corpus selection and combining our approach with ontology. The application of our research will also be the future work in our plans. The hybrid methods to measure the similarity between concepts can be used in natural language processing, information retrieval, semantic Web and is worthwhile for deeper research.

References

1. Agirre E, Rigau G (1994) A proposal for word sense disambiguation using conceptual distance. In: Proceedings of international conference on recent avances in natural language processing, Bulgaria pp 1–5
2. Carpuat M, Fung P, Ngai G (2006) Aligning word senses using bilingual corpora. ACM Trans Asian Lang Inf Process 5:89–120
3. Yang M, Chen J (2008) A new similarity measurement based on distance and correlation test for content-based images retrieval. In: Proceedings of 2008 congress on image and signal processing. IEEE
4. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on artificial intelligence (AAAI'06). American Association for Artificial Intelligence (AAAI Press), Boston, pp 775–780
5. Green SJ (1999) Building hypertext links by computing semantic similarity. IEEE Trans Knowl Data Eng 11:713–730
6. Cimano P, Handschuh S, Staab S (2004) Towards the self-annotating web. In: Proceedings of the 13th international conference on World Wide Web (WWW'04), ACM
7. Mika P (2005) Ontologies are us: a unified model of social networks and semantics. In: Proceedings of international semantic web conference (ISWC'05)
8. Mori J, Matsuo Y, Ishizuka M (2007) Extracting keyphrases to represent relations in social networks from web. In: Proceedings of the 20th international joint conference on artificial intelligence
9. Agrawal R, Imielinski T, Wami AS (1993) Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD conference on management of data. ACM, pp 207–216
10. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of 20th international conference on very large data bases (VLDB'94), ACM
11. Li YH, Bandar ZA, McLean D (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15:871–882
12. Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the 10th international conference on research on computational linguistics (ROCLING X'97), Taiwan, China, pp 8–22
13. Petrakis GME, Varelas G, Hliaoutakis A, Raftopoulou P (2006) X-Similarity: computing semantic similarity between concepts from different ontologies. J Digit Inf Manag
14. Church KW, Hanks P (1989) Word association norms, mutual information, and lexicography. In: Proceedings of the 27th. annual meeting of the association for computational linguistics. Association for Computational linguistics pp 76–83
15. Turney PD (2001) Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In: Proceedings of the twelfth European conference on machine learning (ECML'01), LNCS 2167, Springer
16. Sahami M, Heilman T (2006) A Web-based kernel function for measuring the similarity of short text snippets. In: Proceedings of the 15th international conference on World Wide Web (WWW'06), ACM
17. Carmel D, Farchi E, Petruschka Y, Soffer A (2002) Automatic query refinement using lexical affinities with maximal information gain. In: Proceedings of SIGIR'02, ACM

Chapter 123
Measuring Similarity Between Concepts Based on Ontology

Li Chen, Ying Zhang, Zi-Lin Song and Zhuang Miao

Abstract Similarity denotes the relatedness of two entities. Determining similarity of two entities is an important problem in the domain of software engineering, artificial intelligence, information retrieval, Web service. In this paper, we analyze the construction of HowNet which is a widely used Chinese-English bilingual ontology and propose a novel approach for calculating the similarity of sememes which are the most important element of HowNet. Then we calculate concept similarity in the foundation of sememes similarity. Compared with the other WordNet-based methods, our HowNet-based approach can measure the similarity between concepts precisely as well.

Keywords HowNet · Sememes superposition · Sememes difference · Hierarchy depth · Concept similarity

123.1 Introduction

The word "ontology" has a long history in philosophy known as metaphysics, which deals with the nature of the reality of what exists. But if mentioned in computer science, the ontology is an explicit specification of a conceptualization, that is, ontology is a description (like a formal specification of a program) of concepts and relationships [1, 2].

L. Chen (✉) · Y. Zhang · Z.-L. Song · Z. Miao
Institute of Command Automation, PLA University of Science and Technology,
Nanjing 210007, China
e-mail: ivan.chen@163.com

Y. Zhang
e-mail: zhywl66@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_123,

Research on ontology is becoming increasingly widespread in the computer science community. Currently one of the most active areas of research is calculating similarity between concepts in ontology. The similarity calculation is used in a wide range of fields such as semantic web, search engines, natural language processing, knowledge engineering, information extraction and retrieval. However, how to improve the precision of similarity calculation has been the bottleneck of using the ontology.

Many approaches for similarity calculation based on the ontology have been proposed in the literature. These are classified into three main categories. Edge Counting Methods: Measure the similarity between two terms (concepts) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy [3, 4]. Information Content Methods: Measure the difference in information content of the two terms as a function of their probability of occurrence in a corpus [5–7]. In this work WordNet [8] is used as a statistical resource and information content is computed according to the probabilities of occurrence of terms. This approach is independent of the corpus and also guarantees that the information content of each term is less than the information content of its subsumed terms. This constraint is common to all methods of this category. Computing information content from a corpus does not always guarantee this requirement. Hybrid methods combine the above ideas [9–11]. All the approaches mentioned above use WordNet as the knowledge base. The WordNet which is based on synsets is constructed to allow a user to easily distinguish between different senses of a word and to represent one underlying concept.

Besides WordNet, there is another important ontology called HowNet [12]. It is a Chinese–English bilingual ontology which provides plenty of knowledge for natural language processing applications [13]. As the concepts described in HowNet are represented in both English and Chinese, the methods based on HowNet are able to measure similarity between two concepts in different languages. This contributes a lot to cross-lingual applications such as machine translation and cross-lingual information retrieval. In this paper, we analyze the construction of HowNet and propose a novel approach for calculating the similarity of sememes which are the most important elements of HowNet. In the foundation of sememes similarity, we calculate concept similarity using improved Hungarian algorithm.

The remainder of this paper is organized as follows. In Sect. 123.2 we introduce the construction of HowNet. In Sect. 123.3, we propose a novel approach for calculating the sememes similarity. Then we introduce the Hungarian algorithm and improve it for calculation of concept similarity. The last section presents the conclusions and future work.

123.2 Main Knowledge about HowNet

In contrast to WordNet, HowNet is built using a constructive approach: basic units of meaning which are called sememes are used to build up concept definitions and do not put all of the concepts directly into a tree, but describes them by a set of sememes. The philosophy behind HowNet is based on the idea that all things are in constant motion and are ever changing in a given time and space. Attributes can be used to record the changes as things evolve from one state to another. A concept can, therefore, be defined by setting the values of different attributes for each thing. Given a set of sememes representing things, attributes, and their values, HowNet gives a definition for any word and any of their particular senses.

There are more than 2000 sememes in the current version of HowNet. The sememes are classified into 10 categories. The sememes in each category are organized in a hierarchical tree.

123.3 Calculation of Sememes Similarity

As we have described above, HowNet do not put all of the concepts directly into a tree, so we cannot calculate the similarity between two concepts directly. In HowNet, the concepts are described by a set of sememes. We may calculate the similarity between two sememes in advance.

The sememes in HowNet are organized in a hierarchical tree. So we may calculate the sememes similarity through this tree. Three main factors which influence the calculation are discussed below.

Sememes Superposition which is denoted as $s(a_i, a_j)$, where a_i and a_j represent the testing sememes. The Sememes Superposition is the number of the same upper sememes which are above the testing sememes. The Sememes Superposition indicates the homology of testing sememes. Let $r(a_i)$ be the set of sememes from a_i to root and $||r(a_i)||$ represent the number of all items in the set $r(a_i)$, then$s(a_i, a_j) = ||r(a_i) \cap r(a_j)||$. We give a simple hierarchical tree of sememes as shown in Fig. 123.1 to explain the calculation. In Fig. 123.1, according to the definition above $||r(n3)|| = 3$, $||r(n4)|| = 3$, $||r(n8)|| = 4$ and $s(n3, n4) = 2$, $s(n3, n8) = 3$. The Sememes Superposition is proportional to the similarity between testing sememes.

Sememes Difference which is denoted as $d(a_i, a_j)$. The Sememes Difference is the subtraction between the number of all upper sememes and the Sememes Superposition. The Sememes Difference indicates the difference of testing sememes. Mathematically, $d(a_i, a_j) = ||r(a_i) \cup r(a_j)|| - ||r(a_i) \cap r(a_j)||$. In Fig. 123.1, $d(n3, n4) = 2$, $d(n3, n8) = 1$. The Sememes Difference is inversely proportional to the similarity between testing sememes.

Hierarchy Depth which is denoted as $h(a_i)$. The Hierarchy Depth is the hierarchy of sememes in the sememes tree. The similarity of testing sememes is proportional to the total summation of the hierarchy depth of testing sememes, and inversely proportional to the subtraction of the hierarchy depth of testing sememes, such that in Fig. 123.1, $\text{sim}(n8, n9) > \text{sim}(n3, n4)$ and $\text{sim}(n8, n3) > \text{sim}(n8, n1)$.

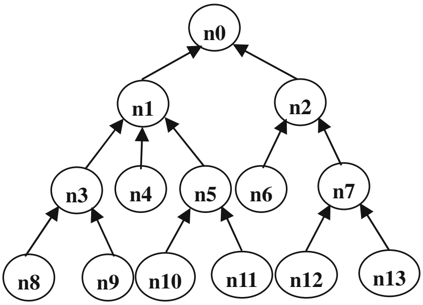


Fig. 123.1 A simple hierarchical tree

Taking consideration of all these factors above, we propose the equation for calculating the sememes similarity as follows:

$$\mathrm{sim}'(a_i, a_j) = \frac{s(a_i, a_j)\tau(h(a_i) + h(a_j))}{d(a_i, a_j) \times (|h(a_i) - h(a_j)| + 1)} \tag{123.1}$$

where $s(a_i, a_j)$ represents the sememes superposition; $d(a_i, a_j)$ represents the sememes difference; $h(a_i), h(a_j)$ represents the hierarchy depth; τ represents the adjustable coefficient.

The next work is to normalize the result of our calculation so that the range of similarity is [0, 1]. The proposed equation for normalization is as follows:

$$\mathrm{sim}(a_i, a_j) = 1 - \omega^{-\mathrm{sim}'(\mathrm{a_i}, \mathrm{a_j})} \tag{123.2}$$

where ω represents the normalization parameter whose value is greater than 1. The greater theω the faster our result close to 1.

123.4 Determining Concept Similarity Using Improved Hungarian Algorithm

123.4.1 Mathematical Model of Concept Similarity

In HowNet, the concepts are described by a set of sememes. The calculation of similarity between concepts can be derived from the similarity between sememes.

Definition: sememes matrix (abbreviated as *SM*). Let $(a_1, a_2, \ldots, a_n)$ represent the set of sememes which construct the concept CP_a, and $(b_1, b_2, \ldots, b_m)$ represents the set of sememes which construct the concept CP_b. In order to calculate the similarity between CP_a and CP_b, the *SM* is constructed as follows, where $\mathrm{sim}(a_i, b_j)$ represents the sememes similarity, i.e. $SM = (\mathrm{sim}(a_i, b_j))_{n \times m}$

$$\begin{array}{ccccc} (\mathrm{sim}(a_i,b_j))_{n\times m} & b_1 & b_2 & \cdots & b_m \\ a_1 & \mathrm{sim}(a_1,b_1) & \mathrm{sim}(a_1,b_2) & \cdots & \mathrm{sim}(a_1,b_m) \\ a_2 & \mathrm{sim}(a_2,b_1) & sim(a_2,b_2) & \cdots & \mathrm{sim}(a_2,b_m) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_n & \mathrm{sim}(a_n,b_1) & \mathrm{sim}(a_n,b_2) & \cdots & \mathrm{sim}(a_n,b_m) \end{array} \tag{123.3}$$

Then the calculation of similarity between concepts can be transformed into finding the optimum solution of assignment problem from the sememes matrix. When we calculate the concept similarity denoted as $\mathrm{sim}(CP_a, CP_b)$, it is required to build a sememes assignment plan which fits one to one matching between a_i and b_j. To build the mathematical model, we import 0-1 variables:

$$x_{ij} = \begin{cases} 1 & a_i \text{ match with } b_j \\ 0 & a_i \text{ not mach with } b_j \end{cases} (i = 1,\ldots,n; j = 1,\ldots,m) \tag{123.4}$$

The mathematical model for calculating concept similarity is shown as follows:

$$\mathrm{sim}(CP_a, CP_b) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\mathrm{sim}(a_i, b_j)$$
$$s.t.\begin{cases} \sum_{i=1}^{n} x_{ij} = 1 & (j = 1,\ldots,m) \\ \sum_{j=1}^{m} x_{ij} = 1 & (i = 1,\ldots,n) \\ x_{ij} = 0\, or\, 1 & (i = 1,\ldots,n; j = 1,\ldots,m) \end{cases} \tag{123.5}$$

The key to this problem is making use of the property of Hungarian algorithm [14]: if we subtract a constant k from each element of some rows (or some columns) in SM and get a new matrix $SM^{(0)}$, then the two matrixes have the same optimum solution of assignment problem.

123.4.2 Improvement of Hungarian Algorithm

However, the Hungarian algorithm is designed to solve minimum assignment problems and may not adapt to the problem in this paper because we need to calculate maximum concept similarity. We make a matrix transformation to solve the problem in this paper. Supposing z is the maximum element of the matrix SM, let $\mathrm{matrixSM}' = (\mathrm{sim}(a_i, b_j)')_{n\times m} = (z - \mathrm{sim}(a_i, b_j))_{n\times m}$, then the minimum assignment problems of SM' has the same optimum solution as the maximum assignment problems of SM.

There is another problem we have met with so we need to propose another approach to solve it. In practice, the number of sememes which construct the concepts is always unequal to each other, i.e. $n \neq m$. In this condition, we need to preprocess the sememes matrix. There are mainly two approaches:

Addition of 0. In the sememes matrix *SM*, supposing$n > m$, we add several vacant sememes. The number of vacant sememes is $n - m$ and let $\text{sim}(a_i, b_k) = 0$ where $k = m + 1, m + 2, \ldots, n$ and $i = 1, 2, \cdots n$. This method imports vacant sememes and sets the similarity between vacant sememes and other sememes is 0.

Subtraction of the last. Supposing $A_n = \{a_1, a_2, \ldots, a_n\}$, there are m sorting schemes for A_n which sort ascendingly by the value of elements in A_n and it can be denoted as $t = \{a_o, a_p, \ldots, a_q\}$. The list of sorting schemes can be denoted as$\{t_1, t_2, \ldots, t_m\}$. For any element a_i of A_n in any sorting scheme t_j, we define its grade as $C_j(a_i)$ equals to the amount of the elements in t_j which follow a_i. For a_i, its finally sorting grade is $C(a_i) = \sum_{k=1}^{m} C_k(a_i)$. Now delete $(n - m)a_i$ at the end of this sorting. In order to reduce the inaccuracy, we set a threshold denoted as σ. If in the a_i which is to be deleted, if there is no data greater than σ, then keep this a_i alive until deleting $(n - m)a_i$. The example is as follows:

$\text{sim}(a_i, b_j)_{n \times m}$	b_1	b_2	b_3	*total*	$\text{sim}(a_i, b_j)_{n \times m}$	b_1	b_2	b_3	*total*
a_1	0.75^2	0.82^3	0.65^2	7	a_1	0.75^2	0.82^3	0.65^2	7
a_2	0.95^3	0.25^0	0.3^0	3	a_2	0.95^3	0.25^0	0.3^0	3
a_3	0.1^0	0.38^1	0.92^3	4	a_3	0.1^0	0.38^1	0.92^3	4
a_4	0.45^1	0.75^2	0.5^1	4	a_4	0.45^1	0.75^2	0.5^1	4

As shown on the left, if we do not set σ, a_2 is deleted. On the contrary, if we set $\sigma = 0.9$, a_4 is deleted. In our example, it is more reasonable to delete a_4 than a_2.

The method of addition of 0 imports vacant sememes and makes the result smaller than the ideal value. The method of subtraction of the last wipes off several sememes which are at the end of the sorting and make the result greater than the ideal value. We use these two methods synthetically. Let $\text{sim}(CP_a, CP_b)_1$ be the result of addition of 0 and $\text{sim}(CP_a, CP_b)_2$ be the result of subtraction of the last, then the finally result can be denoted as:

$$\text{sim}(CP_a, CP_b) = \mu \text{sim}(CP_a, CP_b)_1 + \lambda \text{sim}(CP_a, CP_b)_2 \tag{123.6}$$

Where $\mu + \lambda = 1, 0 < \mu < 1, 0 < \lambda < 1$.

123.5 Conclusions and Future Work

In this paper, we analyze the methods for calculating concept similarity based on ontology and introduce an important bilingual ontology called HowNet. Compared with WordNet, HowNet does not put all of the concepts directly into a tree, but describes them by a set of sememes. In this foundation, we first propose an approach to calculate sememes similarity with the factor of sememes superposition, sememes difference and hierarchy depth. Then we calculate concept similarity using improved Hungarian algorithm. Our HowNet-based approach can measure the similarity between concepts as well as those WordNet-based methods. Our plans for future work include the improvement of our algorithm for more precise similarity calculation and apply our research to cross-lingual applications such as machine translation and cross-lingual information retrieval.

References

1. Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5:199–220
2. Gruber TR (1995) Towards principles for the design of ontologies used for knowledge sharing. Int J of Hum-Comput Stud 43:907–928
3. Rada R, Mili H, Bicknell E, Blettner M (1989) Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybern 19:17–30
4. Li YH, Bandar ZA, McLean D (2003) An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans Knowl Data Eng 15:871–882
5. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene ontology: the relationship between sequence and annotation. Bioinformatics 19:1275–1283
6. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the 21st national conference on artificial intelligence (AAAI'06). American Association for Artificial Intelligence, Boston, pp 775–780
7. Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)
8. Miller GA (1995) WordNet: a lexical database for english. Commun ACM 38:39–41
9. Formica A (2009) Concept similarity by evaluating information contents and feature vectors a combined approach. Commun ACM 52:145–149
10. David J, Euzenat J, Šváb-Zamazal O (2010) Ontology similarity in the alignment space. In: Proceedings of International Semantic Web Conference (ISWC'10), LNCS 6469. Springer, pp 129–144
11. Maguitman AG, Menczer F, Roinestad H, Vespignani A (2005) Algorithmic detection of semantic similarity. In: Proceedings of the 14th international conference on World Wide Web (WWW'05). ACM, Chiba, Japan pp 107–116
12. Dong ZD, Dong Q (2006) HowNet and the Computation of Meaning. World Scientific, Singapore
13. Liu Q, Li SJ (2005) Word semantic similarity computation based on HowNet. In: Proceedings of the 3rd Chinese lexical and semantic proseminar. Taiwan, China
14. Zhao QC, Yan CB (2008) Advances in Assignment Problem and Comparison of Algorithms. In: Proceedings of 27th Chinese Control Conference (CCC'08). IEEE, Kunming, China, pp 607–611

Chapter 124
An Attributes-Based Access Control Architecture within Large-Scale Device Collaboration Systems Using XACML

Feng Liang, Haoming Guo, Shengwei Yi, Xiaoqiang Zhang and Shilong Ma

Abstract Containing multiple domains and a large number of heterogeneous distributed devices, large-scale device collaboration systems require a fine-grained, flexible and secure mechanism for device access control. This chapter presents and evaluates a distributed device access control architecture Multiple Policies supported Attribute-Based Access Control (MPABAC) to support device authentication and authorization among multiple domains. Based on eXtensible Access Control Markup Language (XACML) standard and Attribute-Based Access Control (ABAC) model, this architecture supports cross-domain authentication and authorization, hierarchical policy combination and enforcement, unified device access control and fine-grained attributes-based privilege description. Experiments show that the performance of this implementation is acceptable within the production environment.

Keywords Large-scale device collaboration system · Hierarchical policy decision point · Multiple policies attributed-based access control

F. Liang (✉) · H. Guo · S. Yi · X. Zhang · S. Ma
State Key Laboratory of Software Development Environment,
Beihang University, 100191 Beijing, China
e-mail: liangfeng@nlsde.buaa.edu.cn

H. Guo
e-mail: guohm@nlsde.buaa.edu.cn

S. Yi
e-mail: yisw@nlsde.buaa.edu.cn

X. Zhang
e-mail: zhangxq@nlsde.buaa.edu.cn

S. Ma
e-mail: slma@nlsde.buaa.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_124,

124.1 Introduction

Represented by the "Internet of Things", large-scale device collaboration systems are already applied into fields such as smart area management, disaster detection and analysis, intelligent resource planning, etc., where the environmental data can be obtained through the terminal devices and exchange via the open internet communication.

Large-scale device collaboration systems usually contain large numbers of heterogeneous devices and therefore need to process large-scale real-time tasks and complex collaboration processes. For example, the National seismological precursory network project achieved the collaborative observation of nearly 1,000 seismological precursor devices, which come from the subdomains within 30 provinces, 300 stations in the scope of the nationwide. The landscape lighting control system of the Olympic Central Area needs to process the orchestration of more than 20,000 lightings to reach the artistic lighting and the lighting devices are controlled by different subsystems. In order to guarantee the security of device access, authentication and authorization, the large-scale device collaboration systems need functionalities such as cross-domain authentication, dynamic authorization and universal device description. But there are currently many challenges against these goals, such as existing coarse-grained device access mechanism, device heterogeneity, multiple policies combination and performance issue.

All these deficits require a novel access control mechanism, which the traditional identity-based access control models such as Discretionary Access Control (DAC) [7], Mandatory Access Control (MAC) [6], Role Based Access Control (RBAC) [9] are not effective because cross-domain authentication and authorization and more fine-grained policy description are required. Attribute-based access control(ABAC) [12] is a more flexible and scalable access control model as it is based on the attributes from user and resources. This chapter proposes an attribute-based device access control architecture (MPABAC) to guarantee the device access control. This architecture supports cross-domain authentication and authorization, hierarchical policy combination and enforcement, unified device access control and fine-grained attributes-based privilege description.

This chapter is organized as follows. Section 124.2 describes the related work on device access control models, frameworks, and systems. Section 1243 demonstrates the formalized model of MPABAC. Section 124.4 presents the MPABAC Architecture and its implementation in detail. Section 124.5 gives experiments and performance evaluations and Sect. 124.6 draws the conclusion.

124.2 Related Work

Since the early 1990s, ABAC has appeared with the development of Internet-based distributed application and new security mechanisms such as Public Key Infrastructure (PKI) [3]. In ABAC, access decisions are based on attributes of the requestor and resource, and the authentication can be delayed until necessary because no pre-knowledge of the users is necessary by the resources. In the following, we introduce the ABAC in terms of models and algorithms, framework and systems.

LeMay et al. [5] introduced logic programming theory for modeling attribute-based access control system and policy maintenance, therefore improving the faster policy transformation. Yuan and Tong [12] proposed the ABAC model in terms of policy model and architecture model and presented the mathematical formulation of the policy model. Shen and Hong [10] proposed an attribute-based access control model WS-ABAC to use attributes associated with subject, object and environment, and service parameters for access control measures in Web Services environment. However, these works were not directly relevant to device access control. Lang et al. [4] proposed a flexible ABAC model called ABMAC in Grid Computing. However this work did not consider the fine-grained requirements of devices.

As one of the earliest work, Bonatti et al. [1] proposed an uniform attribute-based access control framework and model to regulate service access and information release in large-scale networks. Damian et al. [2] then presented a privacy-enhanced authorization model and language containing new elements such as Subject expression, Object expression, Actions and Conditions, Purposes and Obligations to provide anonymity, pseudonymity, and therefore improving authorization. However, these works were not directly relevant to device access control. Yu et al. [11] first realized a fine-grained attribute-based data access control framework for wireless sensor network, Fine-grained Distributed Access Control scheme (FDAC). However, FDAC considered only data access within sensor network; device control and monitoring are not considered.

124.3 MPABAC Model Formalization

As current large-scale device collaboration systems usually contain multiple domains, meanwhile, different from other resources, devices require more complex access control description, so it is essential to generate fine-grained access control policies and combine multiple policies from different domains to make a decision. Therefore we propose the MPABAC Model.

In MPABAC Model, access control decisions are made from the policies among multiple domains based on the attributes of entities such as subject, device, device manager, environment and actions. These entities and their attributes are described as below.

The entities of MPABAC Model:

a. Subject *sub*. A subject is the entity that sends the request to the Device and invokes the actions on the Device.
b. Device *dev*. A device refers to a physical device, containing the attributes of that device.

c. Environment *env*. Environment represents the required context information for making a policy decision. It contains information not related to any specific *sub* or *dev*.
d. Action *act*. An action is an operation provided by Device and it can be invoked by *sub*.

Suppose the maximum numbers of Subject, Device, Environment and Action are *A*, *B*, *C* and *D*, the maximum number of the attributes from these entities are *K*, *L*, *M*, *N*, then the sets of these entities and their attributes can be defined as follows:

$$SUB = \{sub_1, sub_2, \ldots, sub_a | 1 < a < A\} \tag{124.1}$$

$$DEV = \{dev_1, dev_2, \ldots, dev_b | 1 < b < B\} \tag{124.2}$$

$$ENV = \{env_1, env_2, \ldots, env_c | 1 < c < C\} \tag{124.3}$$

$$ACT = \{act_1, act_2, \ldots, act_d | 1 < d < D\} \tag{124.4}$$

$$SUBAttr = \{subAttr_1, subAttr_2, \ldots, subAttr_k | 1 < k < K\} \tag{124.5}$$

$$DEVAttr = \{devAttr_1, devAttr_2, \ldots, devAttr_l | 1 < l < L\} \tag{124.6}$$

$$ENVAttr = \{envAttr_1, envAttr_2, \ldots, envAttr_m | 1 < m < M\} \tag{124.7}$$

$$ACTAttr = \{actAttr_1, actAttr_2, \ldots, actAttr_n | 1 < n < N\} \tag{124.8}$$

As each local domain may employ different security mechanisms and therefore has its own policy description method, each policy is encapsulated as an independent atom policy to ensure the compatibility and scalability of MPABAC. The final decision is made of the combination of all these atom policies. What is more, as in some systems, the policies have different priorities for device control, so each MPABAC policy includes a priority PRI = $\{level_1, level_2, level_3, level_4, \ldots, level_o | 1 \le o \le O\}$ (O is the maximum number of the privileges). With all the defined entities and their attributes, the policies can be described as below:

A single policy can be described as $Policy_i = (Sub_i \times Dev_i \times Env_i \times Act_i, pri_i), Sub_i \subseteq SUB, Dev_i \subseteq DEV, Env_i \subseteq ENV, Act_i \subseteq ACT, pri_i \in PRI.$

$$\begin{aligned} Policy_i &\leftarrow (Sub_i \times Dev_i \times Env_i \times Act_i, pri_i) \\ &\leftarrow (f_{canAcces}s(Sub_i, Dev_i, Env_i, Act_i), pri_i) \\ &\leftarrow (f_{AttricanAccess}(SubAttr_i, DevAttr_i, EnvAttr_i, ActAttr_i), pri_i). \end{aligned} \tag{124.9}$$

The combine function:

$$\begin{aligned} Decision &\leftarrow f_{combine}(Policy_1, Policy_2, \ldots, Policy_n) \\ &\leftarrow f_{combine}((f_{AttricanAccess}(SubAttr_1, DevAttr_1, EnvAttr_1, ActAttr_1), pri_1), \\ &\quad (f_{AttricanAccess}(SubAttr_2, DevAttr_2, EnvAttr_2, ActAttr_2), pri_2), \\ &\quad (f_{AttricanAccess}(SubAttr_i, DevAttr_i, EnvAttr_i, ActAttr_i), pri_i)). \end{aligned} \tag{124.10}$$

124.4 Architecture and Implementation

124.4.1 MPABAC Architecture

The expressing, managing and enforcing authorizations for device access policies in a distributed environment require the presence of an architecture that supports distributed policy creation, evaluation and user authentication. eXtensible Access Control Markup Language (XACML) [8] defines a general policy description language and an access decision language. XACML is composed of Policy Administration Point (PAP), Policy Decision Point (PDP) and Policy Enforcement Point (PEP), Policy Information Point (PIP).

Figure 124.1 shows the MPABAC architecture including the XACML architecture. The Architecture can be divided into two layers, the Upper Layer and the Local Domain Layer. The Upper Layer is composed of the Authorization Engine and the MasterPDP, the Local Domain Layer is composed of Device Manager and multiple Local domains. Each domain contains its own LocalPDP, LocalPIP, LocalPAP and Local Attribute Authorities. The Authorization Engine includes a PEP and Interpreter, to receive the user request, Interpret it into authorization request, trigger the authorization request and enforce the authorization. The MasterPDP is responsible for parsing the authorization request into multiple XACML authentication request. Within each domain, the LocalPDP is to process the authentication with the LocalPAP, LocalPIP and Local Attribute Authorities. The Device Manager follows the command from Authorization Engine and generates the corresponding command script to conduct the access. Each time the user sends an access request, an access request is submitted to the MasterPDP from the Authorization Engine. The MasterPDP then analyzes the request and generates multiple XACML authentication request and distributes them into different LocalPDPs according to the domain each request belongs to. After that, the LocalPDP makes an authentication decision based on the attributes collected from all Local Attribute Authorities and the policies generated from LocalPAP within each domain. If LocalPDPs within all related domains permit the Request, then the MasterPDP will authorize the user with proper privileges to access the device via Device Manager.

As the large-scale device collaboration systems usually include more than one administrative domain, we employ a hierarchical structure for cross-domain authentication and authorization. The privilege of this loosely coupled distributed authentication structure is to support multiple policies among different domains.

124.4.2 Priority Description

XACML does not directly support priorities between different policies. But in production environment, because of administrative relationships, policies from

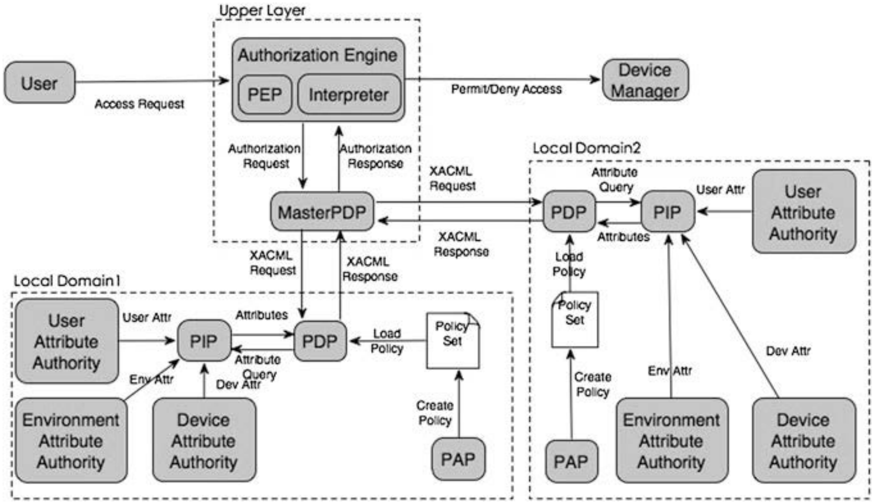


Fig. 124.1 The MPABAC architecture

different PAPs may not be equally important, it requires priority ranking among the multiple related policies when making a decision.

We categorize the policies into two different scopes, including the local domain scope (LocalPolicy) and the Meta layer scope (MetaLayerPolicy), then different priorities are set depending on the administration strategies. For example, in more central-controlled systems, the Upper Layer scope should own higher priority than the Local Domain Layer, so as to enforce the controlling strategies, and probably each local domain should own equal priority. In federation environment where multiple domains are more independent, the Meta layer and other domains should own lower priority than the local domain. We set Priority as the CombinerParameter in the XACML PolicySet description so that each policy is attached with a priority.

124.5 Experiments

Our test bed includes three machines, one is the upper layer server equipped with a Intel Core 2 Duo 2.66 CPU and 2 GB memory, other two are local domain servers with a1.66 GHz AMD processor and 1 GB RAM. All these machines are interconnected via switched gigabit Ethernet. The three machines are running a Debian Linux with 2.6.18 Kernel, with Sun JDK 1.6 as the Java platform.

124.5.1 Authentication Duration Test

The efficiency of our implementation depends largely on the time span of the authentication process. According to our algorithm design, the authentication

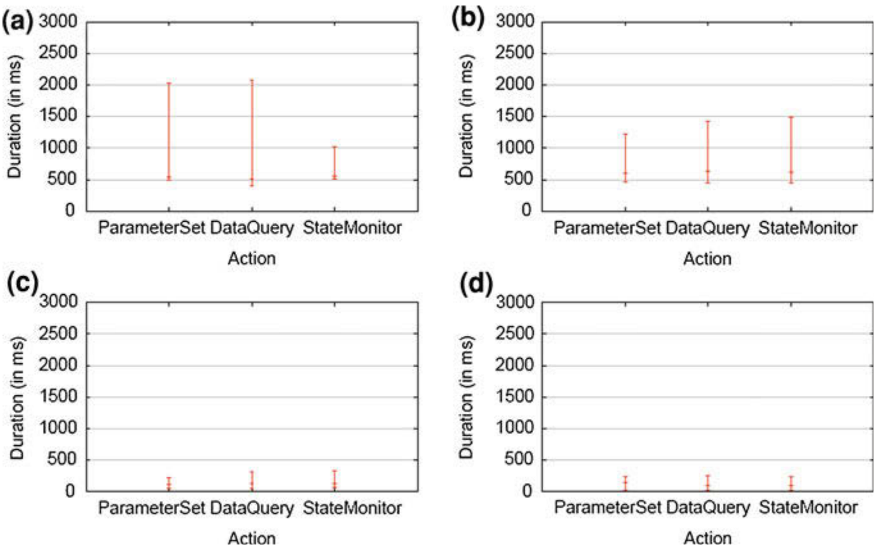


Fig. 124.2 The authentication duration test with MPABAC. **a** Multiple resources with multiple policies, **b** Multiple resources with one policy, **c** One resource with multiple policies, **d** One resource with one policy

process is influenced by both the number of resources and the number of policies, what is more, it takes different times to generate command scripts for different actions (ParameterSet, Data Query and State Monitor), therefore we conducted several experiments to test the duration of authentication process with all these three factors included.

As shown in Fig. 124.2, with one resource and different number of policies (in this test we take 30 policies), every action needs a relatively smaller duration (the mean value at around 120 ms), and the number of policies does not have a very obvious influence. The same rule stands for multiple resources (in this test we take 200 resources) with different number of policies (mean value at around 530 ms). However, there is a large increase in both mean time and standard deviation, for example, compared with the ParameterSet action in Fig. 124.2c), the same action in Fig. 124.2a) has an increase of 410 ms in the duration mean value and an increase of 965 ms in the standard deviation. So it is clear that the time spent for authentication process lasts longer when there are more resources involved.

124.5.2 Scalability Test

From the above test, we conclude that the number of resources has the most obvious impact on authentication duration, therefore we also conducted the scalability test, to test the authentication duration time with different number of resources.

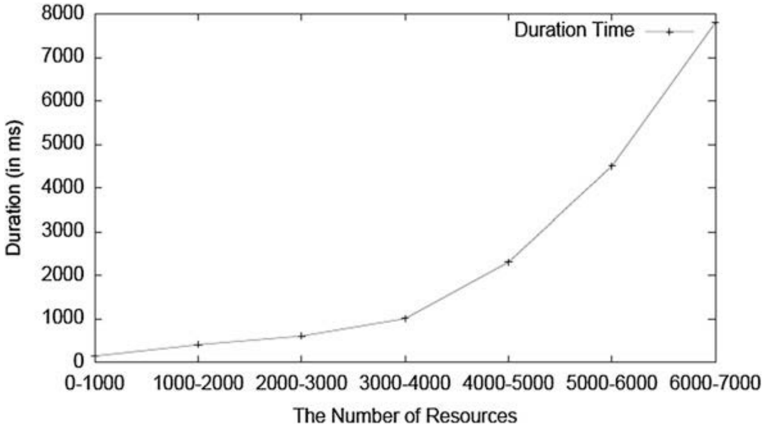


Fig. 124.3 Authentication duration test under different number of resources

In Fig. 124.3, we analyze the correlation between the number of resources and the duration time. Obviously the duration time increases with the number of resources. When the number of resources is under 4,000, the duration time is under 1 s and increases quite slowly, but after 4,000, the duration time increases very quickly. As most large-scale device collaboration systems have around 5,000 devices, this reflects that our results are acceptable in production environment.

124.6 Conclusion and Future Work

Fine-grained authentication and authorization in large-scale multi-domain device collaboration systems are important security issues. While the traditional MAC, DAC and RBAC models are not sufficient for this, ABAC can be a promising approach. In this chapter we proposed MPABAC architecture to realize this by supporting prioritized hierarchical policies combination and enforcement among multiple domains, unified device access control and fine-grained attributes-based privilege description.

Our experiments demonstrate that the overhead exposed by our system is acceptable and that the system scales under load. The duration time of the authentication process depends on the number of resources in the system. Our experiments show that the duration lasts for less than 1 s and scales quite well when the device number is under 4,000.

In the future, we will investigate more algorithms for policy combination and perform experimental assessments when applying it on real large-scale device collaboration system scenarios.

Acknowledgements This research work was supported by both the self-conducted exploratory research program "Green Lighting in Internet of Things " from State Key Laboratory for Software Development Environment in China (No. SKLSDE-2010ZX-06) and the Special Program for Seism-Scientific Research in Public Interest "Research in Online Processing Technologies for Seismological Precursory Network Dynamic Monitoring and Products" (No. 201008002).

References

1. Bonatti PA, Samarati P (2002) A uniform framework for regulating service access and information release on the web. J Comput Secur 10:241–271
2. Damiani E, di Vimercati SDC, Samarati P (2005) New paradigms for access control in open environments. In: Proceedings of the fifth IEEE international symposium on signal processing and information technology, December 2005, pp 540–545
3. ITU-T (2000) ITU-T recommendation X.509-ISO/IEC 9594-8: information technology and open systems interconnection and the directory: public-key and attribute certificate frameworks. Technical report, ITU-T, 2000. http://www.infosecurity.org.cn/content/pki_pmi/x509v4.pdf
4. Lang B, Foster Ian T, Siebenlist F, Ananthakrishnan R, Freeman T (2009) A flexible attribute based access control method for grid computing. J Grid Comput 7(2):169–180
5. LeMay M, Fatemieh O, Gunter CA (2007) Policymorph: interactive policy transformations for a logical attribute-based access control framework. In: Proceedings of the 12th ACM symposium on access control models and technologies, SACMAT '07. ACM, New York, USA, pp 205–214
6. Loscocco PA and Smalley SD (2001) Meeting critical security objectives with securityenhanced linux. In: Proceedings of the 2001 Ottawa Linux symposium, July 2001
7. Loscocco PA, Smalley SD, Muckelbauer PA, Taylor RC, Jeff TS, Farrell JF (1998) The inevitability of failure: the flawed assumption of security in modern computing environments. In: Proceedings of the 21st national information systems security conference, pp 303–314
8. Moses T (2005) eXtensible Access Control Markup Language (XACML) Version 2.0. Technical report, OASIS, Febuary 2005. http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf
9. Sandhu RS, Coyne EJ, Feinstein HL, Youman CE (1996) Role-based access control models. Computer 29:38–47
10. Shen H, Hong F (2006) An attribute-based access control model for web services. In: Proceedings of the seventh international conference on parallel and distributed computing, applications and technologies, PDCAT '06, Washington, DC, USA, 2006. IEEE Computer Society, pp 74–79
11. Yu S, Ren K, Lou FW (2010) Toward fine-grained distributed data access control in wireless sensor networks. IEEE Trans Parallel Distrib Syst 99:255–273
12. Yuan E, Tong J (2005) Attributed based access control (abac) for web services. In: Proceedings of the IEEE international conference on web services, ICWS '05, Washington, DC, USA, 2005. IEEE Computer Society, pp 561–569

Chapter 125
Lung Cancer Analysis of Factors Influencing Hospital Costs

Jianhui Wu, Guoli Wang, Jing Wang and Sufeng Yin

Abstract Health care costs for the control reference reveal the cost of lung cancer patient factors. Hospital cost data have many complex factors in both quantitative and qualitative variables. The use of traditional multiple regression analysis is not appropriate. This study uses BP neural network modeling and sensitivity analysis of factors analysis. In this paper, lung cancer patient factors affecting the cost of BP neural network modeling, through sensitivity analysis found the main factors followed by days of hospitalization, treatment outcome, age, hospital number, occupation, cost categories, admissions, etc., and proposed to shorten the hospital stay time for a breakthrough by the medical insurance system reform, strengthening the role of the medical cost control system to solve.

Keywords Hospital charges · Factors · BP neural network · Sensitivity analysis

125.1 Introduction

Lung cancer is the most common primary malignant tumor. A survery from at least 35 countries showed lung cancer to be the leading cause of death in men second only to breast cancer deaths in women. The disease incidence is of more than 40 years of age, the peak age of onset in the 60–79 years old. Male and female prevalence was 2.3:1. Race, family history and smoking have implications on lung cancer. In cancer deaths in China, lung cancer accounts for common malignant

J. Wu (✉) · G. Wang · J. Wang · S. Yin
Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry,
Division of Epidemiology and Health Statistics, School of Public Health,
Hebei United University, Tangshan 063000, China
e-mail: wujianhui555@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_125,

tumors in men as the fourth and the fifth largest in women [1]. In this study, two levels of Tangshan, Hebei Province of lung cancer patients to inpatient hospital costs analysis to identify the main influencing factors, and propose specific measures and measures to control medical costs, improve efficiency and use of health resources Hospital service quality are of great significance. General information on the hospital costs were skewed distribution, hospital costs influenced by many complex factors and quantitative variables and qualitative variables in both, the use of traditional multiple regression analysis is not appropriate [2, 3]. Therefore, this study uses BP neural network modeling and sensitivity analysis of factors. BP neural network as it can on the linear or nonlinear multivariable without preconditions in the case of statistical analysis, statistical methods with the traditional analysis of the variables that need to be consistent with certain conditions compared with its own advantages. BP neural network essentially implements a mapping from input to output function, and mathematical theory has proved that it has the realization of any complex nonlinear mapping function. This makes it particularly suitable for solving complex problems within the system.

125.2 Principle of the BP Neural Network

The BP neural network is composed of the forward-propagating and the ant propagation. In the process of forward-propagating, the infed information treated with in implicit hierarchical unit passes from input level to the output level. Each neuron's condition only influences next neuron's condition. If it cannot obtain the expected output in the output level, it transfers to the ant propagation—makes the error signal return along the connection way, and makes the error signal be the smallest by modifying connection weights between each neuron [4, 5]. The BP algorithm is a learning algorithm proposed for multi-layered perceptions (MLP). The BP neural network's transmission function among the concealed levels uses continuous and differentiable nonlinear function, and usually is the sigmoid function. The transmission function among the output levels may uses the linear function or the sigmoid function, which is determined by the distribution of the output layer's vector. Basic steps: (1) Initialization weight and threshold value, $w_{ji}(0), \theta_j(0)$ is small stochastic value. (2) Importing the training samples: input vector $x_k, k = 1, 2, \ldots, P$; Expected output $d_k, k = 1, 2, \ldots, P$; Iterating (3) to (5) for each input sample. (3) The condition of computer network's actual output and implicit strata unit:$o_{kj} = f_j\left(\sum_i w_{ji} o_{ki} + \theta_j\right)$. (4) Computation training error: output level $\delta_{kj} = o_{kj}(1 - o_{kj})(t_{kj} - o_{kj})$, Concealed level $\delta_{kj} = (1 - o_{kj}) \sum_m \delta_{km} w_{mj}$. (5) Revision weight and threshold value $w_{ji}(t+1) = w_{ji}(t) + \mu \delta_j o_{ki} + \alpha[w_{ji}(t) - w_{ji}(t-1)]$, $\theta_j(t+1) = \theta_j(t) + \eta \delta_j + \alpha[\theta_j(t) - \theta_j(t-1)]$, η for the length of stride, t for the times of iteration. (6) Each time k experiences from 1 to P, judging whether the target satisfies the requirement of accuracy or not may see, the BP neural network model make the function question of input and output sample's function transfer to a nonlinear optimized question, and has used the ordinaries gradient descent in the optimized techniques. If we regard the neural network as the mapping between the inputs and the output, this mapping is an altitude nonlinear mapping [6, 7]. BP neural network model establishes the sensitivity analysis to determine the input variables on output variables predicting the degree of importance. Sensitivity analysis is a part of changing network input in order to observe changes in network output corresponding to decide this part of the network in the importance of the output prediction. This is done in order to change the sample variable values of each record. For the categorical variables, all possible combinations are to be test class; for continuous variables, use the range of values for four equal portions of the split point, such as the return value one of the [0, 1] range, respectively 0, 0.25, 0.5, 0.75, and 1 of these values. When the value is changed, record the output of one of the largest and the smallest, and calculate the maximum and minimum values of the difference accounted for the largest proportion of output, and ultimately the sensitivity of the variable is the proportion of all records of the mean [8, 9].

125.3 BP Neural Network Model

125.3.1 Example Synopsis

The data come from two tertiary hospitals in Tangshan City, Hebei Province of lung cancer patients in the medical record, excluding incomplete records of key information, a total of 378 cases have valid cases. Survey include: sex, age, frequency of hospitalization, occupation, hospital, treatment outcome, surgical cases, length of stay, cost categories, secondary diagnosis and hospitalization costs and the composition so that the various factors and quantitative methods in Table 125.1 Retrospective methods used to investigate the advantages of this method are able to use the force of law required in the medical files of information collected, analyzed medical cost factors. Using SPSS17.0 package sorting and statistical processing of data, using the package MLP neural network function in the BP neural network modeling and sensitivity analysis.

As the basis of the selected hospital is better, the file is intact. The investigators are specially trained investigators, thus ensuring a high quality of survey data and more accurate information.

125.3.2 BP Neural Network Model Results

BP neural network model established the data set into training set and test set. Training set for network training and the test set is used to test the trained network to check the generalization ability of the network size. The proportion of the data

Table 125.1 Factors and quantitative methods

Code	Factors	Quantitative methods or units
X1	Gender	1 = male; 2 = Female
X2	Age	Years
X3	Hospitalization	Times
X4	Professional	1 = pastoral fishermen; 2 = workers; 3 = students; 4 = education industry; 5 = cadres; 6 = enterprises workers; 7 = medical health personnel; 8 = other
X5	Admissions	1 = general; 2 = acute; 3 = risk disease
X6	Treatment	1 = cured; 2 = improved; 3 = death; 4 = cured; 5 = other
X7	Surgery	0 = none; 1 = there
X8	Length of hospital stay	Days
X9	Cost categories	1 = medical insurance for urban workers; 2 = basic medical insurance for urban residents; 3 = rural cooperative medical care; 4 = retired cadres; 5 = commercial insurance; 6 = expense
X10	Secondary diagnosis	0 = none; 1 = there
Y	Hospital charges	Yuan

Table 125.2 Model parameters

Network information		
Hidden layer(s)	Number of units[a]	151
	Number of hidden layers	1
	Number of units in hidden layer 1[a]	1
	Activation function	Sigmoid
	Dependent variables	y
Output layer	Number of units	1
	Rescaling method for scale dependents	Standardized
	Activation function	Identity
	Error function	Sum of squares

[a] Excluding the bias unit

subsets with no clear division of requirements, generally greater than the training set test set and training set than in the distribution of the total data set which are relatively close to the training sample, representative of the information can be fully utilized (Tables 125.2, 125.3 and 125.4).

125.3.3 Sensitivity Analysis

Sensitivity analysis showed: in decreasing order of number of days of hospitalization, treatment outcome, age, hospitalization, professional, cost category, hospitalization, gender, surgery and secondary diagnosis of the situation. Each factor in the cost of hospitalization in the number of days of hospitalization reached a

Table 125.3 Division of the data set

Case processing summary			
		N	%
Sample	Training	256	69.6
	Testing	112	30.4
Valid		368	100.0
Excluded		10	
Total		378	

Table 125.4 Model fit index

Model summary		
Training	Sum of squares error	58.849
	Relative error	0.462
	Stopping rule used	1 consecutive step(s) with no decrease in error[a]
	Training time	0:00:00.141
Testing	Sum of squares error	30.091
	Relative error	0.841
Dependent variable:		y

[a] Error computations are based on the testing sample

Table 125.5 Sensitivity analysis

Independent variable importance		
	Importance	Normalized importance (%)
x1	0.036	12.8
x3	0.091	32.0
x2	0.174	61.1
x4	0.079	27.7
x5	0.036	12.8
x9	0.068	23.9
x8	0.284	100.0
x10	0.021	7.2
x6	0.178	62.5
x7	0.033	11.5

maximum of 0.284, the longer the hospital stay, drug costs, treatment fees, inspection fees, increased fees and charges fees beds, making the total cost also increase (Table 125.5).

125.4 Conclusion

Positive selection for lung cancer prevention on the one hand and taking effective measures to reduce or avoid the inhalation of carcinogenic substances in the air and dust pollution, such as the mobilization of quitting smoking, ban smoking in public places to prevent air pollution, strengthen the protection of harmful dust and other operations. On the other hand on the high incidence of patient groups to focus on screening, early detection and timely treatment. However, in recent years, medical costs have been on the rise, and a heavy economic burden on society and to the patient's family. The factors that affect the cost of hospital inpatient days as the most important factor, shorter hospital stay, ineffective and inefficient can reduce hospitalization time, on the one hand will help reduce the psychological burden of patients and their families and economic burden, but this can speed up hospital beds turnover, and improving bed utilization, thus contributing to the economic interests of hospitals and departments. With the gradual deepening of health reform and medical insurance system reform, hospital patients are gradually moving to low cost, high quality service goals. However, in recent years, medical costs have gradually increased. Many factors affect the cost of hospitalization. There are controllable factors and uncontrollable factors. To curb the excessive growth of medical costs to be kept under strict management the focus should be on control. Need to reduce the inefficient medical costs, eliminate waste in the health care costs, strengthen medical staff medical ethics, medical units must also take into account the interests of society and patients, mostly for the sake of patients, the patient's medical expenses to a minimum.

References

1. Yang L, Li L, Chen Y, Parkin DM (2005) China's mortality trends of lung cancer and lung cancer incidence and mortality estimates and projection. Chin J Lung Cancer 8(4):274–278
2. Jing W, Man L (2009) Multivariate statistical methods in the study of hospital fees progress. China Health Stat 26(1):91–95
3. Biyao L, Yi S (2006) Based on BP neural network modeling of hospital costs. Zhejiang University, a master's degree thesis
4. Bartfay E, Mackillop WJ, Peter JL (2006) Comparing the predictive value of neural network models to Logistic regression models on the risk of death for small-cell lung cancer patients. Eur J Cancer Care 15(2):115–124
5. Gaoli S, Fangping D (2003) Based on the MATLAB language improved algorithm of BP neural network. Bull Sci Technol 19(2):130–135
6. Lifeng Z, Ersheng G, Pihuan J (1998) Comparison between neural networks and multivariate linear regression method. Mod Prev Med 25(3):272–274
7. Erol FS, Uysal H, Ergiun U et al (2005) Prediction of minor head injured patients using logistic regression and MLP neural network. J Med Syst 29(3):205–215
8. Juying Z, Jian W, Shuqin Y (2002) Neural network model factors in the hospitalization cost analysis. Chin J Hosp Adm 18(3):143–145
9. Yinqi M, Lifeng Z, Peihuan J (1998) Artificial neural network application in statistics. Med Inf 6(11):21–24

Chapter 126
An Improved Differential Evolution Algorithm Based on Mutation Strategy for Dynamic Economic Dispatch

Hongfeng Zheng, Min Hu, Ziqing Xie, Chunchao Shi and Minmin Zhou

Abstract Dynamic economic dispatch (DED), is a method of scheduling the online generators with a predicted load demand over a certain period of time taking into account the various constraints imposed on the system operation. In this chapter, an improved differential evolution (IDE) algorithm was presented for power system Dynamic economic dispatch (IDED). The proposed IDE algorithm was tested on a system consisting 15 generators. The scheduling horizon is chosen as one day with 24 intervals of 1 h each whose cost function took into account the valve-point effects except the prohibited discharge zones. The results indicate that IDE algorithm outperforms GA, PSO and DE algorithms in solving DED problems.

Keywords Dynamic · Dispatch · Improved · Differential evolution

126.1 Introduction

The application of optimization techniques to power system planning and operation has been an active research in the recent past. A wide variety of mathematical optimization techniques have been applied to solve the power system operation and control problems.

Recently, DE has successfully been applied in the optimization techniques to power system planning and operation. According to recent reports [1], DE obtains better performance than genetic algorithms and PSO in terms of convergence speed and the quality of solutions on many global optimization problems.

H. Zheng (✉) · M. Hu · Z. Xie · C. Shi · M. Zhou
Zhejiang Industry Polytechnic College, Shaoxing, China
e-mail: Zhf660729@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_126,

The classical DE conventionally has several mutation strategies, and three control parameters, i.e., many works have been done along these directions. In [2], Qin and Suganthan presented a self-adaptive DE (SaDE) algorithm for numerical optimization, which focused on adaptation for parameter CR and mutation strategies of DE. Brest [3] introduced self-adapting control parameter settings in DE (SADE) to reduce the effects of the parameters. Ali and Torn [4] introduced auxiliary population and automatic calculating of the parameter. Yang [5] introduced a neighborhood search strategy to DE (NSDE), which generates F from Gaussian and Cauchy distributed random numbers instead of predefining a constant F. On the basis of NSDE, Yang [6] proposed another variant of DE, called SaNSDE, which combines the idea of SaDE and NSDE. Besides the above improvements on the control parameters, other new strategies are also introduced to DE. Sun [7] proposed DE/EDA by combining DE and estimate of distribution algorithm (EDA). Rahnamayan [8] presented a novel DE variant (ODE) by applying OBL to DE, in which ODE not only estimates the current search point, but also considers its opposite point. By simultaneously evaluating the current search point and the opposite point, it can get a better approximation to the global optimum.

In this chapter, we propose a novel DE variant by employing an improved mutation strategy called IDE, then we focus on the applications of IDE algorithms to power system Dynamic economic dispatch (IDED).

126.2 DE Based on Improved Mutation Strategy

126.2.1 Differential Evolution

Differential evolution (DE) is a population-based and directed search method [9], it starts with an initial population vector, which is randomly generated when no preliminary knowledge about the solution space is available. There are several variants of DE [9], where the most popular var6led *DE/rand/1/bin*, and often used in [3] and [8].

Let $X_{i,G}(i = 1, 2, \ldots ps)$ are solution vectors in generation G, where ps is the population size. The main idea of DE is to generate trial vectors. Mutation and crossover are used to produce new trial vectors, and selection determines which of the vectors will be successfully selected into the next generation.

For each vector $X_{i,G}$ in generation G, a mutant vector $V_{i,G}$ is defined by

$$V_{i,G} = X_{r1,G} + F(X_{r2,G} - X_{r3,G}) \tag{126.1}$$

where $i = 1, 2, \ldots ps$ and $r1$, $r2$, and $r3$ are mutually different random integer indices selected from $\{1, 2, \ldots ps\}$.

DE employs a crossover operator to build trial vectors by recombining two different vectors. The trial vector is defined as follows:

$$U_{i,G} = \left(U_{1i,G,} U_{2i,G,}, \ldots, U_{Di,G,}\right),$$

where $j = 1, 2, \ldots, D$ (D = problem dimension) and

$$U_{ji,G} = \begin{cases} V_{ji,G}, & \text{if } \mathrm{rand}_j(0,1) \leq \mathrm{CR} \vee j = k \\ X_{ji,G}, & \text{otherwise} \end{cases} \tag{126.2}$$

where CR is the predefined crossover probability, and $\mathrm{rand}_j(0,1)$is a random number within (0, 1) for the ith dimension, and $k \in \{1, 2, \ldots, D\}$ is a random parameter index.

It is an approach which must decide which vector ($U_{i,G}$ or $X_{i,G}$) should be a member of next (new) generation $G + 1$.

126.2.2 Improved Differential Evolution

In this chapter, we focus on the improvement of the "DE/target-to-best/1" strategy. The modified scheme is described as follows.

$$V_{i,G} = X_{r1,G} + F(p\mathrm{best}_{i,G} - X_{i,G}) + F(X_{r2,G} - X_{r3,G}) \tag{126.3}$$

where $r1$, $r2$, and $r3$ are three different random integers within $[1, ps]$, $p\mathrm{best}_i$ is the previous best vector of X_i, and ps is the population size.

We use Eq. (126.3) to generate the mutant vector X_i instead of Eq. (126.1) in Algorithm 1. The specific descriptions of the IDE are presented in Algorithm 2.

126.3 IDE for Dynamic Economic Dispatch

Dynamic economic dispatch (DED) is a method of scheduling the online generators with a predicted load demand over a certain period of time taking into account the various constraints imposed on the system operation.

126.3.1 Problem Objective

The objective of the DED problem is to schedule the committed units economically over a scheduling period T as given by

$$\text{Minimize } F = \sum_{t=1}^{T} \sum_{i=1}^{N_g} \mathrm{FC}(i,t) \tag{126.4}$$

The production cost is expressed as

$$\mathrm{FC}(i,t) = a_i + b_i P_{it}^2 + c_i P_{it}^2 + \left| e_i \times \sin\left\{ f_i \times \left(P_{it}^{\min} - P_{it} \right) \right\} \right| \tag{126.5}$$

where e_i and f_i represent the cost coefficients of ith unit valve point effects.

126.3.2 System Constraints

The power system equality constraint is expressed as

$$\sum_{i=1}^{N_g} P_{it} - P_{Dt} - P_{Lt} = 0 \tag{126.6}$$

where $t = 1, 2, \ldots, T$, P_{Dt} is the forecasted total power demand at time t and P_{Lt} is the total transmission losses of the system at time t. The general form of loss formula using B-coefficients is

$$P_{Lt} = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P_{it} B_{ij} P_{jt} \tag{126.7}$$

The generator capacity constraints are expressed as

$$P_i^{\min} \leq P_{it} \leq P_i^{\max} \tag{126.8}$$

The ramp rate limits are given by

$$P_{it} - P_{i(t-1)} \leq UR_i \tag{126.9}$$

$$P_{i(t-1)} - P_i \leq DR_i \tag{126.10}$$

126.3.3 Dynamic Economic Dispatch Using IDE

The number of decision variables will be the number of generating units multiplied by the number of time intervals. The population size should be 5–10 times the value of the dimension of the problem in order to avoid premature convergence. The Improved Differential Evolution (IDE) employed for DED problem is briefly discussed as follows:

Initialization

The real power generation of ith plant at time t is expressed as

$$P_{it} = P_i^{\min} + \rho(P_i^{\max} - P_i^{\min}) \tag{126.11}$$

where $\rho \in [0{,}1]$ is uniformly distributed random number. The objective function of the DED problem, which is to be minimized, is given by

$$\Psi = \sum_{\mathrm{t}=1}^{T} \sum_{i=1}^{N_g} FC(P_{it}) + \sum_{Z=1}^{N_C} \lambda_Z |\mathrm{VIOL}_Z| \tag{126.12}$$

where λ is the penalty factor, $\mathrm{N_c}$ represents the number of constraints and VIOL is the constraint violation.

Mutation

IDE generates new parameter vectors by adding the weighted difference vector between two population members to a third member. A perturbed individual is therefore generated on the basis of the parent individual in the mutation process by

$$\hat{Z}_b^{G+1} = Z_P^G + F \times (p\mathrm{best}Z_{\mathrm{i}}^G - Z_i^G) + F(Z_{\mathrm{j}}^G - Z_k^G) \tag{126.13}$$

where F is a scaling factor and j and k are randomly selected. The scaling factor F [0,1] ensures the fastest possible convergence and G represents generation number.

If the new decision variable is out of the limits by an amount, this amount is subtracted or added to the limit violated to shift the value inside the limits and appropriate adjustments are made to satisfy the prohibited operating zone constraints.

Acceleration Operation

If the best fitness at the present generation is not further improved by the mutation and crossover operations, then the present best individual is pushed towards a better point. Thus, the accelerated phase is represented as

$$Z_b^{G+1} = \begin{cases} \hat{Z}_b^{G+1}, & \textit{if } \psi(\hat{Z}_b^{G+1}) < \psi(\hat{Z}_b^{G}) \\ \hat{Z}_b^{G+1} - \alpha\nabla\Psi, & \text{otherwise} \end{cases} \tag{126.14}$$

where Z_b^{G+1} is the best individual. The gradient of the objective function $\nabla\psi$ can be calculated with finite variation. The step size $\alpha \in [0,\ 1]$ is determined by the descent property. Initially α is set a value of one to obtain the new individual Z_b^N. If the descent property is satisfied, i.e.,

$$\psi(Z_{\mathrm{b}}^N) < \psi(Z_b^{G+1}) \tag{126.15}$$

then the Z_b^N becomes a candidate in the next generation and is added to this population replacing the worst individual. If the descent property is not satisfied, then step size is lowered a little. The descent method is repeated to search Z_b^N until $\alpha\nabla\psi$ is sufficiently small or a specified number of iterations are performed. This faster decent results in a premature convergence and the migration phase regenerates a new population.

Migration Operation

A migration phase is introduced to regenerate a newly diverse population of individuals to enhance the investigation over the search space, and thus, reduce the pressure of selection from a small population. The new populations are obtained based on the best individual Z_b^{G+1} The *h*th gene of the *i*th individual is regenerated as

$$Z_{hi}^{G+1} = \begin{cases} Z_{hb}^{G+1} + \delta_{hi}\left(Z_{h\min} - Z_{hb}^{G+1}\right), & \text{if } \overset{\sim}{\delta}_{hi} < \frac{Z_{hb}^{G+1} - Z_{h\min}}{Z_{h\max} - Z_{himn}} \\ Z_{hb}^{G+1} + \delta_{hi}\left(Z_{h\max} - Z_{hb}^{G+1}\right), & \text{otherwise} \\ i = 1, \ldots, N_P, h = 1, \ldots, n \end{cases} \tag{126.16}$$

where δ and $\bar{\delta}$ denote uniformly distributed random numbers. This diversified population is then used as the initial decision parameters to escape the local optimum points. The migration operation is performed only if the population diversity P is smaller than the desired tolerance of population diversity ε1.

$$\mathrm{p} = \frac{\left\{ \sum_{\substack{i=1 \\ i \neq b}}^{N_P} \left(\sum_{h=1}^{n} \eta_z \right) \right\}}{n\left(N_p - 1\right)} < \varepsilon 1 \tag{126.17}$$

$$\eta_z = \begin{cases} 1, & \text{if} \left| \frac{Z_{hi} - Z_{hb}}{Z_{hb}} \right| > \varepsilon 2 \\ 0, & \text{otherwise} \end{cases} \tag{126.18}$$

where parameter ε2 expresses the gene diversity with respect to the best individual. η_z is the scale index. Degree of population diversity is between zero and one. A value of zero implies that all genes gather around the best individual. On the other hand, the value of one implies that the current candidate individuals are a diversified population. Therefore, the tolerance of population diversity is accordingly assigned within this region.

With the members of the next generation thus selected, the cycle repeats until there is no improvement in the best individual. In this study also HDE with random vector perturbation and binominal crossover is employed.

126.4 Computer Simulation

In this study, the performance of the IDE-based DED algorithm is implemented using C++ code on a personal computer and is evaluated using an illustrative test system consisting of 15 generators [10]. The scheduling horizon is chosen as one day with 24 intervals of 1 h each. The effect of valve point loading is also included in the fuel cost characteristics. This case study does not consider the prohibited discharge zones. The thermal generator data and the load demand are summarized

Table 126.1 Generating unit data

Unit	P_i^{min} (MW)	P_i^{max} (MW)	ai (MW)	bi ($/ MW)	Ci ($/ MW2)	ei ($/ h)	fi (1/ MW)	URi (MW/h)	DRi (MW/h)
1	150	455	671	10.1	0.000299	240	0.031	80	120
2	150	455	574	10.2	0.000183	240	0.033	80	120
3	20	130	374	8.8	0.001126	140	0.036	130	130
4	20	130	374	8.8	0.001126	140	0.034	130	130
5	150	470	461	10.4	0.000205	230	0.033	80	120
6	150	460	630	10.1	0.000301	240	0.032	80	120
7	135	465	548	9.8	0.000364	220	0.033	80	120
8	60	300	227	11.2	0.000338	200	0.035	65	100
9	25	162	173	11.2	0.000807	150	0.039	60	100
10	25	160	175	10.7	0.001203	150	0.039	60	100
11	20	80	186	10.2	0.003586	140	0.039	80	80
12	20	80	230	9.9	0.005513	140	0.039	80	80
13	25	85	225	13.1	0.000371	150	0.039	80	80
14	15	25	309	12.1	0.001929	100	0.042	55	55
15	15	25	323	12.4	0.004447	100	o.o42	55	55

Table 126.2 Load demand for 24 h

Time (h)	Load (MW)	Time (h)	Load (MW)	Time (h)	Load (MW)	Time (h)	Load (MW)
1	1,741	7	2,658	13	2,989	19	2,777
2	1,847	8	2,777	14	2,930	20	2,989
3	2,017	9	2,930	15	2,777	21	2,888
4	2,251	10	2,989	16	2,463	22	2,569
5	2,369	11	3,051	17	2,369	23	3,238
6	2,582	12	3,142	18	2,582	24	1,970

in Tables 126.1 and 126.2. The *B*-coefficient data are the same as given in the reference.

The comparison of the optimal system costs obtained from the IDE-based approach with that of particle swarm optimization (PSO) and genetic algorithm is given in Table 126.3. The proposed approach yields better results than Ga, PSO and DE.

126.5 Conclusion

This chapter presented the application of DE based on Improved Mutation Strategy for solution of power system DED problem. The algorithms have been devised to efficiently according to the problem dimensionality and the constraints. It is quite evident from the comparison against other evolutionary algorithms that the DE

Table 126.3 Comparison of optimal solution for 15-generator system

Method	GA	PSO	DE	IDE
Total cost ($/24 h)	794,712	788,592	782,136	778,488

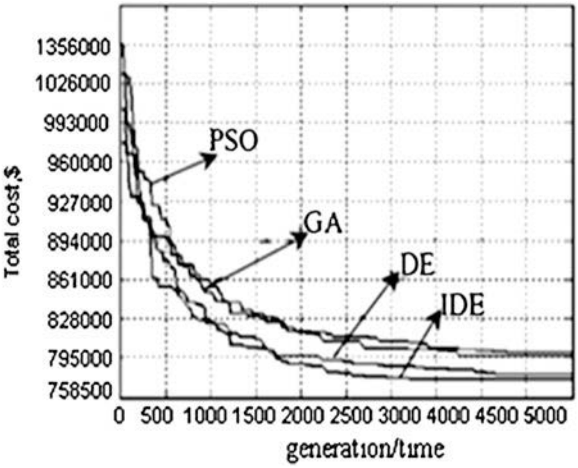


Fig. 126.1 Cost curves of different algorithms

approach based on Improved Mutation Strategy provides a competitive performance in terms of optimal solution as well as computation effort (Fig. 126.1).

References

1. Vesterstrom J, Thomsen R (2004) A comparative study of differential evolution, particle swarm optimization, and evolutionary algorithms on numerical benchmark problems. Proc Congr Evol Comput 2:1980–1987
2. Qin AK, Suganthan PN (2005) Self-adaptive differential evolution algorithm for numerical optimization. Proc Congr Evol Comput 2:1785–1791
3. Brest J, Greiner S, Boskovic B, Mernik M, Zumer V (2006) Self-adapting control parameters in differential evolution: a comparative study on numerical benchmark problems. IEEE Trans Evol Comput 10:646–657
4. Ali MM, Torn A (2004) Population set-based global optimization algorithms: Some modifications and numerical studies. Comput Oper Res 31:1703–1725
5. Yang Z, He J, Yao X (2008) Making a difference to differential evolution. In: Advance in Metaheuristics for hard optimization, pp 397–414
6. Yang Z, Tang K, Yao X (2008) Self-adaptive differential evolution with neighborhood search. In: Proceedings of congress on evolutionary computation, pp 1110–1116
7. Sun J, Zhang Q, Tsang E (2004) DE/EDA: a new evolutionary algorithm for global optimization. Info Sci 169:249–262
8. Rahnamayan S, Tizhoosh HR, Salama MMA (2008) Opposition-based differential evolution. IEEE Trans Evol Comput 12:64–79

9. Storn R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11:341–359
10. Lakshminarasimman L, Subramanian S (2008) Applications of differential evolution in power system optimization. IEEE Trans Evol Comput 14:257–263 (springerlink.compp)

Chapter 127
Study on Power Sensitive Software Framework in E-Paper Device

Qing-Cheng Li, Zhan-Ying Zhang and Jin Zhang

Abstract E-paper screen mobile devices that primarily process multimedia data, such as image, audio and text, are expected to become important platforms for pervasive computing. Diverse mobile devices with wireless network are now growingly used, range over various applications. There is an accelerating trend toward ubiquitous computing. Although there have been some ways available for implementation of mobile applications, this research field is still under development. One challenge that remains is the ability to make allowance for running any burdensome task in mobile devices which is commonly run in desktop devices. Traditionally, E-reader devices can only support limited services for limitations of resource. Ubiquitous computing applications promise user to richer and more complicate behavior, but the current state of research in this field is still far removed from that vision. This paper presented a novel and effective wireless network-based software framework to support developers building applications in E-paper devices. The framework includes 3G and Wi-Fi network manager, service module-based XML, service module based on Webkit and resource scheduler. The proposed framework can provide the technical basis for mobile business applications.

Keywords E-paper · Wireless · Framework · QT · 3G · Wi-Fi

Q.-C. Li (✉) · Z.-Y. Zhang · J. Zhang
College of Information Technical Science, Nankai University,
Weijin Road.94, Tianjin 300071, China
e-mail: liqc@jinke.com.cn

Z.-Y. Zhang
e-mail: zhangzhanying@mail.nankai.edu.cn

J. Zhang
e-mail: sir.zhangjin@gmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_127,

127.1 Introduction

E-paper is a kind of display technology which is designed to display reflected natural light using electrophoretic colored particles [1]. This kind of display material is applicable for protracted reading and brings comfortable reading experience. There are some challenges due to the characteristics of E-paper when network functions are added in the system. They are enumerated as following:

Excessive power consumption. The E-paper device discussed in this paper uses E-ink screen and builds 3G wireless module and Wi-Fi wireless module in. Wireless modules require a larger operating current which may cause excessive power consumption. The battery manager must be adapted to let battery-powered E-paper devices reach the desired lifetime.

Network stability. There is a strong need for high stability network for timely data synchronization. However, the uncertainty of changeable environments and various obstacles reduce the stability of wireless network. Network instability will cause long latency time when users interact with mobile devices and reduce users' satisfaction.

Limited network bandwidth. It is extremely difficult to implement application which needs high network bandwidth. For example, browsing some websites in the Internet with mobile web browser will take a long time for downloading html data.

Providing proper and practical network services. Choosing proper application on E-paper devices is important to achieve satisfaction of users.

Resource competition. There is an inherent conflict in the design goals for providing high quality of service and limitations of resource such as power energy, computing capability, storage capability and network bandwidth. An effective resource scheduler is needed for quality of service provisioning.

Besides, some of these challenges are shared by conventional CRT and TFT screen devices. To address these problems, this paper presented a novel and effective software framework based on wireless network for E-paper device. It gives a practical support for developers to construct wireless network applications. The rest of this paper is organized as follows:

Section 127.2 introduces the framework of the system. Section 127.3 describes the design of individual part of the framework. The implementation and experimental evaluation is shown in Sect. 127.4. Finally, Sect. 127.5 concludes this paper.

127.2 Software Framework

With the perspective of an application designer in mind, we want to provide a conceptual framework that automatically supports all the tasks that are common across applications, requiring the designer to only provide support for the application-specific tasks. The proposed software framework is shown in Fig. 127.1.

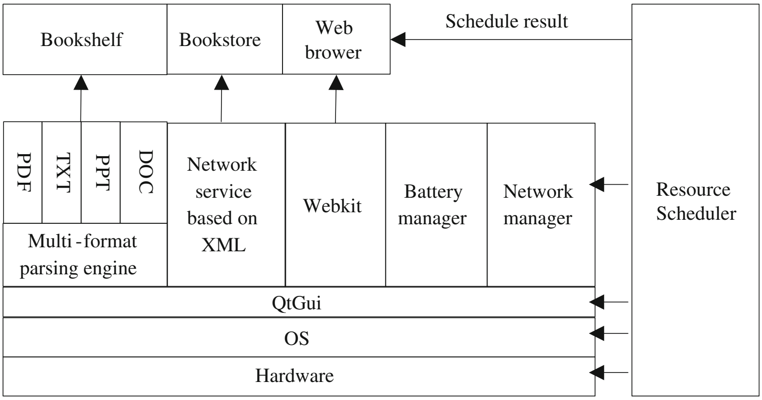


Fig. 127.1 Software framework based on wireless network for E-paper devices

Multi-format parsing engine [2] is used in the system. It is a document browser containing multi-format parsers. It is compatible of many kinds of files. There have been over ten kinds of commonly used document format parsers developed including PDF, DOC, RAR, MP3. There are two kinds of network services in the system: services based on XML and services based on Webkit.

Users are typically facing unpredictable environments where specialized devices outnumber, users now have to deal with wide variety of devices. It is conceptually envisioned that users can do anything with any small device and practicality. We have been witnessing a great step toward ubiquitous computing paradigm [3, 4] since the growingly use of wireless network and mobile devices.

Various applications are now increasingly used in changing device. Yet these applications did not adapt to the changes very well. Rather, users prospect to do various kinds of behavior, such as processing documents, reading, purchasing, communicating with friends and searching for merchandises, through mobile devices in changing environment, but moving away from the desktop bring up many obstacles for the application transplant. To address this problem, we proposed a software framework based on wireless network to fit multiple mobile devices.

The wireless network created a bridge between mobile terminal and remote server. In network service based on XML, some data interfaces in xml format are defined to characterize request/response information related to the data exchange between user and server. Subtle and high-level interpretations of a user command can be accomplished through these xml data. The focus is mainly on understanding and handling xml interface that can be defined according to different requirements to rapidly accomplish the behavior of an user command.

Network service based on XML is designed with C/S structure. Content provider installs application and data on remote server to supply content to clients. The data interfaces are previously customized. At first, client acquires data by sending requests to server. Then client parses and shows the reply data from the server. In addition, client may also upload data to server, such as post the log message to the server. Since XML data can achieve high compression ratio and the size of the XML data can be adjusted according to the actual requirements, this service module is feasible and proper in wireless network. The other kind of network service is based on a modified Webkit in Qt.

127.3 System Design

127.3.1 Battery Manager

Achieving expected lifetime of battery is a key criterion for battery-powered mobile devices. Energy consumption is mainly relevant to the usage of CPU and wireless network. When system is in offline state, energy consumption is in proportion to the use of CPU. While in online state, energy consumption is mainly in proportion to the use of wireless network.

Firstly, it is required to always monitor the battery state. A power daemon is added in the system to monitor the state of battery, user input event and network traffic.

E-paper screen has a special capability of maintaining screen image without power. The system needs electrical energy only when user turns page and it does not need energy while user is reading. Hence, if there is no user input for 2 s in offline state, the system will enter idle state and the electric current of the device will be reduced to about 1 mA. Later, the system will be waked up rapidly and executes the corresponding job when any user input event comes. Therefore, according to user's common reading habits, system will stay in idle state for most of the time in offline state. Furthermore, the lifetime of the battery is not dependent on how long the device is used, but depends on how many times of turning page. In addition, once power daemon found residual energy is lower than baseline required, it will prompt user to recharge.

When network connection is established, the system can never enter the idle state since wireless module needs electrical energy to maintain network connection. Therefore, energy consumption will be increased enormously. For example, for the platform we used, the required electrical current of wireless module is normally 350–400 mA. Attempting to alleviate this problem, different strategies of 3G and Wi-Fi module management have been adopted in this paper. Wi-Fi management will be described in the following subsection since it is more complicated. The state transition diagram of 3G module is shown in Fig. 127.2.

3G module is controlled by monitoring network traffic and user input event. There are four states of the module: run state, power save state, disconnected state and closed state. Run state is the active state while the module is working. Power save state means that the module is in idle state. Disconnected state means network

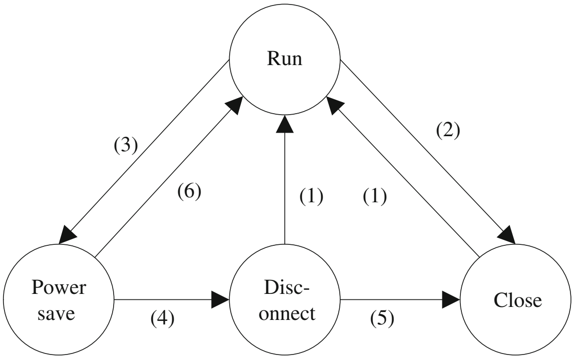


Fig. 127.2 State transition diagram of 3G module

is disconnected but the module is not closed. Since starting 3G module takes relatively long time, the network can be connected quickly in this state for there is no need to start module. Close state means 3G module is closed. The conditions of transition corresponding to the number in Fig. 127.2 are:

(1) When users need to access Internet through wireless communication, network trigger will start 3G module and connect network.
(2) The triggers of this transition are disconnecting network, shutdown and locking screen.
(3) If there is no network traffic, the module will automatically enter power save state.
(4) If there is no network traffic for ten minutes no matter if there is user input event, the module will enter disconnected state. In addition, considering the scenario that users may have been left where they are while downloading jobs are still working, when there have been no user input event for ten minutes, then network traffic will be checked at that moment. If there is no network traffic too, the module will enter disconnected state, else the module will wait for the finish of downloading jobs and then enter disconnected state.
(5) If there is no network traffic for 30 min continuously, then the module will be closed.
(6) The module enters run state if there is network traffic.

127.3.2 Wi-Fi Network Manager

Mobile phones normally have color screens, always-on connectivity and can transfer data at high speed. While E-paper devices normally have gray screens like paper. In order to saving electrical energy, they just connect network when needed. This conflict between wireless communication and energy saving brings a challenge for network management. The state transition diagram of Wi-Fi network management is shown in Fig. 127.3.

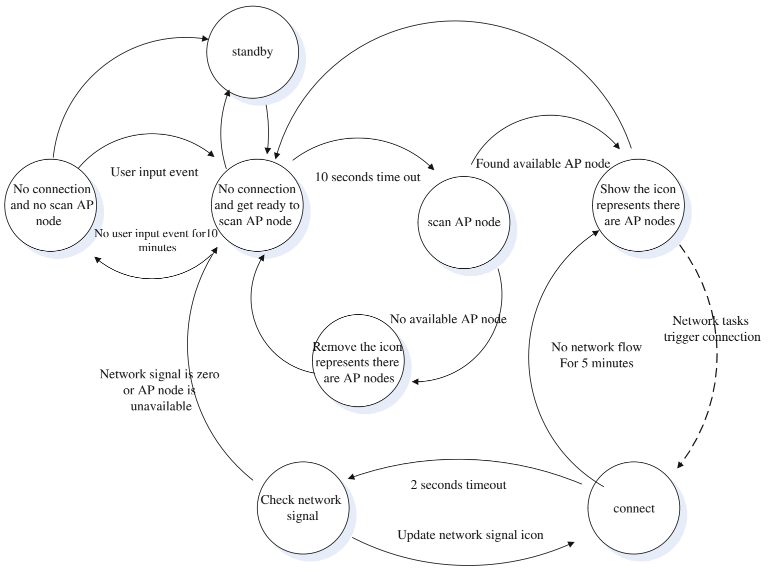

standby
No connection and no scan AP node
User input event
No user input event for10 minutes
No connection and get ready to scan AP node
10 seconds time out
scan AP node
Found available AP node
Show the icon represents there are AP nodes
No available AP node
Remove the icon represents there are AP nodes
Network signal is zero or AP node is unavailable
No network flow For 5 minutes
Network tasks trigger connection
Check network signal
2 seconds timeout
Update network signal icon
connect

Fig. 127.3 State transition diagram of Wi-Fi network management

Previously, Wi-Fi connecting application is only used in the scenario that the user knows there is existing AP node around. Users launch connection manually when they want to use wireless communication. Now, in order to encourage users to use network applications by searching available AP nodes automatically, a Wi-Fi network manager is proposed. The manager is triggered when user inputs. It scans available AP node periodically and notifies the user that there are AP nodes around if any AP node is found. But the network will not be connected immediately. The connecting process is triggered automatically when the user needs to access Internet.

When the user presses button or clicks touch screen, wireless module will be started and the scan process will be triggered. If there is any AP node found around, a notify icon will be shown on the status bar to notify that user network is available. At the same time, the scan results will be saved in disk. Since the user may move out of the coverage of the AP node at any time, the Wi-Fi network manager will scan one time per ten seconds. If there is no available AP node, the notify icon will be removed. If the user use network application, connecting process will directly use those AP nodes have been saved in the disk. In our system, the process of scanning takes 4 s, scanning plus connecting takes 10 s. If the saved AP nodes are used, the process of connection will be finished within 6 s.

In order to save energy, if there is no network traffic continuously for 5 min, network will be disconnected. In addition, Wi-Fi network manager checks the status of the AP node periodically. The network will be disconnected immediately if the AP node is unavailable.

127.3.3 XML-Based Network Service

All operations are completed by the remote server. Terminals take charge of uploading and downloading without local database. Terminal interacts with server in request and response mode. Messages of request and response are programed with XML. Hence, it is costless when user changes switch device from one to another. Users can switch easily between terminals with the same account. Updating software will not cause any loss of user data.

XML messages start with <?xml version = "1.0" encoding = "UTF-8"?>. The format of message is defined with XMLSchema following the international standard in http://www.w3.org/2001/XMLSchem-a. Compression(gzip) is specified in "Accept-Encoding" in http request message header, "Content-Length" contains the compressed data size [5]. The field containing complex information in the XML content of response message, such as station message which contains a link to the content, is represented using CDATA [6]. User-defined tags are customized for special objects.

This approach is more flexible and highly efficient than mobile web browser since it enables users to combine different types of services to carry out specialized functions.

An XML parser (Libireader) designed with Python language is provided in the toolkit. Developers are required to configure Libireader with demo XML files. The script can create c++ code according to the structure of demo XML files.

127.3.4 Resource Scheduler

Researchers have proposed application adaptation which changes parameters to trade off service quality for resource usage [7–9].

Due to the limitations of mobile devices' resource including CPU, network bandwidth, memory and IO. There is a strong need for adjusting the service quality of coexist processes according their priorities and users' preferences. User's preference is acquired by user's input. The coexist processes are divided into two queues: foreground processes and background processes. The adjustment is triggered by three changes: the change of the number of coexist processes, exhaustion of residual resource, weak network signal below the baseline. The principle of adjustment is trying to guarantee the foreground process by reducing the service quality of background processes. When adjustment is triggered, background processes are notified to pause by calling Linux system call "pause()". The paused process can be waked up by sending SIGCONT to it when it becomes foreground process. When storage capacity is exhausted, the background process with the lowest priority will be killed to save memory.

Table 127.1 Electrical current of every state

State	Current (mA)
Standby	0.8
No connection and operate	350
No connection and get ready to scan and operate	370
No connection and get ready to scan and no operation	70
Scan	370
Connected and no operation	70
Connected and operate	450

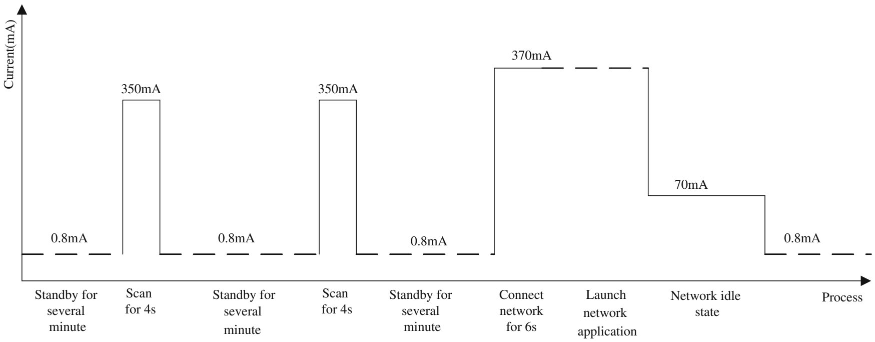


Fig. 127.4 The changing progress of electrical current

127.4 Experimental Evaluation

We have successfully implemented a prototype of the proposed framework. The hardware platform is a E-paper reading device with a single Samsung S3C2416 CPU, 128 MB SDRAM, Epson S1D13521 display controller. This CPU supports speeds of 400 MHz. The GUI is Qt 4.5.0. The Wi-Fi module is marvell sd8686 and the 3G module is EM770 W. The operating system is Linux with a modified kernel 2.6.11.7. The electrical current of every state is tested. The statistic results are shown in Table 127.1.

The changing progress of electrical current in our system can be depicted qualitatively in Fig. 127.4. The experimental results demonstrate that the proposed design can help system to save energy and meet the utility demands.

127.5 Conclusion

The development of wireless network has changed the dissemination of information carriers and modes of transmissions. E-books, web pages, news aggregators and blog networks and other forms of communications are beloved by users for their hard real-time property, rich in content and diverse way to show characteristics. In future, mobile devices might become the default physical interface for ubiquitous computing applications. This paper presented a wireless network-based software framework of in E-paper devices. All challenges are analyzed and efficient approach is proposed for meeting the challenging demands. We believe that our work will enable more benefits in a manner that greatly enhances speed of developing applications. We believe that the principles will be extensible to other embedded system. We hope it can be found useful in more applications.

Acknowledgments The authors would like to thank Tianjin Jinke Electronics Co., LTD for providing the hardware platform, other members of the Embedded System and Information Security Lab for their informative discussions. This work was supported by the China National High Technology Research and Development Program 863 (Granted No. 2008AA03A334).

References

1. Heikenfeld J, Drzaic P, Yeo JS, Koch T (2011) Review paper: a critical review of the present and future prospects for electronic paper. J Soc Inf Disp 19:129
2. Li QC, Zhang L (2008) Platform independent embedded multi-format parsers and pivotal problems. Comput Appl 28(4):1039–1041
3. Weiser M (1991) The computer for the 21st century. Sci Am 265:66–75
4. Weiser M, Brown JS (1997) The coming age of calm technology. In: Denning PJ, Metcalfe RM (eds) Beyond calculation: the next fifty years. Springer, New York, pp 75–86
5. http://www.ietf.org/rfc/rfc2616.txt 2011
6. http://doc.qt.nokia.com/4.7/index.html 2011
7. Li B, Nahrstedt K (1999) A control-based middleware framework for quality of service adaptations. IEEE J Sel Areas Comm 17(9):1632–1650
8. Mesarina M, Turner Y (2002) Reduced energy decoding of MPEG streams. In: Proceedings of SPIE multimedia computing and networking conference, Jan 2002
9. Noble B, Satyanarayanan M, Narayanan D, Tilton J, Flinn J, Walker K (1997) Agile application-aware adaptation for mobility. In: Proceedings of the 16th symposium operating systems prinsciples, pp 276–287

Chapter 128
The Dual Polarized Ceiling-Mounted MIMO Antenna Analysis and its Implementation

Gao Feng, Zhu Wentao, He Jiwei and Liu Xu

Abstract Transmit diversity schemes have been studied for high spectral-efficiency and high bit-rate transmission, such as multi-input multi-output (MIMO) systems. In this paper, a novel dual polarized ceiling-mounted antenna is designed and simulated. In the MIMO systems, forward error correction coding is essential for high quality communications. At the two typical scenes, the single-input single-output system compared with two single polarized antennas and an dual polarized antenna MIMO system is tested. It is shown that the MIMO system have improved throughput level obviously. The dual polarized ceiling-Mounted dual streams MIMO antenna will save much construction cost in the long-term evaluation systems in the future, which can be called the "green" MIMO antenna.

Keywords MIMO antenna · Ceiling-mounted · TD-LTE · Dual polarized

128.1 Introduction

The third generation partnership project (3GPP) has launched the study item of the LTE system. There are two kinds of techniques for LTE:TDD and FDD LTE. As the biggest globe operator, China Mobile Communications Corporation

G. Feng (✉) · Z. Wentao · H. Jiwei · L. Xu
16A, Danling Street, Haidian District, People's Republic of China
e-mail: aa_little@163.com

Z. Wentao
e-mail: zhuwentao@cmdi.chinamobile.com

H. Jiwei
e-mail: hejiwei@cmdi.chinamobile.com

L. Xu
e-mail: liuxu@cmdi.chinamobile.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_128,

(CMCC) focused on the TD-LTE. In Evolved TD-LTE, the target peak data rate is up to 100 Mb/s for the downlink and 50 Mb/s for the uplink. The transmission data rate can be improved by applying orthogonal frequency division multiplexing (OFDM) and multiple-input multiple-output (MIMO) antennas which is tolerant of multi-path interference.

The use of MIMO antennas can provide high spectral efficiency and link reliability for point-to-point communication in fading environments [1, 2]. However, it is difficult to install multiply antennas at one spot in the building. This paper solved the problem by designing a dual polarized ceiling-mounted antenna, whose horizontal polarized element covered the GSM1800 MHz, TD-SCDMA, WCDMA, TD-LTE and WLAN frequency bands, and the vertical polarized element covered GSM, CDMA, WCDMA, CDMA2000, TD-SCDMA, TD-LTE and WLAN systems. The dual polarized ceiling-mounted antenna can be used as a MIMO antenna instead of the two single polarized ceiling-mounted antennas in LTE system, so this antenna can be called "green" antenna using dual streams technique.

128.2 MIMO Network System Modeling

Compared with the single-input single-output (SISO) system, the MIMO system has been developed to increase the spectral efficiency. MIMO can obtain the spatial diversity gain by using channel matrices for recovering transmitted data streams of TD-LTE BS. The system is shown in Fig. 128.1.

The MIMO system is modeled with Nt transmit and Nr receive antennas, where $Nr \geq Nt$. The channel impulse response from the transmit antenna i to the receive antenna j can be represented by $h_{i,j}$. For the system with two transmit and two receive antennas, assuming that the signal $s = [s1\ s2]$ are transmitted from each transmit antenna at time t, the received signal can be represented as follows [3].

$$\begin{bmatrix} r_1 \\ \vdots \\ r_{Nr} \end{bmatrix} = \begin{bmatrix} h_{11} & \cdots & h_{1Nt} \\ \vdots & \ddots & \vdots \\ h_{Nr1} & \cdots & h_{NrNt} \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_{Nt} \end{bmatrix} + \begin{bmatrix} n_1 \\ \vdots \\ n_{Nr} \end{bmatrix} \tag{128.1}$$

where $r_{i,j}$ represents the complex random variable representing an additive white Gaussian noise channel. The channel response matrix can be represented as follows by applying the singular value decomposition.

The upper bound of SINR of the mobile receiver for a cell with K users is obtained as follows [4]

$$\mathrm{SINR}^k = \frac{\det[H_k^{-1} H_k]}{\sigma^2_{v_1^k} + \sigma^2_{v_2^k}} \tag{128.2}$$

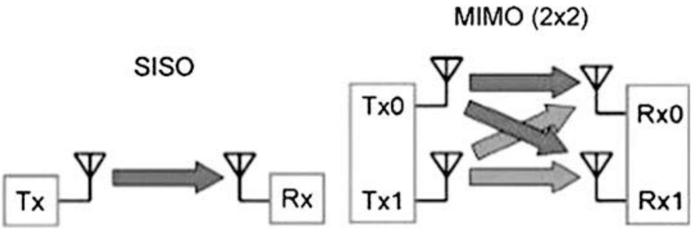
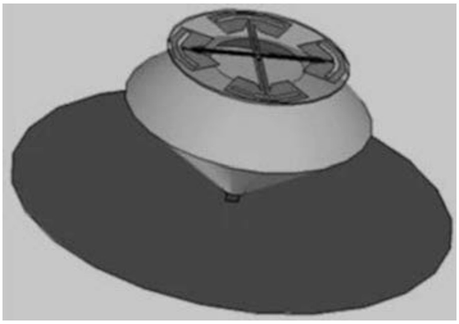


Fig. 128.1 Antenna and channel of SISO and MIMO (2 × 2) system

Fig. 128.2 Inner structure of the dual-polarized antenna



Under the Gaussian approximation [5], the bit error rate (BER) can be written with the SINR as

$$\overline{\mathrm{BER}}^{k} = Q(\sqrt{\mathrm{SINR}^{k}}) \tag{128.3}$$

where Q(·) is Gaussian approximation function.

128.3 Simulation and Analysis of Dual-Polarized Indoor Antenna

Figure 128.2 shows the inner structure of the dual-polarized antenna. The horizontal polarized element is formed with four folded dipoles which are rotating feed, and the vertical polarized element is formed with the inverted-cone monopole and the bottom board. Because the horizontal polarized element and vertical polarized element are coaxial, they can decrease the interplay, reduce the interference, and increase the isolation with each other.

The frequency range of the dual-polarized antenna as shown in Fig. 128.2 covered 824–960, 1710–2700 MHz by vertical polarized element and covered 1710–2700 MHz by horizontal polarized element. The vertical polarized element is equivalent to the traditional single polaried ceiling-mounted antenna, which can cover the frequency range of GSM, CDMA, WCDMA, CDMA2000, TD-SCDMA,

Table 128.1 Radiation parameter of the dual-polarized antenna

Vertical polarization			Horizontal polarization		
Antenna gain (dB)	HPBW (°)	Non-circularity(dB)	Antenna gain (dB)	HPBW (°)	Non-circularity (dB)
5.1–7.7	42–64	±1–±2	3.2–5.9	55–71	±0.33–±1.4

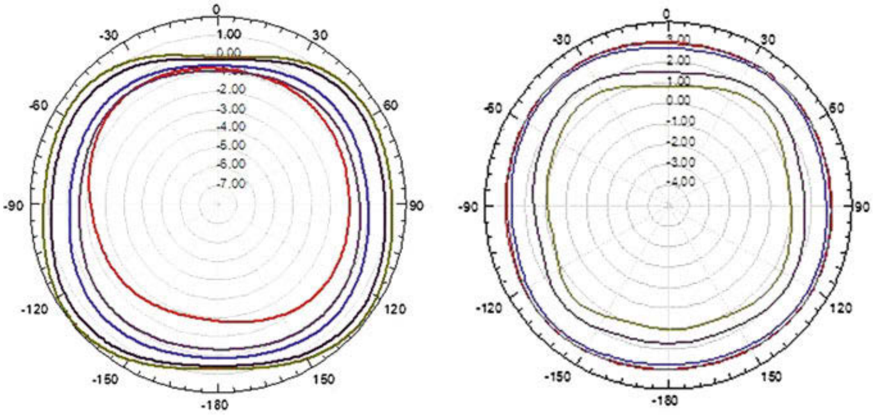


Fig. 128.3 Non-circularity of the dual-polarized antenna. **a** non-circularity of vertical polarization. **b** non-circularity of horizontal polarization

TD-LTE, WLAN and so on, the horizontal polarized element can cover the frequency range of WCDMA, TD-SCDMA, TD-LTE, WLAN systems.

The antenna has been simulated and analyzed by the Ansoft HFSS software. The antenna gain, Half Power Beam Width (HPBW) and non-circularity of the antenna are shown in Table 128.1, Fig. 128.3. Different curves in Fig. 128.3 represent different working frequencies (1.88G, 2G, 2.1G, 2.3G and so on)

From the simulation and analysis, we can see that the dual-polarized antenna has good non-circularity and high gain.

128.4 Application Testing

The indoor TD-LTE system using three-type antennas was built in three environments respectively: one monopole antenna, two monopole antennas and one dual-polarized antenna. By comparing the throughput of TD-LTE system with different antenna types, we can evaluate the impact of the dual-polarized antenna. The configurations of TD-LTE system were listed in Table 128.2.

The test system satisfied the following conditions: (1) Indoor distribution system and corn network equipment operate normally; (2) The testing terminal is in working order, and common indoor ceiling monopole antenna is selected in the comparing test. The frequency is 806–960 and 1710–2500 MHz.

Table 128.2 Configurations of TD-LTE system

Center frequency	Bandwidth	AMC	Harq
2.35 GHz	20 MHz	Enable	Enable
Frame structure	Uplink-downlink	Uplink power control	Adaptive MIMO
FS2	Configuration 1	Enable	Enable
Special subframe	CP	RB	Transmission mode
Configuration 7	Short CP	24 RB	TM3

Table 128.3 Measured results in the near area

Parameters	Two single polarized antennas (0.3 m distance)	Two single polarized antennas (0.9 m distance)	Two single polarized antennas (1.5 m distance)	A dual-polarized antenna
RSRP1	−83.67	−82.832	−83.129	−82.399
RSRP2	−84.92	−86.351	−84.835	−87.152
NTA	2.5254	3.5489	4.8713	4.4384
RI	2	2	2	2
MCS1	26.585	26.563	26.662	27.327
MCS1 variance	1.6296	2.7404	2.6886	1.0238
MCS2	26.57	26.56	26.658	27.325
MCS2 variance	1.6416	2.793	2.7138	1.037
SINR	30.358	31.515	31.525	32.354
DLL1 Thput (kbps)	15821	16867	16911	18038
UL MCS	27.75	27.75	27.75	27.5
UL L1 Thput (kbps)	12794	12690	12691	12691
DL L3 Thput	15447	16437	16406	17450
UL L3 Thput (kbps)	12630	12641	12663	12690

The indoor environments are various. In this testing we choose two representative scenes: one is the opening scene just like corridor, another is the closed scene similar to office. In both scenarios, the throughput of terminal wear measured in the near, mid and far area.

When testing terminal access in the TD-LTE system, we download and upload data from FTP severing for 1 min respectively. The following parameters are recorded: NTA, RI, MCS, RSRP, SINR, Uplink traffic of L1, Downlink traffic of L1, Uplink and downlink traffic of L3.

The measured results in the near area using dual-polarized antenna and using one monopole antenna are listed in Table 128.3. The measurements are the average of sample values.

128.5 Conclusion

From the testing of the TD-LTE system, it is shown that the dual-polarized ceiling-mounted antenna have good performance. Through the actual practice, it can be concluded that:

The dual-polarized antenna worked as good as two single antennas, at the closed scene;
The dual-polarized antenna is preferred than two single antennas at the opening scene, and its throughput is twice as much as the SISO system in the near field;
The throughput of dual-polarized antenna is 1.2–1.4 times than the SISO system in the far field;
The dual-polarized antenna used in this article has good performance. Using this antenna can save much materials and improve the network capability.

References

1. Tarokh V, Seshadri N, Calderbank AR (1998) Space-time codes for high data rate wireless communications: performance criterion and code construction. IEEE Trans Inform Theory 44(2):744–765
2. Foschini GJ, Gans MJ (1998) On limits of wireless communications in a fading environment when using multiple antennas. Wirel Pers Commun 6:311–335
3. Futaki H, Ohtsuki T (2003) LDPC-based space-time transmit diversity schemes with multiple transmit antennas. In: Proceedings of IEEE V TC2003 Spring
4. Xiao Y, Zhao Y, Lee MH (2006) Canceling co-channel interference for MIMO CDMA systems. In: Proceedings of the 8th international conference on signal processing, Beijing
5. Xiao Y, Lu L-Y, Wang Y-C (2004) Space-time spreading in downlink of TD-SCDMA systems. In: Proceedings of the IEEE 6th circuits and systems symposium on emerging technologies: frontiers of mobile and wireless communication, vol 2, pp 635–638

Chapter 129
Research on Embedded Browser Based on Qt WebKit for E-Paper Devices

Qing-Cheng Li and Zhan-Ying Zhang

Abstract In recent decades, there has been an accelerating trend in the use of diverse mobile embedded devices which have been an important part in users' daily life. Since mobile users need to access the rich resource of the World Wide Web, mobile embedded browser has been a necessary part of an embedded system. This paper analyzed the related techniques and framework of Qt WebKit and developed an embedded browser based on Qt WebKit in E-paper mobile devices. In addition, according to the characteristics of E-paper screen, we proposed some key problems about the implementation. Some novel and effective solutions to address these problems were also proposed. Experimental results demonstrate that this embedded browser can reach the desired benefits.

Keywords E-paper · Mobile device · Wireless · WebKit · Qt · Embedded browser

129.1 Introduction

The World Wide Web is one of the most important information sources today. At present, users tend to use diverse mobile devices such as mobile phones and Personal Digital Assistants (PDAs) pervasively.

E-paper is a kind of display technology which is designed to display reflected natural light using electrophoretic colored particles [1, 2]. This kind of display

Q.-C. Li · Z.-Y. Zhang (✉)
College of Information Technical Science, Nankai University,
Weijin Road 94, Tianjin 300071, China
e-mail: zhangzhanying@mail.nankai.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_129,

material is applicable for protracted reading and brings comfortable reading experience. The advantages of E-paper device are not only the capability of storing books, but also the ability of browsing in the internet, subscribing and downloading books. These allow users a lot of behaviors without the need of connecting to a desktop computer. Users can read anytime and anywhere.

Embedded browser is an essential application in mobile devices. There have already been some embedded browsers in existence such as IE, Gzilla, opera and Minimo. However, some of them do not support Chinese, some are not open source code, some do not support Linux platform and some have poor render effects [3, 4]. WebKit is an open source browser kernel engine with high render efficiency, small memory occupancy, portability and compatibility.

WebKit is mainly constructed with three parts: WebCore, JavascriptCore and Ports [5]. WebKit has nearly all functional features of browser in the desktop computer. WebKit has been adopted in many mobile phones, such as Google's Gphone, Apple's iPhone, Nokia's Series 60 browser. WebKit is not a browser with full functions but an engine. Hence, application developers need to develop web browser using WebKit as kernel to provide specified services.

This paper developed an embedded browser based on WebKit for E-paper devices. This browser has the following functions:

Downloading HTML/XHTML file.
Parsing HTML/XHTML file.
Producing DOM tree of document objects.
Composing visible HTML elements by allocating position, height and width with the layout manager.
Displaying the HTML document on the screen.

This program is developed with Qt for Embedded Linux which is a C++ framework for GUI and application development for embedded devices. It runs on a variety of processors, usually with Embedded Linux. Qt for Embedded Linux provides the standard Qt API for embedded devices with a lightweight window system [6].

However, it is difficult to implement embedded browser in E-paper devices for some inherent limitations about special screen refreshing mechanism, computing ability, storage capability and bandwidth. Either flow rate is insufficient or CPU is over-occupied [7]. The following are the requirements for embedded browser:

- Compared to conventional TFT-LCD and STN-LCD, the refreshing speed of E-paper screen is slower. This feature requires that the refresh frequency cannot be too high.
- It requires high efficiency to provide high quality service. But wireless network is more unstable than wired network because of the uncertainty and complexity of various kinds of environments.
- Achieving expected lifetime of battery is a key criterion for battery powered mobile devices. Energy consumption must be controlled in an acceptable range.

Based on this, we proposed the key problems of designing embedded browser in E-paper devices and corresponding solutions.

We have successfully implemented the proposed techniques mentioned above in an E-paper mobile reading device. Experimental results have shown satisfactory results.

The remainder of this paper is organized as follows.

Section 129.2 talks about key problems of the design and respective solutions. Then, in Sect. 129.3 we describe our implementation and preliminary experimental results. Finally, Sect. 129.4 sums up our discussions.

129.2 Key Problems and Solutions

Here, we list the following several key problems of designing embedded browser in E-paper devices.

129.2.1 Refreshing Policy

The refreshing speed of E-paper is relatively slow as mentioned in the previous section. The system performance will be reduced significantly if the screen is refreshed frequently. For instance, when the web browser loads a URL, the page will be shown part by part. Hence, there are many child frames before the whole page is shown. In addition, E-paper has special refresh mechanism, unlike CRT and TFT dynamic active refresh. E-paper adopts a kind of static positive refreshing policy that it refreshes only when the data on the screen updates. Based on this feature, there are many times of refreshments during loading a web page. The latency time of loading is increased since CPU is over-occupied by refreshing tasks.

Another problem is caused by the feature of E-paper's refreshment mode. The E-paper display driver program provides three refreshing functions for upper level applications:

Full screen print: it refreshes the whole screen in a time, it is the slowest of the three methods.

Partial screen print: it refreshes a rectangle block of screen, it supports grayscale print.

Fast partial screen print: it is the fastest in three mode. It is a black–white print mode which means that a grayscale image will be shown as a black–white two-color image.

When loading a web page, the screen will flicker frequently if full screen print mode is used and the users' satisfactions will be declined. But the second and third print modes may cause retained image of the previous frame if the screen is refreshed frequently.

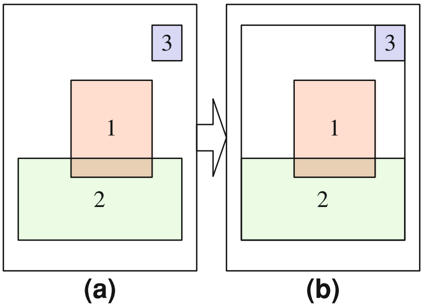


Fig. 129.1 The principle of deferred IO

We adopted the principle of Deferred IO mechanism to address the problem. The principle of Deferred IO is shown in Fig. 129.1.

The display driver does not refresh immediately when refreshing request comes. Instead, the refreshing action will be deferred to execute. It refreshes the screen once per 100 ms. When timer is timeout, all screen refreshing requests are combined into a single request by integrating these refreshing areas together. At the same time, the refreshing mode is decided by the size of the combined refreshing area. There is a previously defined minimum area threshold. If the refreshing area is larger than this threshold, full screen print will be used, else partial screen print will be adopted.

This mechanism can effectively reduce the frequency of refreshment to a certain degree. However, this mechanism still has a problem. There will be a progress bar shown on the screen to notify the progress of loading when the browser loads a URL. The progress bar changes its value every time it receives the load progress signal of QWebView. The signal is emitted every time an element in the web page completes loading and the overall loading progress advances. This signal tracks the progress of all child frames. Since the signal continually comes for every little change, the screen will also be refreshed frequently. Furthermore, the progress bar is located at the bottom of the screen. This causes the combined area to be refreshed often surpassing the threshold. Then the screen will be refreshed with full screen print mode many times. User performance is reduced due to flickering of the screen. In order to address this problem, coarse progress granularity is adopted to constrain the times of refreshment. If the progress distance between adjacent refreshment is less than 30%, the screen will not refresh.

129.2.2 Network Manager

A Qt for Embedded Linux application requires a server application to be running, or to be the server application itself. Any Qt for Embedded Linux application can act as the server. When more than one application is running, the subsequent applications connect to the existing server application as clients. The server and

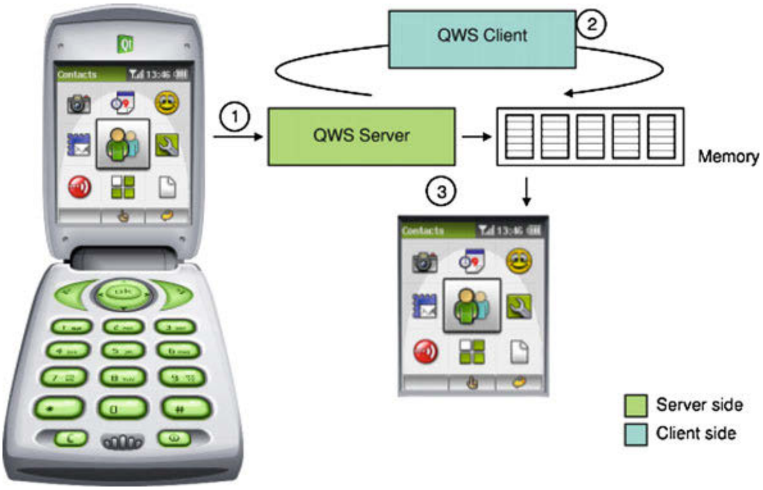


Fig. 129.2 Event processing in Qt [7]

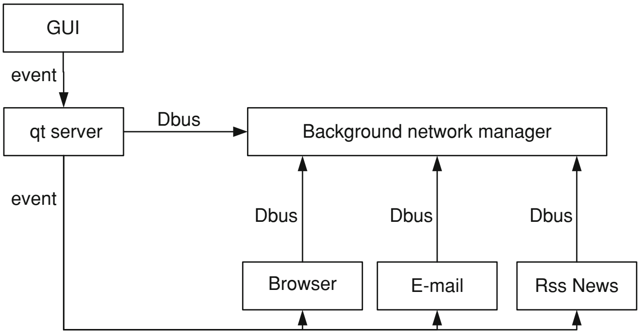


Fig. 129.3 Network connecting management

client processes have different responsibilities: the server process manages pointer handling, character input, and screen output. In addition, the server controls the appearance of the screen cursor and the screen saver. The client process performs all application specific operations [8].

As shown in Fig. 129.2, all system generated events, including keyboard and mouse events, are passed to the server application which then propagates the event to the appropriate client.

There are two network connecting management strategies used in the system. The first is shown in Fig. 129.3.

In the first method, all applications that need to use network are registered in the Qt server. When a received event is sent to network application the Qt server will check if network is available. If not, it will send a Dbus message to notify the background network manager to launch a connecting process.

In the second method, checking network status function is added in all network-related places in Qt network library including socket, ftp, http. In this function, first, it checks whether the network has been connected, because a file will be created as a symbol when connect program successfully connects the network. If this file exists, then the network is available, else a Dbus message of connecting request is sent to notify the background network manager to launch the network connecting process.

129.2.3 Performance Enhancement

Dynamic link libraries can obtain the benefits of good extendibility and saving memory at the expense of slower launching speed. Since the functions' addresses are unknown during compiling, they need to be searched in the function symbol table while the process is launching. This process will take most of the starting time (about 80%). Prelink tool is used to obtain the loading addresses before the process is launched. In addition, a kind of preload method is also adopted. We insert a "pause()" in the main entry function. When the process is launched, it will stop execution there after dynamic link libraries are loaded. When the process needs to response to user, the process continues running by sending a SIGCONT signal to it.

In addition, some necessary URL error correction approaches are adopted. There are usually some minor mistakes during users, input URL. For example, the user may incorrectly input "ww" for "www". These errors are corrected automatically through analyzing the inputted text.

Mobile device's input tools are normally stylus and keystroke. For those devices without touch screen, it is difficult to select a hyperlink of interest with few keys. We developed hyperlink mode in addition to normal browsing mode to switch the focus among hyperlink nodes.

129.3 Experimental Test

We have successfully implemented a mobile web browser to test the proposed methods. The hardware platform is an E-paper reading device with a single Samsung S3C2416 CPU, 128 MB SDRAM, Epson S1D13521 display controller. This CPU supports speeds of 400 MHz. The GUI is Qt 4.5.0. The Wi-Fi wireless module is marvell sd8686. The operating system is Linux with a modified version of Linux kernel 2.6.11.7.

First, the launching time of embedded browser is tested. We used two E-paper reading devices with the same hardware configurations, one of them used Preload and Prelink and the other has not, to launch our embedded browser simultaneously.

Table 129.1 Launching time of embedded browser

Device	Average launching time (s)
With preload and prelink	1.87
No preload and prelink	3.96

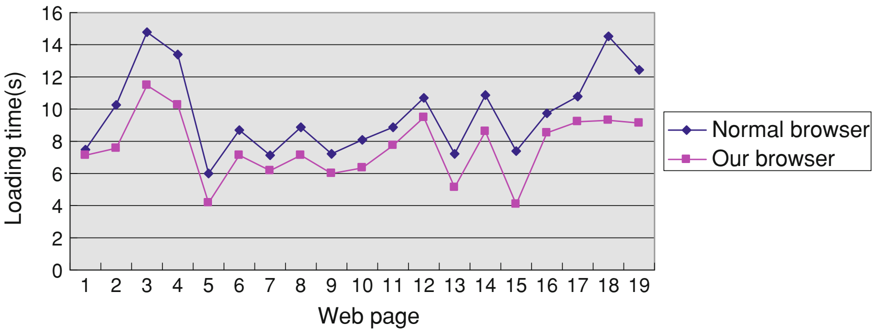


Fig. 129.4 Comparison of loading URL time

We have tested 20 times and their launching time is tested. Experimental results are shown in Table 129.1.

Second, the latency time of loading web pages are tested. Test data are 20 randomly selected web sites. We run our test program to browse these web sites. Test results are shown in Fig. 129.4.

129.4 Conclusion

HTML is the most widely used information format on the internet. Mobile handheld devices are expected to become the default physical interface for ubiquitous computing applications. Embedded browser is significant in mobile handheld devices. We developed an embedded browser based on WebKit for E-paper devices and proposed key problems of the design. Experimental results are stable and reasonable. They demonstrate that the methods in this paper reach the desired benefits. We believe that the principles will be extensible to other mobile applications. We hope it can be found useful in more applications.

In the future, we will continue to explore the way to improve the performance of the embedded browser. For example, since E-paper screen can only show gray-scale image, some useless rendering functions, such as 3D rendering and color rendering, can be removed or simplified in order to improve loading speed. In addition, disk cache can be used to enhance performance. But the content in client may not be consistent with the server. The speed of writing Nand-flash is also a bottleneck.

In addition, we will try to combine with user preference to provide more personalized service. Our work will enable more benefits in a manner that greatly enhances the user experience.

Acknowledgments The authors thank Tianjin Jinke Electronics Co., LTD for providing the hardware platform and other members of the Embedded System and Information Security Lab for their informative discussions. This work was supported by the China National High Technology Research and Development Program 863 (Granted No. 2008AA03A334).

References

1. Heikenfeld J, Drzaic P, Yeo JS, Koch T (2011) Review paper: a critical review of the present and future prospects for electronic paper. J Soc Inf Disp 19:129
2. Li QC, Zhang L (2008) Platform independent embedded multi-format parsers and pivotal problems. Comput Appl 28(4):1039–1041
3. Zhao JW, Zhou Y, Wang ZQ, Du S (2009) Research and impelmentation of embedded browser based on Webkit. Electron Meas Tech 34(3):135–138
4. Yang FM, Li J, Zhou ZY, Hu GR (2003) Design and implementation of embedded browsers. Comput Eng Sci 25(4):39–41
5. http://www.webkit.org/, 2011
6. http://doc.trolltech.com/4.7-snapshot/qt-embedded-linux.html, 2011
7. Satyanarayanan M (1996) Fundamental challenges in mobile computing. In: Proceedings of 15th annual ACM symposium on principles of distributed computing, ACM Press, pp 225–233
8. http://doc.qt.nokia.com/4.6/qt-embedded-architecture.html, 2011

Chapter 130
Research of Network Intrusion Detection System Based on Data Mining Approaches

Xiao-chun Guo, Dong-mei Ma, Ying-juan Sun and Hong-ying Ma

Abstract This paper provides a network intrusion detection system based on data mining approaches. The framework of the intrusion detection system and the function of components are introduced. An anomaly intrusion detection system is implemented based on association rule. This system does not depend on experiences, it has flexibility.

Keywords Network security · Anomaly detection · Data mining · Association rule

130.1 Introduction

As network-based computer systems play increasingly vital roles in modern society, they have become the targets of our enemies and criminals. Therefore, we need to find the best ways possible to protect our systems.

Xiao-chun Guo (1972–)—Female, Shenyang, Master, Associate professor, main research direction for information security and confidentiality.

X. Guo · D. Ma · H. Ma
School of Information Engineering, Shenyang Broadcasting TV University,
Shenyang 110003, Liaoning, China
e-mail: guoxc72@163.com

D. Ma
e-mail: mdm1226@sina.com

H. Ma
e-mail: mahy@sytvu.cn

Y. Sun (✉)
College of Computer Science and Technology, Changchun Normal University,
Changchun, China
e-mail: syj_pyf@sohu.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_130,

Intrusion prevention techniques, such as user authentication, avoiding programming errors, information protection and firewall technology have been used to protect computer systems as a first line of defense. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weaknesses in the systems due to design and programming errors, or various "socially engineered" penetration techniques. Intrusion detection is therefore needed as another wall to protect computer systems.

Currently, many intrusion detection systems are constructed by expert systems or based on statistical methods, which need more experience. Data mining approaches have the advantage that they extract knowledge and rules which people are interested in from a large number of data. Data mining approaches do not rely on the experience. Intrusion detection system based on data mining approaches [1] can help to find knowledge and rules from the system logs, audit data, and network traffic etc. Network security using this technology is a new attempt at home and abroad.

130.2 Intrusion Detection

Intrusion detection can identify people who are not authorized to use computer systems (e.g. hacking), and authorized users who have abused their authority (such as internal attack).

Intrusion detection techniques can be categorized into misuse detection, which uses patterns of well-known attacks or weak spots of the system to identify intrusions; and anomaly detection, which tries to determine whether deviation from the established normal usage patterns can be flagged as intrusions.

According to the source of test data [2], intrusion detection system can be divided into host-based intrusion detection system and network-based intrusion detection system. Host-based intrusion detection system can detect possible intrusions by analyzing audit data and system logs. Network-based intrusion detection can detect possible intrusions system by analyzing network packets.

130.3 Systematic Framework

We want to build a network-based anomaly detection model [1, 3, 4]. The premise of establishing model is that the behavior of hosts and servers can reflect some laws during the long running in the network. For example, what is the server which hosts frequently visit, which ports of the server are frequently visited. In the learning phase of establishing model, we should collect data of normal network conditions, so that the law (anomaly detection model) is the normal state of the behavioral pattern of the host and server. In the detecting phase, if there is some connection that does not meet these rules, we have to think that these connections are abnormal.

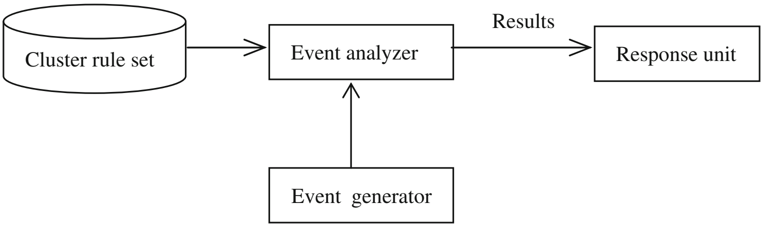


Fig. 130.1 Framework of anomaly intrusion detection system

The framework of anomaly intrusion detection system shown in Fig. 130.1 consists of event generator, event analyzer, response unit and cluster rule set (anomaly detection model). Event generator can capture network packets. Event Analyzer analyzes the network packets which is obtained from the event generator according to the association rules of cluster rule set, and produces results. Response unit can respond to the results of the analysis. Cluster rule set (anomaly detection model) describes the characteristics of the user's normal behavior.

130.3.1 Event Generator

In order to establish a TCP connection, two sides of connecting need to be a three-way handshake, therefore a connection contains multiple IP packets. Multiple packets belonging to the same connection should be merged into a connection record.

A connection record contains the following attributes:
(time, duration, service, src_host, dst_host, src_bytes, dst_bytes, flag)
Time: the start time of connection.
Duration: the time of connection from the beginning to the end.
Service: connection application protocol, such as WWW, FTP, DNS, Telnet, etc.
src_host: source host.
dst_host: destination host.
src_bytes: the bytes which source host sends.
dst_bytes: the bytes which destination host sends.
flag: the connection status, normal end state and the states whose connection requests are rejected etc.

130.3.2 Cluster Rule Set (Anomaly Detection Model)

Cluster rule set describes the characteristics of the user's normal behavior. Therefore, in the learning phase of establishing anomaly detection model, data of user's normal behavior requires collecting. The following describes the process of modeling.

130.3.2.1 Mining Association Rules

The goal of mining association rules is to derive multi-feature (attribute) correlations from a database table. Association rules define as follows:

$I = \{i_1, i_2, \cdots i_m\}$ is called itemset, let D be a transaction database, any subset of I is called $T(T \subseteq I)$, and has a unique identifier ID. Define support(X) as the percentage of transactions (records) in D that contain X. An association rule is the expression $X \to Y, s, c$. Here $X \subset T$, $Y \subset T$, and $X \cap Y = \Phi$. $s = \text{support}(X \cup Y)$ is the support of the rule, and $c = \dfrac{\text{support}(X \cup Y)}{\text{support}(X)}$ is the confidence.

The purpose of mining association rules is to identify credible and representative rules, so a minimum support threshold and a minimum confidence threshold should be given. Mining association rules is to derive association rules which the support and confidence exceed in the specified thresholds. This mining process can be divided into two steps: First, identify all the frequent itemsets that are all itemsets whose support is not less than the minimum support threshold. Second, construct the rules whose confidence is not less than the minimum confidence threshold from the frequent itemsets in the first step.

Here, we introduce the Apriori algorithm of mining association rules [5]. Introduce a number of notation: k-itemset: itemset of size k; L_k: frequent itemset of size k; C_k: Candidate itemset of size k.

(1) Apriori algorithm: find all frequent itemsets

Input: database D; the minimum support threshold min_sup.

Output: the frequent itemsets L in D.

```
    $L_1$=find_freguent_1-itemsets (D);
    for (k=2; $L_{k-1} \neq \Phi$;k++) {
        $C_k$=apriori_gen ($L_{k-1}$, min_sup);
        for each transaction t ∈ D{ //scan D for counts
            $C_t$=subset ($C_k$, t); //get the subsets of t that are candidates
            for each candidate c ∈ $C_t$
            c.count ++;
      }
      $L_k$={c ∈ $C_k$ | c.count≥min_sup}
    }
    return L= $\wedge_k L_k$;
            procedure apriori_gen ($L_{k-1}$:frequent (k−1)-itemsets; min_sup:minimum support threshold)
        for each itemset $l_1$ ∈ $L_{k-1}$
            for each itemset $l_2$ ∈ $L_{k-1}$
            if ($l_1[1]=l_2[1]$) ∧ ($l_1[2]=l_2[2]$) ∧ … ∧ ($l_1[k-2]=l_2[k-2]$) ∧ ($l_1[k-1]<l_2[k-1]$) then {
            c=$l_1 \infty l_2$; //join step: generate candidates
            if has_infrequent_subset (c, $L_{k-1}$) then
```

```
                delete c; //prune step:remove unfruitful candidate
            else add c to $C_k$;
          }
      return $C_k$;
            procedure has_infrequent_subset (c:candidate k-itemset; $L_{k-1}$:frequent(k−1)-itemset)
    for each (k−1)-subset s of c
      if s $\notin L_{k-1}$ then
         return TRUE;
     return FALSE;
```

(2) generating association rules

1) For each frequent itemset l, generate all non-empty subset of l.

2) s is called each non-empty subset of l, if $\frac{\text{support\_count(l)}}{\text{support\_count(s)}} \geq \text{min\_conf}$, generate association rule: "$s \rightarrow (l - s)$, min_sup, min_conf". The min_conf is called minimum confidence threshold.

In analysis network traffic, a connection is called a transaction *T*, transaction database *D* is composed of many connection records. Each transaction *T* is composed of duration, service, src_host, dst_host, src_bytes, dst_bytes, flag. Time is the Unique identifier for the transaction. The following lists an association rule:

$$\text{src\_host} = 202.96.7.5 \wedge \text{dst\_host} = 202.108.35.210 \rightarrow \text{service} = \text{WWW},10,90$$

The association rule indicates that there are 10% of network traffic connections which is consistent with the source host IP 202.96.7.5, destination host IP is 202.108.35.210 and the access service is WWW service. WWW service may be 90% of the access service when the source host IP is 202.96.7.5 and destination host IP is 202.108.35.210.

130.3.2.2 Clustering Rule Set

Clustering rule set is the process of learning and training for anomaly detection model (association rules). The basic idea is:

Initialize cluster rule set, and then collect a certain amount of network data, mine these network packets and get some association rules. These rules are to be integrated into the cluster rule set. Cluster rule set will be updated. Update process is as follows:

(1) For each new association rule, it will be matched with the association rules of the cluster rule set. The meaning of the rule matching is that two rules are exactly the same on both sides.
(2) If a new association rule matches the association rule of the cluster rule set, the counter of the association rule in cluster rule set will add 1, and the support and confidence of the association rule in the cluster rule set will be

updated by the weighted average method. Otherwise, if the new association rule does not match the rule of the cluster rule set, this new association rule will be added to the cluster rule set, and the counter of the rule will be set to 1.

The cluster rule set is trained many times in this way, until cluster rule set is stable (Little or no new association rules are added).

130.3.2.3 Generating Model

The rules of cluster rule set should be compressed. We delete those rules whose counter value is less than the minimum counter value specified by the user. The cluster rule set which will be cut is the anomaly detection model.

130.3.3 Event Analyzer

The captured IP packets that we need to test will pretreat and convert into connection record.

Connection record set that we need to test execute association rule mining, then generated association rule set to be generated compare with cluster rule set (anomaly detection model), and calculate the similarity [3]. If similarity is less than a user-defined a threshold, we consider connection record set that we need to test exist anomaly. Meaning of two rules that match is that rule's left and right sides are equal and the deviation of two rules's support and confidence should be in the range of user defined. similarity $= \frac{p}{n} * \frac{p}{m}$, in which n is cluster rule set's number of rules, m is the number of rules that is mined from testing records. p is the number of rules with two rules that match.

130.3.4 Response Unit

Response unit will process the result of the event analyzer.

When the test results are abnormal, it will alarm the security administrator, and the alarm information will store to a file for later analysis.

When the test results are normal, detection rule set that is generated will integrate into the cluster rule set, and the cluster rule is set to be updated. The process of updating is described in Sect. 130.3.2.2. This makes the intrusion detection system to have a self-learning ability.

130.4 Conclusion

This paper describes the use of association rules algorithm to construct the process of network intrusion detection model, and how to use anomaly detection models have been established. Because data mining approaches [6] can extract rules quickly from a large number of network data, it greatly improves the performance of intrusion detection system.

Acknowledgments This paper is supported by (1) Project of Research on Science and Technology of Jilin Education Ministry of China under Grant No.2007-172 and 2010-383. (2) Science-technology Development Project of Jilin Province of China under Grant No. 20115056. (3) The Natural Science Foundation of Changchun Normal University.

References

1. Lee W, Stolfo SJ (2000) Data mining approaches for intrusion detection. http:\\www.yahoo.com
2. Debar H, Dacier M, Wespi A (1999) Towards a taxonomy of intrusion-detection systems [J]. Comput Netw 31:805–822
3. Lee W, Stolfo SJ, Mok KW (1999) A data mining framework for building intrusion detection models [C]. In: Proceedings of the 1999 IEEE symposium on security and privacy, pp 120–132
4. Lee W, Stolfo SJ, Mok KW (2001) Mining audit data to build intrusion detection models. http:\\www.yahoo.com
5. Han J, Kamber M (2001) Data mining concepts and techniques [M] (trans: Ming F, Xiao-Feng M). Machinery Industry Press, Beijing
6. Liu H, Lin Y, Han J (2009) Methods for mining frequent items in data streams: an overview. Knowl Inf Syst (Online: 11 Nov 2009)

Chapter 131
On the Design of University Network Teaching Resources Database

Zhen Liu, Ning Li, Wenlong Wan and Yulan Li

Abstract All kinds of teaching information resources based on the Internet is the goal of construction and application of education information as well as an important work link in the process of the university information construction, which plays a very important role in the realization of education modernization. Basing on the introduction of the situation and existing problems of current campus network information resource construction, the modules, functions, resources classification and metadata of digital instructional resource library have been analyzed; the design ideas and scheme for the university teaching resource library are put forward in the paper.

Keywords Network education resources · System structure · Module · Resource library · Design

131.1 Introduction

In recent years, with the rapid development of communication and network technology, campus network construction also is propelled, achieving obvious development in network infrastructure and various aspects of application system. But in addition to necessary hardware resources, campus network should also have abundant teaching resources of the network and get enough applications. Otherwise, the university informatization construction is like a car “not available

Z. Liu (✉) · N. Li · W. Wan · Y. Li
Henan Institute of Science and Technology,
Xinxiang 453003, Henan, China
e-mail: liuzhen@hist.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_131,

car" same, whose real function will not achieve. Therefore, the construction of network education information resources is a long-term and arduous foundation project and also the key project of campus net which has become the core of campus net development [1]. The construction of campus network teaching resources, is not simply to digital electronic, a traditional teaching resources, such as books, films, in kind, but to restart the planning, teaching design and construction according to discipline and network teaching characteristics. At colleges and universities teaching resources of the network building generally exist in the following issues:

(1) In the teaching resource construction, many schools overall purchase commercialized database. This, in part, solves the shortage of teaching resources, but overall, effective resource is scarce and more garbage resources without reference value. Part of the development of resource is only the simple material accumulation, such as subject teaching plans and texts, without taking into account the characteristics and the specific needs of academic knowledge teaching. It is the insufficient understanding and analysis of school teaching needs of the manufacturer to provide commercialization of insufficient, especially the current course reform deeply that cause the long term of new resources content, unable to make a new resources complement.
(2) Resource management ease-of-use poorer. As the maintenance and update resource need some specific techniques, the school resources in technical maintenance and update depend only on the manufacturer. The building of commercialization repository makes the interactivity and the communication among the user and between the manufacturers and the users. Most of the resource managements provide only resource adding, deleting, querying, browsing and other functions, not paying close attention to the interaction of the construction of the resources content.
(3) Lack of overall planning, low information resource utilization rate; dispersing, insufficient resource construction standard, not integrated resource construction standard the impeded inter-communication, unable to share the excellent teaching resources; the unestablished of the safeguard mechanism for resource sharing. All of these have serious impacts on the overall effect, of modern teaching and become one of the bottlenecks of the deep propulsion and development of the current education informatization [2].

To sum up, general repository obviously cannot satisfy their needs, so it is more important to construct the teaching resources with the characteristic of its own schools as a supplement for the repository. Simultaneously the teacher and the students should not only be the consumers, but also they should be positive resources participants and builders. This paper mainly puts up some suggestions and opinions based on the main points of campus network sharing of college teaching resource construction.

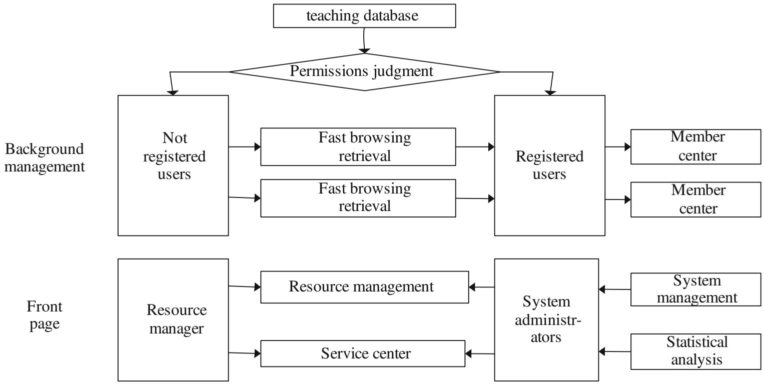


Fig. 131.1 System function structure

131.2 The System Structure Design of the Network Teaching Information Resource Database

Module design is the foundation and key of teaching resource construction not only considering the data structure, storage formats, input/output devices and applications mode selection of software and hardware, but also emphasizing the teacher's use custom, classroom teaching needs, and following the principle of coordination, alternation, flexibility and ease construction [3]. The author thinks that, general colleges and universities teaching resource should include the following five main modules, function and structure design as shown in Fig. 131.1 below.

System uses Microsoft Visual Studio 2005 as a development platform, and chooses to support ODBC interface SQLServer2000 as backend database, chooses ado.net as access web database interface, through the asp.net and ado.net union, establishes and provides the homepage content and includes database information. Use the enterprise manager (SQLServer provide query analyzer enterprise manager) and (inquires, such as a strong parser) tools that can easily design the database, development, deployment and management.

Generally speaking, a complete database system can be divided into three categories: the system administrator, resource user administrators and ordinary users.

(1) *The system administrator*. System administrators in system with the highest authority, for all the functional system provided, all have operation permissions. Its main responsibility is the repository system maintenance and configuration of the general situation of system, statistical analysis, classification of management of resources and the user management (chiefly to resource manager management), and is responsible for the security of the system management.

(2) *Resources administrator*. Resources administrator permissions behind the system administrator, its system administrator permissions are given, has

certain user operations access (is mainly to the common user management). Its main responsibility is to engage in resource evaluation management, resources audit, resource maintenance and resource needs support services [4]. Resource manager specific duties include the following two aspects.

131.2.1 Resource Management

(1) *Resources audits*. The resources to upload auditing and storage; in the management of resources, the users just can see after the administrator audit the resource which exist in the list.
(2) *Resource maintenance*.

Add resources. The server resources from the management of a single file upload, including the compressed package with multiple file upload;

Delete resources. Delete the repeat resources;

Modify resources. To modify the information such as the name of resources, resource types, resources classification, upload the author if necessary.

131.2.2 Service Center

131.2.2.1 User Management

Manage all registered users, including delete illegal users, setting members of integral, etc.

131.2.2.2 Service Management

Service management content includes:

Needs support. Check the users' resources demand situation and give the reply;

Review management. Check the user's comment for resources, timely adjust resources evaluation level;

Suggest feedback. Give feedback to the users' complaints, suggest promptly;

Ordinary users. Ordinary users is a systematic end user and beneficiaries.

Registered users. can browse all resources, but can only download "tourist zone" zero integral resources, cannot enjoy other services.

For the convenience of customers using resources provided by the system, system provides registered users some function to browse, resource download, search for collection, resource demand. In addition, for the convenience of administrators and registered users and convenient communication between registered users for resources, also provides resources extended comment, resources, to achieve functions such as uploading the purpose of resources sharing.

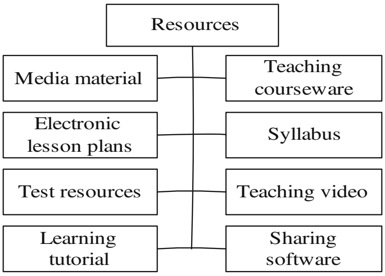


Fig. 131.2 Level 1 classification

Fast browsing with retrieval.

Fast browsing. Formulated according to Ministry of Education discipline and profession education resources type classification and discipline and type respectively in accordance with the layered browsing resources;

The retrieval. To find necessary resources according to the key word and resource category.

Tourist area. Provide zero download integral resources of the unregistered users.

Use resources.

Resources downloaded. Download the needed resource to a local computer.

Resources evaluation. Make subjective and objective evaluation on the quality of resources and obtain rewards.

Resources collection. One can collect the resource to personal favorites for the sake of using it next time.

Needs support. When the required resources are not retrievable, the user can request support by giving stand inside short message to administrator.

Member center.

Personal information modification. Registered users can modify their own basic information and password.

The upload resources. Registered users on local resources and URL resources uploaded to resource center, waiting for resource manager audit and review after the success of the users receive rewards.

The resources uploaded. Registered users to view has uploaded resources, not audit before can be modified.

My favorites. Registered users view, delete resources collected;

Complaints suggestions. Registered users to resource service unsatisfactory and complain or puts forward personal advice.

131.3 The Resources Classification Network Teaching Information Resource Database

Resource classification design should focus on serves for the teaching, satisfied, according to practical teaching needs, convenience and sharing of the principle of design. Resource level classification into eight categories, as shown in Fig. 131.2.

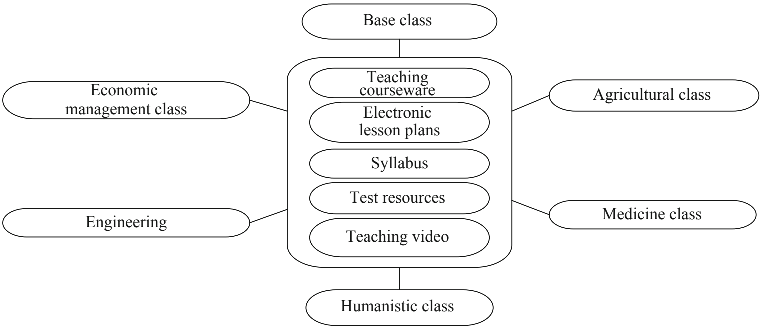


Fig. 131.3 Level 2 classification

Table 131.1 Resource secondary classification

Base class	Mathematics, mechanics, physics, chemistry, astronomy, earth, biology, foreign language
Agricultural class	Agronomy, forestry science, zootechnics, fisheries science
Economic management class	Management, economics, statistics
Medicine class	Basic medicine, clinical medicine, preventive medicine, special medicine, pharmacy,TCM
Engineering	Engineering foundation, surveying and mapping, materiality, mining, metallurgy, machinery, electrical and automation, energy and dynamics, nuclear science, electronics and communications, information Systems, computer, chemical engineering, textile, food science, civil construction, hydraulics, transportation, aerospace, environment, safety
Humanistic class	Literature, Marxism philosophy, philosophy, jurisprudence, education, journalism and communication, politics, history, society, art, sports

This nature of discipline includes five types of resources namely teaching courseware, electronic lesson plan, the teaching outline, exam resources and the teaching video. Therefore the five types of resources were compared with the secondary classification, as shown in Fig. 131.3. A subclass of secondary classification was formulated according to Ministry of Education, discipline and profession, as shown in Table 131.1. Various universities can also be modified according to the actual conditions.

Media material, learning tutorials and sharing software of these three kinds of resources are weak, so according to the own property classification resources. Media material is divided into picture type, audio kind and animation, as shown in Fig. 131.4 and Table 131.2.

Learning tutorial consists of learning tutorial classes, including graphic design, 3D design, animation, web pages and office software, tools application, network security, program design, video processing and other software, etc.

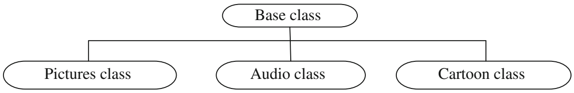


Fig. 131.4 Level 3 classification

Table 131.2 Resource secondary classification

Photo	Scenery, animals, flowers, building, automobile, movies, women, figure, food, system design, game, sports, brand, cartoon, hand paint, vector, Dazzle color, festival, else
Audio class	Animal sound,characters sound, instrument sound, festival voice, bell alarm sound, war fighting sound, transportation traffic sounds, daily life voice, appliances office sound, natural voice, sports voice, collision oscillation voice, humorous voice, Web application voice, titles out music voice, cartoon voice
Animation class	Plot animation, teaching experiment, scripting, 3D effect, interactive button, menu effects, effects demo, game code, interactive code, masks application, mouse effect, download effect, voice, application, text effect, virtual video, else

Sharing software consists of sharing software, including the network software and system tools, application software, contact chat, image, multimedia class, industry software, game recreation, programing development, safety-related and education teaching, etc.

Specific resource list forms including:

(1) *Resource properties*. Media material, teaching courseware, electronic lesson plan, the teaching outline, exam resources, video resource, learning tutorials and sharing software (level classification).
(2) *Application scope*. Picture type, audio, animation kind, video of class, learning tutorial class, sharing software, base class, agricultural class, students of economics, engineering medicine and humanistic classes.
(3) *Download points*. Each resource has the corresponding points, to download the resources will deduct the corresponding user. One cannot download without enough points. Integration settings are from 0 to N.
(4) *Upload date*. Upload time, format for 2008-7-12.
(5) *File size*. Upload file size, with K for the unit.
(6) *Download times*. The system automatically records the number of each resource download.
(7) *Resources evaluation*. Use the star as the symbol for the evaluation of the resource, the highest rank is with five stars.

131.4 Network Teaching Database Data Design

The network teaching resources database both must consider the consistency of the database design on the versatility and file format for these materials, and the secondary development and utilization data provide convenience. Reference "education resource construction technical specifications" and the classification of resources according to above teaching resource, will set the metadata for two ways to design and the five types [5].

131.4.1 Media Material

Media material for classroom teaching and experimental teaching provide auxiliary or directly to the course, it is the basic material spread teaching information unit, divided into three categories: pictures, audio and animation. Pictures of such material can be divided into footage pictures, the illustration for PPT and PSD layered pictures, etc., to ensure that the needs of secondary processing, should save as JPG etc. the compression ratio and distortion of smaller format. PSD, and some layered pictures, shall provide pack downloads.

The network teaching resources database audio resource material mostly through acquisition tapes and become, in the process of collecting as far as possible should pay attention to its acquisition or converted to MP3 format file, the file format can guarantee sound can be also used to be a universal network, the network audio format.

Animation class materials mainly divided into the plane animation and 3D animation. Plane animation mainly adopts Flash format, this format can support network access; 3D animation is into AVI format movie, another part of source files are packaged compression.

131.4.2 Courseware and the Network Courseware

Courseware and network courseware is one or a few points on the implementation of relatively complete teaching education, teaching for the software, according to the operation platform, can be divided into the courseware online courseware and single operation. When the courseware online will run in standard browser, the courseware can be directly put on the network platform. The single operation after download through the network courseware can run on a computer in the position, because courseware varied, such as format, Flash, Author war PPT to wait, to ensure the consistency of the download, should its package compressed into RAR or ZIP format file [6].

The network teaching resources database construction and application for teachers provides a better teaching environment, to provide students with a good self-study platform, greatly improving the campus network application efficiency.

References

1. uplift (2010) He Xiaoping under network environment, digital education resource sharing model j explore jinxing library journal, 40(4):31–34
2. ZhouZhenJun XiaoNa, JiaYong jiangsu (2007) edge Education information resources classification sharing mechanism study [J] audio-visual education research (10):24–27
3. Zhao ZhanYang, Le Xingtong (2007) Some hot spots of information resource database thinking about the problems of disaster prevention technology [J]. J of Institute 9(02):101–103
4. CaoWang rainbow (2010) Teaching resources of the network database management. J Inf Sci 28(4):2462250–2462530
5. FuKeMeng, horse fee (2010) Our information resources into standard system J Struct Inf Sci 28(11):1602–1604
6. ShiRuiFang (2010) Construction of college teaching resource [J]. J Shanxi Econ Manag Cadre Inst 18(02):122–124

Chapter 132
Research on Network Security of Digital Library

Junhui Fu, Wenxian Xiao, Jinna Lv and Guohong Gao

Abstract With the rapid development of modern computer communications and multimedia technology, the traditional library began to change rapidly. Library information resources become digital, networked, and promote the way of library management and service change and development, resulting in the digital library appear. The establishment of digital libraries realizes the sharing and utilization of network resources, provides convenience for the readers, but the attendant issues are also worthy of our attention, that is, the security of digital libraries. Once the library network system has been destroyed, the losses will be incalculable.

Keywords Digital library · Network security · Security policy

132.1 The Concept of Network Security

International Organization for Standardization (ISO) on the computer system security is defined as [1]: data processing system for the establishment and use of technology and management of security, protection of computer hardware, software and data is not the reason by accident and malicious destruction, modification and disclosure. It can be understood as the security of computer networks: By using a variety of technical and management measures, make normal operation of network systems, so that ensure network data availability, integrity and confidentiality [2]. Therefore, the establishment of network security protection measures aimed at ensuring the data exchange and transmission through the network does not increase, modify, loss and leakage and so on.

J. Fu (✉) · W. Xiao · J. Lv · G. Gao
School of Information Engineer, Henan Institute of Science and Technology,
Xinxiang 453003, China
e-mail: 599246483@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_132,

132.2 Digital Library Concepts and Connotations

The digital library research and development work is raging in digital library understanding and knowledge and is in constant in-depth expansion. As to the Internet, as a representative of network technology and popularization of concepts such as digital earth, people gradually tend to put forward link and standing in culture national or even worldwide information resource sharing, the highly connected to the ubiquity of the network environment to know digital library, trying to make the breakthrough in tradition library, toward the broader self-cultural foothold of social culture and even economic cooperation. More representative views are:

China digital library construction and the development of national 863 plan China digital library development strategy group leaders Xu WenBo stressed [3]: to construct the new century from the height of the Chinese culture, think "think digital library construction of digital library, is to develop the network information resources, take the initiative positive effective preemption Internet positions, digital library is the measures in the construction of network content."ace ace He will digital library definition for "website supports multiple search function of magnanimous database".

132.2.1 Digital Library Definition Level

It is divided into broad sense and narrow sense. The narrow sense of digital library hierarchy focuses on the "digital library" concept [4]. One word on the traditional library concept is limited that one entity concepts of digital library is also an independent "entity", no matter this "entity" is virtual or reality. If above XuWenBo, digital library boils down to "magnanimous database", "knowledge center", "libraries and information institutions". The generalized digital library melts into a distributed management thoughts, in no limit of time and space network environment and adopts a tolerant attitude and broad perspective. If above the country and the "is the next generation Internet above the network information resources management mode", "is in the distributed computer network environment information resources organization form from higher level," limited "entity" limitations.

132.2.2 The Characteristics of Digital Library

Digital library problems, can be put forward in different views. Most researchers used the Shanghai library in 2001 [5], the liu wei papaer digital library was in the formulation of the digital library, and there are three points: digital resources, network access and distributed management. These three points from the ontology level basically generalizes the characteristics of digital library. Other characterstics are individualized service modes, convinience and humanization. Jiangxi university of finance and economics mentioned that in the library ChenXiangZhu several problems concerning the understanding and thinking of digital library existed.

132.2.3 The Elements of Digital Library

Shenzhen University Library ZhangDaoYi according to the practice of the construction of digital library, the elements of inductive digital library for five: abundant digital resources, the digital resource integration, network information revealed and the active service mechanism, the overall cooperation mechanism library network information service, the talent team.

132.2.4 Digital Library of Knowledge Management

Knowledge management is a development to a higher stage information management thought, but the digital library is often regarded as a future information management mode, in this sense, the study say digital library knowledge management is a thing well. Therefore, the knowledge management of digital library is study will man. This research mainly in the digital library knowledge management of the main content, the management mechanism, technical support, digital library of knowledge discovery, knowledge generation, the knowledge spread and knowledge organization strategy, knowledge management, digital resources disposition, etc.

132.2.5 Digital Library the Relationship with the Traditional Library

This relationship include theory and practice two aspects, namely the digital library and the library intelligence relationship, the construction of digital library the relationship with the traditional library. The latter mainly refers to the roles and responsibilities of traditional library in digital library construction process.

132.3 Main Factors Threaten the Network Safety of Digital Library

There are a lot of factors that affect computer network security, which can be roughly divided into the following categories [6]:

132.3.1 Hardware Issues

Library network system is composed of network servers, disk arrays and other equipment combination. Its security relies on the reliable operation of the overall device. A part of any error, may affect the normal operation of the entire network. In the hardware configuration, you must fully take into account the compatibility between various devices and whether they are configured correctly and so on. In addition to these, we should also take into account the external environmental factors, the computer running has high requirements on the surrounding environment, such as temperature, humidity, power supply stability, etc., in order to protect the server system in engine room running normally, you should do: fire, moisture, dust, static electricity, lightning, etc., equip with fire, moisture, dust, static electricity, lightning and other devices, in addition to the need for good quality server also need equip with the appropriate power outlet and power UPS uninterruptible power supply to ensure power supply stability, the maximum reduction effects on the hardware devices because of external factors, to guarantee a secure computer system operating environment. Specific standards refer to "People's Republic of telecommunications industry standards" of the communication center room environment conditions.

132.3.2 Human Error

For example, the operator security vulnerabilities caused by improper configuration, user security awareness is not strong, careless user password selection, users free to lend the account to others, will be a threat to network security.

132.3.3 Malicious Attacks

This is the biggest computer network threat. Hacker attack and computer crime belong to this category. Such attacks can be divided into the following two: one is the active attacks, which selectively destroy the effectiveness and integrity of information in various ways; the other is a passive attack, it does not affect the network in the normal work cases, an interception, theft, deciphering the secrets to get important information. The two attacks on computer networks could cause great harm, and cause the leakage of confidential data.

132.3.4 Computer Virus

Computer virus is a man-made program causing damage effects to computer information or systems when the computer is running. This program is not an independent existence, it concealed among other executable program, both destructive, but also infectious and latent. It will affect system performance, so that machines will not run; Severe case will lead to data loss, system crashes, will give users an immeasurable loss.

132.4 Preventive Measures

To protect the digital library to be safe and stable operation, we can take the following several ways to enhance its security.

132.4.1 Improved Safety Consciousness

Network security is the most important person safety concept and consciousness, library leader and decision making. Participants must increase network safety equipment and personnel training input, change hard light into soft, heavy construction light maintenance concept. Librarians should strengthen the relevant laws, regulations and policies of study, consciously accept. Antivirus prevent dark technical training, actively improve the sense of self-protection and ability.

132.4.2 Use Firewall Technology

Network Firewall technology is used to enhance access control between the network to prevent the external network users from illegally entering the internal network via the external network, access to internal network resources, protect the internal network operating environment and the special networking equipment. It is transmitted between two or more network packets according to a certain way, such as links to the implementation of security policy check to determine whether the communication between the networks are allowed, and monitor the network running. The basic idea of the firewall—not on each host system protection, but all access to the system through a point, and to protect it and shield the information and protect the network structure outside the world as much as possible. It is a barrier set between the trusted internal network and not trust outside the world, it can implement a wider security policy to control the information flow, to prevent the invasion of unpredictable potential damage. A new generation of firewall products—monitoring firewall, the network can take the initiative in the data layers, real-time monitoring, based on an analysis of these data, monitoring firewall can effectively determine the various layers of trespass. Also, this test firewall products are generally also with the distributed detectors, these detectors placed in a variety of application servers and other network nodes being not only able to detect external attacks from the network, but also has a strong preventive role in the inside Vandalism.

132.4.3 Using Network Security Technology

The digital library system in the whole network security, can purchase UPS, in power. This is able to send power to each workstation information, and can store information, till a system closes. And, outside, it can use digital closed-circuit monitoring system and other technical measures to guarantee system security.

(1) Data encryption refers to a message with encrypted keys and encrypted conversion, cannot read directly, and the ciphertext recipient will read the ciphertext after decryption function using decryption key reduction Cheng-Ming Wen. Therefore encryption is to protect the data security by effective means. In computer networks, encryption divided into communication encryption and file encryption. Communication encryption and link encryption, node encryption and in end adds three dense kinds.
(2) The user terminal and user authentication are to control access to information of the network for the implementation of the management system. It may prohibit illegal user authentication jehu the operation, can more effectively protect software and data on Whole. The more widespread application is password-based user authentication. This way the user accesses authentication Control union, verified through user input password to determine if users can access system. This comes in Shoesmag E-mail: 99246483-@qq.com net library work research after the system allows operation.
(3) Firewall technology is the currently used online security technology, it is a kind of passive defense access control technology, through border in network set up of the corresponding nets.

Winding communications monitoring system to realize its function. A protected network firewall is established only through software and hardware, and the management measures. The crossing of comprehensive information network boundary provides monitoring and control modified method. The current implementation of the main technical firewall have functions such as packet filtering and the application layer gateway (or proxy server), etc. Internet firewall technology currently has matured more.

Simple firewall provides a reliable network security control method.

132.4.4 Virus Prevention

Library addition has its own web server, data server and the server library automation system. The various departments are also equipped with working machine, coupled with the electronic reading and audio–visual room and other places of the machine. The number of hosts is not few, apparently the maintenance work of the host is not an easy task. We can see that, for each host to install anti-virus software and update virus database in a timely manner is necessary, even if the antivirus software cannot prevent and control all of the virus. But most the virus can still play a role. In addition to using antivirus software to prevent viruses, we must enhance the library staff awareness of virus prevention, in daily work, in viewing Web pages, receiveing mail and downloading files and be careful not to open unfamiliar e-mail or download unknown files. This inadvertently allow viruses to take advantage of the loophole. In the use of removable storage media, we must first carry out its anti-virus before opening it.

132.4.5 Data Backup

The library has a lot of servers, and there are a large number of data files in these servers, in order to prevent in the first place and do the server system backup. Data backup is also a very important. While the server is damaged, we can do the server data recovery on time using the data files backup, and strive to minimize losses. Considering in view of the data capacity for data backup, because it takes up more space, we can use hard drive backup, for dedicated backup of server's hard drive, and for the system backup files, we can store them to hard disk or CD according to its size, ready for use.

132.4.6 Rational Use Computer Management Tools

In the control panel there are security policy management tools and service files. We need rational use of these two system tools to develop reasonable and workable security policy for digital libraries. For planning the level of access control and permissions to stop unnecessary risk of services and enhance system security. For example: Remote registry service, the service is allowed to remotely modify the registry, enable it to provide a path for hackers to host your way. we can close it to get a little more security protection. There are many services which are not necessary to open, we can choose to on or off these services according to our needs.

132.4.7 Strengthening Management

In this rapidly information changing era, especially with changing computer technology, library needs to enhance the training of computer management for the staff, professional ethics education, targeted knowledge and skills in network security training and provide them adequate learning time as much as possible to better serve the library building.

As the system administrators need to constantly learn new knowledge and enrich their brains to adapt to new changes. In the daily work they should always check the system log, to detect intrusion and virus and to monitor the forecasts and timely response.

132.5 Conclusion

With the emergence of new technologies, network security is constantly changing. The digital library will also be faced with more challenges. Strengthening the library network security management is currently a very urgent task. Library needs to increase investment, continue to improve and perfect the network security and minimize the negative effects.

References

1. http://www.edu.cn/ruan_jian_ying_yong_1720/20061108/t20061108_204156.shtml
2. Lu J (2009) Analysis of network security and firewall technology [J]. Inn Mong Sci Tech Econ 6:82–84
3. Preetham, Ran X (2004) Internet Security and Firewall [M]. Tsinghua University Press, Beijing, 2004
4. Han X, Ke Z (2005) Computer network security and practical technology [M]. Electronic Industry Press, Beijing, 2005
5. Zhang X, Zhang Z (2006) St. University Library Information Security Research [J]. Libr Info Serv Compliment :178–181
6. Wang H (2008) Library Network Security and Management [J]. Libr Info Serv (4):29–31

Part XII
Network Design Methodology

Chapter 133
A Study on the Influence Mechanism of Characteristics of the Internet Economy on the Transaction Cost

Wang Ying, Zhang Tongyao and Zheng Hao

Abstract This essay analyzes the typical characteristics of international economy and describes the definition of transaction cost. In addition, we analyze the influence on the transaction cost produced by Internet economy characteristics and point out that the internet economy would reduce the exogenous transaction cost, while at the meantime increase the endogenous transaction cost. At last, we disclose the internal influence mechanism on the transaction cost produced by Internet economy characteristics.

Keywords Internet economy · Transaction cost · Influence mechanism

133.1 Introduction

As a new economic pattern backed up by information and knowledge, international economy significantly impact on the traditional economic theories and on the cost definition, price model of products and the enterprise organization as well as management theories. Based on the Internet economy characteristics and definition of transaction cost, this essay discloses the internal influence mechanism on the transaction cost produced by Internet economy characteristics.

133.2 The Definition of the Internet Economy

The Internet economy actually is a new style of economy where we can achieve the maximum revenues from the extensive application of information technologies

W. Ying (✉) · Z. Tongyao · Z. Hao
Management School, China Univerity of Mining Technology,
Beijing 100083, China
e-mail: wyzxguoguo@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_133,

and Internet. It not only indicates the proliferation of information technology industry, but also signifies the worldwide establishment of high-tech industry and the revolutionary transformation of the traditional industries and economy departments that is led by the popularization and application of the high-tech [1]. Therefore, the Internet economy cannot be viewed as the "virtual" economy purely or even completely as the contrary to traditional economy. In fact, the development of Internet economy is based on the traditional economy and simultaneously distinguishes from the traditional economy in terms of the economy operation, growth and efficiency.

133.3 The Main Characteristics of the Internet Economy

133.3.1 Positive Feedback

In contrast to the decline trend of the marginal utility of the traditional economy, the expansion of the Internet scale is capable of leading to the marginal revenue increase effect [2]. Metcalfe's law states that the value of a telecommunications network is proportional to the square of the number of connected users of the system. Since each new connected user gains much more information communication opportunities because of the interconnection, Metcalf's law points out that the Internet is featured with magnificent externality and positive feedback.

133.3.2 Incompatibility

As a result of the incompatibility of the Internet systems, the oligopoly situations of Internet market appear frequently. Homogeneous Internet products even can bring an oligopoly supply for the reason that they belong to incompatible systems. The incompatibility is probably caused by the essential technology utilized by different systems. Moreover, the incompatibility is formed by the special assets barriers installed by different network organizations with respective financial purpose.

133.3.3 Locked-in Effect

Information products possess a notable quality of "locked-in effect". The feature of "locked-in effect" evolves on the basis of the theories of "incompatibility" and "path-dependence" [3]. The "locked-in effect" connotes that the members and consumers of various Internet systems will be prone to keep loyalty to their original choices of Internet systems because of the high switching cost. The organization members and consumers of the original Internet systems have to not only pay the non-refundable "sunken cost" resulting from the special investment on the original system, but also the new knowledge study cost. The latter cost that is usually called "shift cost" by economists is of more magnitude in value than the former one. When the "shift cost" ascends to some extent that the user would be impeded to choose new systems, we consider the organization members and customers are locked in.

133.4 The Concept of the Transaction Cost

The cost of economic activity formed the final determinant which influences the selection of economic behavior modes, which is decided by the scarcity of economy resources. On the regular basis, the transaction is regarded as a process where the suppliers and the demanders seek for the satisfactory collaborator and eventually settle down and perform the contract to implement the partial or complete transfers of property right. Transaction cost is the transfer cost of property right during the above mentioned process. In his further research on the transaction cost, Williamson categorize the cost into the exogenous transaction cost and the endogenous transaction cost, where the former one includes direct or indirect transaction cost such as the information collection cost, the transaction techniques cost and the negotiation cost, which are tangible and objective; the latter one, however, is an artificial consumption generated by the self-interest decision-making of the transaction subjects, which is intangible and subjective [4].

133.5 The Influence Mechanism of Characteristics of the Internet Economy on the Transaction Cost

According to the standard, established by Williamson, of categorization of transaction cost, it is not difficult to find that Internet economy directly result in the dramatic decrease of the exogenous transaction cost [4, 5]. The concrete influence mechanism is embodied in the following respects:

- Network technique facilitates the collection and the processing of transaction information of enterprises. Through Internet, both sides of the transaction can search for the perfect partner through the whole world in the terribly short time with very low expense. With the updating of technical instruments, the functions of search engine will be enhanced and the expense of information searching will be reduced gradually.
- Network technique could simplify the process of market transactions, resulting in the drastic reduction of the direct cost of transaction, such as technique cost.

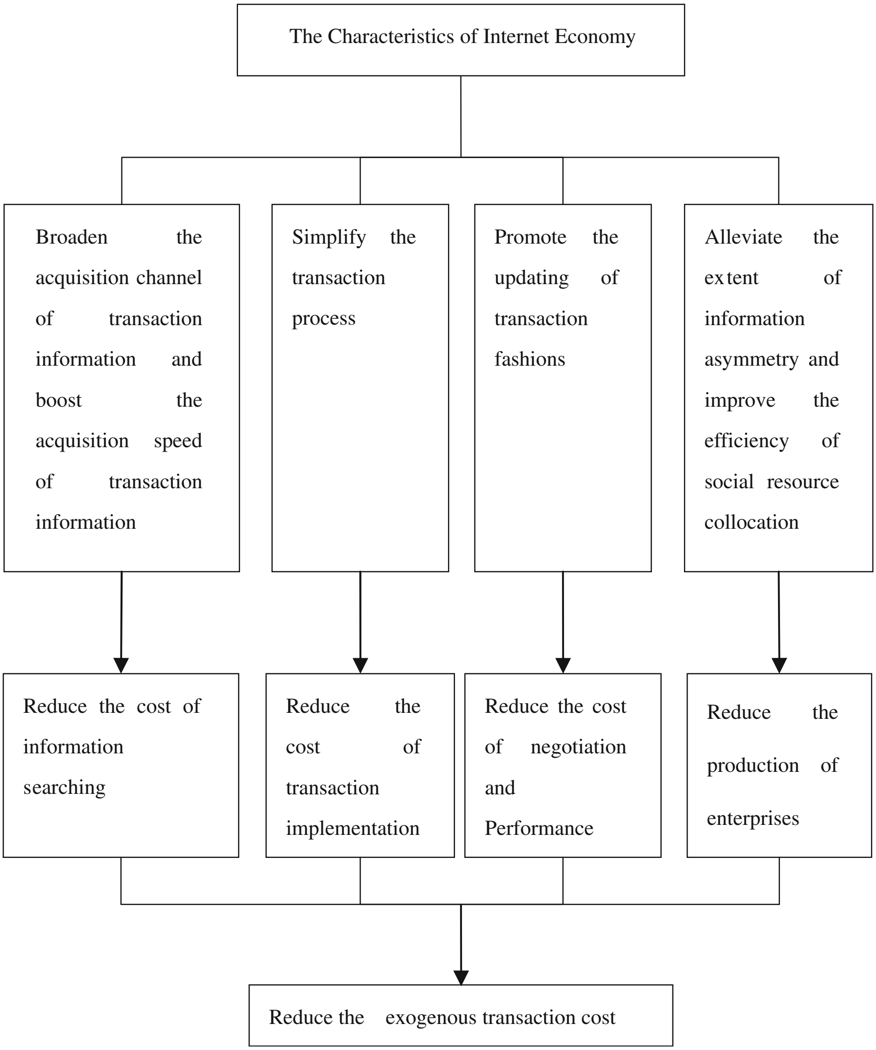

The Characteristics of Internet Economy
Broaden the acquisition channel of transaction information and boost the acquisition speed of transaction information
Simplify the transaction process
Promote the updating of transaction fashions
Alleviate the ex tent of information asymmetry and improve the efficiency of social resource collocation
Reduce the cost of information searching
Reduce the cost of transaction implementation
Reduce the cost of negotiation and Performance
Reduce the production of enterprises
Reduce the exogenous transaction cost

Fig. 133.1 Internal influence mechanism on exogenous transaction cost

Further, network technique could continually promote the renovation of transaction fashions, leading to the decrease of implementation cost. For instance, the emergence of online ordering and e-commerce evidently reduced the cost of negotiation and performance.

- According to the economics theories, information asymmetry will induce the inefficiency of market transactions. The Internet, nevertheless, could alleviate the extent of information asymmetry between both sides of transaction and improve the efficiency of social resources collocation. Therefore, the

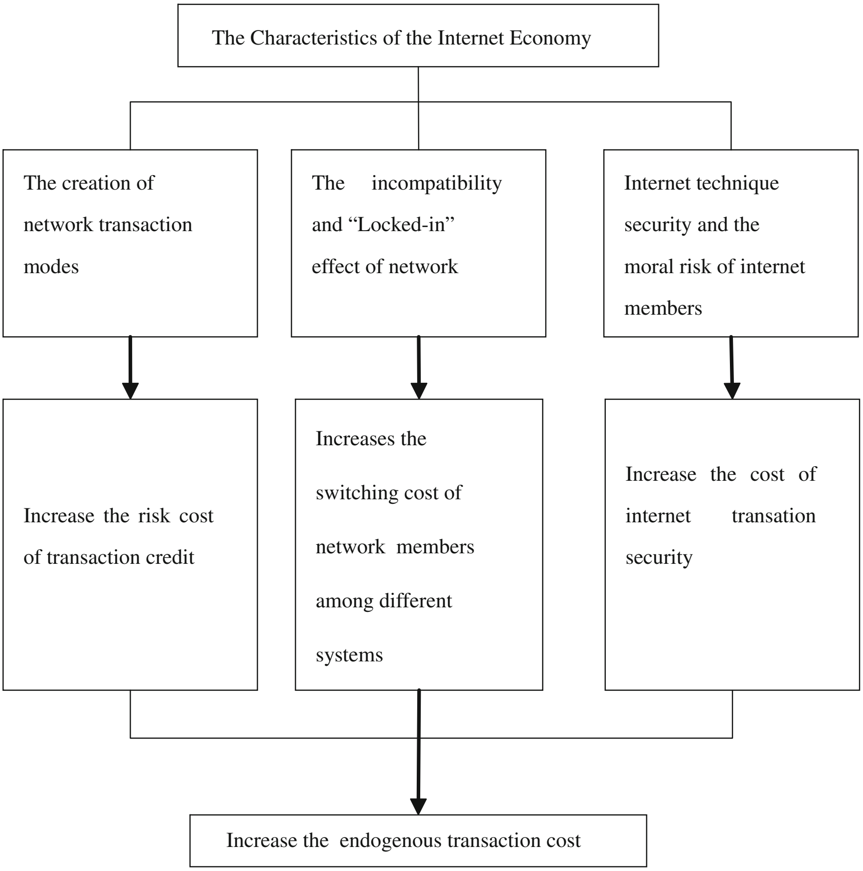

The Characteristics of the Internet Economy
The creation of network transaction modes
The incompatibility and "Locked-in" effect of network
Internet technique security and the moral risk of internet members
Increase the risk cost of transaction credit
Increases the switching cost of network members among different systems
Increase the cost of internet transation security
Increase the endogenous transaction cost

Fig. 133.2 Internal influence mechanisms on endogenous transaction cost

enhancement of mobility and acquisition efficiency of production elements will undoubtedly reduce the production cost of enterprises.

The concrete influence mechanism will be revealed in Fig. 133.1.

At the same time, it is necessary to realize that Internet economy is just as a double-blade sword. The characteristics of Internet economy also lead to the increase of endogenous transaction cost to some extent. The concrete influence mechanism is embodied in the following respects:

- The unisochronism of Internet transactions results in rationality deficiency and opportunism tendency of the transaction performers, which would not only increase the cost of performance of contracts, but also the credit risk in the e-commerce domain.
- The characteristics of incompatibility and "Locked-in effect" promote the enterprise to be absorbed in investing in network constructions and hold them as

special assets, especially the network operators. The investments build a huge quit barrier, which increases the switching cost of network members among different network systems.

- The cost of network security arising from network operation risks is an inevitable issue confronted by enterprises depending on network running. This risks cost includes not only moral risk cost caused by network members, but also the security risk cost caused by the network technique itself, such as the harmful impact on network transactions brought by the network virus attacks etc. (Fig. 133.2).

133.6 Conclusion

Some scholars deem that Internet economy reduces the transaction cost of enterprises, but we consider that this statement is inaccurate. Based on the characteristics of the Internet economy and the transaction cost theory, we analyze and disclose the Internal influence mechanism on the transaction cost produced by Internet economy characteristics. Finally, we concluded that the Internet economy would reduce the exogenous transaction cost, while at the meantime increase the endogenous transaction cost.

References

1. Williamson OE (1975) Markets and hierarchies [M]. The Free Press, New York
2. Shapiro C, Varian H (1999) Information rules: a strategic guide to the network economy [M]. Calarendon Press, Oxford
3. Porter M (2001) ComPetitive advantages and the Internet times. Harvard Business Rev USA 39:151–156
4. Zhang R (2001) The analysis of international economy characteristics and operation rules [J]. Trans Tianjin Univ (1):46–50
5. Wang Z (2002) The principle and law in the Internet era. Study Enterp (4):102–104

Chapter 134
Underground Wireless Multi-Hop Ad Hoc Networks Based on OLSR for Coal Mine

Yanjing Sun, Ruhui Li, Man Yu and Menglong Wang

Abstract In order to resolve the wireless transmission problem underground complex environment, we present wireless multi-hop Ad hoc networks based on OLSR routing protocol and full coverage communication architecture based on 802.11n. To overcome the disadvantage of hop count, expected transmission count (ETX) and expected transmission time (ETT) are used as routing metrics to improve UDP data transmission for coal mine. The intrinsically safe wireless node is designed and implemented. Experiments show that ETT does better than ETX in multimedia transmission capability for underground wireless multi-hop ad hoc networks.

Keywords Wireless ad-hoc networks · Underground coal mine · OLSR · Routing metric

Y. Sun (✉) · R. Li · M. Yu · M. Wang
School of Information and Electrical Engineer,
China University of Mining and Technology,
Xuzhou 221116, China
e-mail: yanjingsun_cn@163.com

R. Li
e-mail: rhlcumt@163.com

M. Yu
e-mail: ymcumt@163.com

M. Wang
e-mail: wmlcumt@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_134,

134.1 Introduction

Traditional coal mine communication is based on cable communications. With the development of wireless network technology, underground wireless mobile communications is showing advantages. Leakage communication is mainly limited by wiring conditions, cannot cover mining operations surface without wiring. Low-power, self-organizing wireless sensor networks (WSN) have become research hotspot in recent years, and get a certain amount of research and application in the safety production of coal mine[1–7]. Li et al. proposed a (SASA) [5] Structure-Aware Self-Adaptive system for early warning and monitoring of coal mining regions collapse. Yu et al. designed a coal mine safety monitoring system capable of collecting temperature, humidity and concentration of methane based on Zigbee [6]. Sun et al. proposed a monitoring system for the mine working face, and it effectively improved the network lifetime [7]. Aniss et al. studied AP arrangement of long-distance communication environment for long-wall mining [8, 9]. Video transportation underground was tested and video data transmission within non-visible was solved by placed AP in the tunnel corner in [10].

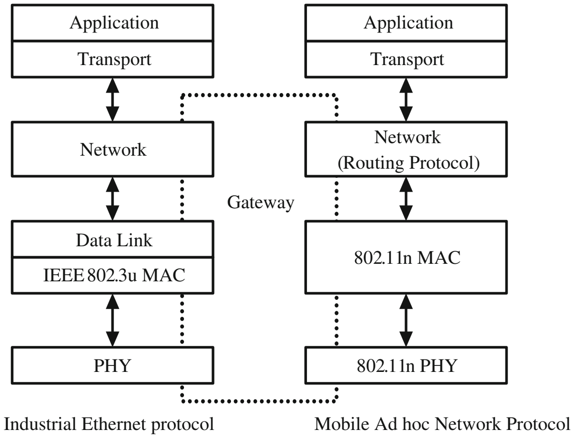
Wireless ad hoc network does not need fixed infrastructure, it can overcome the existing difficulties in the deployment of wireless system, while meeting the requirements of real-time multimedia data transmission and achieve environmental monitoring and reconstruct emergency-aid communication system after disaster.

Therefore, this Chapter proposes full coverage wireless Ad hoc network communication protocol system based on 802.11n to overcome the shortcoming that OLSR protocol calculates routes by counting hops. We used the expected transmission count (ETX) [11] and the expected transmission time (ETT) as the routing metrics to build the underground wireless network based on IEEE 802.11n and OLSR routing protocol, as well as design and implement the 802.11n wireless node on embedded Linux platform. Performance experiments are finished for the underground wireless multi-hop Ad hoc network.

134.2 Underground ad Hoc Network Communication System

Wireless environment such as coal face, roadway and Gob in coal mine is different from on ground. To meet the coal mine environmental interference, dynamic topology and low latency requirements, protect the bandwidth and real-time of network, we designed the communication protocol system compatible with industrial Ethernet network protocol with the development trend of integrated automation backbone communication platform of coal mine fiber industrial Ethernet. The system uses IEEE 802.11n protocol in MAC layer, supports multi-hop topology and end-to-end communications and provides device level wireless connection. Underground Ad hoc network adopts IEEE 802.11n-based communication protocol as shown in Fig. 134.1.

Fig. 134.1 Ad hoc communication protocol of coal mine



IEEE 802.11n network protocol is based on open system interconnection model, PHY layer uses MIMO technology to join parse the received information by multiple antennas, the physical layer data throughput in 40 Hz channel is twice as much as in 20 Hz channel used by 802.11n PHY. In the MAC layer, using frame aggregation technology to pack much MSDUs and MPDUs to reduce costs and increase data transfer rate. Network layer with self-organization and self-healing capabilities, adopts OLSR protocol to undertake formation and maintenance of network topology.

134.3 Underground Wireless Ad Hoc Network Routing Protocol

134.3.1 OLSR Routing Protocol

Each node of proactive routing protocol interacts topology information periodically, the protocol masters all nodes' routing that reached the whole network with the advantages of small transmission delay, while has large routing protocol overhead to reflect the current network topology timely and accurately, such as optimization link state routing protocol OLSR.

If and only if reactive routing protocol needs routing, the source node just creates a routing to reach destination, so it is called on-demand routing, the nodes only have part of network topology and routing information. Routing protocol needs to complete the routing discovery, routing maintenance and routing demolition work. Compared with the proactive routing, data transmission delay will increase with routing discovery process, such as AODV and DSR.

Taking into account the relative stability production environment of coal mine tunnel and routing maintenance is simple, the proactive routing can meet many types of real-time data transmission requirements of coal mine better. According to [1, 4, 12, 13], OLSR protocol has better performance in end-to-end delay, end-to-end throughput, routing protocol, etc., so this Chapter selects OLSR as the routing protocol.

134.3.2 Routing Metric

Routing metric is the core of routing protocol that routing protocol uses to calculate the best path of end to end, so different routing metric has remarkable influence on the performance of network. RFC3626 provided OLSR to select hops as routing metric, but OLSR may choose low-quality link of one hop instead of high-quality link of two hops in the harsh coal mine underground wireless environment.

ETX. ETX can handle asymmetric problem that frequently appears in wireless network and prefers to choose the shorter routing. The probability of packets are transmitted from the sender to the receiver correctly is called forward transmission rate d_f. ETX corresponded to ACK packages of the back link from the receiver to the sender is called back transmission rate d_r. $d_f \times d_r$ is the rate of data is transmitted and received by ACK successfully.The node calculates the sender's forward transmission rate by detecting frame information. ETX is defined as (134.1):

$$\mathrm{ETX} = \frac{1}{\mathrm{d_f} \times \mathrm{d_r}} \tag{134.1}$$

ETX is beneficial to realize the maximum of link throughput performance, but ETX cannot distinguish link bandwidth and packet loss rate of different data frame and does not support multi-rate.

ETT. ETT is used to estimate the time taken to send a data frame successfully to the link. ETT has to calculate the packet loss rate of forward link and back link and the link bandwidth first, determine packet loss rate by detecting frame that used in radio link of ETX and measure link bandwidth by using packet pairs technology. It is defined in (134.2).

$$\mathrm{ETT} = \mathrm{ETX} \times \frac{S}{B} \tag{134.2}$$

In the above formula, S and B represent packet size and link bandwidth. The node measures bandwidth B of each link by packet-pair technology, specifically, each node sends two detecting packet data with different lengths successively to its adjacent nodes by multicast every second (137 and 1137 bytes). The neighboring node immediately calculates the time between two received data packet after received detecting packet data, and feedback the time to sender node. The calculation of bandwidth is shown in (134.3).

$$B = S_L / \min_{1 \le i \le n} \mathrm{d}_i \tag{134.3}$$

Here, B is link bandwidth, S_L is the maximum of detection packet data, d_i is receiving time interval. ETT also take link bandwidth into consideration, though the bandwidth evaluation will bring some routing overhead, it is very much suitable for different situations such as coal mine control and monitor, multimedia data transmission, real-time, etc.

134.4 Performance Experiments

Coal mine safety monitoring data requires high quality for real-time, UDP protocol is usually used in coal mine. We developed wireless test nodes based on embedded Linux and core module of PXA270 processor. The node uses RT3070 chip which meet the 802.11n standard. The node realized function of OLSR protocol based on protocol stack OLSR daemon wrote by Unik University, supported IPv4 and IPv6 addressing and achieved ETT routing metric strategy by link_probe.

The node works in the ad hoc state, uses channel 1 and 802.11n mode, the highest transmission bandwidth of point to point is 130 Mbps. We place laptop A and B running UDP server of Iperf and client separately to test actual network's throughput of the end to end, packet loss rate and delay jitter, the UDP buffer is set to 250 KB.

For the particularity of coal mine data, different underground functional subsystem will use UDP packet with different size. We set UDP packet with different size in 2 M bandwidth to test the actual data throughput, packet loss and delay jitter, the parameters of OLSR routing protocol are set as in Table 134.1.

The test result of ETX and ETT is shown in Figs. 134.2, 134.3 and 134.4. When the size of data packet is close to 1,000 bytes, bandwidth used to transmit data can be fully taken advantage of, the actual throughput is very close to 2 M bandwidth and has the minimal jitter at the same time. By contrast, it is found that the larger data packet (>1,500 bytes) has more great impact on throughput and jitter, this because UDP protocol cannot guarantee the data packet arrived in order by ID, fragmenting and reassembling data packet have certain influence to the performance of system, so the performance of throughput and jitter are worse than the small data packet's. It can be seen, ETT's performance is better than ETX's. ETX may choose low bandwidth link of one hop rather than high bandwidth link of two hops at this moment.

Packets with size of 1,000 bytes have the lowest packet loss rate, if each node sends a large number of small data packet, it will intensify competition of wireless channel and lose much many packets, due to the character of nodes within one hop range of Ad hoc network sharing wireless channel.

Based on the above analysis, we can see that the data packet with 1,000 bytes has the best performance in 2 M bandwidth environment, next, we used data

Table 134.1 Test Parameters for underground wireless ad hoc networks

Routing parameters	Value
HelloInterval	2.0 s
HelloValidityTime	20.0 s
LinkQualityWinSize	12
TcInerval	5.0 s
TcValidityTime	30.0 s
ETX_interval	5 s
ETT_window	6
Emission_interval	5.0 s
ETT_expiration_time	30.0 s

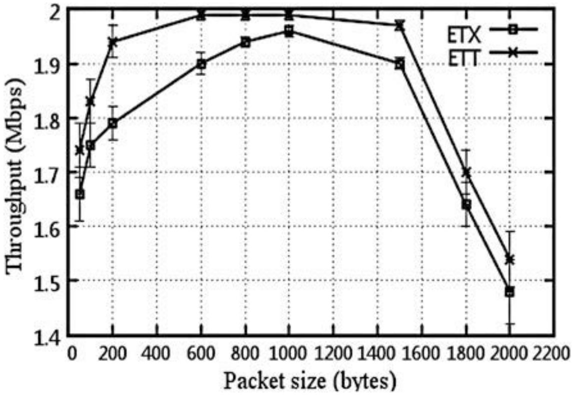


Fig. 134.2 Network throughput

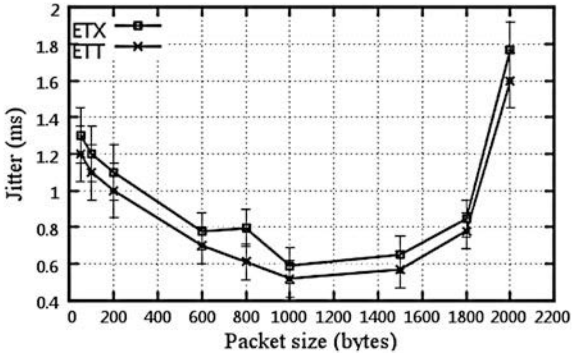


Fig. 134.3 Jitter

packet with 1,000 bytes to test network performance of ETX and ETT in different bandwidth, the results are shown in Figs. 134.5, 134.6 and 134.7.

Figure 134.5 shows the throughput with different bandwidths of end to end. ETT improves the throughput of end to end significantly and plays high bandwidth

Fig. 134.4 Packet loss rate

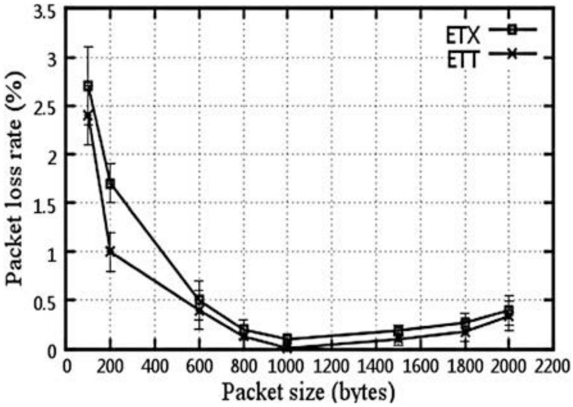
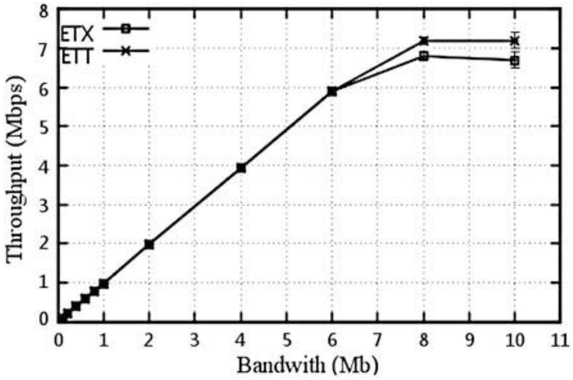


Fig. 134.5 End throughput of different bandwidth



characteristic of 802.11n, the throughput of end to end is up to 7.2 Mbps that can completely satisfy the requirements of audio and video transmission.

Figure 134.6 indicates the packet loss rate with different bandwidths. We can tell that the increase in throughput will bring higher packet loss rate significantly by Fig. 134.5 and 134.6, this is because the competition of wireless channel and buffer overflow can cause a lot of packets lost. Wireless video compression algorithm has some adaptability for packet loss, so it owns larger bandwidth within sustainable packet loss, in fact, if control the packet loss rate below 2%, the actual network throughput of ETT is still about 7 Mbps.

Figure 134.7 shows that the jitter will increase with the increasing bandwidth. This is because the increasing data packet will intensify competition of wireless channel and more data packet will be cached in the buffer of UDP, if data packet cannot be sent on time, it will increase the delay jitter. At that moment, the audio signal transmission is very sensitive to delay jitter while demands lower in bandwidth, so it should select appropriate transmission bandwidth according to the actual situation.

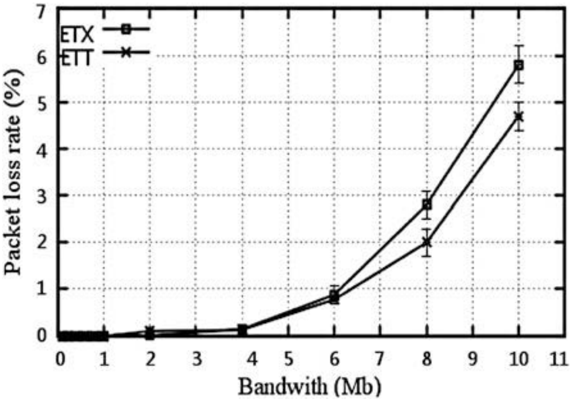


Fig. 134.6 Packet loss rate of different bandwidth

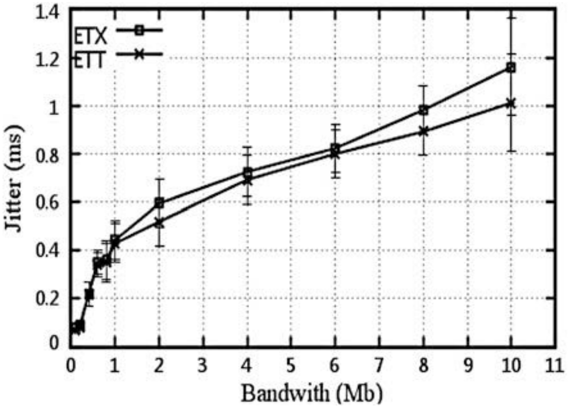


Fig. 134.7 The jitter rate of different bandwidth

134.5 Conclusions

Coal mine wireless ad hoc network based on 802.11n has the characteristics of easy deployment, low cost, high bandwidth and low latency and supports reliable data transmission with different real-time requirements. This Chapter presents the underground wireless ad hoc network architecture which is appropriate for coal mine based on 802.11n/OLSR, designs and realizes wireless test node. The performance difference between ETX and ETT routing metric of OLSR protocol is analyzed and compared by experiment. It has significance to solve the key technology problem of disaster relief communication system in coal mine.

Acknowledgments This work is sponsored by Qing Lan Project, National Science Foundation of China under grant No.50904070, the China Postdoctoral Science Foundation No.20100471412.

References

1. Zhou GB, Zhu ZC, Chen GZ (2010) Hiberaachy topology control of wireless sensor networks in coal min laneway. J Chin Coal Soc 35:333–337
2. Sun L, Xu Z, Zhai WY (2010) Model of ad hoc networks for rescuing in mine. In: Second international conference on networks security wireless communications and trusted computing. IEEE Press, Wuhan, pp 210–213
3. Zhang ZB, Xu XL, Yan LL (2009) Underground localization algorithm of wireless sensor network based on Zigbee. J Chin Coal Soc 34:125–128
4. Yang T, Barolli L, Ikeda M, Xhafa F, Durresi A (2009) Performance Analysis of OLSR protocol for wireless sensor networks and comparison evaluation with AODV Protocol. In: International conference on network-based information systems. IEEE Press, Indianapolis, pp 335–342
5. Li M, Liu YH (2009) Underground coal mine monitoring with wireless sensor networks. ACM Trans Sens Netw 5:1–29
6. Yu LM, Li AQ, Sun Z(2008) Design of monitoring system for coal mine safety based on wireless sensor network. In: IEEE/ASME International Conference on Mechtronic and Embedded Systems and Applications, pp 409–414
7. Sun YJ, Qian JS (2007) Research on safty monitor system for underground unmanned roadway based on WSN. Chin J Sens Actuators. 20:2517–2521
8. Hargrave CO, Ralston JC, Hainsworth DW (2007) Optimizing Wireless LAN for Longwall Coal Mine Automation. IEEE Trans Ind Appl 43:111–117
9. Aniss H, Tardif PM, Ouedraogo R (2004) Communications network for underground mines based on the IEEE 802.11 and DOCSIS standards. In: IEEE 60th vehicular technology conference, pp 3605–3609
10. Beaudoin JJ, Tran G, Tardif PM (2004) Underground experiments of video transmission over an IEEE 802.11 infrastructure. In: IEEE 60th Vehicular Technology conference, pp 3610–3614
11. Douglas SJD, Aguayo D, Bicket J, Morris R (2003) A high-throughput path metric for multi-hop wireless routing. In: Proceedings of the 9th annual international conference on mobile computing and networking, pp 134–146
12. Jamali A, Naja N, El Ouadghiri D (2009) Comparative analysis of ad hoc networks routing protocols for multimedia streaming. In: International conference on multimedia computing and systems, pp 381–385
13. Klein A (2008) Performance comparison and evaluation of AODV, OLSR and SBR in mobile ad-hoc networks. In: international symposium on wireless pervasive computing, pp 571–575

Chapter 135
An Illumination Intensity Collecting System of Citrus Leaf Based on WSN

Xuejun Yue, Tiansheng Hong, Songbin Zhai and Yuanjie Chen

Abstract Illumination intensity plays a crucial part in the growth of healthy citrus plant leaves. Lack of illumination or too much illumination will affect the leaves of citrus plants negatively. A less expensive and minimal energy-saving photosensitive sensor is designed to be used on the leaves of citrus plants. The node hardware consists of an Atmega128L low-power MCU, a CC1000 RF IC and so on, utilizing the network coordinator and the RFD nodes to set up a LAN through wireless communication. The leaf intensity values of illumination collected by nodes will be transmitted to the master node and then transmitted to the remote receiver in order to realize the collecting illumination intensity values of various parts of canopy of the citrus plant by real-time wireless collection. When the illumination intensity value does not match the required illumination intensity value, the system will send out different alarms. This data collection system can be applied in daily management of citrus orchards for the purpose of realizing remote surveillance of the plant's growth condition.

Keywords Intelligent · Wireless sensor · Local area network · Citrus plant · Illumination intensity of leaf

X. Yue · T. Hong · S. Zhai · Y. Chen
Key Laboratory of Key Technology on Agricultural Machine and Equipment, Ministry of Education, South China Agricultural University, Guangzhou 510642, China
e-mail: yuexuejun@scau.edu.cn

X. Yue · T. Hong (✉) · S. Zhai · Y. Chen
College of Engineering, South China Agricultural University, 483 Wushan, Tianhe, Guangzhou 510642, China
e-mail: tshong@scau.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_135,

135.1 Introduction

Affected by external environment and internal factors, the process or the intensity of the citrus plant photosynthesis would change, and the illumination intensity among external environment is the most important factor.

The illumination intensity directly restricts the intensity of photosynthesis: being needed in CO_2 assimilation, ATP and NADPH come from crucial enzymes in the light reactions or the dark reactions including KuBP carboxylase and PEP carboxylase in the light activation, etc [1]. Photosynthesis in the citrus plant increases with the illumination intensity increasing in certain ranges. When one condition of illumination intensity, called light compensation point, is reached, it enables a balance of CO_2 to an absorptive amount and lost amount in the photosynthesis, and the intensity of photosynthesis is mainly restricted by the product of the light reaction. When the illumination intensity increases to a certain level (the light saturation point), the intensity of photosynthesis of plant does not increase anymore or it just goes CO_2 through a small increase, which is restricted by enzyme activities and CO_2 concentrations in the dark reactions.

It is of practical significance in the agriculture field to detect and monitor the illumination intensity value in the process of the growth of the citrus plant.

135.2 Confirmation of Overall Design

WSN nodes consist of one network coordinator (host node) and two secondary nodes and compose star network [2]. Secondary nodes are set in different parts or representative positions of a same citrus plant as well as different citrus plants. Illumination intensity values collected by these sensors are subjected to a process of analog-to-digital conversion, and then sent to the host node through wireless channels. The data can be processed by circuits in host nodes: two threshold values are set according to light compensation point and light saturation point; two LCD display illumination intensity values collected by two nodes, and LED shows indicator lights of corresponding ranges of illumination values. Buzzer decides whether to give an alarm according to whether the illumination value exceeds various threshold values or not. Network coordinator (host node) with serial port can connect to the PC through a serial port line, and PC can not only debug software by the serials, but also can collect data synchronously. Overall design diagram is shown in Fig. 135.1.

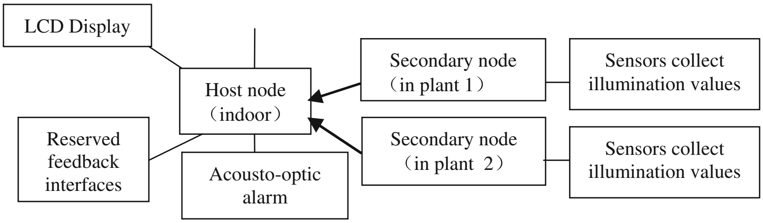


Fig. 135.1 Overall design diagram

135.3 Design and Manufacture of System Hardware

135.3.1 Selection of WSN Components

The system is used to monitor citrus plants in citrus orchards for severe environment and low-power consumption. It does not request extremely high processing capacity. So it is appropriate to use MSP430F14 and Atmgea128L [3]. Considering low cost, Atmega128L is chosen as the processor chip.

Taking RF chip performance, cost and power dissipation into consideration, TR1000 of the company RFM and CC1000 of Chipcon are ideal choices. These two chips have their own advantages. TR1000 saves more energy while CC1000 has higher sensitivity with longer transmitting distance. In addition, CC2420 of Zigbee is more widely applied to wireless sensor network. Finally, CC1000 is used as RF chips of WSN nodes. And a simplified LDR is chosen as the sensor to induce illumination intensity and the LCD display nodes as the host nodes sending data.

135.3.2 Design of Network Coordinator and Secondary Node

Secondary nodes collect illumination values through the LDR circuit, and then transform the collected analog signals into the digital signals stored in the registers. And the data is transmitted to the register by CC1000 and sends out after modulation through wireless signals.

Antennas of the network coordinator receive data signals, and the original signals would be reconstructed through modulation of CC1000, which then leave up to ATmega128L to be processed. Besides, the data will be displayed in a LCD1602 and in upper computer through serial connection. In the meantime, it can adjust upper and lower limits through button operation. When received illumination intensity value has connection with the set values, there are some

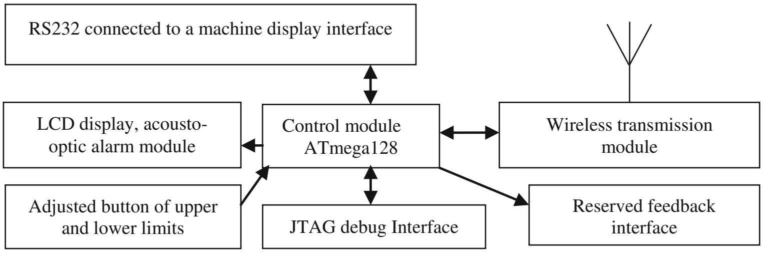

RS232 connected to a machine display interface
LCD display, acousto-optic alarm module
Control module ATmega128
Wireless transmission module
Adjusted button of upper and lower limits
JTAG debug Interface
Reserved feedback interface

Fig. 135.2 Web coordinator diagram

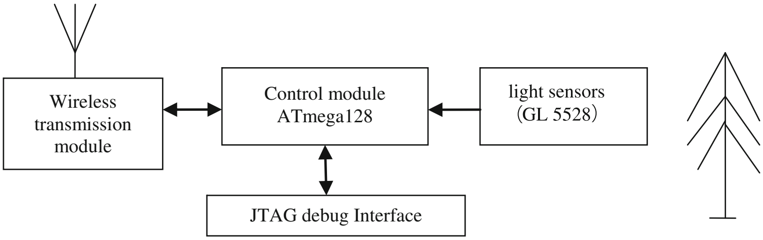


Fig. 135.3 From the node diagram

follow-up movements to show, such as acousto-optic alarm and so on. It is illustrated in Figs. 135.2 and 135.3.

135.3.3 Design of CC1000 Peripheral Circuit

CC1000 supports frequency ranging from 300 to 350 MHZ. Fequency of 433 MHZ is chosen as the transmission band because it is a domestic open frequency band. Component parameters of peripheral circuit are calculated by the SmartRF studio. The circuit is shown in Fig. 135.4.

135.3.4 Processor and its Peripheral Circuit

Processor module of sensor node is composed of a Atmega128L chip as processor, a peripheral decoupling circuit, a filter circuit, an oscillation circuit and various kinds of peripheral interfaces, as illustrated in Fig. 135.5.

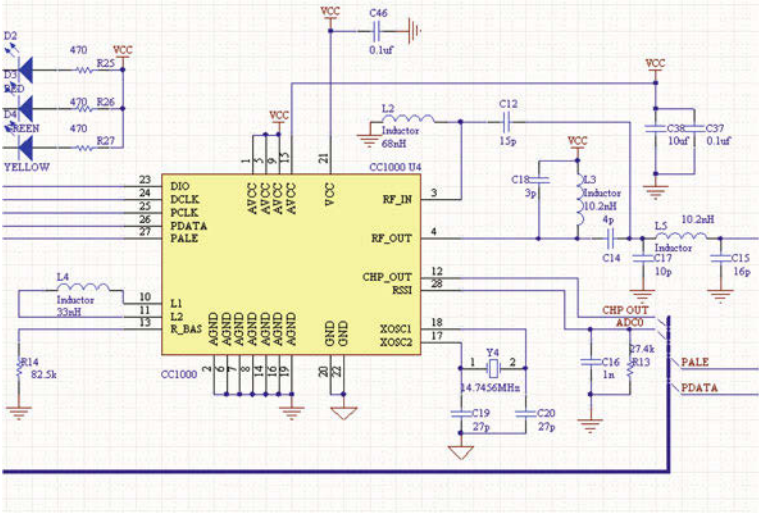


Fig. 135.4 CC1000 peripheral circuit

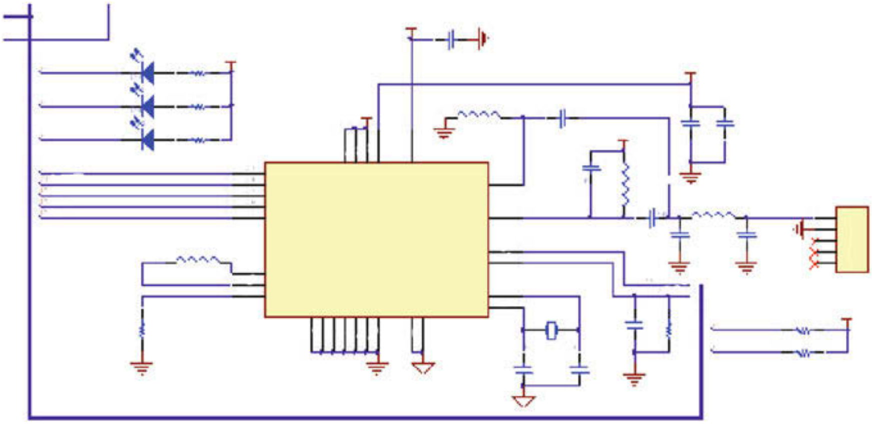


Fig. 135.5 Processor and peripheral circuit

135.3.5 External Interface Circuit of Network Coordinator

External interface circuit of network coordinator includes a LCD1602, a LED (red, green, yellow) display, a buzzer and a serial port circuit. Three external AA batteries are adopted as the power supply. External interface circuit diagram is shown in Fig. 135.6.

External interface circuit PCB of the network coordinator is shown in Fig. 135.7.

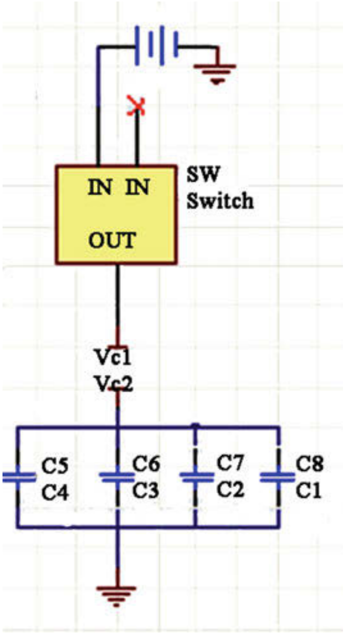


Fig. 135.6 The external interface circuit nodes

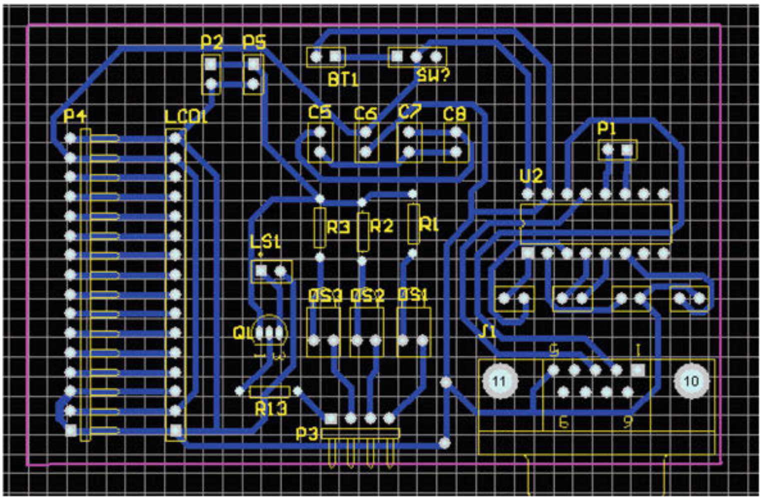


Fig. 135.7 The master node external interface circuit PCB

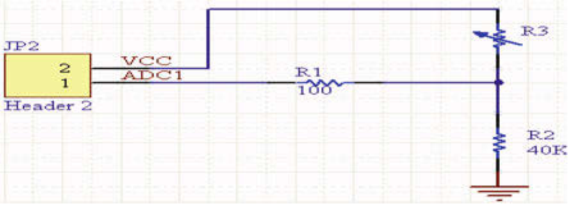


Fig. 135.8 Light intensity acquisition circuit

135.3.6 *Design of Light Intensity Acquisition Circuit*

WSN child node also needs an external interface circuit, which is a light intensity acquisition module. This module is constructed using a LDR and is shown in Fig. 135.8.

135.4 Research and Achievement of Intelligent Control

To achieve the functions, the system needs support of software after WSN network coordinator and child node and attachment module are constructed as the system hardware. Programing, compiling and the creation of final hex file happen in the software environment called WinAVR.

The flowcharts of the system network coordinator and the RFD node are shown in Figs. 135.9 and 135.10, respectively.

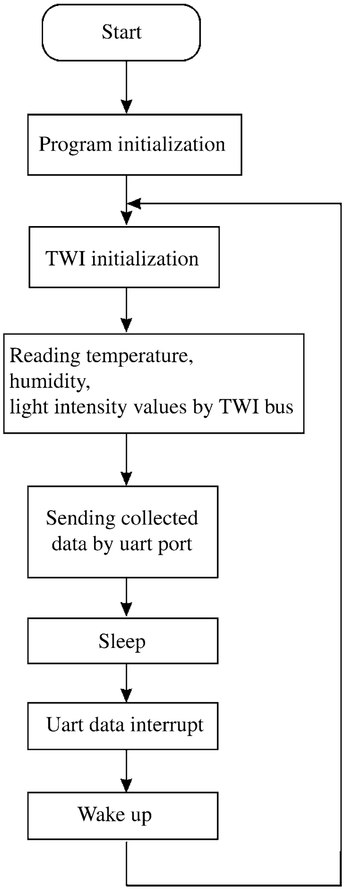
135.5 Experiment and Analysis of System Test

135.5.1 *Tools and Methods of System Test*

An illumination intensity adjustable desk lamp and a multimeter are used in the lab test experiment. And adjustable range of lamp would be as wide as possible.

Methods and steps of system test: (1) Set upper limit value of ADC to 860 mA, and set lower limit value of ADC to 341 mA. According to the analog-to-digital formula, ADC = VinX1024/Vref, Vref = 3 V, we can calculate corresponding upper and lower limit input voltages as 2.5 and 1 V, respectively. Through the calculation of LDR circuit, we also know corresponding values of LDR are 8 and 120 KΩ. By reading illumination-resistance characteristic diagram of GL5528, corresponding illumination intensity is about 80 and 5 lux. (2) Place the WSN node under the desk lamp, then start the system. (3) Adjust the brightness of the desk lamp and finely tune the distance between the WSN node and the desk lamp, then record the data shown on the LCD (ADC switching value). (4) The voltage between pin ADC1 of ATmega128L and ground can be measured by a multimeter, and we can get the ADC value through a formula that is ADC = VinX1024/Vref.

Fig. 135.9 The flow chart of data collection procedure



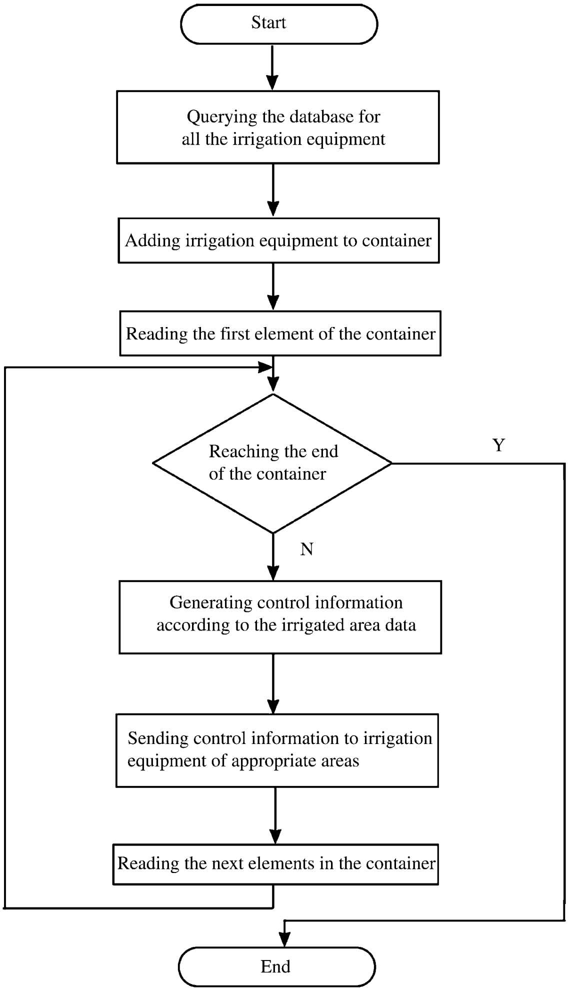
(5) Repeat steps 3 and 4 to get measurements of 5 groups, and observe the situation of acousto-optic alarm and calculate the error of data.

135.5.2 Experiment Result and Analysis of System Test

Experiment data is shown in Table 135.1. System average error is 1.064%. Experiment results show that illumination intensity can be measured in this design and the error value is acceptable. The reasons causing the error: (1) Measurement error of the multimeter. (2) Unstable light source leading to unstable LCD display. (3) The error of visual inspection.

Experiment shows that collected data of illumination intensity value is indirect and there existed a nonlinear relationship between the illumination intensity and

Fig. 135.10 The flow chart of control procedure


Start
Querying the database for all the irrigation equipment
Adding irrigation equipment to container
Reading the first element of the container
Reaching the end of the container
Y
N
Generating control information according to the irrigated area data
Sending control information to irrigation equipment of appropriate areas
Reading the next elements in the container
End

the LDR value, which is not convenient for the direct extraction of illumination intensity that we can only get corresponding value by illumination-resistance characteristic diagram. In daily management of citrus orchards, precision requirement for illumination intensity of plant are not high and it is all right to estimate the illumination range according to the collected data. So this system is of good practicability. We can achieve intelligent measurement of illumination intensity value exactly and rapidly by better illumination sensors so that the system can be developed further.

Table 135.1 Light intensity value data experiment

Exp	LCD	ADC measure value/V	ADC calculation value/V	Error (%)	Acousto-optic alarm
1	751	2.2	750	0.13	Green light is on
2	961	2.8	955	0.62	Red light is on buzzer sounds
3	793	2.3	785	1	Green light is on
4	440	1.7	443	0.67	Green light is on
5	105	0.3	102	2.9	Yellow light is on

135.6 Conclusion

The system uses the Atmega128L low-power AVR single chip and the CC1000 RF IC so that it saves energy. Underlying protocol confirms to the IEEE 802.15.4 standard protocol and the OSI, and it is easy to be developed once again. The system adopts CSMA/CA back off mechanism to keep two child nodes from data collision in the process of sending data to network coordinator and it can change the transmitted power of WSN nodes through programing.

After the setup of the system, it has certain practical significance that it can be realized that the secondary nodes collect the illumination intensity; the host nodes transmit the data to PC wirelessly and succeed in the non-destructive real-time collection of illumination intensity value of citrus.

Acknowledgments The work was supported by grants from Special Fund for Agro-scientific Research in the Public Interest of China (No. 200903023), Earmarked Fund for Modern Agro-industry Technology Research System of China ([2007]431), China national spark program (2010GA780046), Science and technology projects in Guangdong Province (2010B020314007) and Guangdong province agricultural mechanization scientific research issue project scheduling ([2010]605).

References

1. Akyildiz IF (2002) A survey on sensor networks [J]. IEEE Commun Mag 40:102–114
2. Beckwith R, Teibel D, Jones PB (2004) Report from the field: results from an agricultrual wireless sensor network [J]. In: Proceedings of 29th IEEE LCN'04, Tampa, 15–17 November 2004
3. Baggio A (2005) Wireless sensor networks in precision agriculture [J]. In: Proceedings of REALWSN'05, Stockholm, 20–21 June 2005

Chapter 136
An Approach to Construct Coarse-Grained Program Slicing

Lin Du, Yehong Han and Ke Zhao

Abstract On the basis of analyzing the defects present in traditional system dependence graph, the method based on ripple effect is proposed to compute coarse-grained program slicing. The method perfects object-oriented program semantics and to reduce the computation complexity by expanding the signification of coarse-grained and analyzing ripple effect. Object-oriented program semantics are described in detail. At length, the algorithms for analyzing ripple effects, constructing system dependence graph and computing coarse-grained program slicing are designed.

Keywords Program slicing · Coarse-grained · Ripple effect · System dependence graph

136.1 Introduction

Program slicing is a technique for simplifying programs by focusing on selected aspects of semantics. Program slice consists of the parts of a program that potentially affect the values computed at some point of interest. Such a point of

L. Du (✉) · Y. Han · K. Zhao
School of Computer Science and Technology, University of Qilu Normal,
Jinan 250014, Shandong, China
e-mail: dul1028@163.com

Y. Han
e-mail: sdzzhyh@163.com

K. Zhao
e-mail: ke_zhao2008@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_136,

interest is referred as a slicing criterion, and is typically specified by a pair (program point, set of variables). The parts of a program that have a direct or indirect effect on the values computed at a slicing criterion C constitute the program slice with respect to criterion C. As a viable method to restrict the focus of a task to specific subcomponents of a program, program slicing has extensive applications in software engineering. With regard to program slicing, the fine-grained reach the level of statement, while the coarse-grained reach the level of method. Coarse-grained slicing is more suitable for software measurement. We generally use the system dependence graph for static slicing of single-procedure programs. Object-oriented program system dependence graph (OOSDG) can reflect the semantic characteristics of object-oriented programs, deal with data and control flow between the processes, describe parameter transference and carry out inter-process analysis. However, the following questions in the actual construction of system dependence graph are present.

First, the method to construct system dependence graph is complicated, more over by the lack of accuracy. Because of high complexity, on account of different studies, the actual construction ignores some of the semantic characteristics of object-oriented program. This would result in inaccuracy. For example, in the OOSDG, only one parameter node is constructed for each corresponding class member variable. The class member variable has a separate copy in each class instance, respectively. This indicates that a class member variable defined in one place is used in different places, which would result in the wrong data dependence among different class instances. Second, traditional construction method results in the loss of program semantic [1]. System dependence graph based on the analysis of process and lack of semantic association, does not reflect the characteristics of object-oriented language completely [2]. The interactions among the objects constitute the main framework of the object-oriented program. However, this interaction does not base on the order of execution. Hence, the construction of the traditional system dependence graph would not take into account all the relationships present among the objects. Therefore, the semantic of object-oriented is lost more.

In order to solve the above defects, the method based on ripple effect analysis is presented to construct coarse-grained program slicing. First of all, the meaning of coarse-grained is extended in order to make the size of grain come up to object-oriented program's semantic units that are class, instance and member method, and member variable. Ripple effects analysis plays the role of two aspects. First, the results of the ripple effect are mapped to the dependence graph in order to add semantic relationship among different objects. Second, the scope of analysis through ripple effect is narrowed in order to reduce the complexity of constructing a graph.

The rest of the chapter is organized as follows. In Sect. 136.2, the meaning of coarse-grained is extended to simplify system dependence graph. The method of ripple effect analysis is proposed. By constructing coarse-grained system dependence graph based on ripple effect, object-oriented program semantics are described in detail. In Sect. 136.3, this paper designs the algorithms for analyzing ripple effects, constructing system dependence graph and program slicing. Section 136.4 concludes the whole chapter.

136.2 Constructing Coarse-grained System Dependence Graph Based on Ripple Effect Analysis

136.2.1 Expanding the Meaning of Coarse-Grained and Simplifying System Dependence Graph

In order to understand object-oriented programming, analysis of the interactions' relationship among multiple units is better than analyzing a single statement. The meaning of coarse-grained is extended in order to make the size of grain come up to object-oriented program's semantic unit that is class, instance, member method and member variable [3].Coarse-grained expanded is defined as follows

Definition 1 The graph G which meets the following characters is referred to as coarse-grained. (1) Graph G contains statement and predicate in the main(),class, instance, member method and member variable.(2) If a statement which is in the member method M belongs to G, then M also belongs to G. (3) If the instance, member method or member variable in the class A belongs to G, then class A also belongs to G.

On the basis of defining the coarse-grained, system dependence graph is simplified. Describing the process dependence,it need not enter the process inside but indicates process prelude node only. The data dependence which belongs to parameter nodes of different methods is indicated by data dependence among multiple methods. It is achieved by data dependence edge which points to the call directly.

136.2.2 Ripple Effects Analysis

Ripple effects analysis plays the role of two aspects. First, the results of the ripple effect are mapped in the dependence graph in order to add semantic relationship among different objects [4]. Second, the scope of analysis through ripple effect is narrowed in order to reduce the complexity of constructing the graph. Analysis of ripple effect for recording the units involved by the ripple of one unit which is called the source of ripple. The following method is used: First step, the complete ripple graph which reflects corresponding object-oriented programming is constructed. Starting from the source of ripple, the direct and indirect ripple unit can be found through traversal of all the ripple edges. However, the above method has the following problem. The method to construct the complete ripple graph is complicated whose computation complexity is same as constructing system dependence graph. The result does not match our original intention of reducing the analysis scope through ripple effects analysis. Object-oriented program has the following properties. The interaction among the various units is either direct or indirect [5, 6]. In particular, the indirect relationship can be expressed by the direct relationship among multiple units, this is called transitive. We can draw on the experience of the method to process transitive in the cluster. Ripple effect can be recorded through the use of matrix. Thereby the complete ripple effects can be calculated through matrix transitive operations.

136.2.3 Semantic Description

Object-oriented program semantics are described in detail in the system independence graph based on ripple effect as follows.

136.2.3.1 Description of Class, Instance, Member Method, the Relationship among Member Variables

In order to express membership, instance node, member method prelude node and member variable node are connected to the accessory class prelude node. Method and process have the same status. For method and process we do not achieve internal processing but provide a prelude node. The meaning of method prelude node is expanded through hiding data transference among multiple parameter nodes. The expression of data dependence among parameter nodes of different methods relies on data dependence among methods. Then, three kinds of nodes should be increased as follows: Member variable node should be increased because member method refers to member variable; instance node should be increased because member method refers to class instance node; when a method is called by the class instance, instance node expressing message receiver object is increased in the method node. The aim is to reflect the change of object's state.

136.2.3.2 Description of Class Inheritance

In order to express inheritance, different class prelude nodes which have inheritance are connected. When one class interacts with the other class, it is convenient to couple each other through class prelude node and class member edge. In order to reflect the inheritance hierarchy clearly and reduce backtracking, we take the following approach. If a virtual method in the child class which is inherited from the parent class is modified, the method is described only in the child class. Meantime, associated edge should be increased between class prelude node of the parent class and method prelude node of the child class. Thus making the expression of inheritance mechanism and virtual method not to require to increase associated edge between method of the parent class and method of the child class. Only the associated edge is increased between class prelude node and method prelude node.

136.2.3.3 Description of Polymorphism and Dynamic Binding

Polymorphism can be expressed completely by the virtual method prelude node and polymorphism call edge. Through multiple call edge, call node can be connected to each method node which is called by object possibly. The dynamic selection can be expressed by multiple polymorphism nodes which have the same protocol. This method can express all the possibilities.

136.3 Algorithm Design

In the following, the algorithms for analyzing ripple effects and constructing system dependence graph are designed.

136.3.1 Computing Ripple Effect

Algorithm 1 Ripple effect analysis.

Input: The source of the ripple
Output: The units influenced by the ripple source
Step 1: The range of analysis which consists of the source of the ripple and all the units which participate in the ripple effect analysis is determined.
Step 2: The matrix REA and the matrix REO are defined. REA recorded the direct ripple effect of all the units. REO is defined as follows: the value of the unit which is the source of the ripple is 1.Another unit's value are 0. It is easy to find that the unit whose value is 1 must located in the diagonal of the matrix REO.
Step 3: The transmission of ripple effect is calculated. REO = REO*REA. For the matrix REO, all the units' value is modified as 1 if the value is not 0.
Step 4: If the matrix REO is different from the initial REO, then the algorithm returns to step 3.Otherwise, the algorithm would carry on with the next step.
Step 5: If the matrix REO no longer varies, then the ripple effect analysis is over. For the final REO, the units whose value is 1 are the required units.

136.3.2 Constructing System Dependence Graph

The scope of the description of the system dependence graph is reduced to the statement, the predicate in the main ( ) and class, instance, member method and member variable are achieved by the ripple effect analysis.The object-oriented program is represented as a two-tuples (M, C) in which M is the main ( ) and C is a collection of classes. The algorithm calls the function connect ( ) which connects call node of process dependence graph and method prelude node [7, 8]. Meantime the class graph and the main program are connected.

Algorithm 2 Constructing system dependence graph.
Input: the abstract syntax tree of P = (M, C)
Output: the OSDG of P

```
void ConstructOSDG ()
      { for (class Ci of C)
         { for (method m defined in Ci)
            { if (m is ''marked'')
                    make Ci and the ''marked'' method of
                    base class connected as membership;
              else{
                    calculate the PrDG of m;
                    make m as''marked'';
                   }
              }//end for1
            }//end for2
            connect();
         }//end ConstructOSDG
```

136.3.3 Program Slicing

Algorithm 3 Computing coarse-grained program slicing.
Input: the OSDG of P
Output: coarse-grained slicing

```
void SliceNodeObject (Node node, EdgeSet includeEdges,
NodeSet visitedNodes, BOOL back)
      {
         if (node is not marked)
         {
            mark node as visited;
            insert node into visitedNodes;
            if (back)
             {
              for (all edges e leading from other node n to node)
              {
               if (kind of e is in includeEdges)
               SliceNodeObject (n, includeEdges, visitedNodes,
               back);
               }//end for1
              }
            else
              for (all edges e leading from node to other node n)
            {
             if (kind of e is in includeEdges)
             SliceNodeObject(n, includeEdges, visitedNodes,
                                                     back);
           }//end for2
         }//end if2
      }//end if1
    }//end SliceNodeObject
    void ComputeSlice (Node node)
      // phase 1
      SliceNodeObject (node, {control dependency edge, data
           dependency edge, polymorphic call edge,call edge},
                                     visitedNodes, FALSE);
      // phase 2
      for (all nodes n in visitedNodes)
           SliceNodeObject (n, {call edge,instance edge },
                                      visitedNodes, TRUE);
    } //end ComputeSlice
```

136.4 Conclusions

This chapter analyzes the defects present in traditional system dependence graph. The defects include high computation complexity, deficiency of accuracy and loss of program semantic. The method based on ripple effect is proposed to compute coarse-grained program slicing. Coarse-grained is extended and defined in order to make the size of grain come up to object-oriented program's semantic units ,that are class, instance and member method, and member variable. Ripple effects analysis plays the role of two aspects. First, the results of the ripple effect are mapped to the dependence graph in order to add semantic relationship among different objects. Second, the scope of analysis through ripple effect is narrowed in order to reduce the complexity of constructing the graph. Finally, the algorithms for analyzing ripple effects, constructing system dependence graph and program slicing are designed. The recent years have witnessed the increase of software size and complexity. Thus, several software engineering tasks required to reduce the size of programs or to decompose a larger program into smaller components. As a viable method to restrict the focus of a task to specific subcomponents of a program, program slicing has extensive applications in software engineering. Recent works about program slicing are concerned on improving the precision of slicing methods and computing object-oriented program slicing. We strongly believe that, in the near future, this research field will be paid much more attention by the researchers and promote the fundamental theories' research in the related fields, for example program debugging [9, 10], program testing [11–13], software measurement [14, 15] and software maintenance [16, 17].

Acknowledgments This work is partially supported by the Shandong Province High Technology Research and Development Program of China (Grant No. 2011GGB01017); Research Foundation of Qilu Normal University; Project of Shandong Province Higher Educational Science and Technology Program; Soft Science Research Program of Shandong Province.

References

1. Ap Xu BW, Zhou YM (2001) Comments on a cohesion measure for object-oriented classes. Softw Pract Experience 31(14):1381–1388
2. Chen ZQ, Zhou YM, Xu BW, Zhao JJ, Yang HJ (2002) A novel approach to measuring class cohesion based on dependence analysis. IEEE international conference on software maintenance, pp 377–383
3. Chen ZQ (2001) Slicing object-oriented Java programs. ACM SIGPLAN Not 36(4):33–40
4. Xu BW, Chen ZQ, Zhou XY (2001) Slicing object-oriented Ada95 programs based on dependence analysis. J Softw 12(12):208–213
5. Chae HS, Kwon YR (1998) A cohesion measure for classes in object-oriented systems. In: Proceedings of the 5th international software metrics symposium. IEEE Computer Society Press, pp158–166
6. Briand LC, Morasca S, Basili VR (1999) Defining and validating measures for object-based high-level design. IEEE Trans Softw Eng 25(5):722–743
7. Korel B, Tahat L, Bader A (2003) Slicing of state based models. In: Proceedings of the IEEE international conference on software maintenance, pp 34–43
8. Harrold MJ, Jones JA (2001) Regression test selection for Java software, OOPSLA 2001, pp 313–326
9. Jiang S, Zhang C (2010) A debugging approach for Java runtime exceptions based on program slicing and stack traces. In: Proceedings of the 10th quality software international conference
10. Horwitz S, Liblit B, Polishchuk M (2010) Better debugging via output tracing and callstack-sensitive slicing. Softw Eng 36(1):7–19
11. Li H, Peng Y, Ye X, Yue J (2010) Test sequence generation from combining property modeling and program slicing. In: Proceedings of the 34th annual computer software and applications conference, Seoul
12. Majumdar R, Xu R-G (2009) Reducing test inputs using information partitions. Lect Notes Comput Sci 5643:555–569
13. Tao C, Li B, Sun X, Zhang C (2010) An approach to regression test selection based on hierarchical slicing technique. In: Proceedings of 2010 IEEE 34th annual computer software and applications conference, Seoul
14. Gupta N, Rao P (2001) Program execution-based module cohesion measurement. In: Proceedings of the 16th IEEE international conference on automated software engineering, San Diego
15. Meyers TM, Binkley D (2004) Slice-based cohesion metrics and software intervention. In: Proceedings of the 11th working conference on reverse engineering
16. Hill E, Pollock L (2007) Exploring the neighborhood with dora to expedite software maintenance. In: Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering
17. Hoang Viet MS (2009) Software maintenance: a program slicer using cross referencer. ProQuest Diss Theses 48(4):72–76

Chapter 137
Study on Intrusion Detection Model Based on Improved Genetic Algorithm and Fuzzy Neural Network

Jinguang Chen, Yuesheng Gu and Zhixiong Li

Abstract The network intrusion is one of the most important issues for the security of the internet. The internet intrusion may lead to terrible disaster for network users. It is therefore imperative to detect the network attacks to protect the information security. However, the intrusion detection rate is often affected by the structure parameters of the fuzzy neural network (FNN). Improper FNN model design may result in a low detection precision. To overcome these problems, a new network intrusion detection approach based on improved genetic algorithm (GA) and FNN is proposed in this chapter. The improved GA used energy entropy to select individuals to optimize the training procedure of the FNN, and satisfactory FNN model with proper structure parameters was then attained. The efficiency of the proposed method was evaluated with the practical data. The experiment results show that the proposed approach offers a good intrusion detection rate, and performs better than the standard GA-FNN method with respect to the detection rate. Thus, the proposed new intrusion detection method is efficient for practice applications.

Keywords Intrusion detection · Improved genetic algorithm · FNN

J. Chen (✉)
School of Teacher Education, Huzhou Teachers College,
Huzhou 313000, People's Republic of China
e-mail: cjg2003@hutc.zj.cn

Y. Gu
Henan Institute of Science and Technology,
Xinxiang 453003, People's Republic of China
e-mail: hz34567@126.com

Z. Li
Wuhan University of Technology,
Wuhan 430074, People's Republic of China
e-mail: bingheku6@sina.com.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_137,

137.1 Introduction

Network security is a hot topic in computer security and defense worldwide. Intrusion may lead to absence of the internet service and even cripple the whole system for weeks. Hence, it is very important to detect the intrusion in time to prevent breakdowns. Advanced machine learning algorithm, including evolution algorithm, intelligent artificial neural network (ANN) and support vector machine (SVM) and so on, all appear in the intrusion detection of the networks [1]. Among them, ANN [2] is the most extensively used method. However, ANN detection performance is mainly determined by its structural parameters. It is often difficult to determine the ANN parameters without a large number of trials. Although Xie [3] and Qiao et al. [4] use GA to tune the ANN structures improve the network attack detection accuracy, but without considering the GA individual selection adjustment, and just using the KDD data set to validate their methods, and not for practical applications. Therefore, to improve the GA optimization process and to test the real data set will have important significance for the ANN based intrusion detection [5–8].

In order to solve the above problems, this paper proposed a new intrusion detection method. This method has been marked to achieve fuzzy-ANN parameter optimization using improved GA. By using the practical data set for experimental analysis, the results show that the new method can detect the network attack efficiently and the detection rate is higher than the standard GA-based method.

This chapter is organized as follows. In Sect. 137.2, the proposed method for intrusion detection based on the combination of improved GA and FNN is described. The application of the proposed method is presented for intrusion detection in Sect. 137.3. The performance of the proposed new method is described. The effectiveness of the proposed method is valued by analyzing the real attack data. Conclusions are drawn in Sect. 137.4.

137.2 Improved GA-ANN Intrusion Detection Model

Due to the interference of inside and external excitations, the intrusion is a kind of typical non-stationary signal. The different signal components of intrusion data exhibit various characteristics. The uncertainties of the signal make the actual attack present fluctuations which are difficult to identify. ANN is an intelligent approach to deal with non-stationary signal. With its strong learning ability, the ANN is quite suitable for practical intrusion detection applications. So the hidden intrusion pattern can be easily recognized by ANN. However, its identification efficiency relies mainly on the structural design of the ANN. Thus, the GA optimization is applied to the ANN design. By doing so, satisfactory ANN detection model can be obtained, and hence the detection rate could be improved.

In the analysis of intrusion detection, the FNN is first used to learn the actual link between the recoded data and the intrusion. Then, the improved GA is applied to the model optimization. Finally, the intrusion detection model is tested by new samples. The detail of the new method is described as follows.

137.2.1 Improved GA

Standard GA [9] involves the following procedures: (137.1) coding, (137.2) selection, (137.3) crossover, (137.4) mutation. In the individual selection process, standard GA only searches individuals with adaptive value. The inner link among the selected individuals is usually neglected. Therefore, the probability distribution of the population is hard to form, leading to a low convergence speed, so as to make the GA optimization fall into premature convergence [10]. To overcome this shortcoming, we used the energy entropy-based selection to increase the diversity of population. As GA selects according to the individuals' energy, it needs to calculate each individual x. The individual energy entropy can be expressed as [10]

$$E(t) = \mathrm{In}(\frac{1}{p_t}) \tag{137.1}$$

where, p_t denotes energy probability at lever t. By employing the annealing selection method [10], we can derive the individual selection probability [11]

$$P(x_i) = e^{-(E(x_i)+\beta f(x_i))} / \sum\nolimits_{i=1}^{n} e^{-(E(x_i)+\beta f(x_i))} \tag{137.2}$$

where, $P(x_i)$ denotes the probability of individual x in new population, $(E(x_i) + \beta f(x_i))$ denotes fitness value of individual x.

After the energy entropy based individual selection procedure, the link between individuals is connected to maintain the diversity of population.

137.2.2 Fuzzy Neural Network

Since the integrated fuzzy logic and ANN provides more powerful learning ability, we use the integration of the fuzzy-artificial neural network (FNN) to detect the intrusion in this chapter.

Fuzzy method is used in various problems. However, the determination of membership functions depends on human experts and experiences, resulting in time consumption and lack of self-adaptation. To overcome this problem, the artificial neural network (ANN) has been applied to auto-tune the membership functions of the fuzzy inference [12]. The structure of fuzzy neural network is shown in Fig. 137.1.

The FNN consists of four layers. The input layer connects with input feature vector $P = [p_1, p_2, \ldots, p_i]^T$. The fuzzy layer is used to fuzz each input p_i to get

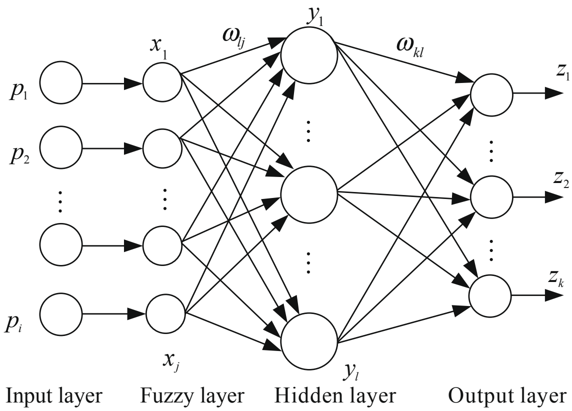


Fig. 137.1 Structure of fuzzy neural network

corresponding fuzzy membership values $x_j = \mu_{Aij}(p_i) = [q_{1j}, q_{2j}, \ldots, q_{ij}]^T$. The Gaussian function was adopted as the fuzzy membership function, that is

$$\mu_{Aij}(p_i) = \exp\left[-\left(\frac{p_i - a_{ij}}{b_{ij}}\right)^2\right] \tag{137.3}$$

where a_{ij} is the center of membership function and b_{ij} is the width of the function. Hence, the output of the fuzzy layer is $\mathbf{X} = [x_1, x_2, \ldots, x_j]$. In the hidden layer, for the l neuron nodes, the weights ω_{lj} were used as the fuzzy relation matrix to perform fuzzy inference rules. Then the singleton output of the lth fuzzy rule is

$$y_l(x_j) = \omega_{lj} x_j \tag{137.4}$$

The fourth layer outputs the fuzzy decision of the FNN. The weighted average method for inverse fuzzy was used in this paper. It can be noted that the main purpose of the FNN is to optimize the coefficients of a_{ij}, b_{ij}, ω_{lj} and ω_{kl}. The traditional way is to train the FNN by back propagation (BP) algorithm. However, the BP algorithm may lead to local optimal solutions, declining FNN performance. For this reason, the improved GA was applied to train the FNN because of its good global searching ability and strong robustness. It first encodes the coefficient values to form chromosomes, and then performs replication, crossover and mutation to evolve the chromosomes. Finally, use the optimal chromosome to represent network weight and membership function coefficients. In the GA searching processing, the fitness function plays an important role in the optimization result. To ensure the searching precision, the following fitness function was used:

$$J = \frac{1}{e(q)}, \quad \text{and} \quad e(q) = \frac{1}{2}\sum_{p}\sum_{k}(z_k - T_k)^2, \tag{137.5}$$

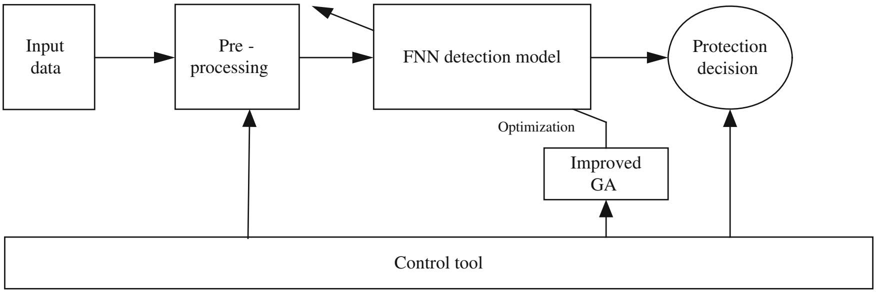


Fig. 137.2 The network intrusion detection system based on improved GA-FNN

where $q(q = 1,\ldots,N)$ is the chromosome number, p is training sample number, T_k is desired output, and z_k is the real output of FNN. The intrusion detection flowchart of improved GA-FNN is shown in Fig. 137.2.

137.3 Experimental Analysis

In order to validate the performance of the proposed algorithm, the intrusion experiments were carried out in real practice application in this paper. The DoS and probe attack types were investigated. The recorded data describes 10 main attributes of the test network connection, including duration, service type, the bytes issued from source to destination, the bytes from destination to source, etc. In the intrusion detection, 10,000 normal samples, 5,000 DoS attack samples and 1,000 probe attack samples were used to evaluate the proposed approach.

As mentioned above, the improved GA-FNN is employed to identify the network attacks. The input feature vector of the FNN is 10 network features. The output adopted decimal coding, i.e., $T = [1\text{–}3]$, where $1-3$ indicated the normal, DoS and Probe in the experimental tests, respectively. Hence, the structure of 10-15-25-3 was chosen for the fuzzy neural network in this work. In the GA optimization, the population size is 250, the crossover probability is 0.9, and the mutation probability is 0.01.

The GA optimization procedure is illustrated in Fig. 137.3. The evolution performance of the improved GA and standard GA is compared. One can note that the improved GA has faster convergence speed and better fitness value, which is benefited from the energy entropy-based annealing selection process.

The intrusion detection performance of improved GA-FNN model and standard GA-FNN model is shown in Table 137.1. From Table 137.1, the proposed method for intrusion detection performance is better than the standard GA-FNN. By optimization of the individual selection, the GA calculation speed is increased when compared with the standard method, and the detection rate is enhanced by 2.5%, as well as the false alarm rate by 0.66%. For this experiment result,

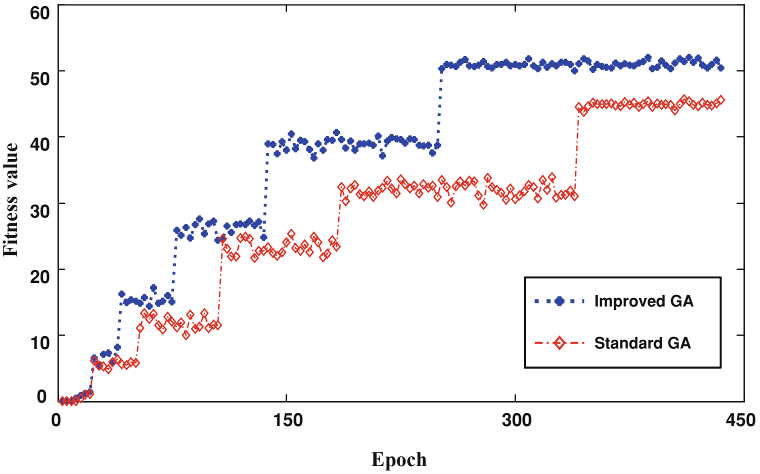


Fig. 137.3 The performance for FNN optimization using improved GA and standard GA

Table 137.1 The intrusion detection performance of the FNN-based model

Improved GA-FNN model			Standard GA-FNN model		
Detection rate	False alarm	Time (s)	Detection rate	False alarm	Time (s)
95.3	1.12	1.27	92.8	1.78	1.91

Table 137.2 The intrusion detection performance of improved GA-FNN and FNN

FNN model			Improved GA-FNN model		
Detection rate	False alarm	Time (s)	Detection rate	False alarm	Time (s)
88.7	2.36	1.57	95.3	1.12	1.27

the intrusion detection performance of improved GA-FNN has obviously enhanced the detection rate. One can note that the improved GA is more effective in the FNN optimization than the standard GA in this specific case.

The intrusion detection performance of improved GA-FNN model and FNN model is shown in Table 137.2. It can be seen from Table 137.2 that the GA optimization plays an important role in the intrusion detection busing FNN model, which can improve system detection efficiency significantly. Compared with FNN method, the new approach increases the detection rate by 6.6%, as well as the false alarm rate by 1.45%.

137.4 Conclusions

Intelligent method has been widely used in intrusion detection, especially for the fuzzy logic and ANN. However, reasonable structural parameters of the intelligent model play an important role in satisfactory detection performance. Therefore, this paper proposed a new intrusion detection method based on improved GA-FNN. The innovation is that the new method uses the energy entropy-based selection method to ensure the diversity of the generations of the GA, and hence the optimization of the FNN structure could be enhanced when compared with standard GA. The real practice data was applied to the validation of the proposed approach. The analysis results verify the effectiveness of this method. The intrusion detection rate and false alarm have been improved when compared with the standard GA-FNN approach, and hence the proposed method is of application importance.

References

1. Zhao X, Jing R, Gu M (2008) Adaptive intrusion detection algorithm based on rough sets. J T singhua Univ (Sci Tech) 48:1165–1168
2. Li Z, Yan X, Yuan C, Zhao J, Peng Z (2011) Fault detection and diagnosis of the gearbox in marine propulsion system based on bispectrum analysis and artificial neural networks. J Marine Sci Appl 10:17–24
3. Xie Z (2010) Support vector machines based on genetic algorithm for network intrusion prediction. J Comput Simul 27:110–113
4. Qiao L, Peng X, Ma Y (2006) GA-SVM-Based feature subset selection algorithm. J Electronic Meas Instrum 20:1–5
5. Xiong W, Wang C (2009) Hybrid feature transformation based on modified particle swarm optimization and support vector machine. J Beijing Univ Posts Telecommun 32:24–28
6. Li Z, Yan X, Yuan C, Peng Z (2010) Gear faults diagnosis based on wavelet-AR model and PCA. Proc SPIE 7820:138–141
7. Li Z, Yan X, Yuan C, Li L (2010) Gear multi-faults diagnosis of a rotating machinery based on independent component analysis and fuzzy k-nearest neighbor. Adv Mater Res 108–111:1033–1038
8. Li Z, Yan X (2011) Application of independent component analysis and manifold learning in fault diagnosis for VSC-HVDC systems. Hsi-An Chiao Tung Ta Hsueh 45:46–51
9. Ahn CW, Ramakrishna R (2002) A genetic algorithm for shortest path routing problem and the sizing of populations. IEEE Trans Evol Comput 6:566–579
10. Fu X, Zhengjian Z (2010) Research on complex electronic equipment fault location based on improved genetic algorithm. In: Proceeding of 2010 International Conference on Computer Engineering and Technology, vol 1, pp 1454–1457
11. Yi Z, Yang X (2005) A fast genetie algorithm based on energy-entropy. Syst Eng Theory Pract 25:123–128
12. Huang Z, Yu Z, Li Z, Geng Y (2010) A fault diagnosis method of rolling bearing through wear particle and vibration analyses. Appl Mech Mater 26–28:676–681

Chapter 138
A Non-orthogonal and Multi-Width RBF Neural Network for Chaotic Time Series Prediction

Peng Zhou and Dehua Li

Abstract A non-orthogonal and multi-width learning algorithm of radial basis function (RBF) neural network is presented for chaotic time series prediction. It is based on an adaptive algorithm, which takes advantages of the good selection capability of the non-orthogonal method for assigning an appropriate number of hidden units for the network and the ability of the multi-width model for guaranteeing a natural overlap between kernel functions. The proposed algorithm may specify the locations and widths of kernels simultaneously. For known and unknown noise chaotic dynamical systems, the novel algorithm can predict them well and its effectiveness is illustrated by results from experimenting on some examples such as Honen chaotic time series.

Keywords Non-orthogonal · Multi-width · Radial basis function neural network · Chaotic time series

138.1 Introduction

There are many research results about chaos time series prediction due to their wide applicability in many practical systems such as secure communication, chemical reactions, biological systems and information processing. It can be see

P. Zhou (✉) · D. Li
Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China
e-mail: zhoupengsirok@163.com

D. Li
e-mail: lgf1007@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_138,

from them that the key problem of prediction lies in the formation of prediction models. Some classical models, such as auto regressive moving average (ARMA), recurrent neural networks (RNN), and support vector machine (SVM), have been developed for the prediction. Radial basis function (RBF) neural network have also been employed independently or as an auxiliary tool to predict chaotic time series. Chng et al. proposed a gradient RBF (GRBF) model for chaotic time series prediction, Rosipal et al. constructed a resource-allocating RBF network for chaotic time series prediction. Rojas et al. proposed a normalized PG-RBF network to analysis chaotic time series.

However, these algorithms intensively depend on the distribution of the input–output samples; they may break down or have improper number of centers when its corresponding parameters or thresholds can not be specified correctly. A significant contribution to construct RBF networks is made by Chen et al. through development of orthogonal subset selection algorithms. Although they all have good robustness, the Achilles heel is that the final model obtained by the forward selection method is not optimal [1]. Moreover, the default kernel width of orthogonal method is uniform.

In this letter, we proposed a new method of non-orthogonal and multi-width technique assisted by a differential evolution (DE) algorithm for chaotic time series prediction. The number of hidden units can be obtained by the forward and backward recursive method in the non-orthogonal space. The multi-width technique offers the advantage of taking the distribution variations of the different chaotic time series data into account and guarantees a natural overlap between Gaussian kernels. Thus we can obtain a unique and optimal prediction model of RBF neural network.

The rest of the Chapter is organized as follows. In Sect. 138.2, we consider the problem of chaotic time series prediction using an RBF prediction and effect of its model architecture and width on the prediction. Section 138.3 presents the non-orthogonal and multi-width method which is developed to construct an optimal RBF predictor. Section 138.4 describes the Henon mapping simulations carried out to evaluate the performance of the proposed technique.

138.2 Chaotic Time Series Prediction Using an RBF Network

According to the Takens [2] embedding theorem, we have

$$x_i = f(x_{i-1}, \ldots, x_{i-L}), \tag{138.1}$$

where $\{x_\mathrm{i}\}$ is a chaotic time series generated by a D-dimensional nonlinear system. Therefore, reconstructing a chaotic system from its time series measurements is specified by two ingredients. First, a suitable embedding dimension L may be determined by many techniques. Second, a good representation f makes chaotic time series predicting work well. A RBF neural network has a simple topological structure and university approximation ability and can be used as a predictor to model the unknown mapping f.

An RBF predictor, f_K, is a linear combination of K RBFs that is

$$\widehat{x}_i = f_K(x_{i-1}) = \sum_{j=1}^{K} w_j \phi(\|x_{i-1} - c_j\| / \sigma_j), \tag{138.2}$$

where $\phi(\|\boldsymbol{x}_{i-1}\text{-}\boldsymbol{c}_j\|/\sigma_j)$ is the RBF with center c_j and widthσ_j, $\|.\|$ denotes the Euclidean norm. To make f_K close to f, the distance between f and f_K is minimized with respect to the parameters of f_K, that is, $\boldsymbol{c}_j$s, σ_js and w_js. Here the coefficients w_js (j = 1,2,…,K) are chosen by the linear least squares (LS) or recursive least squares (RLS) method.

The prediction errors decrease monotonically with increasing number of hidden units. For a noisy chaotic time series, the RBF net does not require an infinite number of hidden units for an optimal prediction. The reason is that although using a large number of hidden units tends to make f_K close to f, it also results in a larger approximation error contributed by noise. Moreover, the chaotic time series is not uniformly distributed and must take the distribution variations of the data into account to obtain better prediction performance.

138.3 The Non-orthogonal and Multi-width Algorithm for Optimal RBF Prediction

A forward and backward recursive method in the non-orthogonal space is employed to realize an optimal RBF network, in which the multi-widths method is incorporated to consider the real distribution of chaotic time series fully.

138.3.1 Forward Selection

First, a forward regression method is considered here, given the noisy chaotic time series $\{y_i\}$, the first center of the RBF predictor can be selected as follows:

For 1 $\leq$ i $\leq$ M (M = 1, 2,…, N-L), compute

$$[Q]_i = ((q_i)^T Y)^2 / ((q_i)^T q_i), \tag{138.3}$$

where $Y = [y_{L+1},\ldots,y_N]_{\mathrm{T}}$, q_i is data matrix with $\phi(^{\|}\boldsymbol{y}_i\text{-}\boldsymbol{c}_j^{\|}/\sigma_j)$ as it ijth element for i = 1, 2,…, N-L and j = 1,2,…,N -L, N is number of points used for training. Note that since the value of $[Q]_{\mathrm{i}}$ represents the contribution of q_i to the network output, the corresponding vector q_i is selected when $[Q]_i$ takes the maximum value $[Q]_{i\max}(\max\{[Q]_i\}_{\mathrm{i}=1,\ldots,M})$. In a way, the $i$maxth x and q becomes the first selected center and regressor vector.

To obtain a good prediction performance of network, we must avoid redundant information as well as grasp necessary information of the hidden units. As a consequence, the concept of new information is proposed to achieve this target. Assume that the k regressor terms $\{q_k^{(1)}q_k^{(2)},.., q_k^{(k)}\}$ are added to the model and then let $R_k = [q_k^{(1)}q_k^{(2)},... q_k^{(k)}]$. The corresponding selected centers are $c_k^{(i)}$ $(x_k^{(1)}, x_k^{(2)},... x_k^{(k)})$. The remaining $M-k$ centers $c_{M\text{-}k}$ and regressor terms $R_{M\text{-}k}$ denoted by $\{x_{M-k}{}^{(i)}\}_{i=1,..,\mathrm{M-k}}$ and $\{q_{M\text{-}k}{}^{(i)}\}_{\mathrm{i=1,..,M-k}}$, respectively. To formulate forward selection iteration process, the projection of $q_{M\text{-}k}{}^{(i)}$ onto R_k is defined as

$$u_i = R_k(R_k^T R_k)^{-1} R_k^T q_{M-k^{(i)}}. \tag{138.4}$$

And then the new information expressed by the residual regressor $q_{M\text{-}k}{}^{(i)}$ can be written as

$$e_i = q_{M-k^{(i)}} - u_i. \tag{138.5}$$

Now, e_i spans a new non-orthogonal information subspace and the square modulus of projection of Y onto e_i is given by

$$[Q]_{k^{(i)}} = (e_i^T Y)^2 / (e_i^T e_i), i = 1, \ldots, M - k, \tag{138.6}$$

where $[Q]_k^{(i)}$ indicates the contribution of each residual regressor term $\{q_{M\text{-}k}{}^{(i)}\}_{\mathrm{i=1,..,M-k}}$ which is used by the network to approximate Y. The larger $[Q]_k^{(i)}$, the more information the residual regressor terms $\{q_{M-k}{}^{(\mathrm{i})}\}_{\mathrm{i=1,..,M-k}}$ contain. If one selected term achieves the maximum contribution among all remaining candidates and then the corresponding input vector is added to the model. If the maximum contribution $[Q]_k^{(i\max)}$ is defined as $\max\{[Q]_k^{(i)}\}_{\mathrm{i=1,\ldots,M-k}}$, then the $(k$+1)th center becomes the $i$maxth column of $\{x_{M-k}^{(\mathrm{i})}\}_{\mathrm{i=1,..,M-k}}$ and $q_{M-k}{}^{(i\max)}$ becomes the $(k$+1)th selected regressor vector. The selected regression matrix updates into R_{k+1}=$[R_k q_{M-k}{}^{(i\max)}]$. After the latest selected center which can mostly reduce the sum of square errors (SSE) is added to the model, whether or not the current model is to be expanded, must be determined by the cost function which can detect an excessive fit of the model to noisy or noiseless data. An RBF network with small number of basis functions yields a high bias and a low variance, whereas network with large number of basis functions yields a low bias but high variance estimator. So the Bayesian information criteria (BIC) [3] is an important instrument in balancing the accuracy and the complexity of the final network and can be used to detect the hidden unit number of RBF predictor. Suppose a subset of previously selected model contains k hidden units, the cost function BIC$_k$ is obtained by

$$\phi_k = 1/R_k^T R_k,\ e_k = Y - R_k(\phi_k(R_k^T Y)),\ S_k^T = e_k^T e_k/M,\ \mathrm{BIC}_k = M \ln S_k^2 + k \ln M. \tag{138.7}$$

Similarly, the cost function (BIC$_{k+1}$) of the (k+1)th intermediate model is acquired by applying R_{k+1} to (138.7). If BIC_{k+1}< BIC_k, then c_{k+1}=$[c_k x_{M-k}{}^{(i\max)}]$. $c_{M-(k+1)}$=$\{x_{M-k}{}^{(i\max)}\}_{\mathrm{i}=1,\ldots,(i\max\text{-}1),(i\max+1),\ldots,(\mathrm{M}-k)}$. The cost function is gradually reduced with increasing the model size. The forward selection procedure continues until the cost function satisfies the following condition $\mathrm{BIC}_{k+1} > BIC_k$.

In order to reduce the computational burden, a fast recursive algorithm is used to update the key step ϕ_{k+1} in the process of computing BIC_{k+1}. The augmentation ϕ_{k+1} of ϕ_k is implemented as follows [4]:

$$\alpha = \phi_k R_k^T q_{M-k^{(i\max)}},\ \bar{e} = q_{M-k^{(i\max)}} - \phi_k \alpha,$$
$$\beta = \bar{e}^T \bar{e}, \phi_{k+1} = \begin{bmatrix} \phi_k + \alpha\alpha^T/\beta & 38; -\alpha/\beta \\ -\alpha^T/\beta & 38; 1/\beta \end{bmatrix}. \tag{138.8}$$

138.3.2 Backward Refinement

If the RBF original basis is non-orthogonal, the energy contributions of different basis vectors are mixed. After one latest selected regressor term is added to model, some previously added units may consequently become very little to the network output. So the stepwise forward selection procedure performs a series of constrained optimization. Even though each step yields a regressor with the largest possible contribution, the final selected regressor terms are not the most compact for the hidden layer. To solve the problem resulted by the stepwise forward method, a backward refinement procedure is proposed. The backward refinement is to review the contribution of any selected regressor term (i.e., compare the contribution of any selected term with all the remains in the full selected terms) and then some selected regressor terms which are found non-contributing or less is removed. The process continues until no insignificant regressor term exists in the selected regressor terms.

To evaluate the contribution of each selected term $\{q_{k+1}{}^{(\mathrm{i})}\}_{\mathrm{i}=1,..,(k+1)}$, and locate the regressor vector of the minimum contribution to net output, first, construct a new matrix $R_{k,i}$ $(\{q_{k+1}{}^{(\mathrm{i})}\}_{\mathrm{i}=1,\ldots,(\mathrm{i}-1),(\mathrm{i}+1),\ldots,\mathrm{k}})$, which is equal to R_{k+1}without the ith term of it. And then the cost function ($\mathrm{BIC}_{k,i}$) of the hypothetical models without the ith term of the selected regressor terms is acquired by applying $R_{k,i}$ to (138.7). If $\mathrm{BIC}_{k,i\mathrm{min}}$ is the minimum value of $\mathrm{BIC}_{k,i}$ (i = 1,…,k+1) and satisfies the criterion $\mathrm{BIC}_{k,i\mathrm{min}} < \mathrm{BIC}_k$. Then the *i*minth selected term is removed from the selected regressor terms R_{k+1} and then put back into the candidate pool $R_{M-(k+1)}$. Similarly, the corresponding *i*minth selected center is also removed and put back into the candidate pool $c_{M-(k+1)}$. After the current round of iteration ends, let k = k-1, BIC_k= $\mathrm{BIC}_{k,i\mathrm{min}}$, $\phi_{k+1} = \phi_{k,i\mathrm{min}}$, R_{k+1}= $R_{k,i\mathrm{min}}$,, $c_{k+1} = c_{k,i\mathrm{min}}$.. The refinement iteration procedure is terminated and then the forward selection process restarts if $\mathrm{BIC}_{k,i\mathrm{min}} > \mathrm{BIC}_k$.

To speed up calculation of $\phi_{k,i}$, we can also use matrix transformation skills. The ith row and ith column of matrix ϕ_{k+1} move to the end and then new matrix $\phi_{k+1}{}'$ can be written as

$$\phi_{k+1}{}' = \begin{bmatrix} \phi & \gamma \\ \gamma^T & r \end{bmatrix}, \tag{138.9}$$

where ϕ is the $k \times k$ matrix, γ is k-dimensional vector, r is scalar. Then $\phi_{k,i}$ is given by

$$\phi_{k,i} = \phi - \gamma\gamma^T / r. \tag{138.10}$$

138.3.3 Multi-Width Method

Most real-life problems including the chaotic time series prediction show non-uniform data distributions. In the subsection, we give a multi-width method to solve the problem. A pilot density estimate is first computed as [5, 6]:

$$p(x_i) = \sum_{j \neq i} k((x_i - x_j)/\sigma_0)/n\sigma_0^d, \tag{138.11}$$

where σ_0 is a global width and the index d is the dimension of the data space $\{x_i\}_{\mathrm{i=1,\ldots,n}}$. The kernel, K, is taken to be a Gaussian function centered at zero and integrating to one. Based on Eq. 138.11 local widths are calculated as:

$$\sigma(x_i) = \sigma_0[\lambda/p(x_i)]^\theta, \tag{138.12}$$

where $p(x_i)$ is the estimated density at point x_i, θ is the sensitivity parameter, a number satisfying $0 \leq \theta \leq 1$ (a suggest value for θ is 1/2 [7]), λ is a proportionality constantn λ, which has an effect on the local width. If $p(x_i) < \lambda$, $\sigma(x_i)$ increases relative to σ_0 implying more smoothing for the point x_i, for data points that verify $p(x_i) > \lambda$, the local width becomes narrower. A good choice [7] is to take λ as the geometric mean of $\{p(x_i)\}_{\mathrm{i=1\ldots n}}$. It can be written as

$$\log \lambda = n^{-1} \sum \log(p(x_i)). \tag{138.13}$$

A particular choice of $\sigma(x_i)$ is able to perform much better than fixed-widths methods, as they offer a greater adaptability to the data.

To address the above mentioned problems of global width, this Chapter presents a new optimal scheme such that the global width, which is employed in the selection of the network architecture (numbers and selections of nodes) and the model parameters (centers and widths), can be evolved by the differential evolution (DE) algorithm based on an exhaustive search. In the Chapter, an individual is a global width σ_0. Like other evolutionary optimization algorithms, DE starts with an initial population and then utilizes crossover operation to increase diversity of the population. The minimal BIC, provided by the forward and backward recursive algorithm, is used as the fitness to measure the performance of individual in the population.

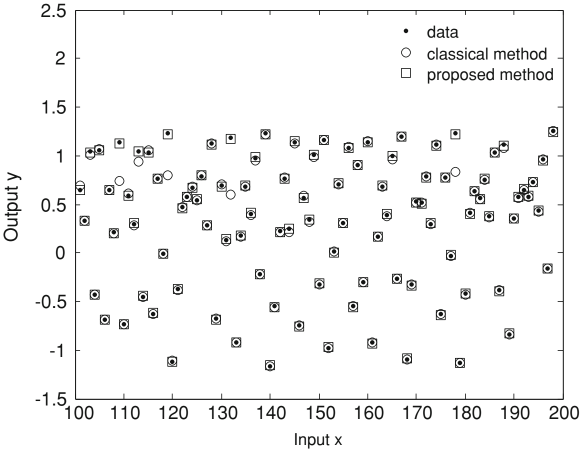


Fig. 138.1 Approximation of Henon mapping under no noise

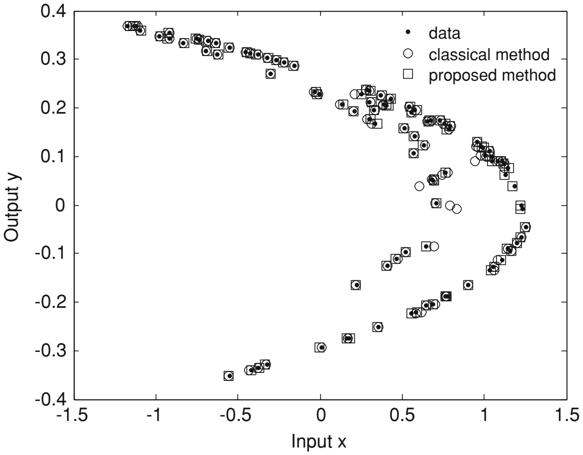


Fig. 138.2 Attractors of Henon mapping under no noise

138.4 Computer Simulation

In the following, Henon mapping [8] is used to test the above algorithm, which is defined as follows:

$$x(k+1) = 1 + y(k) - ax^2(k),\ y(k+1) = bx(k) \tag{138.14}$$

The resulting times series for a = 1.4 and b = 0.3 presents a chaotic behavior, and is recognized as a reference problem in the study of neural networks prediction ability. The task of the neural network is to predict the value of the times series at point $x(k)$, given n earlier points $x(k\text{-}n)$, $x(k-n+1)$, $x(k-n+2)$,…, $x(k-1)$. For all the experiments, the initial point is set at $x(0)$=0, $\mathbf{y}(0)$=0. A data set of noiseless

Fig. 138.3 Number of hidden units under different noise levels

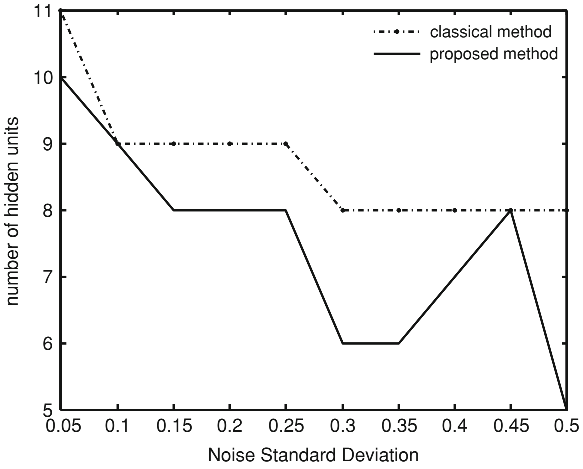
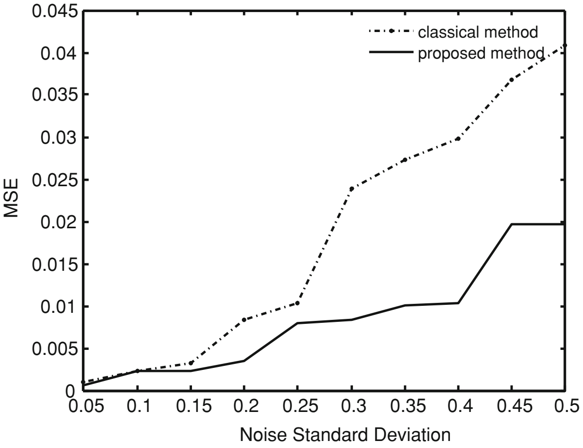


Fig. 138.4 Comparison of MSE under different noise levels



200 samples$\{x(k)\}_{k=1,\ldots,200}$ was obtained with the first 100 samples used as the training set and the last 100 samples as the validation set. The input vector to the Gaussian RBF predictor k (i.e., n=3) is $\boldsymbol{x}$=[x(k–3 x(k–2) x(k–1)]^{T}.

To test the performance of the new method clearly, we conducted a simple comparative analysis between the proposed method and the orthogonal and uniform width method (classical method). The RBF predictor constructed with the proposed method has 15 centers (the proposed method (□)), the predictor obtained with the classical method has 40 centers (the classical method (o)) under no noise. The predictor accuracies and error over the validation set was then computed and the results are plotted in Fig. 138.1. It can be seen from the figure that the proposed method achieves better prediction performance. From Fig. 138.2, it is clear that the proposed method produces a better correct attractor.

To analyze the effect of noise on the performance of the algorithms, the training sets have been corrupted by noise with standard deviation, which is uniformly sampled in the interval (0.05 0.5). Then both of them are applied to these data samples. It can be seen from the Figs. 138.3 and 138.4 that the proposed algorithm always realized the better prediction with smaller network for different noisy chaotic time series.

Acknowledgment The authors wish to thank the editor and anonymous reviews, and are also grateful to Hong Wu and Feng Cheng for writing.

References

1. Sherstinsky A, Picard RW (1996) On the efficiency of the orthogonal least squares training method for radial basis function networks. IEEE Trans Neural Netw 7(1):195–200
2. Takens F (1981) Detecting strange attractors in turbulence. Lect Notes in Math 898:366–381
3. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464
4. Li K, Peng J, Irwin GW (2005) A fast nonlinear model identification method. IEEE Trans Autom Control 50(8):1211–1216
5. Comaniciu D, Ramesh V, Meer P (2001) The variable bandwidth mean shift and data-driven scale selection. Proc Eighth Int Conf Comput Vis 1(1):438–445
6. Yang Z, Zhao Q, Liu W (2009) Energy based evolving mean shift algorithm for neural spike classification, 31st Annual international conference of the IEEE EMBS Minneapolis, Minnesota, September, vol 2(6), pp 966–969
7. Hall P, Hui TC, Marron JS (1995) Improved varible window kernel estimates of probability densities. Ann Stat 23(1):1–10
8. Henon M (1976) A two-dimensional mapping with a strange attractor. Commun Math Phys 50:69–77

Chapter 139
Study on Distributed Denial of Service Attack Detection Model Based on PCA and GA-Artificial Neural Network

Hai Yang

Abstract The Distributed Denial of Service (DDoS) attack is one of the most common intrusions in the network violence. The failure for the DDoS detection and prevention may cause terrible destruction and make huge loss to the network users. As a result, the early detection and protection for the DDoS is imperative for the internet and computer security. The intelligent artificial neural network (ANN) can provide effective DDoS detection performance and make it reliable for the efficient operation of the network. However, the DDoS detection precision is heavily influenced by the structure parameters of the ANN. Inadequate design of the ANN detection model may lead to a low detection rate. To overcome these problems, a new DDoS detection approach based on PCA, improved genetic algorithm (GA) and ANN is proposed in this paper. The Principle Component Analysis (PCA) was firstly used to obtain the most important characteristics of the attack data to sweep away the redundancies. Thus, the input of the ANN is concise enough. Then the improved GA based on the energy entropy selection was used to optimize the structure of the ANN, and efficient ANN detection model with proper structure was hence attained. The efficiency of the proposed method was tested with the practical data. The analysis results show that the proposed new method can give satisfactory DDoS detection rate, and outperforms the standard GA-ANN method with respect to the detection accuracy. Hence, the proposed new DDoS detection approach has application importance.

Keywords DDoS · PCA · Improved genetic algorithm · ANN

H. Yang (✉)
School of Computer and Information Engineering, Luoyang Institute
of Science and Technology, Luoyang 471023, China
e-mail: bingheku6@sina.com.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_139,

139.1 Introduction

Network security is an important research area in information security. Intrusion may lead to absence of the internet service and even cripple the whole system for weeks. As one of the most hard-handing attack mode, distributed denial of service attack (DDoS) is very difficult to prevent due to the occupied large network bandwidth and mainframe resources through illegal request. The normal operation of the internet is threatened terribly by DDoS. This is because DDoS is usually recessive, and has very long latency. The famous Yahoo and CNN have been aimed by DDoS attack and the sites are forced to shut down for weeks [1, 2]. Therefore, it is so important to establish reliable detection methods of rapid reaction to DDoS for normal running of the network [3].

Advanced detection algorithm, including evolution algorithm, intelligent ANN and support vector machine (SVM) and so on, are all used in the intrusion detection of the networks [4]. Among them, ANN [5] is the most promising method. However, ANN detection performance is largely determined by its structural design. It is particularly requied to determine the ANN parameters with a large number of trials. However, the urgent problem at present is that the invading data is very huge, and characteristics of the data make it impossible to train the ANN in an time-saving manner. In addition, there lies a large number of redundant feature among the characteristic space. Such as the US defense advanced research (DARPA) plan for intrusion detection system evaluation of the KDD dataset, each DDoS data contained 41 characteristics, but just 10 or less characters is useful to the DDoS detection. Therefore, ANN is difficult to learn adequate information about the DDoS data and often obtains unsatisfactory detection accurately [6]. In order to reduce the useless characteristics of the input data, principal component analysis (PCA) have been applied to dimension reduction for attack data in [7, 8], experimental results show that PCA can eliminate redundant features effectively, improve the detection rate and the ANN training efficiency. Thus, it is wise to use PCA to eliminate the inference of the DDoS data and enhance the ANN detection ability.

On the other hand, the ANN parameters are very sensitive for the DDoS detection. Although [6, 7] using GA to tune the ANN structures to improve the network attack detection accuracy, but without considering the GA individual selection adjustment. Therefore, there is important significance to improve the GA optimization process for the ANN based DDoS detection [9–11].

In order to tackle the problems mentioned above, a new DDoS detection method is proposed in this paper. The new detection method is based on the integration of PCA and ANN with the improved GA optimization. Through the practical experiments, the analysis results show that the proposed method can detect the DDoS efficiently and the detection rates is higher than the methods without PCA processing and GA optimization.

This work is organized as follows. The motivation of this work is introduced in Sect. 139.1. The proposed method for DDoS detection based on the combination of

PCA-GA-ANN is described in Sect. 139.2. In Sect. 139.3, the validation of the proposed method is discussed. The performance of the proposed new method is investigated. The effectiveness of the proposed method is valued by analyzing the practical data. Conclusions are drawn in Sect. 139.4.

139.2 The Proposed DDoS Detection Model

Due to the high dimension of the DDoS data, the inner connection of the features and the network operation states is very complex. Accurate detection is not easy to put forward. The high feature dimension of the DDoS data apparently not only gives detection system a huge amount of calculation but also decreases the detection rate. The reason is that there are certain dependent relationships among features, and some characteristics are redundant. Only a few features can depict the essential characteristic of attacking data. Fortunately, PCA is a powerful dimension reduction method. Therefore, the PCA has been adopted to improve the feature extraction ability of the original data. Meanwhile, ANN is an intelligent approach to deal with non-stationary signal. With its strong learning ability, the ANN is quite suitable for practical DDoS detection. However, its identification efficiency depends on the structure design. This is why the GA optimization is applied to the ANN design. By doing so, satisfactory ANN detection model can be obtained, and consequently the detection rate can be improved.

In this paper, the PCA is firstly employed to sweep away the redundant features of the sample data, and then ANN is used to learn the patterns of the data. Additionally, the improved GA is adopted to optimize the ANN model.

139.2.1 Principal Component Analysis

Given sample space $X = \{x_1, \ x_2, \ \ldots, \ x_n\}, x_k \in R^m$ is an m dimension column, n is total sample number. Suppose the linear transform for X is [12]

$$F_i = a_i^T X = a_{1i}X_1 + a_{2i}X_2, \cdots + a_{ni}X_n, \ i = (1, \ 2, \cdots n) \tag{139.1}$$

Then the covariance matrix for F is that

$$\bar{C} = a_i^T \sum a_j \, (i, j = 1, \ 2, \ldots n) \tag{139.2}$$

According to $\lambda V = \bar{C}V$, it can calculate the eigenvalue λ and eigenvector V for Eq. 139.2, so

$$V = \sum\nolimits_{i=1}^{n} \alpha_i x_i, \tag{139.3}$$

$$\lambda = x_k \cdot \bar{C}V \tag{139.4}$$

Arrange eigenvalue λ in the descending order, then the high dimension space X can be transformed in a linear space Y:

$$Y = V^T X, \tag{139.5}$$

According to the 85% criteria, select the first p components ($p < m$) in Y as principal components, thus realize the dimension reduction for data X.

139.2.2 Improved GA

Standard GA [13] involves the following procedures: coding, selection, crossover and mutation. In the selection, standard GA only searches individuals with adaptive value. The diversity of the population is constrained, which may lead to premature convergence of the GA optimization [14]. To deal with this situation, the energy-entropy-based selection is employed to increase the diversity of population. As GA selects according to the individuals' energy, it needs to calculate each individual z. The individual energy entropy can be expressed as

$$E(t) = \mathrm{In}\left(\frac{1}{p_t}\right), \tag{139.6}$$

where, p_t denotes energy probability at lever t. By the annealing selection [10], we can derive the individual selection probability [15]

$$P(z_i) = \frac{e^{-(E(z_i)+\beta f(z_i))}}{\sum_{i=1}^{n} e^{-(E(z_i)+\beta f(z_i))}}, \tag{139.7}$$

where, $P(z_i)$ denotes the probability of individual z in new population and $E(z_i) + \beta f(z_i)$ denotes fitness value of individual z.

After the energy-entropy-based individual selection procedure, the link between individuals is connected and to maintain the diversity of population.

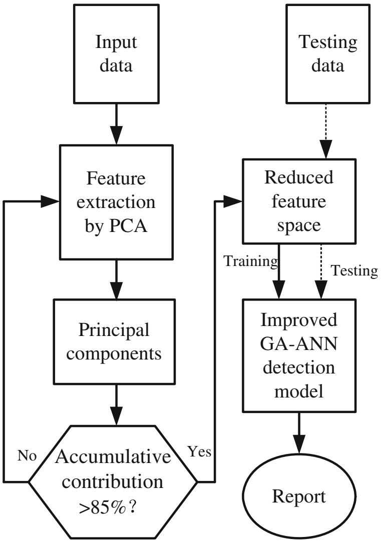
139.2.3 Artificial Neural Network Detection Model

Since the ANN provides powerful learning ability, the RBF neural network is used to detect the DDoS in this paper. The detailed theory of RBF NN is present in Ref. [2]. The DDoS detection model based on PCA and GA-SVM is shown in Fig. 139.1.

The processing steps are as follows:

Step 1: pretreatment input DDoS data, uniform the data format.
Step 2: extract principal components of the input data by PCA.

Fig. 139.1 The DDoS detection model based on PCA and GA-ANN



Step 3: train ANN using the reduced feature space, while optimizing ANN parameters via improved GA.

Step 4: test DDoS detection capability of the proposed model.

139.3 Experimental Analysis

In order to validate the performance of the proposed algorithm, the intrusion experiments were carried out in real practice application in this paper. The recorded DDoS data describes 30 main attribute of the test network connection, including duration, service type, the bytes issued from source to destination, the bytes from destination to source, etc. In the DDoS detection, 5,000 normal samples and 5,000 DDoS samples were investigated.

In experiments, PCA was adopted to reduce the 30 dimension of the original data to 3, 10 and 15 principal components, respectively. The contributions of different numbers of principal component are shown in Fig. 139.2.

The improved GA-ANN is employed to identify the DDoS. The input feature vector of the ANN is 3, 10 and 15 principal components, respectively. In the GA optimization, the population size is 300, the crossover probability is 0.95 and the mutation probability is 0.01.

The GA optimization procedure for the ANN with ten principal components is illustrated in Fig. 139.3. The evolution performance of the improved GA and standard GA is compared. It can be seen in Fig. 139.3 that the improved GA is faster and better than standard GA.

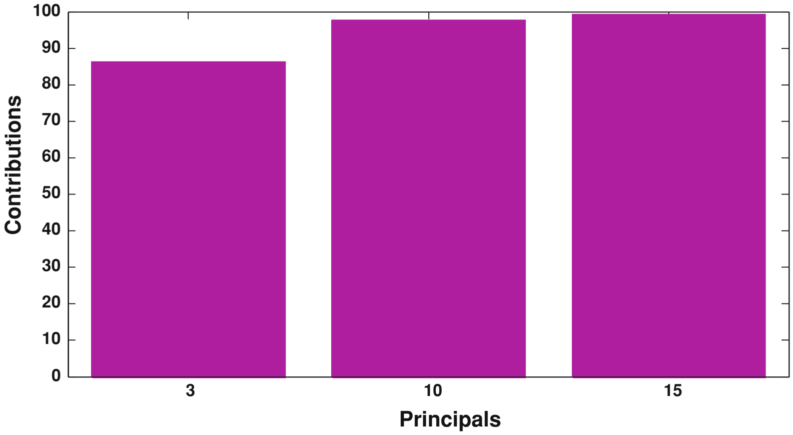


Fig. 139.2 The contribution rate in different number principal components

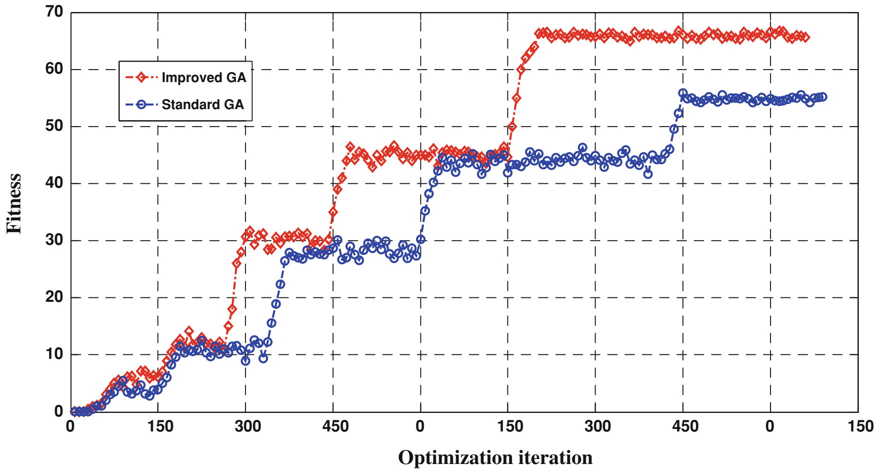


Fig. 139.3 The performance of improved GA optimization and standard GA

The DDoS detection performance of improved GA-ANN model with different principal components is shown in Table 139.1. From Table 139.1, the detection model with ten principal components can provide better performance than the 3 principal components and 15 principal components. Additionally, the number of the principal component plays an important role in the detection rate. Improper input characteristics can drop the detection rate significantly. Hence, it should be comprehensive consideration of speed and the detection rate for the system to determine the reasonable reduced-dimension number. For this experiment result, choosing 10 principal components can obtain good result, and the detection rate is up to 97.5%.

Table 139.1 The DDoS detection performance of the proposed model

Principal components	Improved GA-ANN model	
	Detection accuracy (%)	False alarm (%)
3	77.1	5.61
10	97.5	0.82
15	95.7	1.02

Table 139.2 The DDoS detection of improved GA–ANN and standard GA–ANN model

Standard GA–ANN model			Improved GA–ANN model		
Detection rate	False alarm	Time (s)	Detection rate	False alarm	Time (s)
95.1%	1.53%	1.57	97.5%	0.82%	1.19

The intrusion detection performance of improved GA-ANN model and standard GA-ANN model is shown in Table 139.2. The inputs of the models are 10 principal components. From Table 139.2, the proposed method for intrusion detection performance is better than the standard GA-ANN. The detection rate is enhanced by 2.4%, as well as the false alarm rate by 0.34%. The comparison results show that the improved GA is more effective in the ANN optimization than the standard GA.

139.4 Conclusions

Intelligent method is useful for DDoS detection; however, its structure is very sensitive play for the detection of performance. Hence, a new DDoS detection approach based on PCA and improved GA-ANN is proposed in this paper. One advantage of the new method is that the use of PCA can extract most distinguished characteristics of the DDoS data. The other advantage is the use of energy-entropy-based selection to increase the diversity of the generations in the GA optimization. The analysis of practice DDoS data was implemented. The analysis results show the effectiveness of this new method via high detection accuracy. Thus, the proposed method is reliable for the DDoS detection.

References

1. Zhao X, Jing R, Gu M (2008) Adaptive intrusion detection algorithm based on rough sets. J T singhua Univ (Sci Tech) 48:1165–1168
2. Vapnik V (1995) The nature of statistical learning theory, 1st edn. Springer, Berlin
3. Ren X, Wang R, Kong Q (2009) Principal component analysis and support vector machine based anomaly detection. Appl Res Comput 26:25–27

4. Gu Y, Zhou B, Zhao J (2008) PCA-ICA ensembled intrusion detection system by Pareto-Optimal optimization. Inf Technol J 7:510–515
5. Li Z, Yan X, Yuan C, Zhao J, Peng Z (2011) Fault detection and diagnosis of the gearbox in marine propulsion system based on bispectrum analysis and artificial neural networks. J Marine Sci Appl 10:17–24
6. Qiao L, Peng X, Ma Y (2006) GA-SVM Based feature subset selection algorithm. J Electr Meas Instrum 20:1–5
7. Zhiqiang X (2010) Support vector machines based on genetic algorithm for network intrusion prediction. J Comp Simul 27:110–113
8. Xiong W, Wang C (2009) Hybrid feature transformation based on modified particle swarm optimization and support vector machine. J Beijing Uni Posts Telecommun 32:24–28
9. Li Z, Yan X, Yuan C, Li L (2010) Gear multi-faults diagnosis of a rotating machinery based on independent component analysis and fuzzy k-nearest neighbor. Adv Mater Res 108–111:1033–1038
10. Li Z, Yan X (2011) Application of independent component analysis and manifold learning in fault diagnosis for VSC-HVDC systems. Hsi-An Chiao Tung Ta Hsueh 45:46–51
11. Ahn CW, Ramakrishna R (2002) A genetic algorithm for shortest path routing problem and the sizing of populations. IEEE Trans Evol Comput 6:566–579
12. Li Z, Yan X, Yuan C, Peng Z, Li L (2011) Virtual prototype and experimental research on gear multi-fault diagnosis using wavelet-autoregressive model and principal component analysis method. mech syst signal process. doi: 10.1016/j.ymssp.2011.02.017
13. Fu X, Zheng Z (2010) Research on complex electronic equipment fault location based on improved genetic algorithm. In: Proceeding of 2010 international conference on computer engineering and technology, vol 1 pp 1454–1457
14. Zhang Y, Xiuxia Y (2005) A fast genetie algorithm based on energy-entropy. Syst Eng Theory Pract 25:123–128
15. Zhongyu H, Zhiqiang Y, Zhixiong L, Geng Y (2010) A fault diagnosis method of rolling bearing through wear particle and vibration analyses. Appl Mech Mater 26–28:676–681

Chapter 140
A New UCON Model for Web Industrial Control

Yanpei Liu, Yuesheng Gu and Junhui Fu

Abstract Based on the control core model UCONABC and traditional access control model TRBAC, it introduces the concept and the mission role elements to the characteristics of industrial control independency. A new access control model named UCONTR model, which is applicable to Web industrial control, is proposed in this paper, then the advantages of the application of this model in Web industrial control environment are compared. Finally, this model in distributed environment of multiple domain implement frame is given.

Keywords Usage control · TRBAC · UCONTR · Industrial control · Access control

140.1 Introduction

UCONABC model is the core concept of UCON model [1, 2]. The model is comprehensive authorization rules, obligations and conditions of three decision factors, and proposed access control continuity and variability of the two properties, greatly enriched and improved access control model of the content and means, fully embodies the basic thoughts of the use of control, However, there are still many shortcomings of the model and needs further improvement and perfection, the most prominent being: (1) UCONABC model for the modern access control research provides unified framework, using a highly abstract definition, but in practice the

Y. Liu · Y. Gu (✉) · J. Fu
Henan Institute of Science and Technology,
xinxiang 453003, China
e-mail: 32933415@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_140,

elements of the model needs further refinement, such as attribute definitions, volatility restrictions and dynamic separation of duty. (2) Concurrency issues; In UCONABC model, attribute updates will affect other access authorized in decision-making, In order to ensure effectiveness, attribute updates and the authorization visit must be concurrent, this greatly increases the difficulty of the model. (3) unified management model; as the model definition is highly abstract, for access control, trust management and DRM providing a unified management model is very difficult. (4) The right of the authorization issues. The UCONABC model did not involve entrusting authorized problem, but to entrust authorization in practical applications exists generally, take entrust authorized introducing UCON is essential for the application of modern access control requirements, has a very important practical significance.

According to Web environment industrial control independent characteristics, the UCONABC model based on the combination of TRBAC model [3], in elements and its definition form and other aspects of the expansion, put forward a kind that can adapt to industrial control [4–6] access control model—UCONTR model. This paper from the basic idea, formalized definitions, model analysis and distributed environment to implement frame of this model analysis design.

140.2 The Basic Thought of UCONTR Model

The basic thought of UCONTR model summed up two points. Figure 140.1 shows UCONTR model structure. On one hand is attribute management, including attribute static management and dynamic management. Attributes of the static management is the constant subject and object attribute management, by security administrator access object before the subject. Such as: role—field distribution, users—unchanged character distribution, the same role—unchanged task allocation and personal basic information management etc. Attributes of dynamic management of subject and object has variable attributes to carry on the management, this part of the system control, without need for administrator operater such as: users—variable role distribution, variable role—variable task allocation, task state management etc. On the other hand is the use of user control. Users in use of certain permissions by session tried the process, using decision continuously will authorize, obligations and conditions as decision factors, judge whether the special right to access object.

140.3 The Formal Description of UCONTR Model

UCONTR model shown in Fig. 140.2, it is defined by the following components:

Definition 1 The core elements of the UCONTR include: UCONTR = {(core element, authorization function) | core element ∈{Subjects (S), Subject Attributes

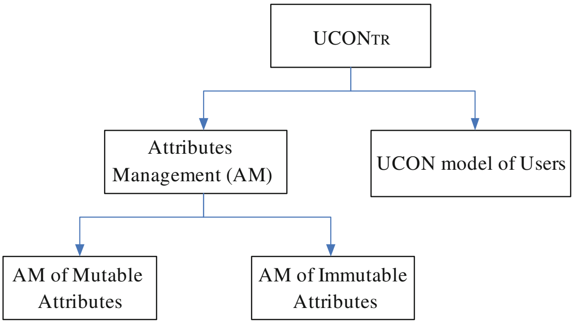


Fig. 140.1 UCONTR model structure

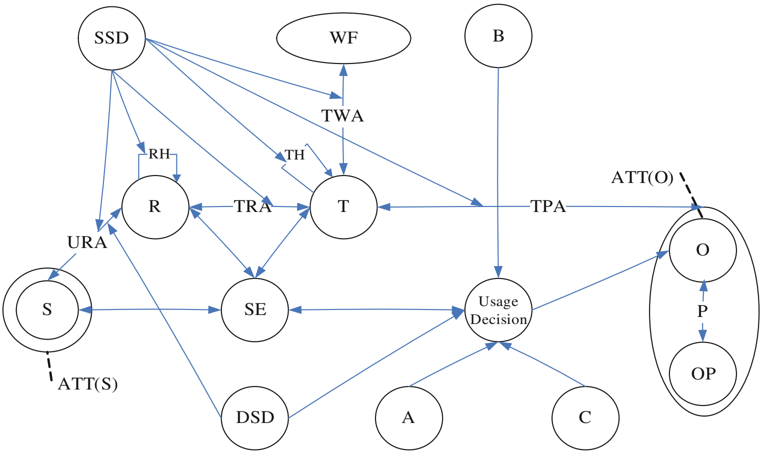


Fig. 140.2 UCONTR model

(ATT(S)), Objects (O), Object Attributes (ATT(O)), Roles (R), Sessions (SE), Permissions (P), Operations (OP), Tasks (T), Work Flows (WF), Static Separation of Duty Relations (SSD), Dynamic Separation of Duty Relations (DSD), Authorizations (A), Obligations (B), Conditions (C), Usage Decisions}, authorization function ∈ {URA, TRA, TWA, TPA, RH, TH, UD}}

Definition 2 Principal attribute

ATT(S) = {RU, Act_RU, DU, AllowedT, StartT, I},

where, RU is the user's role attribute U, Act_RU the user during a session in a role is activated, DU is the main domain property, AllowedT Act_RU the user assumes the role of the time interval, StartT is to visit the starting time, I is the main general attributes, such as: basic user information.

For multi-domain, distributed modern information system, the user is dynamic, the variability of its properties may lead to changes in the role, so that access by the administrator in the manner authorized role before there is a big limitation. UCONTR role model attributes the main role and can not be changed into the role of the variable, the same role in the main by the administrator to give access to the object before the subject, while the variable main role of property by the system dynamics under the given subject.

Definition 3 Objective attribute

ATT (O) = {T, DO, T_status}

Among them, T is the task, DO is the set of domain properties of the object, a task corresponds to a set of domains, which provides that only the group domain users to perform the task, T_status is a task in which the state has created state, activated state, effective state, hung state and failure state. Task state is constantly changing in the entire work task flow, a task state changes directly affect the implementation of other workflow tasks. Thus, the user permissions awarded and task state changes have direct connection.

Definition 4 Task

$T = \{t \mid t \in ATT\ (O) \wedge (t \in TWF \vee t \in TUWF)\}$,

where, TWF says workflow task, TUWF says non-workflow task. Task is defined as the object attribute, Divided into workflow task and no workflow task. Workflow task for handling the workflow affairs, non-workflow task for handling independent affairs, such as users of the basic information of the management.

Definition 5 Constraint relationship

(1) the inheritance

①Task inherited:

$TH \subseteq T \times T, \forall\ t1 \in T, t2 \in T, (t1, t2) \in TH \Rightarrow$ t1 inherit t2

②Role inheriting:

$RH \subseteq R \times R, \forall\ r1 \in R, r2 \in R, (r1, r2) \in RH \Rightarrow$ r1 inherit r2

(2) SSD:

Static mutexes defines three mutexes forms, including permissions mutexes SMP, task mutexes SMT and role mutexes SMR, defined as follows:

SSD = {SMP, SMT, SMR}

Permissions mutexes:

$SMP \subseteq P \times P, \forall\ p1 \in P, p2 \in P, (p1, p2) \in SMP \Rightarrow GetT\ (p1) \cap GetT(p2) = \varnothing$

Task mutexes:

$SMT \subseteq T \times T, \forall\ t1 \in T, t2 \in T, (t1, t2) \in SMT \Rightarrow GetR\ (t1) \cap GetR(t2) = \varnothing$

Role mutexes:

$SMT \subseteq R \times R, \forall\ r1 \in R, r2 \in R, (r1, r2) \in SMR \Rightarrow GetS(r1) \cap GetS(r2) = \varnothing$

GetT (p): p → 2T, Return all permissions p attachment task, GetR (t): t → 2R, Return all the role, with task t, GetS (r): r → 2S, Return all the subject to assume the role of r. The task of containing mutexes permissions are mutually exclusive tasks, the role of containing mutexes tasks are mutually exclusive role.

(3) DSD:

$DSD \subseteq R \times R$

$\forall se \in SE, \exists r1 \in session\_R(se), \exists r2 \in session\_R(se) \Rightarrow (r1, r2) \notin DSD$,

where session_R(se): se 2R, return session se process activated role. Main role in a dynamic mutexes provisions in the process cannot activate session of a dynamic mutexes relationship between the two characters.

Definition 6 User assigned

(1) The role based assigned

$UAUR = \{(ui, ri, uj, rj, di, tspan) \mid ui, uj \in S, ri \in R, di \in DU\}$,

In the domain of di, users can own role ri will be assigned to a user uj, Uj can perform the task of role ri in tspan time range. However, should pay attention that given to the role of the ri cannot with user uj role concentrated any role rj mutexes, this ensures the safety of user appointed strategy.

(2) Based on task assigned

$UAUT = \{(ui, ri, ti, uj, rj, tj, du, tspan) \mid ui, uj \in S, ri, rj \in R, di \in DU\}$,

In the domain of di, users can own role of the task in ri will be assigned to a user uj, with use period tspan, delegate tasks at the same time, ri also delegated, and ri is different from its own role. However, User uj can only carry out the task ti which belong role ri, and task ti cannot with user uj task concentrated any task tj mutexes.

The difference based on the role of the delegate and on task of the delegate are delegated different particle size. Based on the role of the delegate roles have all tasks assigned to another user, based on task assigned will only a single task assigned to another user. Whether based on the role of the delegate or based on task of the delegate, assigned subject and object must belong to the same volume, cross-realm cannot be user assigned.

Definition 7 Authorized rules define

(1) The unchanged attribute management authorization rules defined

① The authorization of constant role attribute management rules defined

$Authorize\_R(sm, r) \Rightarrow Act\_R(sm) \neq \emptyset \wedge Act\_R(sm) \in Authorized\_R(sm) \wedge r \in Manage\_R(Act\_R(sm))$.

$Allowed\_URA(u, r) \Rightarrow authorization =$

(A) $Get\_SSD(r) \cap Authorized\_roles(u) = \emptyset$

(B) $Descent(r) \cap Authorized\_roles(u) = \emptyset$

Get_SSD(r): r → SSD(r), the role of SSD set back. Descent(r): r → 2R, returns the role subset r. Authorized_roles(u): u → 2R: return to the role of the customer sets. User role assigned and undo the subject must activate a security administrator role, and the corresponding characters must assign process in the security administrator is within the scope of jurisdiction to assigned process must satisfy the static mutexes and inherited role constraints.

② the authorization of constant task attribute management rules defined

Authorize_T(sm, r, t) $\Rightarrow$ Act_R(sm) $\neq \varnothing \wedge$ Act _R(sm) $\in$ Authorized_R(sm) $\wedge$ r $\in$ Manage_R(Act_R(sm)) $\wedge$ t $\in$ Manage_T(Act_R(sm)).

Allowed_TRA(r, t) $\Rightarrow$ authorization =

(A) Get_SSD(t) $\cap$ Authorized_tasks(r) = $\varnothing$

(B) Descent(t) $\cap$ Authorized_tasks(r) = $\varnothing$

Get_SSD(t):r → SSD(t), the task of SSD set back. Descent(t): t → 2R, Returns the task subset r. Authorized_tasks(r): r → 2T: returns the task set role. The difference between the rules of the definition and definition ① is management of the object is unchanged task.

③ rights management TPA authorized rule defined

Authorize_P(sm, t, p) $\Rightarrow$ Act_R(sm) $\neq \varnothing \wedge$ Act _R(sm) $\in$ Authorized_R(sm) $\wedge$ t $\in$ Manage_T(Act_R(sm)) $\wedge$ p $\in$ Manage_P(Act_R(sm)).

Allowed_TPA(t, p) $\Rightarrow$ authorization = Get_SSD(p) $\cap$ Authorized_powers(t) = $\varnothing$

Get_SSD(p): p → SSD(p), the rights of SSD set back. Authorized_powers(t): t → 2P, permission to return to the task contains, On the subject of task system permissions to activate a security administrator role, and assigned the task of process corresponding to the must be in and authority within the jurisdiction of security administrator can be. Assigned process mutex constraint must meet permissions static.

(2) Variable attributes definition administrative authority

① the authorization of variable role attribute management rules defined

Allowed_DURA(u, se, r) $\Rightarrow$ authorization =

(A) The same to Allowed_URA

(B) Get_DSD(r) $\cap$ Authorized_roles(u) = $\varnothing$

Get_DSD(r): r → DSD(r), the role of DSD set back.

②the authorization of variable task attribute management rules defined

Allowed_DTRA(r, se, t) $\Rightarrow$ authorization =

(A) The same to Allowed_TRA

(B) Get_DSD(t) $\cap$ Authorized_tasks(r) = $\varnothing$

Get_DSD(t): t → DSD(t), the task of DSD set back.

① and ② that the management not only variable attributes to satisfy unchanged attribute authorized, still need to meet the rules of dynamic mutexes constraint.

Users in the session se, because of the change of variable attributes the role of add and withdrawn is a system of automatic control, no need to administrator executive management behavior. Variable task as well. Variable task t is the same.

140.4 UCONTR Model Analysis

140.4.1 Retained UCON and TRBAC Respective Advantages

Due to the introduction of the concept of roles and tasks, the traditional access control model special TRABC model more convenient, flexible and easy to manage, have diversified accredit way, and UCONTR model inherited TRBAC model of the concept of roles and tasks, retention the TRBAC application in the workflow advantages of affairs. UCONTR model will be the concept of roles associated with the domain, very good place in distributed more show model for users under the environment domain authorized control; UCONTR model lasted use the control principal-objective attribute of variability and the continuity of decision-making authority, actively industrial control processes adapted to the Web and other elements in the task state changing dynamic characteristics, and to authorize the continuity of decision-making process to achieve a dynamic industrial control authority to realize the principle of minimum authorized.

140.4.2 Solve the Model Attributes Such as Role and Task Management

UCONTR model updates and continuous variable property authorized by the monitor automatically, conditions and obligations belong to system external demand, also do not need security administrator for the management, security administrator management objects including subject, object, power limits and authorization rules. This is the same as traditional access control models. In the traditional workflow access control model, same as the principal-objective, role and tasks are defined as independent entities, there are no application problems in a closed environment, but in a distributed, highly dynamic information environment, the apparent lack of flexibility, such as: the multi-domain environment, when a user is removed from the field, the system must repeal this user in the role of the domain, and then transferred to the role of domain re-distribution, that leads to the heavy workload of security administrators. To solve this problem, in UCONTR model, the role is defined as the main property, to achieve inter-domain access. At the same time makes the task of the state property of the task easy to monitor the variability of the dynamic and authorized to achieve continuity. Roles and tasks attributed to enrich the definition of the use of controlled content, is defined using the control mode of highly abstract concrete.

140.4.3 Enrich User Assigned Way in Distributed Workflow Environment

Users assigned that users have the ability to give you permission to others who meet system condition in a short time. UCONTR model based on the definition in characters and based on task two levels of user assigned, the former assigned object is role, the latter's delegate object is mission, rich delegate authorized way, enhance the user's flexibility in appointed, reduced the security administrator management burden, thus further enhance the authorized the dynamics and flexibility. Of course, users are assigned still by the system of strict control, such as the SSD constraints, DSD constraints and domain constraint, etc. So UCONTR model overcame independent access control strategy of runaway happening jurisdiction.

140.5 The UCONTR Model Frameworks in Distributed Environment

Figure 140.3 shows the UCONTR in a distributed multi-domain environment, the implementation of the overall framework.

Domain quoted monitor is within the domain of realizing access control core part, by using the decision-making facilities UDF and use executive facilities UEF composition, UEF issued intercepted on the main access to the object of every request, and will ask to UDF on access control decision-making.

Domain license database stores access control information, includes user, resources, operation type, domain information, etc. Domain quoted monitors inquires, after authorized database. Authorized the database will query information back to the reference monitors, according to the information generated reference monitors access control rules.

Inter-domain authentication among multiple domains distributed authentication, to ensure that users from outside the domain of attribute information provided, the role of information is a safe and reliable.

Role mapping device mainly provides cross-realm visit the operation, Establish the role to local role in outlands map. According to the system's security strategy, guarantee in role mapping process before implementing a compulsory subject must fulfil obligation components requirements. In mapping, the implementation process must satisfy the system components security needs authorization.

The central database records workflow executing events, including task state, main body role of missions, time and workflow current flows through steps and so on.

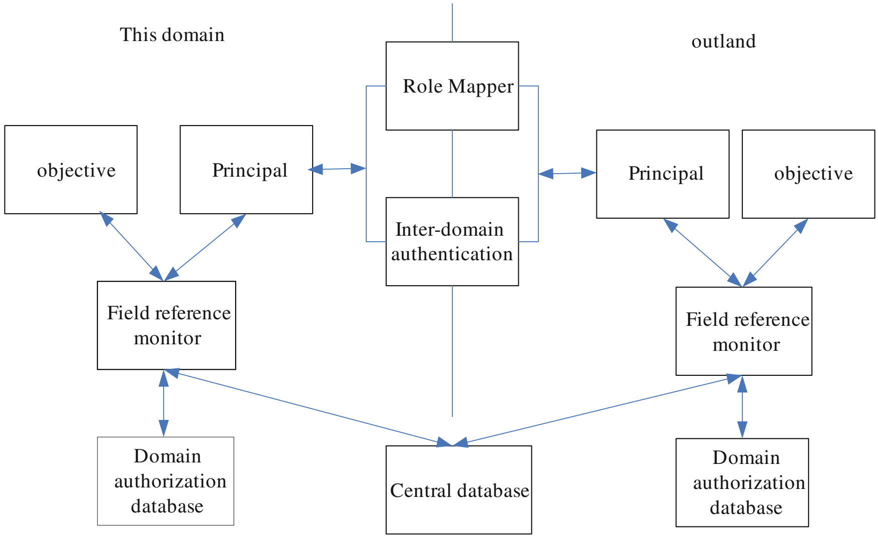


Fig. 140.3 UCONTR model of the overall framework in a distributed environment

140.6 Conclusion

On the basis of the use of UCONABC control model, according to Web environment industrial control independent characteristics, introducing the role concept and modify the definitions role way. Introduction of the task elements, defining the role and task of two permissions assigned task mechanism in a rich delegate way. Also, the analysis and comparison of the model in Web applications, industrial control environment benefits. The model is presented at last in distributed multiple domain environment of the enabling framework.

References

1. Min X, Xuxian J, Sandhu R (2008) Towards a VMM-based usage control framework for OS kernel integrity protection. In: proceedings of ACM symposium on access control models and technologies, Sophia antipolis, France, pp 71–80
2. Park J, Sandhu R (2004) The UCONABC usage control model. ACM Trans Inf Syst Secur 7(1):128–174
3. Hou S, Zng J, Lu X (2008) Research of TRBAC model and its application in enterprise MIS. Comp Eng Des 3106–3108
4. Hu Z, Jin R,Yu W (2007) Survey on attribute mutability of usage control. Comp Eng Appl 66–68
5. Zhang X, Park J, Parisi-Presicce F, Sandhu R (2004) A logical specification for usage control. ACM Press, New York, 2–4 June
6. Peng L, Yang P, Peng Y (2007) Survey of usage access control model. Appl Res Comput 24(9):121–123

Chapter 141
Research of Vertical Handover Algorithm in WLAN and UMTS Heterogeneous Network

Xiaobin Li, Ziwen Guo and Juan Peng

Abstract Aiming at addressing the problems in vertical handover in heterogeneous network, this chapter presents a new vertical handover algorithm based on policy-enabled. It uses a decision mechanism based on the following metrics: the received signal strength (RSS), cost, the available bandwidth, and the velocity of mobile node without adding the computational complexity on mobile node. Simulation results show that in comparison to the traditional vertical handover algorithm based on RSS, the proposed algorithm can drastically decrease the times of handover and the packet loss rate (PLR), and then improves the success ratio of vertical handover and the user's degree of satisfaction.

Keywords Heterogeneous network · Vertical handover · Policy-enabled · Seamless handover

141.1 Introduction

With the fast development in communication technology, there are more and more access networks, especially wireless network, which leads to much more areas have overlapped wireless network, it is called heterogeneous network. It is worth noting that many different networks have complementary features, such like Wi-Fi and WCDMA, Wi-Fi supports short-distance but high speed access services, meanwhile WCDMA supports long-distance but low speed access services.

X. Li (✉) · Z. Guo · J. Peng
College of Information Engineering, Shenzhen University,
Shenzhen 518060, China
e-mail: lixbsz@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_141,

Besides nowadays in life there are more and more mobile devices that support multi-interface which means it has only one way to access the network, take 3G Smartphone for instance it could access network either by Wi-Fi interface or one of the 3G technology [1]. But the situation here is the user could only change its network of attachment manually and the service such as streaming media has to be reset or disconnected. So the vertical handover technique has become one of key techniques in the future wireless network.

In order to achieve the seamless handover, people first follow the principles in horizontal handover algorithm which is based on RSS [2, 3], but this algorithm has serious ping-pong effect and high latency. Eastwood and Migaldi [4] proposes a new algorithm based on RSS and Hysteresis value, which mitigate the ping-pong effect but increase the latency and reduce the handover efficiency. Qingyang and JamaliPour [5] presents a novel algorithm based on analytic hierarchy process (AHP). After the complex configuration, it could mitigate the ping-pong effect and reduce the latency at the same time, but it needs to input lots of parameters to complete complex computing, therefore it is not suitable for mobile devices with low computing capability and limited power. This Chapter proposes a new vertical handover algorithm based on policy-enabled which avoiding the complex computing but still could assure the low latency and high efficiency in handover.

The rest of this Chapter is organized as follows. Section 141.2 introduces some details of existent vertical handover algorithms. Section 141.3 presents our novel contributions. Section 141.4 provides a performance evaluation of the described algorithm and finally Sect. 141.5 concludes this work.

141.2 Related Work

In this Chapter, we present three different existent algorithms, first introducing their basic principles, and then elucidating their respective advantages and disadvantages.

141.2.1 Traditional Algorithm Based on RSS

Basically, the main principle of the traditional vertical handover algorithm derives from horizontal handover algorithm [6]. The main idea is to set an appropriate threshold value $\mathrm{RSS}_{\mathrm{threshold}}$ for each candidate access network (CAN), respectively. If a CAN has a higher RSS value than it's threshold, it will be considered as preferred CAN, contrarily if a node attached network (NAN) has a lower RSS value than its threshold, the mobile node will begin to search for preferred CAN, and then attach to it, but if no preferred CAN found, the deteriorated signal will disconnect the node from its NAN.

This algorithm is simple and efficient, but it can not effectively avoid ping-pong effect, and it has high latency and high PLR.

141.2.2 Algorithm Based on RSS and Hysteresis Value

Algorithm based on RSS and Hysteresis value [7] mitigate the ping-pong effect effectively. Its main principle is to add a Hysteresis value on traditional algorithm, and then if a CAN is considered as a preferred CAN not only in the condition of its RSS value higher than threshold but also in the condition of higher than the $\mathrm{RSS}_{\mathrm{threshold}}$ + Hysteresis Value.

Besides, this algorithm also uses another parameter called dwelling time, by this way mobile node has to wait a period (dwelling time) before it begins to attach to preferred CAN, which avoids the ping-pong effect further.

This algorithm mitigates the ping-pong effect very effectively but increases the latency and reduces the handover efficiency.

141.2.3 Algorithm Based on AHP

AHP is a structured technique for dealing with complex decisions. It was developed by Thomas L. Saaty in the 1970s. It provides a comprehensive and rational framework for structuring a decision problem, for representing and quantifying its elements, for relating those elements to overall goals, and for evaluating alternative solutions. In the Algorithm based on AHP [8], first it needs to decompose the decision problem into a hierarchy of more easily comprehended sub-problems, each of which can be analyzed independently. A numerical weight should input for each element of the hierarchy, allowing diverse and often incommensurable elements to be compared to one another in a rational and consistent way. In the final step of the process, numerical priorities are calculated for each of the CAN, and then output the optimum CAN.

This algorithm could quantify the needs of users in detail and according to which find out the optimum CAN, but it needs to input lots of parameters to complete complex computing, therefore it is not suitable for mobile devices with low computing capability and limited power.

141.3 Vertical Handover Algorithm Based on Policy-enabled

This Chapter includes two parts, first part introduces the main principle of the algorithm, second part illustrates the functions used in the algorithm and the process flow of choosing an optimum CAN.

141.3.1 Main Principle of the Algorithm

The process of vertical handover algorithm includes three stages; first stage is to trigger the handover process. It is worth noting that this stage would filter out the high speed node aimed for avoiding the unnecessary handover. In the second stage: network decision stage, the mobile node will choose the optimum CAN based on following parameters: RSS, cost, and available bandwidth. The last stage is to execute the handover which means the mobile node will change its attached network from NAN to optimum CAN.

141.3.2 Implementations of the Algorithm

In the heterogeneous wireless network environment, RSS directly reflects the quality of transmission channel, so the RSS is the most commonly used indicator in handover decision. On the premise of not taking shadow fading with static zero mean gaussian white noise into consideration, UMTS and WLAN RSS model as follows:

$$\mathrm{RSS}_{\mathrm{UMTS}} = L_{1U} - L_{2U}\mathrm{Log}(\mathrm{d}_U)\,. \tag{141.1}$$

$$\mathrm{RSS}_{\mathrm{WLAN}} = L_{1W} - L_{2W}\mathrm{Log}(\mathrm{d}_W) \tag{141.2}$$

d_U and d_W denote the distance between mobile node and node B in UMTS and the distance between mobile node and WLAN AP. L_{1U},L_{2U}, L_{1W},L_{2W} are parameters for the path loss. Since the coverage area of UMTS is much bigger than WLAN, we can assume that as long as the UMTS signal can be detected, the RSS of UMTS can be default as a constant.

Handover can be divided into two directions in UMTS and WLAN heterogeneous network: one is from UMTS switch to WLAN; the other is from WLAN switch to UMTS. Assuming that while the moving mobile node is communicating with UMTS, it detects a new available WLAN, if the mobile node is moving too fast, it will just handover to WLAN and then handover to UMTS again, so in order to avoid this unnecessary handover, the velocity of mobile node should be an indicator of triggering the handover.

Specific algorithm of triggering the handover is as follows:
Mobile node switch from UMTS to WLAN:

If $((\mathrm{RSS}_{\mathrm{WLAN}} - \mathrm{RSS}_{\mathrm{UMTS}} > ;\mathrm{Threshold}_{\mathrm{RSS}})$
$\&\&$ (The MT's Velocity $<;\mathrm{Threshold}_v))$ then trigger the handover.

Mobile node switch from WLAN to UMTS:

If $(\mathrm{RSS}_{\mathrm{UMTS}} - \mathrm{RSS}_{\mathrm{WLAN}} > \mathrm{Threshold}_{\mathrm{RSS}})$ then trigger the handover

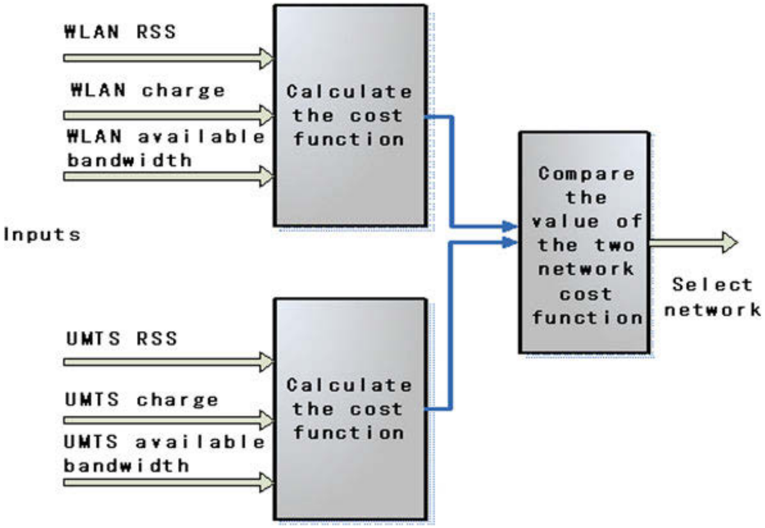


Fig. 141.1 Network decision flow chart

$\text{Threshold}_{\text{RSS}}$ involves the Hysteresis value considered, and Threshold_v is the threshold of velocity.

After triggering the handover, it will be in the second stage: network decision stage, by comparing the cost function value of CAN, find out the optimum CAN. Network decision flow chart is shown in Fig. 141.1. Cost function is defined as follows:

$$f_n = w_B L_n(1/B_n) + w_C L_n(C_n) + w_R L_n(1/\text{RSS}_n) \tag{141.3}$$

In the formula (141.3) f_n is the cost function value of CAN n, B_n is the available bandwidth of CAN n, C_n is the charge of CAN n, RSS_n is the mobile node's RSS from access point. $w_B w_c$, w_R is the weight factor corresponding to above parameters, and it satisfies $w_{B,} + w_c + w_R = 1$.

According to the definition of cost function, the smaller the value of the cost function of CAN indicating the better network condition, that means the bigger the avalible bandwidth, lesser the network cost, and stronger the RSS, more preferred the CAN will be.

141.4 Simulation Results

The simulation scenario is shown in Fig. 141.2. As a result of signal attenuation and shadow effect in the overlaid areas of the two networks, RSS fluctuates near the threshold leading to the mobile node switches between two network many times.

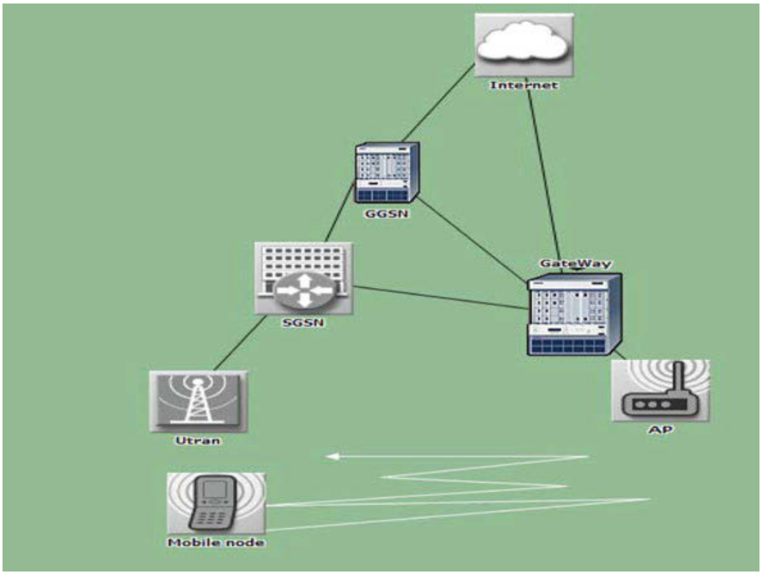


Fig. 141.2 UMTS and WLAN interconnection system

The ping-pong effect in cellular network can be prevented by adding a Hysteresis value or setting a dwelling time.

141.4.1 Handover Times

The proposed algorithm simulation result is compared with the traditional algorithm shown in Fig. 141.3, where the horizontal abscissa is the moving time of mobile terminal and the longitude coordinate is the cumulative handover times in the corresponding time point. The figure shows that the traditional algorithm handover six times, while the proposed algorithm is four times. The proposed algorithm avoids unnecessary handover by adding Hysteresis value and filtering out the high speed mobile node when triggering the handover.

141.4.2 Packet Loss Rate

The PLR in the simulation of vertical handover is shown in Fig. 141.4, the horizontal abscissa is the moving time of the mobile node, the longitudinal coordinates is the PLR in the corresponding time point, it can be seen that the proposed algorithm has better performance, after a stable time the proposed algorithm always has a lower PLR, and with the times of handover increases, the cumulative number of packets not lost will be the assurance of continuous connection between

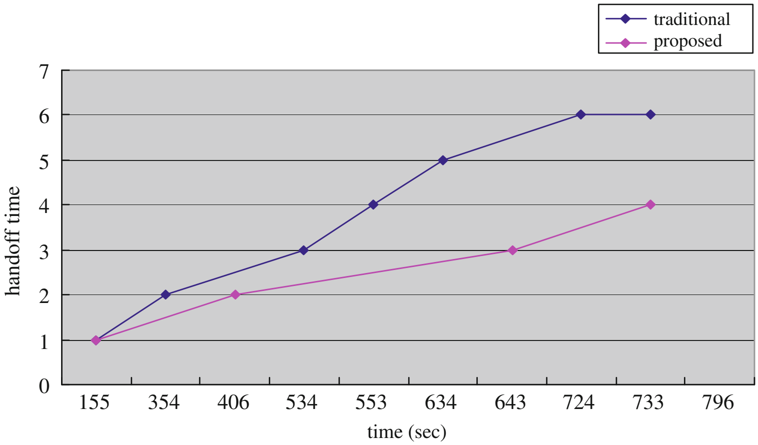


Fig. 141.3 Handover times

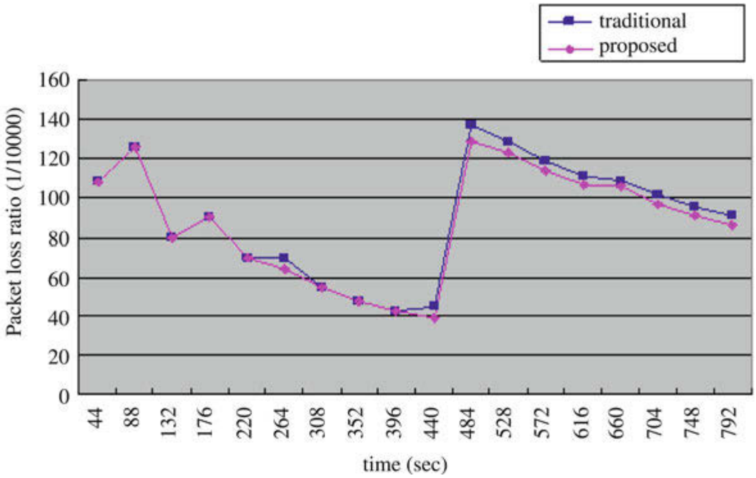


Fig. 141.4 Packet loss ratio of vertical handover

mobile node and NAN, that will let the proposed algorithm outperform the traditional handover algorithm.

141.5 Conclusion

The vertical handover algorithm based on policy-enabled is proposed based on the summarization of many vertical handover algorithms [9–11], the main content of the proposed algorithm as follows: mobile node of high speed is filtered out in the triggering stage of the vertical handover in order to effectively reduce the "Ping-Pong Effect" and decrease the total handover times; In the network decision stage of vertical handover, parameters such as RSS, cost, and bandwidth are involved and synthetically evaluated through the cost function and finally avoiding the computational complexity and large latency at the same time.

Acknowledgments This work is jointly supported by National Natural Science Fund of China (Grant No. 60977051), Science and Technology Project of Shenzhen, China (Grant No. 200730). Authors thank to the two organizations.

References

1. The Internet Protocol Journal. In: IEEE 802.21 12(2) http://www.cisco.com/web/- about/ac123/ac147/archived_issues/ipj_12-2/122_ieee.html
2. Roberto C, Andrea Z (2006) Experimental performance of the handover procedure in a Wi-Fi network
3. Mishra A, Min-Ho S, Arbaugh WA (2003) An empirical analysis of the IEEE 802.11 MAC layer handoff process
4. Eastwood L, Migaldi S (2008) Mobility using IEEE 802.21 in a heterogeneous IEEE 802.16/802.11-based. IEEE Wirel Commun 15:26–34
5. Qingyang S, JamaliPour A (2005) Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques. IEEE Wirel Commun 12:42–48
6. ChengTa C, ChingYao H (2006) Capacity-based compressed mode for intersystem handover in UMTS
7. Ciccarese G, De Blasi M, Marra P (2009) Vertical handover algorithm for heterogeneous wireless networks
8. Ling T, PeiGen B (2007) A new multi-attribute decision-making vertical handover algorithm. In: IET conference, pp 517–520
9. Kun G, Hong J, Xi L (2009) A speed sensitive vertical handoff algorithm based on fuzzy control
10. Bellavista P, Corradi A, Foschini L (2007) Context-aware handover middleware for transparent service continuity in wireless networks. In: Pervasive and mobile computing, pp 439–466
11. Javed K, Saleem U, Hussain K (2010) An efficient approach for vertical handover between WLAN and EVDO 1–5

Chapter 142
An Advance Resource Allocation Algorithm in Network Virtualization

Yuyang Xu, Juan Luo and Lei Chen

Abstract Network virtualization is a promising solution to diversify internet and prevent internet ossification. The allocation of resources is one of the key challenges in network virtualization. An advance resource allocation algorithm was proposed, which focuses on saving bandwidth and combining the advantages of central manner and distributed manner. An advanced VN mapping protocol is designed to communicate and exchange information between central coordinator and substrate nodes. Simulations results showed that the algorithm has better delay performance, which can effectively reduce the communication costs and improve resource utilization.

Keywords Network virtualization · Resource allocation · Resource utilization

This work is partially supported by the National Natural Science Foundation of China (60903019); Hunan Science and Technology Plans to Focus the Project (2009GK2008); The Project-Sponsored by SRF for ROCS, SEM. Hunan young core teacher project.

Y. Xu · J. Luo (✉) · L. Chen
School of Information Science and Engineering, Hunan University Changsha, Hunan, China
e-mail: juanluo@sina.com

Y. Xu
e-mail: xuyueyang2009@sina.com

L. Chen
e-mail: chenleixyz123@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_142,

142.1 Introduction

Network virtualization has been proposed as the alternative to face up the ossification of internet, and a potential solution for diversifying the Future Internet architecture into separate virtual networks (VN) [1, 2]. Every VN owns itself protocols and all the characteristics of one physical network, so the VNs can support simultaneous multiple application architectures on the top of the substrate network. A VN is a set of Virtual Nodes connected to each other via dedicated virtual links over a substrate network. The virtual nodes and virtual links of a VN should be assigned/mapped to a specific set of Substrate Nodes and Substrate Paths, respectively. A Substrate Path is a logical path among two substrate nodes which may be a single substrate link or a set of substrate links. A major challenge in network virtualization is the assigning/mapping of substrate resources to virtual networks (VN) efficiently and on-demand [3, 4]. The current proposed resource allocation algorithm can be classified into two manners: central manner and distributed manner. The central manner [5–7], can easily be designed, and avoid the conflicts during resource allocation. However, the central manner is hard to maintain the global information, so the central manner suffers from scalability limitation, high latency and serious delays in making decisions. The distributed manner [8–10] has advantages such as fewer communication cost, runtime and automatic reparation, strong robustness, and high-speed parallel processing, but the distributed algorithm must make use of effective communication protocol to ensure synchronization of processing among multiple substrate nodes, so it is relatively hard to design.

The remainder of this paper is organized as follows. Section 142.2 summarizes related work. Section 142.3 presents the network model and the VN mapping problem. The design of a distributed VN mapping algorithm is addressed in Sect. 142.4. The implementation and evaluation of the proposed algorithm based on multi-agent-based approach is reported in Sect. 142.5.

142.2 Related Works

Resource mapping in the network virtualization environment aims to fairly and efficiently share physical resource among VNs. It is obvious that the sharing of bandwidth has the most direct impact on the performance of VNs, so it is important to save and efficiently share bandwidth.

Yu et al. [5] proposed path splitting and path migration method to improve utilization of substrate links. Some substrate links with fewer free resource can be allocated to VNs through the path splitting, while the substrate network can dynamically adjust resource allocation through path migration so as to receive more VN requests. But the path splitting may lead out-of-order packet delivery, and the path migration may result in side performance impact on data traffic in VN.

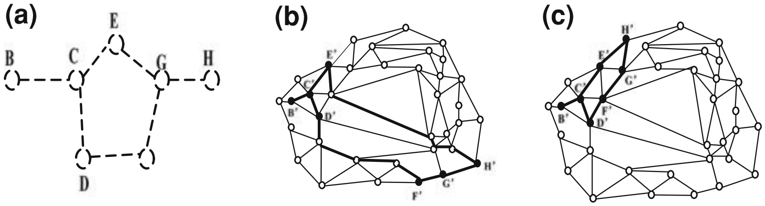


Fig. 142.1 **a** VN topology that needs to map, **b** mapping result that compute according to the previous distributed algorithm, **c** mapping result that save the bandwidth. The mapping result is present by the *thick lines* and *solid circles*

The game theory has been introduced in [6], though the bandwidth of substrate link can be dynamically reallocated among multiple VNs, but the algorithm encourages the efficient behavior among multiple VNs, which may cause the VNs to compete for resource. An adaptive resource allocation scheme is proposed in [7], the architecture proposes objective function for every VN, and maximizes the performance objective independently through distributed and customized protocols. However, if any of the VNs exhibit greedy behaviors, it would lead to unfair results.

The proposal in [8, 9] introduced a virtual network architecture and an associated self-organizing algorithm to equalize the bandwidth, and storage consumption on the physical nodes. But the algorithm is based on live node migration, which may lead interrupting application in VNs. These algorithms focus on efficiently using bandwidth. They are dynamic or runtime. Because of the number of reconfigured VNs, the frequent and the order of reconfiguration affects the VN reassignment performance. The dynamic and runtime designing is very complicated. So we focus on saving bandwidth during the initial resource allocation for VNs in our paper.

142.3 Problem Descriptions

For quite large VN topology, mapping it to the entire substrate at once is not feasible for latency and complexity reasons. One well-known solution is to subdivide the entire VN topology into a set of elementary clusters (e.g. star clusters). This decomposition may reduce the complexity of mapping the entire VN topology. But the methods may cause VN topology to become sparse in the substrate network.

The following example shows this problem. Figure 142.1a shows a VN topology that construct on the substrate network. The topology is divided into two clusters base on VN decomposition in [10]. In the first cluster, the hub is C, the spokes contain B, D, E. In the second cluster the hub is G, the spokes contains F, H. As Fig. 142.1b shows, the first cluster is mapped on substrate nodes such as C′, B′, D′ and E′, when mapping the second cluster, according to the algorithm proposed in [10], suppose that the substrate node with maximal free resource is G′, so choose this substrate node as root to map G, making use of the shortest path algorithm. F and H is mapped to substrate nodes F′ and H′, respectively. As the two clusters connect each other in VN topology, so choose the shortest path between D′ and F′ to map the link connecting D and F, and choose the shortest path between E′ and G′ to map the link connecting E and G, so nodes D and F, E and G become not adjacent in substrate network. The two shortest paths totally occupy nine hops in the substrate network. When mapping the second cluster, if we choose all substrate nodes that own free resource more than G needs to as root, then further choose the result that occupy hops is least as the finally result. As Fig. 142.1c shows, the links connects the two clusters and only occupies two hops in the substrate network. Suppose every link needs 10 M bandwidth, the latter can release 70 M traffic. So the problem of sparse leads wasting bandwidth.

In the distributed manner, once the resource of every substrate node changes, the substrate node must notify other substrate nodes to up data the resources information about it. For example, to map a VN topology that contains 100 nodes, if the substrate network contains 1,000 substrate nodes; the message exchanged among substrate nodes to update resource information is 100,000 that is really huge volume. Moreover, for every cluster, only the substrate nodes with maximal free resource computes the resource allocation, there is no any other result to compare with it, we think it is not always desirable from the standpoint of the whole substrate network.

So our algorithm focuses on saving bandwidth and reduce message volume. To maintain the connect information among clusters, we classify the neighbors of hub three types: Mneighbors, LMneighbors and NMLneighbors. Mneigbors has been mapped; LMneigbors connects the nodes that has been mapped; NMLneigbors has not been mapped and has no relative with the mapped nodes. For example, when mapping the second cluster, E is Mneigbors, F is LMneigbors, and H is NMLneigbors.

142.4 The Design of Algorithm

142.4.1 The Advanced VN Mapping Protocol

In our algorithm, we suppose that it has a central coordinator as in central manner. The advanced virtual network mapping protocol is defined in this section, to indirectly exchange messages and information between all substrate nodes and central coordinator. The protocol is based on six types of messages: REQREGI, ACKREGI, VNTOPO, RESULT, FRESUT and STOP. REQREGI is sent from central coordinator to all substrate nodes to achieve the free resource information; ACKREGI is sent from substrate node to central coordinator and contains the free resource information about substrate node itself.; Central coordinator sends resource information and topology information about VN request (e.g. list of virtual nodes and virtual links, capacities, etc.) to all substrate nodes by VNTOPO message; RESULT is sent from the substrate node to the central coordinator and contains a mapping result for a cluster; the central sends the final result to all substrate nodes via FRESULT; once the entire VN topology is successfully mapped, the central coordinator notifies all substrate node to stop work for this VN topology through STOP.

142.4.2 The Advanced VN Mapping Algorithm

The whole advanced VN mapping algorithm consists of three algorithms: LRDA, CRDA, and SMA. At the beginning of the algorithm, the central coordinate sends REQREGI message to all substrate nodes for knowledge about their free resource information, then the substrate node tells the coordinator free resource values by message ACKREGI. Once coordinator received resource information about all substrate nodes, it broadcasts the resource information and VN topology information to substrate nodes. CRDA is located in the central coordinator, and is started upon by receiving the mapping result for a cluster from a substrate node. The algorithm compares the mapping results with results previously recorded in central coordinator, then discard result with more hops and keep the result with fewer hop; finally, the result with least hop for a cluster is obtained. The final result is sent to all substrate nodes through FRESULT message. LRDA and SMA are located in the substrate node. Once the substrate node receives the FRESULT message, the LRDA is started. LRDA updates the resource information about other substrate nodes according to the received final result, which avoids increase message volume by other message to notify substrate node update information, then check the final result. If it is nominated in the final result, the substrate node would not run the SMA for the next cluster, but still receive the final result for other clusters from central coordinator so as to update free resource information about other substrate node. This is really useful to reduce the times of Resource synchronization when the substrate network finds out mapping result for multiple VN topologies. If it is not in final result, the substrate node calls SMA to compute resource allocation for next cluster. The SMA consists of MAPM, MAPL and MAPNML, and is triggered by receiving a VNTOPO message from central coordinator or by LRDA. The SMA is to finish mapping a cluster and choose the substrate node with free resource is more than hub needs as root. MAPM is responsible for finding the shortest path between root and the host of Mneighbors to map the link connecting the hub and Mneighbors. In MAPNML, there are two resource allocations: mapping the Mneighbors, mapping the link connecting the hub and Mneighbors and mapping. The MAPNML is similar to the algorithm in [9], so there we would introduce more about it. In MAPL, there are three types of resource allocations: mapping Lneighbors, mapping the links connecting

Lneigbors and hub, mapping the links is between Lneigbors and Lneighbors'neighbors that has been mapped. The MAPL is described as follows:

```
Repeat
 1. mapedneiglist: = the mapped neighbors of head(Lneighbors)
 2. tmapedneiglist: = mapedneiglist
 3. clear(colist)
 4. repeat
      a. hoster: = host(head(tmapedneiglist))
      b. list: = shortestpath(root, hoster)
      c. colist: = the intersection of colist and list
      d. tmapedneiglist\head(tmapedneiglist)
      e. clear(list)
 5. Unitl tmapedneiglist is empty
 6. midhoster: = mcapacity(colist)
 7. if C(head(Lneigbors)) > C(midhoster)then Exit else
 8. list: = shortestpath(root,midhoster)
 9. if C(list) < C(link(head(Lneigbors),hub))then Exist else
10. MAPL(hub,head(Lneigbors)): = list
11. tresultlist: = MAPL(hub,head(Lneigbors)
12. MAPN(head(Lneigbors)): = midhoster
13. hop = hop + size(list)-1
14. clear(list)
15. Repeat
      f. list: = shortestpath(mid hoster,host
           (head(mapedneiglist)))
      g. l: = link(head(Lneigbors, head(mapedneiglist))
      h. if C(list) < C(l) then Exit else
      i.   MAPL(head(Lneigbors),
           head(mapedneiglist)): = list
      j.   hop = hop + size(list)-1;
      k.   tresultlist: = MAPL(head(Lneigbors),
           head(mapedneiglist))
      l.   mapedneiglist\head(mapedneiglist)
      m.   clear(list)
      n.   endif
16. endif
17. endif
18. Until mapedneiglist is empty.
```

For every node in Lneighbors, the MAPL finds all mapped neighbors of its (line 1), and the shortest paths from root to the host of every mapped neighbor (line a,

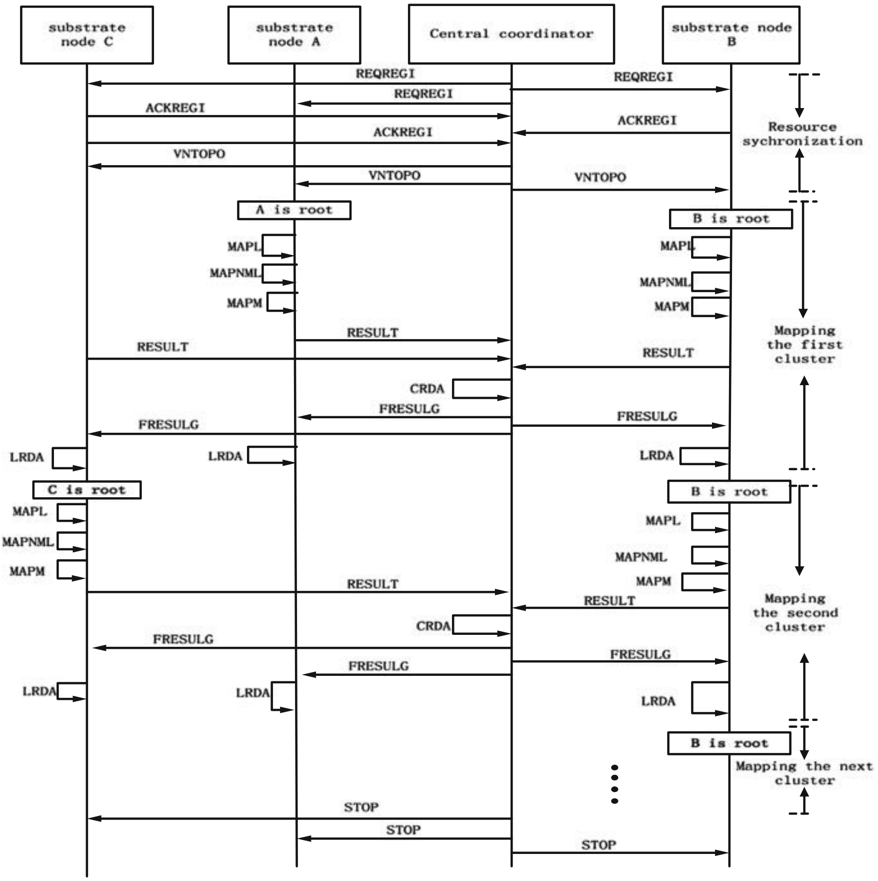


Fig. 142.2 The process of advance VN mapping algorithm

b), then the algorithm tries to find a intersection among these shortest paths (line c). The substrate node with most free resource in intersection, called *midhoster*, is choose as the host of the Lneighbors (line 6 and 12), so the link connecting the Lneighbors and hub is mapped onto the shortest between the root and midhoster (line 10). For every link connecting the Lneighbors and the mapped neighbors, the shortest path between midhoster and the host of the mapped neighbors is choose as host (line g and i). The *hop* is a global variable, it records the amount of substrate links MAPM, MAPNML and MAPL consume. The *tresultlist* collects the result for every cluster. The value of *hop* and tresultlist would be sent to central coordinator via RESULT message. The process of advanced VN Mapping Algorithm is described in Fig. 142.2.

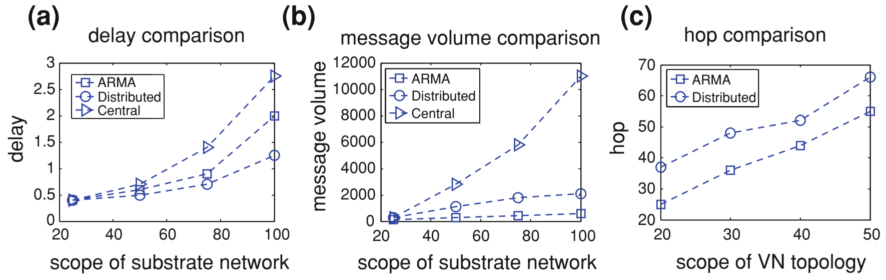


Fig. 142.3 The unit of delay is second, the unit of message volume is packet

142.5 The Implementation and Evaluation

The advance resource mapping algorithm (ARMA) has been implemented and tested over the CloudSim2.0 [11] platform. The simulation platform provides a good support for virtualization. This paper makes use of the Datacenter and Datacenterbroker to simulate substrate nodes and central coordinator, respectively. The topology generator—Brite [12] is needed to build the substrate network. We generated four substrate networks via Brite, containing 25, 50, 75 and 100 substrate, respectively, to allocate resource for a VN request containing 20 nodes. The delay and message volume of the algorithm is evaluated. We also evaluate the consumption of bandwidth in light of hop for different VN topology.

Figure 142.3a shows the comparison of delay generated by central manner, distributed manner and ARMA for a VN topology containing 20 nodes in different substrate network. With the scope expanding, the delay increased in the three algorithm that is because more substrate nodes mean more huge search space for central manner, and more time overhead for exchanging information among substrate nodes in distribute manner and ARMA. The central manner needs to search the whole space for every node mapping and link mapping, while the distributed manner tacks actions based on the local information in substrate nodes, although the ARMA allocates resource based on local information as that in distribute manner. The exchanging of information among substrate nodes is indirect via central coordinator. As a result, the delay of the central manner consumed is most, while the distributed manner is least, and the MRMA is among them in the same substrate network.

Figure 142.3b shows the comparison of message volume generated by central manner, distributed manner and ARMA for a VN topology containing 20 nodes in different substrate network. As the substrate network expands, the message volume also increases in three algorithms. However, the change of central manner is drastic, the tendency of distributed manner is moderation and the AMRA is slightly changing. The central manners need to periodically exchange message with substrate nodes for maintaining the topology information. Only the substrate node maintains the local topology information in the distribute manner and ARMA that is why the changing is different for message volume in the three algorithms. In ARMA, the message in notifying substrate nodes to update resource information is avoided. The substrate node can automatically update resource information according to the final result. So the message volume is less than that the distribute manner generates.

Figure 142.3c shows the comparison of average hops generated by distribute manner based on VN decomposition and ARMA for different VN topology. The hop increases with the VN topology expanding in both distribute manner and ARMA. In ARMA, the central coordinator can choose the result with least hops for every cluster, while the distributed manner achieves only one result generated by substrate node with most free resource in substrate network, so the average hops in ARMA is fewer than that in distributed manner.

142.6 Conclusion

This paper proposed an algorithm combined the advantages of central manner and distribute manner, focusing on saving bandwidth resource by overcomming the sparse problem of VN mapping. Our algorithm obtains better delay performance, less message volume and hops. However, the robustness in our algorithm and choosing hop as metric for result is still worth researching. So we will design the scheme to dynamically choose the central coordinator to ensure the robustness of our algorithm,and choose better metric for result in light of load balancing in the future work.

References

1. Niebert N, Khayat IE, Baucke S, Keller R, Rembarz R, Sachs J (2008) Network virtualization: a viable path towards the future internet. J Wirel Pers Commun 45:511–520
2. Turner J, Taylor D (2005) Diversifying the internet. IEEE Glob Telecommun Conf, Missouri 6–760
3. Haider A, Potter R, Nakao A (2009) Challenges in resource allocation in network virtualization. In: 20th ITC specialist seminar, Hoi An
4. Chowdhury N, Boutaba R (2009) Network virtualization: state of the art and research challenges. IEEE Commun Mag 47(7):20–26
5. Yu M, Yi Y, Rexfor J (2008) Rethinking virtual network embedding: Substrate support for path splitting and migration. In: ACM SIGCOMM CCR 17–29
6. Zhou Y, Li Y, Sun G, Jin D, Su L, Zeng L (2010) Game Theory Based Bandwidth Allocation Scheme for Network Virtualization. In: Proceedings of GLOBECOM 1–5
7. He J, Zhang-Shen R, Li Y, Lee C, Rexford J, Chiang M (2008) DaVinci: dynamically adaptive virtual networks for a customized internet. ACM CoNEXT
8. Marquezan C, Nunzi G, Brunner M, Granville LZ (2010) Distributed autonomic resource management for network virtualization. In: Proceedings of 12th IEEE/IFIP network operations and management symposium, Osaka, pp.19–23

9. Marquezan C, Nobre J, Granville L, Nunzi G, Dudkowski D, Brunner M (2009) Distributed reallocation scheme for virtual network resources. In: IEEE international conference on communications
10. Houidi I, Louati W, Zeghlache D (2008) A distributed virtual network mapping algorithm. In: Proceedings of IEEE ICC, pp. 5634–5640
11. A framework for modeling and simulation of cloud computing infrastructures and services, http://www.cloudbus.org/cloudsim
12. Boston University representative Internet topology generator. http://www.cs.bu.edu/brite/

Chapter 143
Development of LBS Technology Based on Multimedia Broadcast and Multicast Services in 3G Mobile Networks

Lu Lou, Xin Xu, Zhili Chen, Juan Cao and Jun Song

Abstract This chapter presents a new location based service that can be provided in third generation partnership project Multimedia Broadcast and Multicast Services system (MBMS) economically and effectively. We propose a point of interest message embedded into transport protocol experts group (TPEG) and describe the implementation of TPEG message broadcasted over MBMS using the stream delivery method and download delivery method.

Keywords Point of interest · Location based service · Broadcast · TPEG · MBMS · UTMS · FLUTE

143.1 Introduction

A Location based Service (LBS) is an information and entertainment service, accessible with mobile devices through the mobile network and utilizing the ability to make use of the geographical position of the mobile device. On the

L. Lou (✉) · J. Cao · J. Song
College of Information Science and Engineering, Chongqing Jiaotong University,
400074 Chongqing, China
e-mail: cloudlou@163.com

X. Xu
Library of Chongqing Jiaotong University, 400074 Chongqing, China
e-mail: xx1771@163.com

Z. Chen
Faculty of Information and Control Engineering, Shenyang Jianzhu University,
110168 Shenyang, China
e-mail: chenzhili@sjzu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_143,

driving or walking around, consumers want to find an optimal path to destination or information of point of interest (POI) such as park, restaurant, hotel, cinema, gas station, traffic jam, and so forth. Location based services usually also rely on real-time traffic information that be often encoded in Transport Protocol Expert Group (TPEG) protocol. TPEG standard designed by European Broadcasting Union (EBU) is a new traffic information transfer protocol which has three major characteristics, language independent, bearer independent, and multi-modal application [1–3]. Broadcasting based transmission technology is one of main methods to provide dynamic traffic information or public emergency service in recent years, which is used by countries all over the world.

The Third Generation Partnership Project (3GPP) suggests an enhancement of current cellular networks to support Multimedia Broadcast and Multicast Services (MBMS). MBMS is a broadcasting service that can be offered via cellular networks using point-to-multipoint links instead of the usual point-to-point links. Because of its operation in broadcast or multicast mode, MBMS can be used for efficient streaming or file delivery to mobile phones [4, 5].

In this chapter, we present a nevol POI application specification based on TPEG over MBMS, which is satisfied to LBS applications in mobile devices, and explain the practicability and feasibility of that.

143.2 Preparation of Information for TPEG and MBMS

143.2.1 TPEG Overview

TPEG technology has been designed to provide a twentyfirst century multimodal Traffic and Travel Information (TTI) data protocol for delivering content to the end-user, regardless of location or client type in use. The TPEG standards therefore cover the data formats and protocols required for sending the TTI to the broadcaster and the protocols required for broadcasting this to the end users by DAB.

TPEG is a protocol to provide TTI and just two applications has been developed. One is to transfer the road traffic status message (RTM) caused by accident,weather, out-of-door gathering, and so forth. The other is to convey the public transport information message (PTI) such as schedule and route of bus, train, flight, ship, and so forth. As shown in Figs. 143.1 and 143.2, both applications have their own message structure and they contain above mentioned information in it [1–3].

143.2.2 MBMS Overview

MBMS is an IP datacast type of service that can be offered via existing GSM and UMTS cellular networks, which has been standardized in various of 3GPP, and the first phase standards are to be finalized for UMTS release6 [5].

Fig. 143.1 The Structure of TPEG-Message

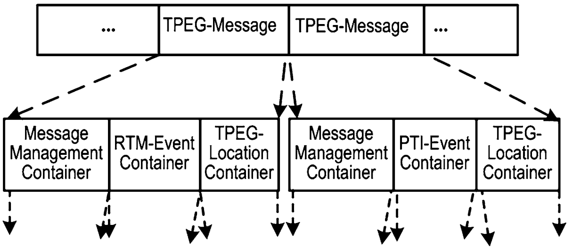
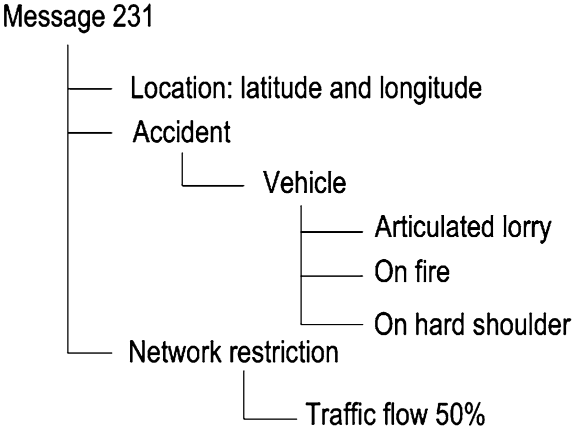


Fig. 143.2 A example of TPEG Road Traffic Messages



The MBMS is a unidirectional point-to-multipoint bearer service in 3GPP cellular network in which data are transmitted from a single source entity to multiple mobiles. Various MBMS user services can be made up of these MBMS bearer services. To support the MBMS, Broadcast and Multicast-Service Center (BM-SC) is newly added to the network, and MBMS controlling functions are added to the existing network entities such as UMTS Terrestrial Radio Access Network (UTRAN), Serving GPRS Support Node (SGSN), and Gateway GPRS Support Node (GGSN). For user equipments (UEs) (or mobiles) to support the MBMS, much additional functionality are also required to be added. Some of them, for example, the interaction among protocol entities which is beyond of the standardization should be defined in detail. New protocols such as File Delivery over Unidirectional Transport (FLUTE) and media codecs are also needed to be implemented. Figure 143.3 shows the 3GPP Release-6 network architecture for the MBMS [5–7]. As shown in Fig. 143.4, the content is distributed via unicast (point-to-point) connections, forcing the network to process multiple requests for the same content sequentially and therefore wasting resources in the radio access and core networks. With the expected increase of high bandwidth applications, especially with a large number of User Equipments (UEs) receiving the same high data rate services, efficient information distribution becomes essential. Broadcast and

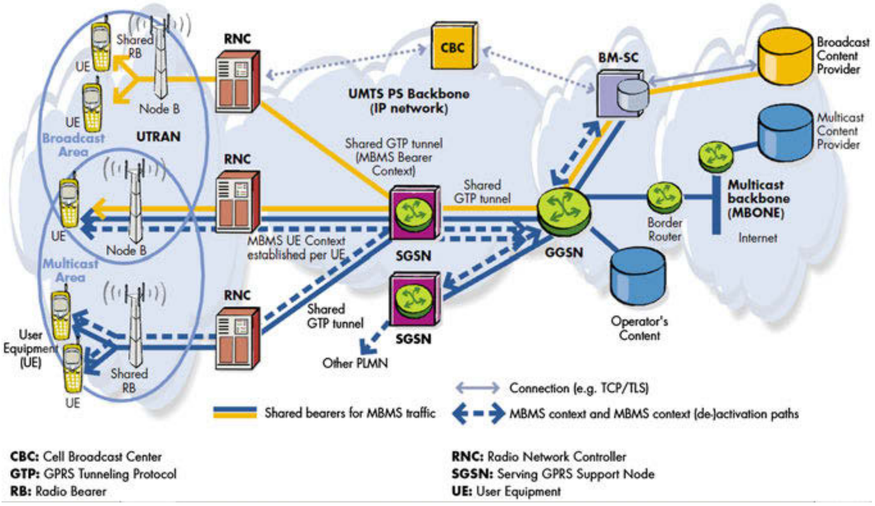
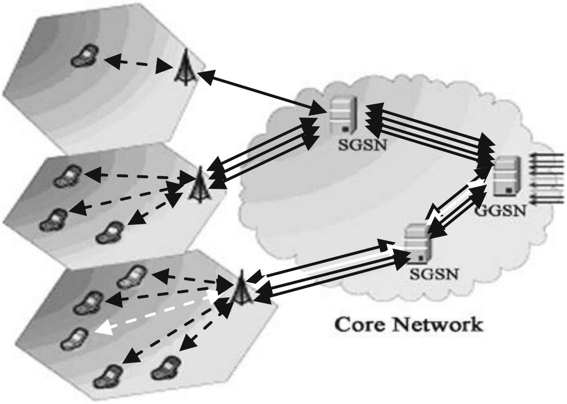

Shared RB
UE
Node B
UE
Broadcast Area
UTRAN
RNC
CBC
UMTS PS Backbone (IP network)
BM-SC
Broadcast Content Provider
Multicast Content Provider
Shared GTP tunnel (MBMS Bearer Context)
Shared GTP tunnel
RNC
UE
Node B
Multicast Area
MBMS UE Context established per UE
SGSN
GGSN
Border Router
Multicast backbone (MBONE)
Internet
RNC
Shared GTP tunnel
SGSN
Other PLMN
Operator's Content
User Equipment (UE)
Shared RB
UE
Connection (e.g. TCP/TLS)
Shared bearers for MBMS traffic
MBMS context and MBMS context (de-)activation paths
CBC: Cell Broadcast Center
GTP: GPRS Tunneling Protocol
RB: Radio Bearer
RNC: Radio Network Controller
SGSN: Serving GPRS Support Node
UE: User Equipment

Fig. 143.3 The MBMS architecture designed in 3GPP

Fig. 143.4 Broadcasting over unicast connections without MBMS



multicast are methods for transmitting datagrams from a single source to multiple destinations. Thus, these techniques decrease the amount of transmitted data within the network. This results in a dramatic cost reduction. Figure 143.5 illustrates the advantage of MBMS multicast distribution.

MBMS provides two modes of services—broadcast and multicast. The broadcast mode enables multimedia data (e.g. text, audio, picture, video) to be transmitted to all users within a specific service area. The broadcast mode needs neither service subscription nor charging. The multicast mode enables multimedia data to be transmitted from one source entity to a specific multicast group. For a user to receive a multicast mode service, it should be subscribed to the multicast group in advance. Multicast mode services are charged, and therefore should be protected from illegal users [8].

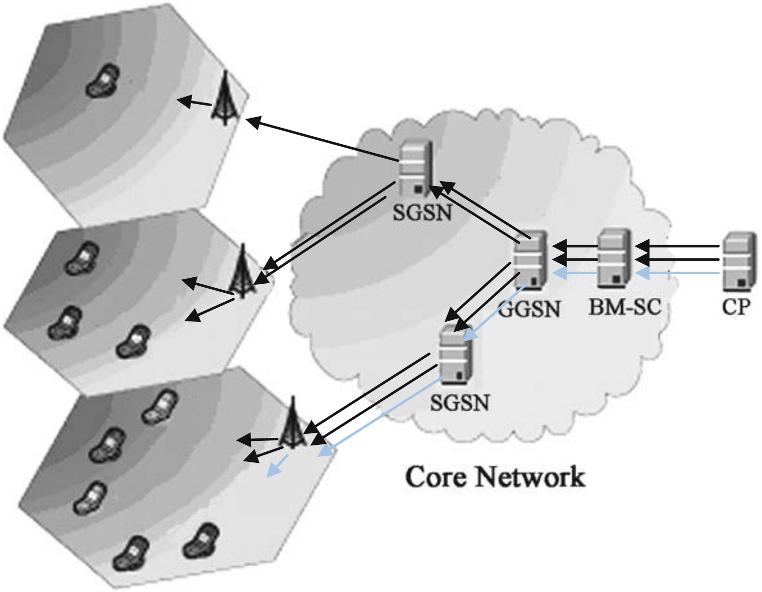


Fig. 143.5 Broadcasting over multicast connections with MBMS

143.3 The Implemention of TPEG Over MBMS

143.3.1 The Design of POI Message Using TPEG

In order to provide greatest flexibility, the TPEG system was designed to allow a number of different service providers to deliver a number of different types of information for ITS, without needing to rely on any facilities provided by the bearer system. This was done to void compromising the bearer independence of the protocol. For this reason, the first level of TPEG framing—*the 'Transport Frame'*—carriers a multiplex of TPEG *'Application Frames'* carrying potentially different types of information, and a TPEG stream is constructed from a sequence of transport frames from potentially different service providers [2, 3].

We design a new POI application specification, which is satisfied to be interoperable with TPEG protocol. The hierarchical transport frame structure including the POI message made up three data fields is shown in Fig. 143.6 and it is embedded into the existing TPEG applications. Each data field is called container, and the first one of the message management container is used to manage the POI information in the receiving side. As shown in Fig. 143.7, the second one, POI event container, consists of the four items such as classification, description, reservation, time information. The POI information is divided into more than ten categories and each category is also classified into several sub-categories. For example, restaurant category is consist of Chinese, western and fast-food

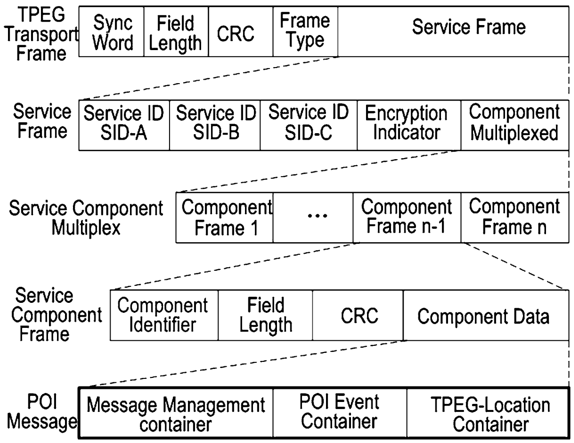


Fig. 143.6 Transport frame structure for POI application

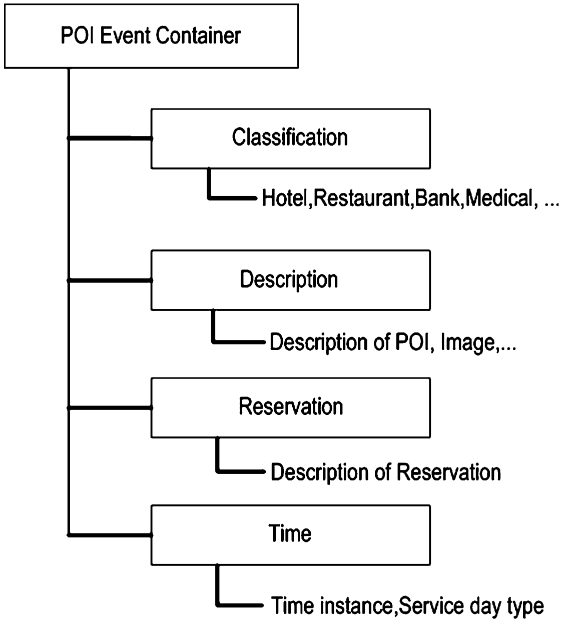

POI Event Container
Classification
Hotel,Restaurant,Bank,Medical, ...
Description
Description of POI, Image,...
Reservation
Description of Reservation
Time
Time instance,Service day type

Fig. 143.7 Structure of POI event container

restaurant, and so forth. The last one, TPEG-location container, represents the exact position of POI by using the WGS84 co-ordinate or descriptor.

The TPEG specifications comprise two separate message representations, TPEG Binary and tpegML. TPEG Binary is a space-efficient description used for digital radio delivery, while tpegML is an XML implementation developed for Internet services. The hierarchical nature of XML fits perfectly to TPEG.

We use standard XML tools to generate tpegML that in the final phase of stream generation will be encoded for the stream. An example message: *A hotel can accommodate up to fifty peoples.it's located on A10 and apart from only 5 km* *of Chongqing international airport. Rate for standard room is 180 yuan per day. discounting only on 8:30AM–18:00PM*. Expressed in tpegML, this message looks like:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE tpeg_document PUBLIC "ITS/tpegML/EN" "tpegML_05.dtd">
<tpeg_document>
<POI message_id="123"
 version_number="1"
 message_generation_time="2009-02-03T13:03:00Z"
 severity_factor="&rtm31_4;">\\
<!-- Location is on A10, Chongqing international airport.-->
<tpeg_loc_container language="&loc41_30;">
<location_coordinates location_type="&loc1_5;">
<WGS84 longitude="105.6212" latitude="29.3864"/>
<descriptor descriptor_type="&loc3_7;"
 descriptor="A10"/>
<descriptor descriptor_type="&loc3_8;"
 descriptor=" Chongqing "/>
<direction_type direction_type="&loc2_2;"/>
</location_coordinates>
</tpeg_loc_container>
<!-- A hotel can accommodate up to fifty peoples. -->
<hotel number_of="1">
<position position="&rtm10_37;"/>
<people number_of="50">
</hotels>

......

</POI_message>
</tpeg_document>
```

The values of the tables are naturally transformed into XML entities, also the document type definition (DTD) of tpegML is divided so that all of these entities are defined in separate files.

143.3.2 The Accessing of MBMS Services

Broadcast mode enables the unidirectional point-to-multipoint transmission of multimedia data and enriched multimedia services. Broadcast mode necessitates neither subscription nor joining by the user. All users located in the broadcast service area as defined by the mobile network operator can receive the data. However, since not all users may wish to receive those messages, the user has the possibility to disable their reception by configuring the UE. Unlike the broadcast mode, the multicast mode requires the user to subscribe for the general reception of MBMS services. End users monitor service announcements and can decide to join one or more available services. Charging is possible either on subscription or on purchasing keys enabling access to the transmitted data. When announcing a particular service, the BM-SC informs the devices about the nature of the service and about the parameters required for the activation of the service (e.g. IP multicast address).

MBMS defines two methods for service announcements: pull (the initiative comes from the receiving UE) and push (the initiative arises from the service itself). In the case of a pull method, the devices fetch the announcements (HTTP or WAP) from a web server. As push method, Short Message Service (SMS) cell broadcast, SMS-PP (point-to-point), WAP-PUSH, MBMS broadcast, MBMS multicast and Multimedia Message Service (MMS) are used [5, 6].

143.3.3 MBMS Protocols and Codecs

Essentially, MBMS offers a scalable mechanism for delivering multimedia content over 3G networks. MBMS User services use the protocol stack shown in Fig. 143.8. A general distinction between two delivery methods exists: the download delivery method and the streaming delivery method. Streaming uses Realtime Transport Protocol (RTP) is used, which in turn uses User Datagram Protocol (UDP). For downloading, the FLUTE protocol applies. FLUTE bases on Asynchronous Layered Coding (ALC) and thus inherits its requirements including massively scalable multicast distribution. ALC itself is a protocol instantiation of Layered Coding Transport building block (LCT) providing in-band session management functionality. The most important feature of FLUTE is to provide the properties of the files in-band together with the delivered files. FLUTE builds on the unreliable UDP but offers a strong Forward Error Correction (FEC) which however gives no absolute delivery reliability. Therefore, MBMS offers the possibility of error correction after finishing the transmission with dedicated channels to a file repair server. Raptor Codes are in use for FEC [9].

143.3.4 Delivery of TPEG and Geodata Using MBMS

Location based services rely on information that can directly be mapped to a real world location.This information, often encoded in points of interest (POIs) and geo-referenced map representations,can, in the broadest sense, be called ‘geodata’. Such geodata can then be mapped and displayed by a location-based application that integrated into the mobile service.

As mentioned above, we have already encoded POIs with TPEG stream. The TPEG stream can be delivery by MBMS using stream multiplexing method. In addition to the POIs, the most common geodatas, such as roads, buildings, vehicles, lakes, forests, and countries along with their extended information, also need be broadcasted as well using the file push delivery method.

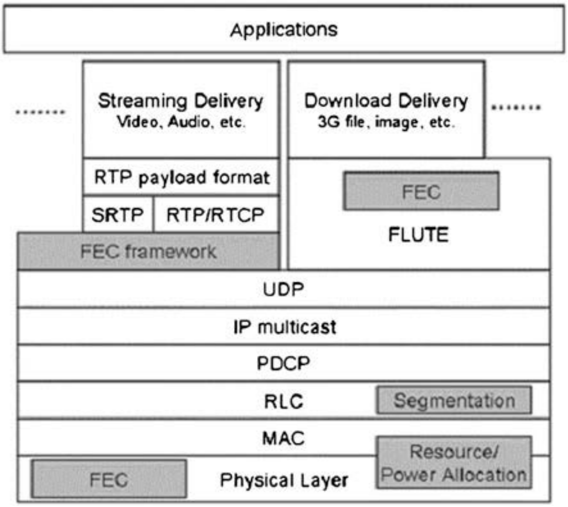


Fig. 143.8 MBMS protocol stack

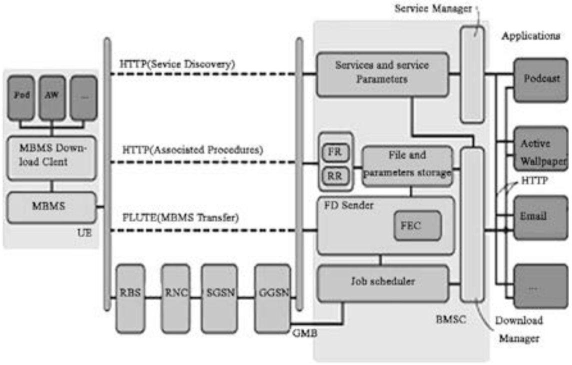


Fig. 143.9 MBMS Content Delivery Architecture

Figure 143.9 schematically depicts a MBMS content delivery architecture. The BM-SC hides the transmission complexity by providing a web service API to application servers. The service and download managers handle the abstraction of internal mobilenetwork procedures to web service methods. The service manager registers new applications and generates the corresponding service attributes and description files: user service description (USD), session description (SDP), and associated delivery procedure description (ADP). The download manager handles transmission requests issued by the applications. The file delivery (FD) sender partitions the files to fit into UDP packets. FEC redundancy is optionally added to increase transmission reliability [10].

In the receiving side, the mobile device (UE) has to correctly manage and process the datas when receiving broadcasted TPEG stream. The TPEG decoder has been implemented in a mobile phone, or a PDA which is connected to the MBMS receiver through USB or SDIO, so that the decoded POI information can be used in the embedded GIS or navigation software. Firstly the received TPEG stream is parsed and decoded before it goes into the GIS or navigation software,then combined with the digital maps,the decoded POI information is displayed easily on the screen of mobile device.

143.4 Conclusions

In this chapter, we propose a novel POI application based on the TPEG over MBMS, and explain the implement of that using the stream delivery method and download delivery method. Not only for LBS application, MBMS services allow for the a lot new attractive applications for mobile communication users. Mobile network operators can introduce their customers to broadband multimedia applications with mobile broadcast services and create distinguishing characteristics in order to differentiate themselves from competitors.

References

1. Transport Protocol Experts Group (TPEG) Specifications (2006) Part 4: Road
2. Transport Protocol Experts Group (TPEG) Specifications (2006) Part 5: Public
3. Transport Protocol Experts Group (TPEG) Specifications (2006) Part 6: Location referencing for applications
4. Digital Audio Broadcasting (2005) Data roadcasting transparent data channel. ETSI TS 101 759 v1.2.1
5. Multimedia broadcast/multicast service (mbms): architecture and functional description (2007) 3G TS 23.246
6. Multimedia broadcast/multicast service (mbms): Protocols and codecs (2007) 3G TS 26.346
7. de Vriendt J, Gomez Vinagre I, Van Ewijk A (2004) Multimedia broadcast and multicast services in 3g mobile networks. Alcatel Telecommun Rev 1
8. Shin J, Park A (2006) Design of MBMS client functions in the mobile. Proc World Acad Sci Eng Technol 18
9. Luby M, Watson M, Gasiba M, Stockhammer T, Wen Xu T (2006) Raptor codes for reliable download delivery in wireless broadcast systems. 3rd IEEE Consumer Commun Netw Conf 1:192–197
10. Thorsten Lohmar U (2009) Scalable push file delivery with MBMS. Ericsson Rev 1:12–16

Chapter 144
Direct Torque Control for Induction Motor Based on Fuzzy-Neural Network Space Vector Modulation

Cai Binjun, Ming XiaoBo and Li Chunju

Abstract In order to improve direct torque control (DTC) system dynamic performance and low-speed performance for space vector modulation (SVM) were analysed. The two PI controllers which were used to generate reference voltage vector in conventional SVM–DTC were analyzed. The parameters of PI controller are difficult to determine. A novel control strategy of DTC induction motor based on fuzzy-neural was proposed. The design process of the neural network controller which generates the flux reference voltage vector and fuzzy controller by applying the torque reference voltage vector was represented. Simulations and experiments were carried out to verify the proposed strategy, and the results were compared with conventional SVM–DTC. The simulation and experiment results verify whether the fuzzy-neural network SVM–DTC is capable of effectively improving the control performance, especially improving SVM–DTC system's low-speed performance.

Keywords Induction motors · Direct torque control · SVM · Fuzzy control · Neural-network control · Low-speed performance

A project supported by scientific research fund of Hunan Provincial Education Department (09C261/09C262).

C. Binjun (✉) · L. Chunju
Hunan Institute of Engineering, Xiangtan 411104, China
e-mail: Cbj12@163.com

M. XiaoBo
Department of Information Engineering, ShangRao Vocational
Technical College, ShangRao 334109, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_144,

144.1 Introduction

Direct torque control (DTC) is a novel high-performance control strategy in the field of AC drives. Compared with complex coordinate transformation in vector control, DTC can be realized easily in digital with simple structure, while problems exist as well. Firstly, the switching frequency of voltage souse inverter is Non-Fixing. Non-Fixing switching frequency causes switching capability of the inverter not to be used fully [1, 2]. Secondly, there is sharp increase or decrease of torque because only one voltage vector works in a sampling period and the options of vector are limited [3].

Many scholars have put forward a lot of solutions for the inherent problem in DTC system. One of the methods is to apply algorithm of space vector modulation (SVM) in DTC system [4]. The key issue for SVM algorithm is how to obtain the reference voltage vector. In conventional SVM–DTC, two PI controllers are used to generate the two components of reference voltage vector. In theory, the reference voltage vector can accurately compensate error of Torque and flux. But in practice conventional SVM–DTC cannot achieve a precise control for two reasons. One is that the determination of the PI controller parameters is subject to repeated try. The other is that control performance of PI controller depends on exact observation on torque and flux [5]. The inaccurate observation occurred in motor controlling when motors work in a low speed.

Fuzzy logic control has manifested its robustness, and has been extensively researched and used as one of the intelligent control methods in control field [6, 7]. To further improve the performance of torque control and to enhance the system robustness, a novel SVM–DTC strategy of induction motors has been proposed. The new SVM–DTC strategy uses fuzzy-neural-network controller to substitute the original PI controller.

144.2 Theory of Direct Torque Control

Direct torque control system applies mathematical analysis about space vector, and is stator flux orientated. The flux–linkage equations of induction machines in the stator stationary reference frame as follows:

$$\psi_{\alpha s} = \int (v_{\alpha s} - R_s i_{\alpha s})\mathrm{d}t \tag{144.1}$$

$$\psi_{\beta s} = \int \left(v_{\beta s} - R_s i_{\beta s}\right)\mathrm{d}t \tag{144.2}$$

where $\psi_{\alpha s}$ and $\psi_{\beta s}$ are the α-axis and β-axis component of $\overrightarrow{\psi}_s$, respectively; $v_{\alpha s}$ and $v_{\beta s}$ are the α-axis and β-axis component of $\overrightarrow{v}_s$, respectively; $i_{\alpha s}$ and $i_{\beta s}$ are the α-axis and β-axis component of $\overrightarrow{i}_s$,respectively.

The electromagnetic torque can be expressed using the following equation:

$$T_e = \frac{3}{2} n_p (\overrightarrow{\psi}_s \times \overrightarrow{i}_s) = \frac{3}{2} n_p (\psi_{\alpha s} i_{\beta s} - \psi_{\beta s} i_{\alpha s}) \tag{144.3}$$

where T_e is electromagnetic torque and n_p is the number of rotor pole pairs.

144.3 The Fuzzy-Neural Network SVM–DTC Scheme

144.3.1 Theory of Voltage Space Vector Modulation

Voltage space vector modulation can synthesize reference voltage vector with arbitrary size and direction by using adjacent basic voltage space vector

$$\overrightarrow{U}_s = \frac{t_1}{T_0}\overrightarrow{U}_1 + \frac{t_2}{T_0}\overrightarrow{U}_2 + \frac{t_3}{T_0}\overrightarrow{U}_0 \tag{144.4}$$

where $\overrightarrow{U}_1$ and $\overrightarrow{U}_2$ are the basic voltage vectors; $\overrightarrow{U}_0$ is the zero vector; and $\overrightarrow{U}_s$ is the reference voltage vector. $T_0 = t_1 + t_2 + t_3$, T_0 is a control cycle.

Formula (144.4) is transferred to two-phase static coordinate system as follows:

$$u_{\alpha s} = \frac{t_1}{T_0} u_1 \cos\theta_1 + \frac{t_2}{T_0} u_2 \cos\theta_2 \tag{144.5}$$

$$u_{\beta s} = \frac{t_1}{T_0} u_1 \sin\theta_1 + \frac{t_2}{T_0} u_2 \sin\theta_2 \tag{144.6}$$

where θ_1 is the angle between vector $\overrightarrow{U}_1$ and the positive direction of α-axis. θ_2 is the angle between vector $\overrightarrow{U}_2$ and the positive direction of α-axis. $u_{\alpha s}$ and $u_{\beta s}$ are the α-axis and β-axis component of $\overrightarrow{U}_s$, respectively.

The effect time of basic voltage vector is answered by (144.5) and (144.6). The example of applying basic voltage vector $\overrightarrow{U}_1$ and $\overrightarrow{U}_2$ to synthesize reference voltage vector $\overrightarrow{U}_s$ is illustrated in Fig. 144.1. Substituting $\theta_1 = 0°$ and $\theta_2 = 60°$ in (144.5) and (144.6) yields

$$t_1 = \frac{(3u_{\alpha s} - \sqrt{3}u_{\beta s})T_0}{3u_4} \tag{144.7}$$

$$t_2 = \frac{2\sqrt{3}T_0 u_{\beta s}}{3u_6} \tag{144.8}$$

$$t_3 = T_0 - t_1 - t_2 \tag{144.9}$$

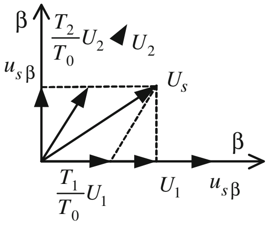


Fig. 144.1 Systhesize reference voltage vector

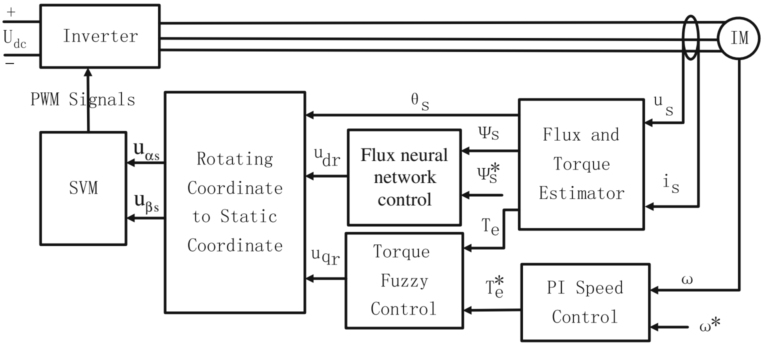


Fig. 144.2 Fuzzy-neural network SVM–DTC system block diagram

144.3.2 Fuzzy-Neural Network SVM–DTC Scheme

The voltage vector can compensate the flux linkage error and torque error is named reference voltage vector. The core issue of SVM–DTC algorithm is how to obtain the reference voltage vector. Figure 144.2 is principle block diagram of improved SVM–DTC.

In Fig. 144.2, the d-axis component of reference voltage vector in the rotor frame is generated by using flux neural-network controller to tackle the flux error, and the q-axis component of reference voltage vector in the rotor frame is generated by using torque fuzzy controller to tackle the torque error.

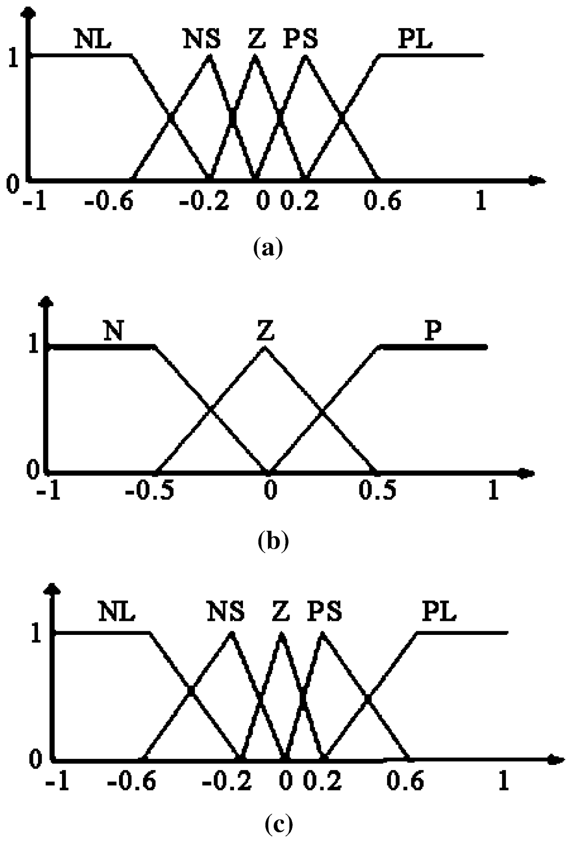
The two components of reference voltage vector in the stationary frame are input into SVM module and generate PWM signal controlling switch state of the inverter.

144.4 Design Fuzzy Controller of Torque

144.4.1 Fuzzy Variables and Membership Functions

The torque fuzzy controller also has two input variables and one output variable. Input variables: torque error E_T and change rate of torque error ΔE_T. Output variable: q-axis component of reference voltage vector u_{qr}. E_T has five fuzzy

Fig. 144.3 The fuzzy membership functions of torque controller **a** Torque error. **b** Change rate of torque error. **c** Q-axis component of reference voltage vector



(a)

(b)

(c)

subsets: PL, PS, Z, NS, and NL. ΔE_T has three fuzzy subsets: P, Z, and N. u_{qr} has five fuzzy subsets: PL, PS, Z, NS, and NL (Fig. 144.3).

144.4.2 Fuzzy Control Rules

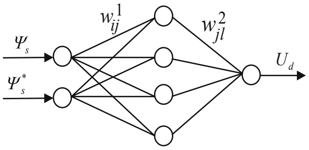
Fuzzy control rules apply IF–THEN form. The rule of torque fuzzy controller R_i can be written as R_i : If $E_T = A_i$ and $\Delta E_T = B_j$ then $u_{\mathrm{qr}} = C_{\mathrm{ij}}$; where: $A_i, B_j, C_{\mathrm{ij}}$ is some fuzzy subset of $E_T, \Delta E_T$ and u_{qr}, respectively.

The total number of rules of torque fuzzy controller is 15 as shown in Table 144.1.

Table 144.1 Fuzzy control rules of torque

u_{qr}	E_T				
$\triangle E_T$	*NL*	*NS*	*Z*	*PS*	*PL*
N	PL	PL	PS	Z	Z
Z	PL	PS	Z	NS	NL
P	Z	Z	NS	NL	NL

Fig. 144.4 Neural-network structure of stator flux



144.4.3 *Fuzzy Reasoning and De-Fuzzy*

The reasoning method is mamdani's procedure based on min–max decision. The firing strength of the rule of torque fuzzy controller $\mu_{R_{\mathrm{ij}}}(u_{\mathrm{qr}})$ can be obtained by considering

$$\mu_{R_{\mathrm{ij}}}(u_{\mathrm{qr}}) = \mu_{A_i}(E_T) \wedge \mu_{B_j}(\Delta E_T) \wedge \mu_{C_{\mathrm{ij}}}(u_{\mathrm{qr}}) \tag{144.10}$$

Fuzzy quantity must de-fuzzy before being sent to control object, and use the center of gravity method for defuzzification.

144.5 Design Neural-Network Controller of Torque

144.5.1 *The Neural-Network Structure of Flux*

The flux reference voltage vector U_{dr} was realized by BP neural network. Its structure is show in Fig. 144.4.

144.5.2 *The Neural-Network Learning Algorithm of Flux*

The neural network can signify arbitrary nonlinear function. Three layer BP neural network contains input layer, hide layer and output. The relationship among three layer is as follows:

$$v_j^1(n) = \sum \omega_{\mathrm{ij}}^1(n) u_i(n) \tag{144.11}$$

$$v_k^2(n+1) = \sum \omega_{\mathrm{jk}}^2(n) v_j^1(n) \tag{144.12}$$

$$y_k(n+1) = f\left(v_k^2(n+1)\right) \tag{144.13}$$

where k is the output layer variable; j is the hide layer variable; v is the neural-network unit; y is the neural-network output; $\omega_{\mathrm{ij}}(n)$ is the neuron weight from i to j; and f is the activate function. Suppose $d(n)$ is the expectation output of neural output, then transient error vector can expressed as:

$$e(n) = d(n) - y(n) \tag{144.14}$$

The target function can be defined as:

$$E(n) = \frac{1}{2}e(n)^T e(n) \tag{144.15}$$

According to the shortest down rules we can obtain the amendment quantity $\omega_{\mathrm{lm}}(n)$

$$\Delta\omega_{\mathrm{lm}}(n) = -\eta\frac{\partial E(n)}{\partial\omega_{\mathrm{lm}}(n)} \tag{144.16}$$

In order to keep the stability of the algorithm, momentum factor α is quoted in the weight:

$$\Delta\omega_{\mathrm{lm}}(n) = -\eta\frac{\partial E(n)}{\partial\omega_{\mathrm{lm}}(n)} + \alpha\Delta\omega_{\mathrm{lm}}(n-1) \tag{144.17}$$

Theoretical analysis verifies this network structure, and mapping arbitrary nonlinearity function only hide layer quantity is enough. The input parameter is flux given value ψ_s^* and flux calculated value, the output variable is the reference voltage vector U_d. Activated function adopts to tansig. Hide layer unit is 4, according to learning rate and target error adjust the quantity of hide layer. The training result shows that the target error can receive less than 0.01 when hide layer quantity is 4, learning rate is 0.2 and training frequency.

144.6 System Simulation Results

144.6.1 Simulation and Results

In this section, the software Matlab/Simulink is used to simulate the whole DTC system to examine the performance of the novel SVM–DTC system. The parameters of the induction machine used in the simulation experiment are:

$$P_n = 2.2\,\mathrm{k}, U_n = 380\,\mathrm{V}, R_s = 4.35\,\Omega, R_r = 0.43\,\Omega,$$
$$L_s = 2\,\mathrm{mH}, L_r = 2\,\mathrm{mH}, L_m = 69.31\,\mathrm{mH}, J = 0.089\,\mathrm{kg}\cdot\mathrm{m}^2, P = 2.$$

The simulation conditions are given as: speed is 50r/min; simulation time is 0.8 s.

As shown in Figs. 144.5a, c, and e are simulation results of conventional SVM–DTC system at low speed, flux is basically round, but the ripple of flux is large;

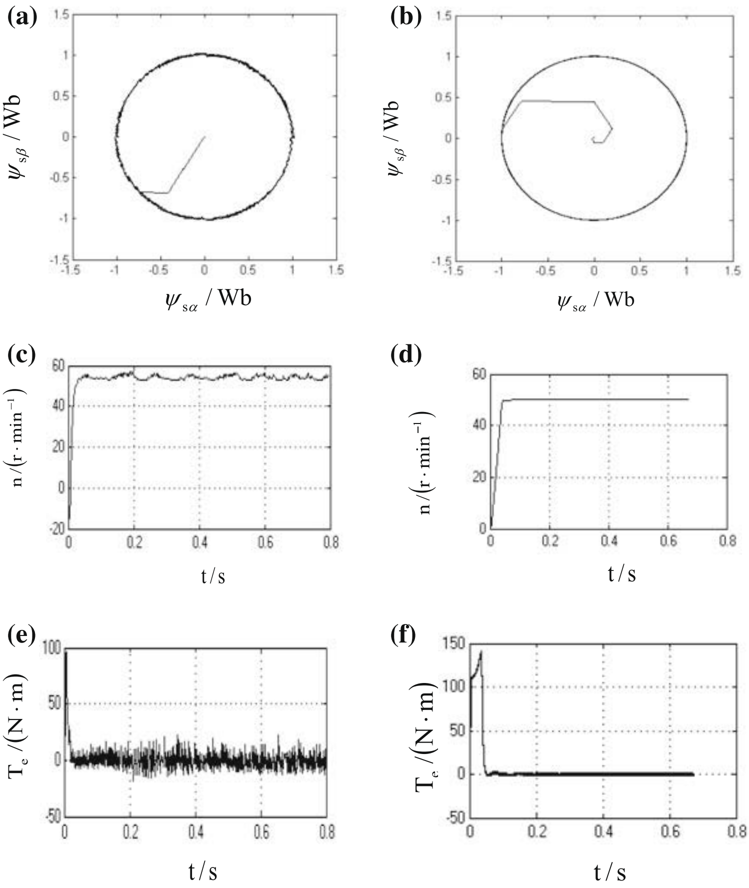


Fig. 144.5 Waveform graphs of simulation. **a** Conventional SVM–DTC flux. **b** Fuzzy-neural SVM–DTC flux. **c** conventional SVM–DTC speed.**d** fuzzy-neural SVM–DTC speed.**e** Conventional SVM–DTC torque. **f** fuzzy-neural SVM–DTC torque

speed response is slow and large overshoot is existed; there is large torque ripple at different jumping change time. Figures. 144.5b, d, and f are simulation results of fuzzy-neural network SVM–DTC system at low speed, waveform of flux is a standard round; speed response is fast and speed ripple decreased sharply; torque can achieve stability quickly and without ripple.

144.6.2 Experiment and Results

In order to verify the correctness of new strategy, the experiments were carried out in DSP2812 experiment platform. The parameters were same as to simulation. The experiment results are follows:

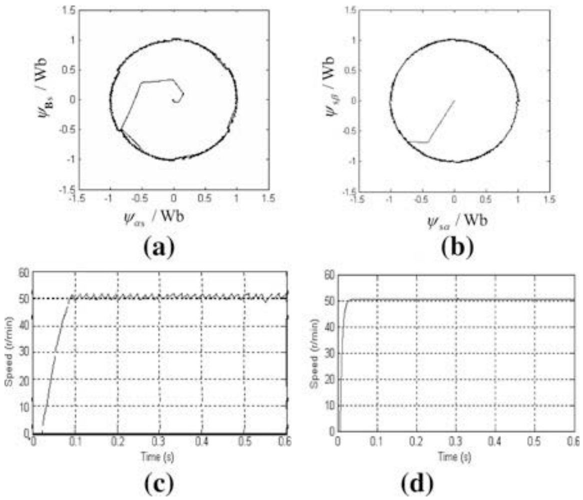


Fig. 144.6 Waveform graphs of experiment. **a** Conventional SVM–DTC flux. **b** Fuzzy-neural SVM–DTC flux. **c** Conventional SVM–DTC speed. **d** Fuzzy-neural SVM–DTC speed

It can be seen from Fig. 144.6 that the experiments results are similar to the simulation results. The new strategy has better torque and speed characteristic compared to the conventional VM–DTC. It can improve the control system's low-speed performance and the new method's validity is verified.

144.7 Conclusion

To resolve the problem in conventional SVM–DTC system of induction motors, a new strategy of direct torque control based on fuzzy-neural network space vector modulation is proposed and applied in induction motor speed control system. By system simulation and experiment, the results illustrate that the new strategy can sharply reduce torque, flux and speed ripple of SVM–DTC system and the system has a better dynamic and steady performance in low-speed.

References

1. Casadei D, Profumo F, Tani A (2002) FOC and DTC: two viable schemes for induction motors torque control. IEEE Trans Power Electron 17(5):779–787
2. Buja GS, Kazmierkowski MP (2004) Direct torque control of PWM inverter-fed AC motors—a survey. IEEE Trans Indus Electron 51(4):744–758

3. Wei X, Chen D, Zhao Chunyu (2005) A new direct torque control strategy of induction motors based on duty ratio control technique. Proc Chin Soc Electr Eng 25(5):93–97
4. Habetler TG, Profumo F, Pastorelli M, Tolbert LM (1992) Direct torque control of induction motor using space vector modulation. IEEE Trans Ind Appl 28:1045–1053
5. Xu Z, Rahman MF (2005) An improved stator flux estimation for a variable structure direct torque controlled IPM synchronous motor drive using a sliding observer. IEEE Trans Ind Appl 38(1):2484–2490
6. Mir S, Elbuluk ME (1995) Precision torque control in Inverter-Fed induction machines using fuzzy logic. Proc 26th IEEE Power Electr Specialists Conf (PESC) 1:396–401
7. Ding X, Liu Q, Ma X, He X, Hu Q (2007) The fuzzy torque control of induction motor based on space vector modulation. Third international conference on natural computation

Part XIII
Semantic Web and Intelligent Search

Chapter 145
A Survey of System Comprehension Based on Semantic Analysis

Yuanxun Yu, Lin Du and Daming Li

Abstract System comprehension is an important activity in software maintenance, as software must be sufficiently understood before it can be properly modified. The study of semantic analysis has become a common technique in this respect and has received substantial attention, particularly over the last decade. This paper reports on a systematic survey aimed at the identifying and structuring of research on system comprehension through semantic analysis. Three kinds of comprehension methods based on semantic analysis are reviewed: latent semantic analysis, program slicing and denotational semantics. The resulting overview offers insight in what constitutes the main contributions of the field, supports the task of identifying gaps and opportunities and has motivated more attentions in the future.

Keywords System comprehension · Latent semantic analysis · Program slicing · Denotational semantics

145.1 Introduction

Traditional software engineering is primarily focused on the development and design of new software. However, most programmers work on software that other people have designed and developed. Up to 60% of a software maintainers time

Y. Yu · L. Du (✉) · D. Li
School of Computer Science and Technology,
University of Qilu Normal,
Shandong 250014, China
e-mail: dul1028@163.com

Y. Yu
e-mail: yuyuanxun@163.com

D. Li
e-mail: deff_lee@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_145,

can be spent on determining the intent of source code. Recent years have witnessed the growing demand to re-evaluate and re-implement legacy software systems. One of the most important aspects of developing legacy software is to understand the software at hand. Understanding a system's inner workings implies studying such as artifacts source code and documentation in order to gain a sufficient level of understanding for a given maintenance task. In this paper, three kinds of system comprehension methods based on semantic analysis are discussed: latent semantic analysis, program slicing and denotational semantics.

The rest of the paper is organized as follows to survey system comprehension based on semantic analysis. Section 145.2 introduces some related works about latent semantic analysis. Section 145.3 presents some pioneering works in program slicing. In Sect. 145.4, we survey the works in denotational semantics. In Sect. 145.5, we conclude the whole paper.

145.2 Latent Semantic Analysis

Latent semantic analysis (LSA) is a technique for creating vector-based representations of texts which are claimed to capture their semantic content. The primary function of LSA is to compute the similarity of text pairs by comparing their vector representations. This relatively simple similarity metric has been situated within a psychological theory of text meaning and has been shown to closely match human capabilities on a variety of tasks. The related works are listed as follows, showing the developmental path of LSA, describing how it computes and uses its vector representations and then giving methods of the theoretical and empirical support for LSA and its current research directions.

In this paper [1], a novel Bayesian LSA framework is presented. The authors focus on exploiting the incremental learning algorithm for solving the updating problem of new domain articles. This algorithm is developed to improve document modeling by incrementally extracting up-to-date latent semantic information to match the changing domains at run time. By adequately representing the priors of LSA parameters using densities, the posterior densities belong to the same distribution so that a reproducible prior/posterior mechanism is activated for incremental learning from constantly accumulated documents. An incremental PLSA algorithm is constructed to accomplish the parameter estimation as well as the hyper parameter updating. Compared to standard LSA using maximum likelihood estimate, the proposed approach is capable of performing dynamic document indexing and modeling.

Martijn et al. [2] propose a hybrid recommender system that combines some advantages of collaborative and content-based recommender systems. While it uses ratings data of all users, as do collaborative recommender systems, it is also able to recommend new items and provide an explanation of its recommendations, as content-based systems do. The approach is based on the idea that there are communities of users who find the same characteristics important to like or dislike a product. This model is an extension of the probabilistic latent semantic model for collaborative filtering with ideas based on cluster wise linear regression. On a movie data set, it shows that the model, at the cost of a very small loss in overall performance, is able to recommend new items and give an explanation of its recommendations to its users.

Thomas [3] describes a model-based algorithm designed for collaborative filtering, which is based on a generalization of probabilistic latent semantic analysis to continuous-valued response variables. The observed user ratings can be modeled as a mixture of user communities or interest groups, where users may participate probabilistically in one or more groups. Each community is characterized by a Gaussian distribution on the normalized ratings for each item. The normalization of ratings is performed in a user-specific manner to account for variations in absolute shift and variance of ratings.

In the paper [4], the authors formulate a standard LSA model for behavior correlation modeling without considering any semantic context of a given scene. A novel two-stage hierarchical LSA model based on semantic scene decomposition is developed in order to improve the robustness of behavior modeling against noise resulting in reduced false alarms in anomaly detection. Specifically, local behavior correlations within each region are modeled. The inferred local behavior patterns are then fed into the second stage for global behavior inference and anomaly detection.

In the paper [5], text categorization models using back-propagation neural network (BPNN) and modified back-propagation neural network are proposed. An efficient feature selection method is used to reduce the dimensionality as well as improve the performance. It constructs a conceptual vector space in which each term or document is represented as a vector in the space. It not only greatly reduces the dimensionality but also discovers the important associative relationship between terms.

Jen et al. [6] propose two novel approaches to extract important sentences from a document to create its summary. The first is a corpus-based approach using feature analysis. The second approach combines the ideas of latent semantic analysis and text relationship maps to interpret conceptual structures of a document. Both approaches are applied to Chinese text summarization.

Zhang et al. [7] propose a approach structural LSA to model explicitly word orders by introducing latent variables. Specifically, the authors develop an action categorization approach that learns action representations as the distribution of latent topics in an unsupervised way, where each action frame is characterized by a codebook representation of local shape context. The effectiveness of this approach is evaluated using both the WEIZMANN dataset and the MIT dataset. Results show that the proposed approach outperforms the standard LSA. Additionally, the approach is compared favorably with six existing models including GMM, logistic regression and HCRF given the same feature representation.

145.3 Program Slicing

Program slicing is a technique for simplifying programs by focusing on selected aspects of semantics. The idea behind all approaches to program slicing is to produce the simplest program possible that maintains the meaning of the original program with regard to this slicing criterion. The process of slicing deletes those parts of the program which can be determined to have no effect upon the semantics of interest. As a viable method to restrict the focus of a task to specific subcomponents of a program, program slicing has extensive applications in software engineering, for example program debugging, program testing, software measurement and software maintenance. The recent works about program slicing are as follows.

Richard et al. [8] define a program semantics that is preserved by dependence-based slicing algorithms. It is a natural extension, to non-terminating programs, of the semantics introduced by Weiser (which only considered terminating ones) and, as such, is an accurate characterization of the semantic relationship between a program and the slice produced by these algorithms. Unlike other approaches, apart from Weiser's original one, it is based on strict standard semantics which models the normal execution of programs on a von Neumann machine and, thus, has the advantage of being intuitive. This is essential since one of the main applications of slicing is system comprehension.

The traditional method of program slicing was based on reach ability algorithm of program dependence graph (PDG) and system dependence graph (SDG). However, to construct PDG and SDG, some data dependence which were irrelevant to the slicing may be computed. The redundant computing wasted time and memory, and reduced slicing efficiency. To address this problem, the authors [9] present a slicing algorithm based on reverse program flow. It firstly constructed reverse flow of the program, then scanned the program along reverse flow from the slicing point, and only computed the data dependences which were relevant to slicing. So it improved slicing efficiency.

Horwitz et al. [10] present callstack-sensitive slicing, which reduces slice sizes by leveraging the series of calls active when a program fails. It describes a set of tools that identifies points of failure for programs that produce bad output and apply point-of-failure tools to a suite of buggy programs and evaluate callstack-sensitive slicing and slice intersection as applied to debugging. Callstack-sensitive slicing is effective. On average, a callstack-sensitive slice is about 0.31 time the size of the corresponding full slice, down to just 0.06 time in the best case.

Rupak et al. [11] present an algorithm that combines test input generation by execution with dynamic computation and maintenance of information flow between inputs. It iteratively constructs a partition of the inputs, starting with the finest (all inputs separate) and merging blocks if a dependency is detected between variables in distinct input blocks during test generation. Instead of exploring all paths of the program, it separately explores paths for each block (while fixing variables in other blocks to random values).

In the paper [12], the authors propose a dynamic path slicing technique for object-oriented programs. Given an execution trace of an object-oriented program and an object created during the execution, a path slice per object with respect to the object, or PSPO, is a part of the trace such that (1) the sequence of public methods invoked on the object in the trace is same as the sequence of public methods invoked on the object in the slice, and (2) given a method invocation in the slice, the state of all objects accessed by the method is same in both the trace and slice.

In the paper [13], a method for generating a slice from a plan-based representation of a program is provided. The method comprises constructing a plan representation of a program, wherein the plan representation comprises a plurality of nodes, edges and ports; and receiving one or more slicing criteria from a user. The slicing criteria comprise one or more variable occurrences or statements from the program, according to which a slice is generated from the plan representation.

Mastroeni and Zanardini introduced semantics-based data dependency both at concrete and abstract domain. This semantics-based data dependency is computed at expression level over all possible states appearing at program points. In the paper [14], the authors strictly improve this approach by (1) considering semantic relevancy of statements (not only expressions), and (2) adopting conditional dependency. This allows us to transform the semantics-based PDG into a semantics-based dependence condition graph that enables to identify the conditions for dependence between program points. The resulting program slicing algorithm designed this way is strictly more accurate than the Mastroeni and Zanardini's one.

During debugging processes, breakpoints are frequently used to inspect and understand runtime behaviors of programs. Although most development environments offer convenient breakpoint facilities, the use of these environments usually requires considerable human efforts in order to generate useful breakpoints. Before setting breakpoints or typing breakpoint conditions, developers usually have to make some judgments and hypotheses on the basis of their observations and experience. To reduce this kind of efforts, Zhang et al. [15] uses three well-known dynamic fault localization techniques in tandem to identify suspicious program statements and states, through which both conditional and unconditional breakpoints are generated.

Jiang et al. [16] present a new approach for locating faults that cause runtime exceptions in Java programs due to error assignment of a value that finally leads to the exception. The approach first uses program slicing to reduce the search scope, then performs a backward data flow analysis, starting from the point where the exception occurred, and then uses stack trace information to guide the analysis to determine the source statement that is responsible for the runtime exception.

In the paper [17], a kind of property extraction method is presented. Property model and dynamic slicing are combined to generate test sequence. As an example, the system structure of Minix3 is introduced. Exec, one of key system callings of Minix3, is modeling, slicing and its test sequences are generated.

Minix3 provides open interfaces and modular. The results of slicing can be used to improve the process of software reuse.

Seoul [18] applies hierarchical slicing technique to regression test selection in order to improve the precision of regression test selection and address the problem of level. The approach computes hierarchy slice on the modified parts of program, then selects test cases from different levels in terms of test case coverage. This approach can select test cases from high level to low level of program.

In the paper [19], base-line values for slice-based metrics are provided. These values act as targets for re-engineering efforts with modules having values outside the expected range being the most in need of attention. The authors show that slice-based metrics quantify the deterioration of a program as it ages. This serves to validate the metrics: the metrics quantify the degradation that exists during development; turning this around, the metrics can be used to measure the progress of a reengineering effort. "head-to-head" qualitative and quantitative comparisons of the metrics identify which metrics provide similar views of a program and which provide unique views of a program.

145.4 Denotational Semantics

Denotational semantics is an approach to formalize the meanings of programing languages by constructing mathematical objects (called denotations) which describe the meanings of expressions from the languages. Denotational semantics is concerned with finding mathematical objects called domains that represent what programs do. For example, programs might be represented by partial functions, or by actor event diagram scenarios or by games between the environment and the system. An important tenet of denotational semantics is that semantics should be compositional: the denotation of a program phrase should be built out of the denotations of its subphrases. The recent works about denotational semantics are as follows.

In the paper [20], a new method is presented to describe real-time process algebra (RTPA). Because some key RTPA processes cannot be described adequately in conventional denotational semantic paradigms, a new framework for modeling time and processes is sought in order to represent RTPA in denotational semantics. Within this framework, time is modeled by the elapse of process execution. The process environment encompasses states of all variables represented as mathematical maps, which project variables to their corresponding values. Duration is introduced as a pair of time intervals and the environment to represent the changes of the process environment during a time interval. Temporal ordered durations and operations on them are used to denote process executions.

Rodriguez-Lopez [21] presents a new mathematical model for the study of the domain of words. The authors do it by means of the introduction of a suitable balanced quasi-metric on the set of all words over an alphabet. It is shown that this construction has better quasi-metric and topological properties than several classical constructions. The authors also prove a fixed point theorem which allows us to develop an application for the study of probabilistic divide and conquer algorithms.

In this paper [22], the authors present a model and a denotational semantics for hybrid systems. The model is designed to be used for the verification of large, existing embedded applications. The discrete part is modeled by a program written in an extension of an imperative language and the continuous part is modeled by differential equations. The authors give a denotational semantics to the continuous system inspired by what is usually done for the semantics of computer programs and show how it merges into the semantics of the whole system. The semantics of the continuous system is computed as the fix-point of a modified Picard operator which increases the information content at each step.

Marcel Oliveira et al. [23] represent a denotational semantics for circus. Circus specifications define both data and behavioral aspects of systems using a combination of Z and CSP. Previously, a denotational semantics has been given to Circus; however, as a shallow embedding of Circus in Z, it was not possible to use it to prove properties such as the refinement laws that justify the distinguishing development technique associated with Circus. This work presents a final reference for the Circus denotational semantics based on Hoare and He's Unifying Theories of Programing (UTP). The authors discuss the library of theorems on the UTP that was created and used in the proofs of the refinement laws.

145.5 Conclusions

With the rapid development of software industry, there has been an amount of increasing of legacy software. One of the most important tasks of developing legacy software is to understand the software at hand. System comprehension is a difficult task of recovering design and other information from a software system. It is difficult to perform because there are intrinsic difficulties in performing the mapping between the language of high level design requirements and the details of low level implementation. What is more, system comprehension process is known to be very time-consuming. In this survey, we have discussed theoretical and empirical system comprehension methods based on semantic analysis (latent semantic analysis, program slicing and denotational semantics).The resulting overview offers insight in what constitutes the main contributions and pioneering works of the field. From above, we strongly believe that, in the near future, this research field will be paid more and more attentions and will promote the fundamental theories research in the related fields.

Acknowledgments This work is partially supported by the Shandong Province High Technology Research and Development Program of China (Grant No. 2011GGB01017); Soft Science Research Program of Shandong Province.

References

1. Jen T, Meng S (2008) Adaptive bayesian latent semantic analysis. Audio Speech Lang Process 16(1):198–207
2. Kagie M, van der Loos M, van Wezel M (2009) Including item characteristics in the probabilistic latent semantic analysis model for collaborative filtering. AI Commun 22(4):249–265
3. Hofmann T (2003) Gaussian latent semantic models for collaborative filtering. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval
4. Jian L, Shao G (2008) Global behavior inference using probabilistic latent semantic analysis. In: Proceedings of British machine vision conference
5. Bo Y, Zong B (2008) Latent semantic analysis for text categorization using neural network. Knowl Based Syst 21(8):900–904
6. Jen-Yuan Y, Ke H-R, Yang W-P (2002) Chinese text summarization using a trainable summarizer and latent semantic analysis. Lect Notes Comput Sci 2555:76–87
7. Zhang J, Gong S (2010) Action categorization by structural probabilistic latent semantic analysis. Comput Vis Image Understanding 114(8):857–864
8. Barraclougha RW, Binkley D, Danicic S (2010) A trajectory-based strict semantics for program slicing. Theor Comput Sci 411(11):1372–1386
9. Bai W, Xian Y (2009) Algorithm of program slicing based on reverse program flow. Appl Res Comput 26(3):920–922
10. Horwitz S, Liblit B, Polishchuk M (2010) Better debugging via output tracing and callstack-sensitive slicing. Softw Eng 36(1):7–19
11. Rupak M, Xu R-G (2009) Reducing test inputs using information partitions. Lect Notes Comput Sci 5643:555–569
12. Sudeep J, Jacob B, Koushik S (2009) Path slicing per object for better testing, debugging, and usage discovery. Technical report no. UCB/EECS-2009-132
13. Abadi (2011) Plan-based program slicing. In: Proceedings of united states patent application
14. Agostino C (2010) Dependence condition graph for semantics-based abstract program slicing. In: Proceedings of the tenth workshop on language descriptions
15. Zhang C, Yan D (2010) An automated breakpoint generator for debugging. In: Proceedings of the 32nd ACM/IEEE international conference on software engineering
16. Jiang S, Zhang C (2010) A debugging approach for java runtime exceptions based on program slicing and stack traces. In: Proceedings of the tenth quality software international conference
17. Seoul, Korea (2010) Test sequence generation from combining property modeling and program slicing. In: Proceedings of the 34th annual computer software and applications conference
18. Seoul (2010) An approach to regression test selection based on hierarchical slicing technique. In: Proceedings of 2010 IEEE 34th annual computer software and applications conference
19. Meyers TM, Binkley D (2004) Slice-based cohesion metrics and software intervention. In: Proceedings of the 11th working conference on reverse engineering
20. Tan Xinming (2008) A denotational semantics of real-time process algebra. Int J Cognit Inform Nat Intell 2(3):57–70
21. Rodriguez-Lopez J (2008) Denotational semantics for programming languages, balanced quasi-metrics and fixed points. Int J Comput Math 85:3–4
22. Olivier B, Matthieu M (2008) A hybrid denotational semantics for hybrid systems. Lect Notes Comput Sci 4960:63–77
23. Oliveira M, Cavalcantib A, Woodcock J (2007) A denotational semantics for circus. Electron Notes Theoret Comput Sci 187(15):107–123

Chapter 146
Information Fusion and Intelligent Pattern Recognition for Network Intrusion in Industrial Network Systems Based on ICA and PSO–ANN

Anqing Wu and Li Feng

Abstract The network intrusion may break down the local area network (LAN) and hence makes the industrial network system out of work, leading to terrible loss in the industry. Therefore, it is critical to detect the intrusion at the early stage in order to protect the security of industrial network system. The intelligent artificial neural network (ANN) provides an effective way for intrusion detection. However, the detection rate is influenced greatly by the structure design of the ANN. Insufficient design of the ANN-based intrusion detection model may decrease the detection accuracy. To deal with this situation, this paper presents a new intrusion detection model based on independent components analysis (ICA), particle swarm optimization (PSO) and ANN for industrial applications. The ICA was used to fuse the complex intrusion input and hence attain distinguished characteristics (that is, *independent components*, *ICs*) about the original data. By the use of *ICs*, the complex of the ANN structure design could be reduced. Then, the PSO was employed to optimize the structure parameters of the ANN. Experiments were carried out to evaluate the efficiency of the proposed approach. The ANN-based search engine was applied to import original data sets from the huge database in the industrial application system. The analysis results show that the proposed new method can offer satisfactory detection performance and thus, can be used in practice.

Keywords Network intrusion · ICA · PSO · ANN

A. Wu (✉)
School of Computer, Hubei University of Technology,
Wuhan 430068, China
e-mail: hgwuanqing@126.com

L. Feng
Wuhan Real Estate Board, Wuhan 430015, China
e-mail: xiaoanfengli@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_146,

146.1 Introduction

Network security has been studied very much in the field of information security. Terrible network intrusion may break down the internet and cripple the industrial network system. The loss and maintenance is very huge to the employer. Therefore, it is very essential to detect the network intrusion in a timely manner to prevent break-downs. Advanced machine learning algorithms, include evolution algorithm, intelligent artificial neural network (ANN) and support vector machine (SVM), have been researched for their application in network intrusion detection [1]. Among them, ANN [2] is the one of the most promising method. However, ANN detection performance relies on proper structure design. It is often difficult to determine the ANN parameters without a large number of trials. Although [3, 4] used PSO to optimize the ANN structure, they did not adopt information fusion to process the original data, and the redundant information might distort the detection results. Therefore, elimination of disturbed features and selection of elite ones have great significance for the ANN-based intrusion detection [5–9].

Due to the influence of the background noise, reliable intrusion data is often hard to acquire. Fortunately, the independent components analysis (ICA) is a useful tool to find a suitable representation of noisy data [10]. By the use of ICA, the distinct characteristics of the intrusion data can be separated effectively, i.e., by feature selection [11]. By doing so, the redundant features in the original data can be eliminated effectively. The applications of the ICA for the signal processing were represented in the literature [11]. However, reports seldom have addressed the network intrusion detection. Since the protection and prediction of the network intrusion is largely based on the ANN performance, it is essential to select distinguished features to reduce the complexity of the ANN model.

In order to tackle the problems mentioned above, a new network detection model based on the integration of ICA and PSO–ANN is proposed in this paper. In addition, we also provide an intelligent way for the retrieval of the intrusion data sets from the industrial database. Experiment results indicate that the proposed method can detect the intrusion efficiently.

This work is organized as follows. Literature review is described in Sect. 146.1. The proposed method for intrusion detection based on the combination of ICA–PSO-ANN is introduced in Sect. 146.2. In Sect. 146.3, the validation of the proposed method is investigated. The performance of the proposed new detection model is discussed. The effectiveness of the proposed method is valued by analyzing the practical data. Some conclusions are drawn in Sect. 146.4.

146.2 The Proposed Intrusion Detection Model

The potential link between the features and the network intrusion is very complex. It often presents high dimension property, and difficult to get accurate detection results. The high dimension of the feature space definitely decreases the detection rate. The reason is that there are certain dependent relationships among features, and some characteristics are redundant. Fortunately, ICA is a powerful dimension reduction method. ICA can extract mostof the important features from the feature space. Therefore, the ICA has been adopted to fuse the feature space with high dimension and map it into a low dimensional space. Meanwhile, ANN is an intelligent approach to deal with non-stationary signal. With its strong learning ability, the ANN is quite suitable for network intrusion detection. However, its identification efficiency depends on the structure design. This is the reason for the PSO to be applied for the ANN structure optimization. In addition, to enhance the possibility of practical use, the ANN-based search engine was employed to import original data sets from the huge database in the industrial application system.

In this paper, the PCA is firstly employed to sweep away the redundant features of the sample data, and then ANN is used to learn the patterns of the data. Additionally, the improved GA is adopted to optimize the ANN model.

146.2.1 Independent Components Analysis

For the intrusion detection, there are many characteristics that need to observed. The feature space sometimes has almost 50 dimensions to monitor the network condition. Due to the interference and noise, the measured input signals were distorted to some degree. All of these add to the difficulty, for one to detect the intrusion precisely. However, the independent component analysis (ICA) can recover the original sources in mixed observations with high dimensions. The distinct characteristics in the original sources could then be obtained. The ICA model was defined as follows [10, 11].

$$x = As, \tag{146.1}$$

where, $x = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}^T$ was an observation vector with m dimensions, $s = \begin{bmatrix} s_1 & s_2 & \cdots & s_m \end{bmatrix}^T$ was the unknown statistical independent of size n, and n was the ICs number of original sources. A was mixing matrix of the sources. The task of ICA is to estimate the matrix A to obtain its inverse matrix W. Then the independent components could be obtained by $s = Wx$. However, there is no direct way to get W because prior information about the mixing matrix is unknown [10]. The objective function was introduced to realize the estimation of ICA. Several objective functions include maximization of non-gaussianity, maximum likelihood estimation and high-order cumulant tensorial methods etc. [11]. Based on these objective functions, several ICA algorithms, such as FastICA algorithms, nonlinear PCA algorithms, infomax algorithms etc. [10], were developed to estimate independent components. FastICA was one among the fast algorithms for ICA estimation [11]. The convergence speed of FastICA was 10–100 times faster than the conventional gradient descending algorithm [4, 10]. Thus, the FastICA was adopted in this paper.

The FastICA algorithm using neg-entropy presented the following iteration [11]:

$$w(k) = E\left\{xg(w(k-1)^T x)^3\right\} - E\{g'(w(k-1)^T x)\}w(k-1) \tag{146.2}$$

where the weight vector W is normalized to unit after every iteration, and the function $g(\cdot)$ is the derivative of nonlinearity function. The RBF nonlinear function was used for ICA in this work.

146.2.2 PSO-ANN

PSO is similar to genetic algorithm and iterated optimization. However, PSO does not update offspring through the crossover and mutation, but according to the flying experience of the particles themselves and companions to adjust its flight. The best position experienced for each particle in flying is the optimal solution to find the particle itself,which is called the individual extreme (P); the global extreme (G_b) experienced for the whole populations is the optimal solution at present. The algorithm for updating Speed $\upsilon$and position Pin the PSO is that [3]:

$$\upsilon_i(k+1) = \omega_i \upsilon_i(k) + c_1 r_1 [P_b - P_i(k)] + c_2 r_2 [G_b - P_i(k)], \tag{146.3}$$

$$P_i(k+1) = P_i(k) + \upsilon_i(k+1) \tag{146.4}$$

where, ω_i is the inertial factor, r_1, $r_2$are the random among [0,1], c_1, c_2 are the learning factors, usually select 2.

The ANN usually uses BP NN and RBF NN. The details of the ANN are not given in this paper. The theories present in BP NN and RBF NN can be refered to the literature [12]. The ANN's network structure and weight parameters play critical roles in the intrusion identification. Hence, the PSO is used to optimize these parameters in this work.

146.2.3 ANN-Based Search Engine

With the rapid development of computer science, the web-based industrial database technique has been gradually used in practice. By intent, the industrial network system can offer data to every client at the same time. The database is very huge and complex in the industrial network systems. It is not convenient and reliable to import data responding to personalized search. The ANN based search engine can provide advantages to avoid the influences generated by human factors and get accurate search results. So it is very reasonable to establish ANN-based search engine to choose network intrusion data.

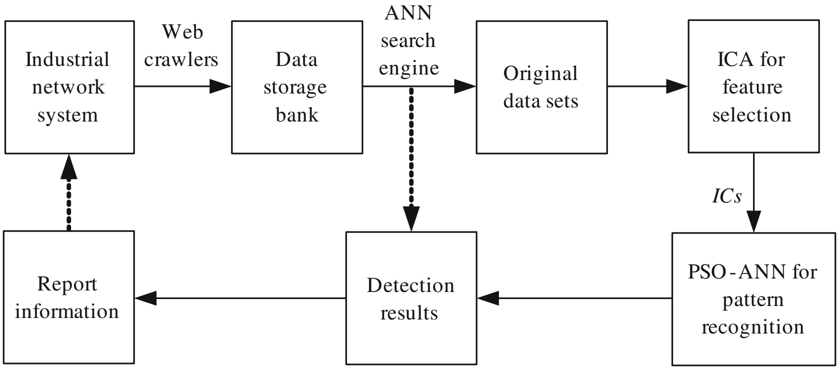


Fig. 146.1 The proposed network intrusion detection system

ANN is used to optimize the personalized intelligent search engine for building user characteristic model and the index characteristic model. It first calculates the variable weights of each link page for personal user, and then integrates the variable sorting weights with traditional fixed sorting weights to derive the new sorting weights. By doing so, the retrieving results can be optimized, and thus provide more precise retrieval results to realize the personalized search engine optimization. Figure 146.1 shows the proposed network intrusion detection system based on ICA and PSO–ANN.

The processing steps are as follows:

(1) The web crawlers find the url in web-based industrial network system and establish the data storage bank.
(2) The ANN-based search engine is used to get the original data according to the characteristics of the intrusion.
(3) The ICA is applied to the information fusion.
(4) The PSO–ANN detection model is established to identify the intrusion.
(5) The ANN-based search engine is used to search the detection results, and extract readable information to feed the industrial network system.

146.3 Experimental Analysis

In order to validate the performance of the proposed algorithm, the intrusion experiments were carried out in real practice application in this paper. The ANN-based search engine was used to get the original intrusion data with 35 attributes of the test network connection, including duration, service type, the bytes issued from source to destination, the bytes from destination to source, etc. 2000 normal samples and 2000 intrusion samples were used to evaluate the proposed detection model.

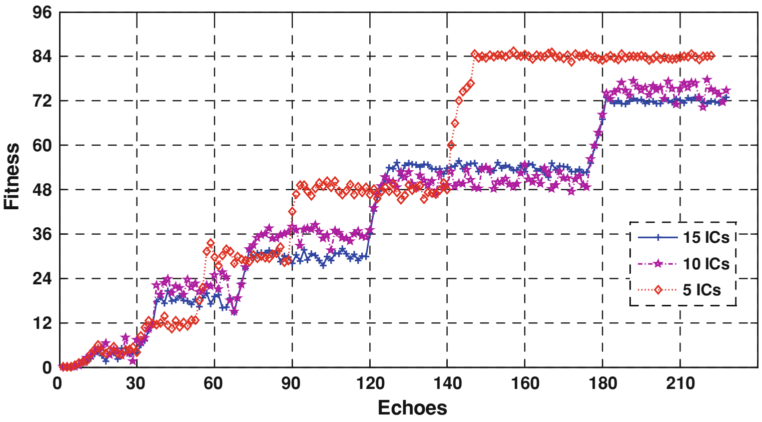


Fig. 146.2 The fitness value of PSO optimization with different *ICs*

Table 146.1 The intrusion detection performance of BP NN

Features or *ICs*	PSO-BP NN model		BP NN model	
	Detection accuracy (%)	False alarm (%)	Detection accuracy (%)	False alarm (%)
5	95.8	1.15	90.3	1.39
10	92.1	1.61	88.5	1.86
15	90.2	1.94	86.3	2.09
35	88.6	2.33	85.8	2.57

In experiments, ICA was adopted to reduce the 35 dimensions into 5, 10 and 15 independent components, respectively.

The PSO–ANN is employed to identify the DDoS. We conducted three kinds of tests. That is, the input feature vector of the ANN is 5, 10 and 15 independent components, respectively. The PSO optimization procedure for the ANN with 5, 10 and 15 independent components is illustrated in Fig. 146.2. As it can be seen in Fig. 146.2, the optimization of 5 *ICs* is faster and better than 10 and 15 *ICs*. Hence, it is reasonable to use 5 *ICs* for intrusion identification.

The intrusion detection performance of ICA-PSO-BP NN model using different input features is compared with ICA-BP NN in Table 146.1. From Table 146.1, the detection model with 5 *ICs* performs best compared to the other input feature numbers. In addition, the BP NN optimized by PSO can offer higher detection rate. The analysis results indicate that the ICA fusion can improve the BP NN detection efficiency, and the PSO optimization further enhances the identification accuracy.

Table 146.2 compares the detection performance of ICA-PSO-RBF NN model and ICA-RBF NN. Similar detection results can be observed in Table 146.2. The ICA processing can provide more accurate detection rate than that without ICA, and the PSO can optimize the RBF NN structure to improve its learning

Table 146.2 The intrusion detection performance of RBF NN

Features or *ICs*	PSO-RBF NN model		RBF NN model	
	Detection accuracy (%)	False alarm (%)	Detection accuracy (%)	False alarm (%)
5	96.5	1.03	91.1	1.22
10	92.8	1.53	89.7	1.78
15	91.3	1.76	87.1	1.95
35	89.4	2.16	86.6	2.27

ability. One can note from Tables 146.1 and 146.2 that the RBF NN outperforms the BP NN with respect to the detection rate and false alarm.

146.4 Conclusions

The web-based industrial database technique can provide sub-departments with the convenient operation environment. However, once intrusion comes, the system may crash down, leading to serious economic loss. To protect the safety and security of the industrial system, a new network intrusion detection model based on ICA and PSO–ANN is proposed in this paper. By taking full advantage of ICA, distinct characteristics of the original intrusion data can be extracted. The analysis of industrial system was implemented to verify the proposed model. The analysis results show the effectiveness of this new detection model. Thus, the proposed detection model based on the integration of ICA and PSO–ANN is feasible for practical applications.

Acknowledgment This project is sponsored by the grants from the Natural Science Foundation of Hubei Province of China (No. 2007ABA017).

References

1. Bace R (2000) Intrusion detect ion. Macmillan Technical Publishing, New York
2. Jiang J, Ma H, Ren D et al (2000) A survey of intrusion detection research on network security. J Softw 11:1460–1466
3. Xiong W, Wang C (2009) Hybrid feature transformation based on modified particle swarm optimization and support vector machine. J Beijing Univ Posts Telecommun 32:24–28
4. Anderson A, Marcelo D, Roberto S (2009) Particle swarm optimization applied to the nuclear reload problem of a pressurized water reactor. Progr Nucl Ener 51:319–326
5. Li Z, Yan X (2011) Application of independent component analysis and manifold learning in fault diagnosis for VSC-HVDC systems. Hsi-An Chiao Tung Ta Hsueh 45:44–49
6. Li Z, Yan X, Yuan C, Zhao J, Peng Z (2011) Fault detection and diagnosis of the gearbox in marine propulsion system based on bispectrum analysis and artificial neural networks. J Marine Sci Appl 10:17–24

7. Li Z, Yan X, Yuan C, Peng Z, Li L. Virtual prototype and experimental research on gear multi-fault diagnosis using wavelet-autoregressive model and principal component analysis method. Mech Syst Signal Process. 10.1016/j.ymssp.2011.02.017
8. Li Z, Yan X, Yuan C, Zhao J, Peng Z (2010) New method of nonlinear feature extraction for multi-fault diagnosis of rotor systems. Noise Vibr Worldwide 41:29–37
9. Li Z, Yan X, Yuan C, Zhao J, Peng Z (2010) The fault diagnosis approach for gears using multidimensional features and intelligent classifier. Noise Vibr Worldwide 41:76–86
10. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley, New York
11. Hyvärinen A (1999) Survey on independent component analysis. Neural Comput Surveys 2:94–128
12. Zhou K, Kang Y (2005) Neural network model and its simulation design suing MATLAB. J T Singhua University Press, Beijing

Chapter 147
Agent-Based User Model for Personalized Retrieval System

Lijun Xu, Xianfeng Yang and Jun Chen

Abstract Information retrieval system for the current over-emphasis on the recall of information, the information retrieved and the user needs not related to most of this limitation, combined with Agent technology has the autonomy, responsiveness, cooperation, learning and other characteristics, to construct an Agent-based personalized search system user model, to provide users with a personalized search service.

Keywords Agent · Retrieval system · Information retrieval

147.1 Introduction

With the rapid development of the Internet and the World Wide Web, Internet content and information has a dramatic growth. The quality of search results in users' increasing demand. The current information service system to retrieve the results of a considerable part of the user need not be related to the fundamentals and cannot meet the needs of the user information retrieval. How to provide users with high quality and efficient personalized information services, information retrieval systems has become an urgent problem [1].

L. Xu (✉)
Institute of Computer and Information Engineering, Xinxiang University,
Xinxiang 453003, Henan, China
e-mail: xljwork@126.com

X. Yang · J. Chen
School of Information Engineer, Henan Institute of Science and Technology,
Xinxiang 453000, Henan, China
e-mail: yangxianfeng@hist.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_147,

To solve the above problem, the design of a retrieval system is based on agent model. System user interest model is to realize personalized information service premise and foundation. A sound retrieval system gives users not only some sort of demand, but also with users who search for interesting content to the user needs. Agent autonomies for itself, response, cooperation and learning and other characteristics, and its technology into the search engine personalized service system to record the user's personalized information. This information retrieval system provides a completely different mode of information retrieval which is expected to greatly increase the efficiency of information retrieval.

147.2 Agent Technology

The concept of Agent [2–4] comes from the artificial intelligence field, which is the hotspot and frontier personalized service research. There is no general opinion on the definition of Agent, who is a software entity which can continually complete autonomy, goal-oriented behavior in a heterogeneous computing environment [5].

The description of Agent includes state description and function description. The state description gives the behavior parameters of Agent; function description gives the Agent's behavior model; resilience describes the Agent's ability to learn new knowledge, the higher the resilience, the faster the speed of Agent to update their knowledge, the lower the resilience, the slower the speed; credibility is an evaluation of the Agent, the higher the credibility, the greater the impact of Agent on the returned results, otherwise smaller. Stability refers to a description of Agent's preference for stability, Agent has a user vector in the user template, which describes some of the user's constant preferences, and stability can be seen as the weight of this vector; function description is made up of the template vector, the variable vector and vector weight. Template vector is the vectorization of a user template initialized by the user, including the template of the user's basic interests; variable vector will reflect the preference model of the user's changes [6].

Through the above analysis, we can see that Agent has the following features:

(1) Autonomy: Autonomy is the core of Agent's concept, which means that when the main body runs, it can create realistic goal-related plans by itself on the condition of the absence of people and other Agents' direct interference or guidance, and complete the task initiatively as requested.
(2) Reactivity: Agent can sense its environment, and can respond to the environment-related events occurred, making it possible for active service.
(3) Cooperation: Agent can interact with other Agents through a kind of Agent Communication Language, the task that a single Agent cannot complete can be completed through multi-Agent's collaboration.
(4) Learning: Agent has the ability to learn, and can take advantage of the available information about the external environment to adjust and modify its behavior to improve the retrieval precision.

Agent technology embodies a new software developing model, which fully covers all aspects of the definition of user requirements, analysis and storage, information input, the needs matches and the results transmission etc., known as the intermediary for user to access information resources, and seen as a "future search engine" of the Media Lab of Massachusetts Institute of Technology (MIT) in US [7].

147.3 System Framework

Personalized user interface is the key link to achieve personalized information services in the future, mainly including the following two aspects: to build user interest model, which tracks user‘s behavior, learning and remembering use's interests; and to build a personalized service model, which is to extract and filter out personalized information from the global information space, providing users with personalized information services [8–10]. Therefore, the personalized user interface model designed in this article includes three components: the user interest model based on the Multi-agent System (MAS), personalized service module and personal manager, the three interrelate with each other through user interest model. The specific framework is shown in Fig. 147.1.

147.3.1 Introduction of the Function Modules

(1) The user interest model based on MAS

The construction of user interest model [6, 7] is an important aspect of designing personalized user interface. It is actually a function set to store the user's interest, to store and manage the user's historical behavior, to store the knowledge of learning the user's behavior and to carry on related derivation. It can be divided into two kinds: explicit and implicit.

(a) Explicit method

Constructing an explicit user interest model is usually by interacting with the user, using key words to get the user's interest, or require users to give the corresponding evaluation of searching results in order to get user information to build models. This method can accurately get the needs information of users, simply to build model, and easy to realize; the disadvantage is the bad initiative.

(b) Implicit method

The construction of implicit user interest model is built on the process of the interaction between the system and user, by analyzing the historical pages the user has browsed and the system's log files, mining the content of user models, this approach has strong initiative, and does not need the user to input interest information explicitly, but the building of the model is complex. This paper introduces

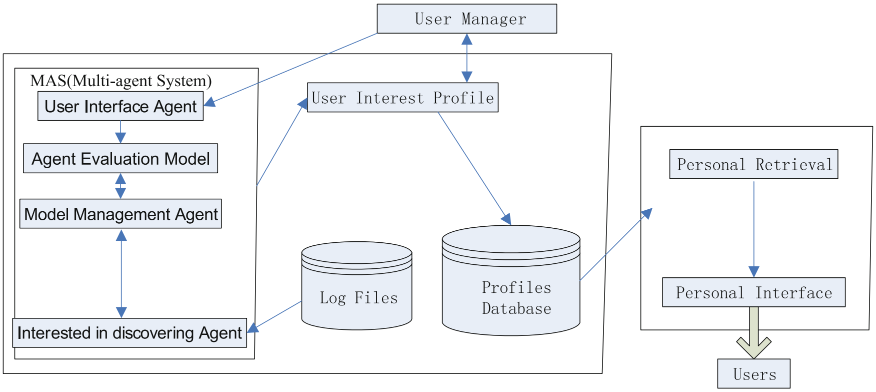


Fig. 147.1 Model for personalized user interface, the overall structure

a multi-agent technology to solve the problem, namely forming user interest models based on MAS.

MAS is a loosely connected network of many Agents working together to solve the problem that a single Agent cannot solve, and is of robust and high efficiency [11, 12]. MAS designed in this paper, includes five Agents which are the user interface, model evaluation, model management, model generation and interests discovering, that every Agent coordinates and cooperates with each other. The specific functions are described as follows.

User Interface Agent: Sending request to the system and receiving system's services, the purpose is mainly to provide a friendly interactive interface to the user and allow the user to choose the interface content provided by the system, introduce or customize related contents, adjust the interface structure, set interface style, and also learn from the user's appliance. It collects users' evaluation information and appliance conditions of the model, and transmits the information to the model evaluation Agent [13].

Model Evaluation Agent: Based on certain rules, evaluates the user's interest degree of the model and the interactive influence with the model management Agent.

Model Management Agent: It is the control center of the entire system. The specific functions are as follows: (1) Delete the interest model whose interest degree is lower than the specified threshold; when some models with high interest degrees get together, it can find the interest center of users based on clustering algorithms and calls model to generate Agent which generates a new model near the interest center, meanwhile continuously adjusting users' interest center according to users' feedback information on the model, making the new model more truly reflect users' interest. (2) when data sources come about more changes, model management Agent calls interest discovering Agent to find the user's new interests and introduces new knowledge to users. (3) When two users have similar interest-center, model management Agent model can push a user model to another user.

Model Generation Agent: It can provide users with a wizard and guides the user to build models; also can build models initiatively. Specifications is embodied in three aspects (1) When system initializes, provides wizards to guide users to build interest models for users' evaluation. (2) Build models near the users' interest center. (3) Build models in the weight vector's direction, of which the user has no attempt to build models yet, in order to find a new interest center which users are not aware of yet. The intelligence of model generation Agent is mainly reflected in the modeling algorithms.

Interests Discovering Agent: It has packaged a number of mining algorithms, such as Web log mining and bookmarks mining. It makes use of Web Mining to analyze users' historical accesses to discover useful user interest models. Basic methods used include clustering, association rules, sequential patterns, statistical analysis etc. In the process of using the user interest model, the user's interest is constantly changing, the key words that the user use are often changed, and the same keyword has multiple synonyms. It makes use of interests discovering Agent to track users' information behavior, from which to find information changes of users' interest needs, and dynamically adjusts the weight of user interest [14, 15].

(2) Personalized Service Module

Personalized service module includes three sub-modules which are personalized retrieval, personalized recommendation and personalized interface [16].

(a) Main functions of personalized retrieval module: according to the personalized retrieval algorithm, carries on retrievals to the outside information world, and separates the information needed from the global information, outputs personalized retrieval results. Specific retrieval mechanism belongs to search engine areas, and is not the main content that this interface model studies, the specific work is to be in future research [17].
(b) Main functions of personalized recommendation module: according to the current users' access situation, understands users' needs and interests, and finds out the user interest model which matches the current user's user model from the user interest model, carries on optimal reorganization of the collected information according to the matching results and recommends the user of the information which they are interested in, but has not been accessed to.
(c) Main functions of personalized interface module: according to the interface information described by the interest model of the current accessing user, gives personalized user interface, which includes the structure and layout, displaying the color and arrangement of the displaying contents and so on, to display the results of personalized recommendations.

(3) Personal Manager

Personal manager is the human interface device of the system, providing users with a self-management platform. Users can use it to manage their personal information, personal interests and personal bookmarks, etc., is mainly used to build an explicit model of user interest.

147.3.2 The Main Advantages of the System

(1) With users' needs as the center

The system puts users in the first place, gives full consideration to the users' needs information, and can according to the information provided by the user interest model, actively provide the information which users are interested in, but has not been accessed to the retrieval system together, which embodies the concept of "customer first" service purposes.

(2) With personality as the characteristic

It allows the user to fully express the information of their individual needs, and provides users with friendly interface, in order to allow users to clearly describe their needs, and facilitate human–computer interaction. The personalized user interface model which is designed according to the characteristic of different users' different needs of information can be "tailored" to achieve the personalized information services basing on specific users' needs.

(3) Having the intelligence and self-learning function

The system introduces multi-Agent technology, which can learn and record the user's habits and preferences, using the user's interests for reference in the process of information retrieval; on the other hand, when resources are updated MAS can respond to the changes of the environment, enable themselves to independently conduct dynamic maintenance and timely updates of users' interests.

147.4 Conclusion

This chapter will refer to the search engine agent technology personalized service system so that information retrieval can meet users with different backgrounds, different purposes and different times of the query requirements, system uses user interest model to provide users with intelligent, personalized information service, system representation of user interest model and update the personalized service to achieve the key technology. The system model of information retrieval services reflects the trend, with a strong theoretical and practical significance.

References

1. Chen S-P, Shan DS, Chengmei H (2004) Personalized web information agent research and development [J]. Shanghai Univ Technol 26(6):575–579
2. Lulu W, Zhongmin W, Zhanbo D (2004) Personalization based on agent active KDD system [J]. Comput Eng 30(13):130–132
3. Quinlan JR (1986) Induction of decision of decision tree. Mach Learn (6):101–105
4. http://www.siriusoft.net/products/netspider.htm
5. Nwana H (1996) Software agent: an overview. Knowl Eng Rev 11(3):205–224
6. Foster I, Kesselman C, Tuecke S (2001) The anatomy of the grid: enabling scalable virtual organizations. Int J Super Comput Appl 15(3):200–222

7. Zhongming MA, Gautam P, Sheng OR (2007) Interest based personalized search [A]. ACM Trans Inform Syst [C] NY
8. Hu S, Li C, Liu Y (2011) Design and implementation of the system of image retrieval based on mobile cloud computing. Value Eng (2):0345–349
9. Liang B, Wang G, Deng X (2011) Research and application of full-text retrieval model based on Lucene. Microcomput Appl (1):010–014
10. Zhang D-Z, Zhang M (2010) The retrieval system of the website based on XML. Comput Knowl Technol (2):046–054
11. Zhang P, Nie G (2010) Design and implementation of full-text searching system based on Lucene. Comput Knowl Technol (1):010–015
12. Song G (2010) Research on the evaluation of information retrieval system based on the behavioral model of information-searching. Document Inf Knowl (1):083–087
13. Zhang J (2010) Design of library's information retrieval system based on mobile agent. Res Libr Sci (1):25–29
14. Song M-H, Cao Y-X, Huang J-M (2010) Research on distributed personalized retrieval system model based on multi-agent. Inform Sci (4):035–038
15. Han Z, Ma W, Sun Y, Han Y, Cui S (2010) Design and implementation of the framework of small-scale information retrieval system based on SQL Server 2008. Comput Programm Skills Maintenance (10):111–113
16. Liu D (2010) Design and implementation of chinese thesis retrieval system based on XML. New Technol Libr Inform Service (5):023–028
17. Luo S (2010) The retrieval strategy mechanism for database group retrieval system based on Internet. Inform Sci (8):045–061

Chapter 148
Approach of Classification Based on Rough Set

Ying-juan Sun, Ying-hui Sun and Dong-bing Pu

Abstract Propose a novel approach of classification based on rough set. Fully consider each condition attribute significance and classification object during making policy. It gets higher classification accuracy and less decision rules without attribute reduction by this approach. Experiments have proved its validity.

Keywords Rough set · Attribute significance · Discretization · Classification

148.1 Introduction

Rough set theory, which was originated by a Poland mathematician Pawlak, is a kind of mathematic theory on analyzing imprecise data. Its characteristic is to find out the law of problem with similar fields, ascertained by indiscernible relations and classes, directly from data rather than their characteristics or descriptions given in advance [1].

Y. Sun
College of Computer Science and Technology Changchun
Normal University, Changchun, China
e-mail: syj_pyf@sohu.com

Y. Sun (✉)
College of Computer Jilin Normal University, Siping, China
e-mail: sunyh178@163.com

D. Pu (✉)
College of Computer Science and Information Technology
Northeast Normal University, Changchun, China
e-mail: pudb@nenu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_148,

Machine learning is to extract knowledge from data. Knowledge discovery, based on rough set, mainly utilizes the information system to represent knowledge and creates the knowledge system with data preprocessing, attribute reduction, rule generation, etc. [2]. Generally, each of records in a decision system is regarded as a decision rule. But there is no wide practical meaning for training samples. Therefore, we can find something in common among samples and obtain a few meaningful decision rules nothing but attribute discretization [3].

148.2 Relative Conceptions

148.2.1 Knowledge Representation System and Decision System

A decision knowledge system, also called a decision table, is a knowledge representation system in the form of $S = (U, A, V, f)$ along with condition attributes and decision attributes [4]. Here, U is a universe of non-empty finite set. $A = C \cup D$ indicates a non-empty finite set too. $C \cap D = \Phi$, $D \neq \Phi$, where C represents the condition attribute set and D is the decision attribute set. $V = \bigcup\limits_{a \in A} V_a$, where V_a is the value function of attribute a and V is a codomain. Function $f : U \to V_a$ is a single mapping. For $\forall x \in U, f : U \to V_a$ enable x own a unique value in V_a when x is given the attribute a [5, 6].

Definition 1 For $B \subseteq A$ in decision table S, if there are two different objects x, y with the same condition attribute value in the attribute set B and they belong to different classification respectively, then x and y are inconsistent in relation to B, else they are consistent in relation to B.

Definition 2 For any attribute subset $B \subseteq A$, the indiscernibility relation $\mathrm{IND}(B)$ is defined as

$$\mathrm{IND}(B) = \{(x, y) | (x, y) \in U \times U,\ \forall_{b \in B} \forall_{x \in U} \forall_{y \in U} (f(x, b) = f(y, b))\}.$$

Definition 3 The indiscernibility relation $\mathrm{IND}(B)$ divides U into such X_1, $X_2, \ldots, X_t$ as t equivalence classes. The dividing is denoted as $U/\mathrm{IND}(B)$. $[x]_B$ is a set in which all of elements exist in U and are equivalent to x by the action of $\mathrm{IND}(B)$[7, 8].

148.2.2 Description of Discretization

It is highly important for rule generation to discretize the continuous attributes in a decision system.

Assuming $V_a = [l_a, r_a]$, l_a and V_a are given by

$$l_a = c_0^a < c_1^a < \cdots c_{k_a}^a < c_{k_a+1}^a = r_a \tag{148.1}$$

$$V_a = \left[c_0^a, c_1^a\right) \cup \left[c_1^a, c_2^a\right) \cup \cdots \cup \left[c_{k_a}^a, c_{k_a+1}^a\right) \tag{148.2}$$

For any breakpoint set $\left\{\left(a, c_1^a\right), \left(a, c_2^a\right), \ldots, \left(a, c_k^a\right)\right\}$ in V_a, a classification P_a in V_a is defined as

$$P_a = \left\{\left[c_0^a, c_1^a\right), \left[c_1^a, c_2^a\right), \ldots, \left[c_{k_a}^a, c_{k_a+1}^a\right)\right\} \tag{148.3}$$

A new decision table $S^p = (U, A, V^P, f^P)$ can be defined by any $P = \bigcup_{a \in A} P_a$, where $f^P(x_a) = i \Leftrightarrow f(x_a) \in \left[c_i^a, c_{i+1}^a\right)$. For $x \in U$ and $i \in \{0, \ldots, k_a\}$, the old decision system will be replaced by a new one after discretization [9, 10].

148.2.3 Attribute Significance

The relevance, between condition attributes and decision attributes, reflects the significance of condition attributes in a decision table. Therefore, the number of potential values of decision attributes indicates the significance of condition attribute relative to a decision attribute when a condition attribute has a value. If potential values of decision attributes are unique when a condition attribute gets a value θ, the value of condition attribute can ascertain the decision attribute uniquely. As a result, we need not take into account other condition attributes whenever condition attributes own the value θ in the process of rule generating.

Definition 4 $M_a = 1/n \sum_{i=1}^{n} 1/l_i$ is the significance of the attribute a in a decision system, where $a \in A$, n is the radix of $V_a\left\{V_{a,1}, \ldots, V_{a,n}\right\}$, l_i is the number of potential values of the decision attribute when a is $V_{a,i}$.

148.3 Algorithm Descriptions

First, compute the significance of attribute for each attribute and order condition attributes by their significance. Second, descretize each attribute according to their descending order. At last get the rule set. Approach of classification based on rough set fully considers the significance for every condition attribute and the core of decision classification in the process of discretization. The algorithm is as follows.

148.3.1 Descriptions of Algorithm

Algorithm of classification based on rough set

Input: training sample set D;

Output: the decision rule table Dtable;

Let $S = (U, A, \{V_a\}, f)$ be a decision knowledge system, n is the number of condition attributes, $\{d\}$ is a decision attribute set and Dtable is empty initially, $S' = (U', A, \{V'_a\}, f')$ is the same as S in structure;

Step 1 Organize the sample set D into the decision system S after preprocessing;

Step 2 $S_0 = S$, $i = 1$, S' is empty and B is empty;

Step 3 Discretize every condition attribute in S with FCM. k is the number of cluster centers which is the number of decision classifications;

Step 4 Calculate the significance for every attribute in S. Sort all condition attributes in descending order on their significances. Suppose the order of condition attributes is $C_1, C_2, \ldots, C_n$;

Step 5 Assign the column of attribute C_1 and decision attribute in S to S';

Step 6 Add C_i to B. Let $X_1, X_2, \ldots, X_t$ be the partition of U' divided by $U'/\mathrm{IND}(B)$. Remove all partitions, which have no inconsistent objects, from S' and add all objects in them to Dtable. Suppose the remained partition is $X'\{X'_1, X'_2, \ldots, X'_{t'}\}$. At the same time, find out those objects corresponding to partitions removed from S' and remove them from S_0. Go to 15 if S' is empty;

Step 7 $\mathrm{i} = \mathrm{i} + 1$, go to 15 when $i > n$;

Step 8 Add the column of C_i in S_0 to S';

Step 9 Let the partition for U divided by $U/\mathrm{IND}(\{d\})$ be $XD\{Xd_1, Xd_2, \ldots, Xd_k\}$;

Step 10 Calculate intersections between each subset $X'_j (j = 1, 2, \ldots, t')$ in X' and that in $XD\{Xd_1, Xd_2, \ldots, Xd_k\}$.Get the partition $\mathrm{Int}_j\{\mathrm{int}_{j1}, \mathrm{int}_{j2}, \ldots, \mathrm{int}_{jk}\}$ $(j = 1, 2, \ldots, t')$.Suppose the partition set to all subsets of X' is INT $\{\mathrm{Int}_1, \mathrm{Int}_2, \ldots, \mathrm{Int}_{t'}\}$;

Step 11 For every $\mathrm{Int}_j\{\mathrm{int}_{j1}, \mathrm{int}_{j2}, \ldots, \mathrm{int}_{jk}\} (j = 1, 2, \ldots, t')$do step 12 to step 15

Step 12 Calculate the maximum value and the minimum value of all objects in $\mathrm{int}_{j1}, \mathrm{int}_{j2}, \ldots, \mathrm{int}_{jk}$, respectively and then generate partition intervals, where the minimum value is the left end point and the maximum value is the right end point in every partition interval. Suppose the codomain of C_i is divided into $\mathrm{Part}_j\{\mathrm{Part}_{j1}, \mathrm{Part}_{j2}, \ldots, \mathrm{Part}_{jk}\}$. We can get k intervals totally;

Step 13 Sort the interval sequence $\mathrm{Part}_{j1}, \mathrm{Part}_{j2}, \ldots, \mathrm{Part}_{jk}$ in ascending order on the position of left end point belonging to each interval.

Step 14 Select two intervals which have intersection but common endpoint and lie foremost in order of sorted intervals. Redivide intervals front-to-rear. That is to say, every two interfacing end points can ascertain an interval. Sort the interval sequence in ascending order on the position of left end point belonging to each interval again.

Step 15 Repeat 14 until any two intervals have no intersection but common endpoint

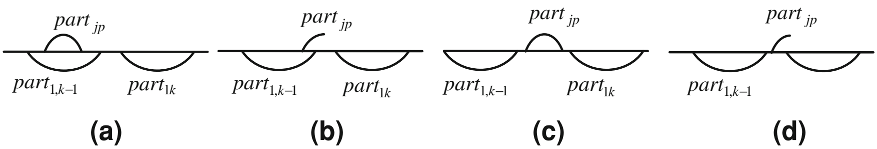


Fig. 148.1 Situations of combination part$_{jp}$ into part$_1$

Step 16 For every Part$_j$$(j = 2, \ldots, t')$ select every interval according on the left endpoint of the interval in ascending order. Insert the selected interval to Part$_1$ with the interval combination method in *C* section.

Step 17 If there is no common end point between two neighbor intervals of Part$_1$, then treat the mean value of neighbor end points as the common end point. Replace discretized intervals of C_i with discretized values. Update the value of attribute C_i in *S*' with a corresponding discretized value according to its interval and go to step 6;

Step 18 Delete all repetitive objects in Dtable.

148.3.2 Interval Combination Method

Combine Part$_j$$(j = 2, \ldots, t')$ to Part$_1$. Select one interval in Part$_j$. Suppose the interval is Part$_{jp}$. There are four cases for Part$_{jp}$ and Part$_{1k}$which is conterminous with Part$_{jp}$. How to combine Part$_{jp}$ with Part$_1$ is described as follows. There are four different situations (shown in Fig. 148.1) when descript the relationship between Part$_{jp}$ and a pair of intervals Part$_{1k}$ and Part$_{1,k-1}$ in Part$_1$. Firstly set the label not to be updated for each interval in Part$_1$ before Part$_{jp}$ being combined. Suppose Part$_{1,\mathrm{end}}$ to be the last interval of Part$_1$. This is to say Part$_{1,\mathrm{end}}$ has the biggest value of left end point in all Part$_1$'s intervals. Each interval is described as Part$_{jp} = [l_p, r_p]$, Part$_{1,k-1} = [l_{k-1}, r_{k-1}]$,Part$_{1,k} = [l_k, r_k]$ andPart$_{1,\mathrm{end}} = [l_{\mathrm{end}}, r_{\mathrm{end}}]$. We label every interval withflag(e.g., Part$_{1k}$ is not updated if flag(Part$_{1k}$) = 0, or else flag(Part$_{1k}$) = 1). Four different situations of combining Part$_{jp}$ into Part$_1$ are described as follows.

(a) Let flag(Part$_{1,k-1}$) = 1 if flag(Part$_{1,k-1}$) = 0, or else divide Part$_{1,k-1}$ into $[l_{k-1}, l_p)$ and $[l_p, r_{k-1}]$, number all intervals of Part$_1$ again sorting by the left end point of each interval in ascending order. The interval $[l_p, r_{k-1}]$ will be the interval k and set flag(Part$_{1k}$) = 1. There is not Part$_{1,k-1}$ when $k = 1$, so update *Part*$_{1k}$ with $[l_p, r_k]$ and set flag(Part$_{1k}$) = 1

(b) Update Part$_{1,k-1}$ with $[l_{k-1}, r_p]$, that is $r_{k-1} = r_p$, if $r_p < = l_k$. At the same time, divide Part$_{1,k-1}$into $[l_{k-1}, l_p)$ and $[l_p, r_{k-1}]$ with the left end point of Part$_{jp}$ if flag(Part$_{1,k-1}$) = 1, number all intervals of Part$_1$ again sorting by the left end point of each interval in ascending order. If $r_p > l_k$ and r_p belong to some one interval of Part$_1$, label the interval to be updated.

Table 148.1 Description of experimental data

Data	Number of condition attribute	Number of classification	Number of sample
Diabetes	8	2	768
Ionosphere	34	2	351
Iris	4	3	150
Wine	13	3	178
Glass	9	6	214

(c) If $\mathrm{flag}(\mathrm{Part}_{1,k-1}) = 0$, update $\mathrm{Part}_{1,k-1}$ with$[l_{k-1}, r_p]$, or else update Part_{1k} with $[l_p, r_k]$ and set $\mathrm{flag}(\mathrm{Part}_{1k}) = 1$.

(d) Update Part_{1k} with$[l_p, r_k]$. If the right end point of Part_{jp} belongs to some one interval of Part_1, label the interval to be updated, or else label the left neighbor interval to the right end point of Part_{jp}in Part_1 to be updated. If $k - 1 =$ end(i.e., $\mathrm{Part}_{1,k-1}$ is the last interval of Part_1), insert Part_{jp} into the last interval position of Part_1 directly.

148.4 Experiment Results

148.4.1 Descriptions of Data Set

In order to validate the performance of our algorithm, we have carried out a series of experiments with data samples in UCI which is a standard data set on machine learning. The description of experimental data set is shown in Table 148.1.

148.4.2 Experiment Analysis

We have completed ten experiments independently and obtained the average value through 10-fold cross validation technology. And we have verified the validity of our algorithm from five aspects, shown in Table 148.2, including the accuracy, average breakpoints, and the average length of rule, average rules and the standard deviation. Experiment results indicate that our algorithm can obtain 99.71% recognition rate in ionosphere data set. We have got better result from other data sets too.

148.5 Conclusions

We propose a novel approach of classification based on rough set. Fully consider each condition attribute significance and classification object during making policy. It gets higher classification accuracy and less decision rules without attribute

Table 148.2 Experiment result of algorithm in this paper

Data set	Accuracy	Average breakpoints	Average length of rule	Average rules	Standard deviation
Iris	0.9553	6.4200	2.8651	14.7500	4.99
Wine	0.8829	21.27	3.1584	63.7700	7.23
Ionosphere	0.9971	48.81	17.1401	54.2700	1.03
Diabetes	0.8282	111.3600	6.0148	512.1100	4.61
Glass	0.9400	68.0400	6.6548	129.3400	5.38

reduction by this approach. Experiments with data samples in UCI have proved its validity.

Acknowledgments This paper is supported by (1) Project of Research on Science and Technology of Jilin Education Ministry of China under Grant No.2007-172 and 2010-383. (2) Science-technology Development Project of Jilin Province of China under Grant No. 20115056. (3) The Natural Science Foundation of Changchun Normal University. (4) The master and doctor launched project of Jilin Normal University.

References

1. Li Y, Zhu S, Chen X (1999) Data mining model based on rough set theory [J]. J T singhua Univ (Sci Tech) 39(1):110–113
2. Feng L, Wang G, Li T (2009) Knowledge acquisition from decision tables containing continuous-valued attributes [J]. Acta Electron Sinica 37(11):2432–2438
3. Wang L, Wu G (2008) Measurements of Discretization Schemes [J]. PR AI 21(4):494–499
4. Zhao J, Ni C, Zhan Y, Du Z (2009) Efficient discretizathion algorithm for continuous attributes [J]. Syst Eng Electron 31(1):195–199
5. Pawlak Z (1982) Rough sets [J]. Int J Inf Comput Sci 11(5):341–356
6. Zhang D, Wang Y (2009) A new ensemble feature selection and its application to pattern classification [J]. J control Theory Appl 7(4):419–426
7. Yang M (2010) A novel algorithm for attribute reduction based on consistency criterion [J]. Chin J Comput 33(2):231–239
8. Wang J, Liang J, Qian Y (2008) Uncertainty measure of rough sets based on a knowledge granulation for incomplete information systems. Int J Uncertain Fuzziness Knowl Based Syst 16(2):233–244
9. Yue H, Yan D (2010) New algorithm for discretization based on information entropy [J]. Comput Sci 37(4):231–237
10. Zhang J, Qiu D (2009) Discretization method based on breakpoint importance in roughness set theory [J]. Meteorol Hydrol Marine Instrum 1:44–47

Chapter 149
Research on Search System Based on Agent and Personal Knowledge Ontology

Zenan Chu, Xinzhi Guo and Xinfa Wang

Abstract There is a lack of semantic comprehension in the traditional information search models. Now, Ontology provides a semantic-based knowledge representation and sharing which is a formal, explicit specification of a shared conceptual model. This paper introduces personal knowledge ontology, comprehensively utilizes various characteristics of an intelligent Agent such as initiative, collaboration, mobility and so on, and adopts the basic principle of collaborative filtering, and then puts forward the search system based on Agent and personal knowledge ontology, which greatly improves the search efficiency. Meanwhile, the paper presents an algorithm which is used to revise the concept of personal knowledge ontology through constant study of the feedback information from users, and also explores the application of multi-Agent collaborative information filtering in search system.

Keywords Personal knowledge ontology · Agent · Collaborative filtering

149.1 Introduction

Today, the stored information on the Internet has increased exponentially as the indexes, which brings about the problem of information overload. However, in searching information online, the traditional Internet search system, lacking the ability to understand the users' intention and filter the network information,

Z. Chu (✉) · X. Guo
Anyang Institute Of Technology, Huanghe Rd. 73,
Anyang 455000, Henan, China
e-mail: chuzenan@yahoo.com.cn

X. Wang
Henan Institute of Science and Technology,
Xinxiang 453000, Henan, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_149,

has increased users' burden of processing information. More often, the Information content provided by search engine is not relevant to customers' expectations, thus users have to spend a lot of time reading all sorts of information given by the search engineer and trying to locate any useful information. Therefore, it is currently an important development trend in information science to combine Semantic Web and the intelligent Agent technology in artificial intelligence so as to enhance the ability of processing and searching information, which will greatly improve users' efficiency in retrieving information.

Intelligent Agent, usually is based on intelligent technology, can work independently and has the capability of semantic interoperability and coordination. The features of Agent are as follows: first, Agent itself is an independent intelligence entity with reasoning and intelligent computing functions, which may independently discover and utilize all kinds of information resources and services according to users' individualized description, and then actively solve the problem and provide services for users. Besides, Agent has the collaborative property, which means different Agent can share and exchange information among each other and multi-Agent cooperates to accomplish users' tasks. Therefore, as the crystallization of distributed computer technology and artificial intelligence technology, intelligent Agent technique brings new solution to solve the problem of Web information retrieval. This article starts from individual requirement of retrieving information, and takes ontology as a mode of expression of personal knowledge to form personal knowledge ontology and make it as the basis of search system. Through the feedback from users, Agent may independently produce and modify the definition of each domain concept in personal knowledge ontology, and what's more, cooperates flexibly with multi-Agent to promote retrieval efficiency.

149.2 Personalized Search Model Analysis

As for researches on personal search model, there are many proposals both at home and abroad, the paper attempts to analyze the information search model based on Agent in Fig. 149.1 [1].

The traditional information search model based on Agent mainly includes the following parts:

(1) User

The user can search for any information he wants on the Internet. In such model framework, the user needs to describe his taste to the agent, and then the agent recommends information to the user according to his taste.

(2) Proxy-like gateway

The user may browse information through proxy-like gateway, and meanwhile the proxy-like gateway classifies URL and information content for the user and sends the message the user has browsed to the agent.

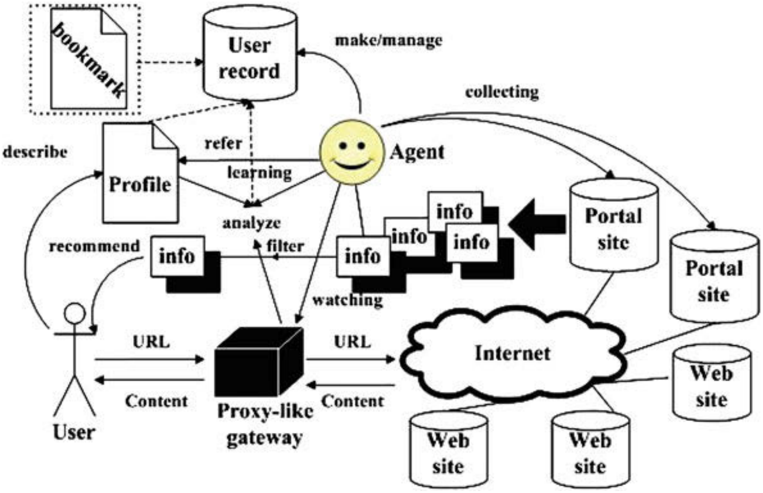


Fig. 149.1 Information search model based on agent

(3) Agent

The agent acts as if an information recommender. According to the pre-made preferences of the user as well as the information accessible to the user which is sent by the proxy-like gateway, the agent will analyze them statistically, classify, and then recommend the proper information to the user.

The traditional information search model framework exists the following problems: As for the message the users are interested in, the model will select information according to the classified catalog designed in advance, which makes it lacking the flexibility of the user-defined classification; Moreover, if the user's interest is constantly changing, such search model also lacks good adaptability; Since one agent is only responsible for the specific search requirement of a single user, the agent fails to provide services for other users through sharing mechanism.

To settle these problems above, we improve the traditional search engines, propose the search system model based on agent and personal knowledge ontology, and take personal ontology as the integral architecture. The knowledge ontology is composed of many concepts and the relations between different concepts, with each concept kept maintaining in the charge of different agents. Through constant revision of the concepts, it helps users search information and also classify the users with the similar demands, thus recommending the more appropriate agent for them.

149.3 Information Search Analysis Based on Agent and Personal Knowledge Ontology

149.3.1 Personal Knowledge Ontology

In order to improve the quality of information retrieval and introduce ontology into information retrieval system so that users can use the terms of formal definitions in ontology to express the needs for information in different domains, inference engine may provide users with incremental information service according to the relationship between the concepts which are represented by terms [2]. Studies have put forward that knowledge ontology provides a clear method for describing concepts of the knowledge in knowledge base. It includes a group of knowledge of terminology, such as the vocabulary of specific topics, the semantic interlinks and simple inference mechanism and logic [3].

However, when browsing web pages in different areas, they may find that the expressions in different pages are varied to refer to the same concept, or the expressions in different pages remain the same to present different concepts. This may be a big trouble for users. In fact, as for different users, their definitions of concepts differ from each other. Therefore, if the users possess a set of knowledge ontology of their own to express their understanding, it is not only convenient to them, but also can enhance the accuracy of the search.

At present, there are different kinds of knowledge ontology on the network, but they fail to gain wide application, the main reason of which may include the following [4]:

(1) Since different people may take on different viewpoints toward things, the knowledge ontology will thus be affected. Using pre-designed knowledge ontology may only present an one-sided concept.
(2) Usually users do not bother to maintain multiple intellectual ontology. But formulating a unified, typical ontology will reduce the credibility of ontology.

Thus personalized knowledge ontology is in urgent need for users to express their needs. But there is a discrepancy between personal knowledge ontology, just as that between ontology. It is mainly reflected in the following two aspects:

(1) On language level: inconsistency in syntax, divergence in logical representation, as well as problems in the description of the knowledge ontology, etc.
(2) On Ontology level: during the process of combining knowledge ontology, different people may have semantic differences towards the same concept, such as the conceptual issues [5], the model scope issues and so on. Studies have also proposed an alternative method, through which Reference Ontology and personal knowledge ontology are in one-to-one correspondence and meanwhile make Reference Ontology into personal knowledge Ontology. However, such method also has its problems, for the factors such as the

credibility and extent established by Reference knowledge Ontology will restrict the constitution of personal knowledge ontology.

This paper proposes an information retrieval method to construct personal knowledge ontology. This way can simplify the corresponding problems between the knowledge ontology, and also enable the users to build their personal knowledge ontology according to their own search requirements.

149.3.2 Definition and Revision of Concept in Personal Knowledge Ontology

We put forward a new method for defining the concept of personal knowledge ontology. Since people's understanding of concept is uncertain and it is difficult to define the concept range, we chose to adopt a feedback mechanism to define the concept. The basic ideas are as follows: in the initial stage, users will first define a concept according to their current requirements, represented by vector set:

$$((c_1, q_1), (c_2, q_2), (c_3, q_3), \ldots, (c_{n-1}, q_{n-1})(c_n, q_n)) \tag{149.1}$$

c: concept vector, q: weight of the concept vector

Users search information through the concept vector as required, and then send back to user's Agent what the user has browsed as well as the search number for other Agent intending to retrieve similar concepts to have browsed this document. After receiving feedback, Agent first deconstructs and analyses such documents, adopts the TF-IDF method to generate all their keywords and weight, and then revises user's concept vector according to them, and finally normalizes the weight of the revised concept vector. Therefore, as for the search for the same concept, the more users choose to browse a certain document, greater the effect will be made by its keyword on users' concept vector. Through constant feedback and modification, users' definition of this concept is consequently formed.

The concept in ontology also needs constant revision and extension so as to define the concept in a more complete and accurate way. Through analysis of the feedback from users' browsing, the arithmetic to revise the definition of the concept vector can be worked out as follows:

C: Vector set of one concept related to active user

$$((c_1, q_1), (c_2, q_2), (c_3, q_3), \ldots, (c_{n-1}, q_{n-1})(c_n, q_n)), \quad -1 < q_i < 1, i = 1 \ldots n. \tag{149.2}$$

D: concept vector selected from the feedback document

$$((d_1, q'_1), (d_2, q'_2), (d_3, q'_3), \ldots, (d_{m-1}, q'_{m-1})(d_m, q'_m)), \quad -1 < q'_i < 1, i = 1 \ldots m. \tag{149.3}$$

After getting the agent's feedback, the revised concept vector N is:

$$((N_1, Q_1), (N_2, Q_2), (N_3, Q_3), \ldots, (N_{n-1}, Q_{n-1})(N_n, Q_n)) \qquad (149.4)$$

N_i is combination set of c_i and d_i, $Q_i = (q_i + q_i' \times \alpha) \times \beta$, where α is the weight of the document generated by users' attention when browsing this document and the search number for other Agent intending to retrieve similar concepts to have browsed this document. When users and the agent of all relative concepts pay attention to this document, α will increase, thus may reflect users' demand effectively. β is a normalized constant which ensures that the weight of the revised concept vector range between 0 and 1. After a period of revision, we can abandon those vectors whose weight was smaller than a certain value (usually these concept vectors are considered as the deviation of the conceptual understanding), and in this way can we express clearly the definition of personal concept.

Agents can form and revise the concept automatically according to the feedback, analyze the information users have browsed, and search based on the revised concept to recommend information to users. It should be stressed that personal knowledge Ontology include multiple concepts, which are interrelated to constitute personal knowledge ontology of the user. A personal search agent will store the revised concept vector automatically, so users may make use of personal knowledge ontology when searching for information at any time.

149.3.3 Application of Multi-agent Collaborative Filtering Technology

Collaborative filtering is used for users to find the content that they are really interested in. First, find other users who have the same interests with this user, and then recommend to him what other users are interested in. The biggest advantage of Collaborative Filtering is that there is no need to analyze the feature attributes of the objects and no special requirements to recommend objects. Besides, it can handle unstructured complex object [6]. As for collaborative filtering, there are basically two kinds of methods:

(1) Cooperative filtering based on memory. First, get neighbor customers who have similar interests by adopting the statistic method, and then calculate them based on the neighbors, thus this method also called cooperative filtering based on customer.
(2) Cooperative filtering based on model. First, get a model with the historical data, and then use this model to forecast. At present, researches in this area are mainly focused on the improvement of cooperative filtering recommendation algorithm. For example, scholars such as Li Yu put forward the personalized recommendation algorithm under the hypothesis that users' multiple interests have similarity.

Based on the theory of cooperative filtering recommendation system, this paper proposes a method which is similar to cooperative filtering to search similar users, and make use of the search result of multi-agent to reduce the users' amount of information processing.

In the framework of Personal Knowledge Ontology based on agent, every concept has its name and concept vector, with which agent can accordingly compute the similarity between each agent. So these inter-correlated agents constitute agent network, through which a new search way is available for users. The basic idea is as follows: Based on the agent he already has, the user seeks out other similar agents through agent network, and then further finds the owners of these agents, thus forming a user group. So the user can use the agency owned by other users. While different agents represent different users' understanding of the similar concept, therefore, the multi-agent filtering can retrieve useful information more accurately.The formula of calculating agents' similarity is as follows:

$$\text{similarity}(A_i, A_j) = \alpha \times C(N_i, N_j) + \beta \times \left(\frac{\sum_{k=1}^{f} (W_{i,k} \cdot W_{j,k})}{\sqrt{\sum_{k=1}^{f} W_{i,k}^2} \sqrt{\sum_{k=1}^{f} W_{j,k}^2}} \right), \quad (149.5)$$

where $C(N_i,N_j)$ Calculates the similarity between two concept; N_i is a name of concept i; N_j: is a name of concept j.

The calculation method is (the common longest String length of two name)/(the longest String length of two name).

$\left(\frac{\sum_{k=1}^{f} (W_{i,k} \cdot W_{j,k})}{\sqrt{\sum_{k=1}^{f} W_{i,k}^2} \sqrt{\sum_{k=1}^{f} W_{j,k}^2}} \right)$: Inner product value of two concepts vector.

α and β are constant, $\alpha + \beta = 1$, $\alpha > 0$, $\beta > 0$.

This paper put forward a formula which is similar with cooperative filtering recommendation.

$$P_{a,j} = \overline{V_a} + k \sum_{i=1}^{n} W_{a,i} * (V_{i,j} - \overline{V_i}), \quad (149.6)$$

where $P_{a,j}$ is a interestingness of agent a's user to document j; V_a is a result of agent a dealing with feedback; $W_{a,i}$ is a similarity of agent a and agent i, it is calculated by formula (149.1); V_i is a result of agent i dealing with feedback; $V_{i,j}$ is a interestingness of agent i to document j

The relation beween k and $W_{a,i}$ is as follows:

$$\frac{1}{k} = \sum_{i=1}^{n} W_{a,i}. \quad (149.7)$$

Through this method, user agent can get the part of the search result that is concerned by the agent which has the similar concepts.

149.4 Conclusion

Based on the personal knowledge ontology, this paper adopts intelligent agent and cooperative filtering recommendation technology to put forward a search system based on Agent and personal knowledge ontology. By making use of agent's characteristics such as independence, initiative, coordination and mobility, the system constantly revise the concept by analyzing and studying the users' feedback so as to improve the efficiency of searching. In addition, by adopting cooperative filtering recommendation technology, the system may recommend to users the right agent which is better able to express personal interests to search information. However, the method of revised concept proposed in this paper can be further improved so as to adapt to high standard needs of information query.

References

1. Abe K, Taketa T, Nunokawa H (2000) An idea of the agent-based information recommendation system using the statistical information. ICPADS, IEEE
2. Xu H, Wang F (2005) Research on information retrieval mechanism based on ontology. J Inf 10:044–048
3. Hendler J (2001) Agents and the Semantic Web. IEEE Intell Syst 16(2):30–37
4. Lacher MS, Groh G (2000) Facilitating the exchange knowledge through ontology mappings. AAAI, Menlo Park
5. Chaffee J (2000) Personal ontology for web navigation. In: 9th International conference on Information and knowledge management 11:456–460
6. Yu L, Liu L, Li X (2004) Research on personalized recommendation algorithm for user's multiple interests. Comput Integr Manuf Syst 12:1105–1109

Chapter 150
Application Analysis on Network File System

Qingxiu Wu and Jun Ou

Abstract This paper describes the role of the network file system and application status. Then, on the basis of understanding and analyzing the implementation principle of network file system, it was installed and configured on the system based on Linux environment. Finally, the trend perspective of the network file system is also presented.

Keywords NFS · Handle · Server

150.1 Introduction

Sun Microsystems launched a widely remote file access mechanism accepted throughout the computer industry in 1984. It is known as the Sun's network file system (NFS). This mechanism allows a computer to run on a server, for some or all of the files on which it can remotely access and also allows other applications to access these files on the computer.

It can achieve the sharing of files. When users want to use remote files as long as the "mount" command on the remote file systems can be attached in its own file system, allowing remote file operations and file no different than the local machine. An application can open to access a remote file, and it can read from the file (Read) data, write to the file (Write) data, positioning (Seek) to a specified

Q. Wu · J. Ou (✉)
Department of Network Engineering, Hainan College of Software Technology,
Qionghai 571400, Hainan, People's Republic of China
e-mail: xhogh@hotmail.com

Q. Wu
e-mail: hncst0898@yahoo.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_150,

location in the file (start, end or elsewhere), the last when used closes (Close) the file. These operations are transparent to programming, methods of operation are exactly the same operations on the local file method. A client/server application designed by Sun Microsystems allows all network users to access shared files stored on computers of different types. NFS provides access to shared files through an interface called the virtual file system (VFS) that runs on top of TCP/IP. Users can manipulate shared files as if they were stored locally on the user's own hard disk.

With NFS, computers connected to a network operate as clients while accessing remote files, and as servers while providing remote users access to local shared files. The NFS standards are publicly available and widely used [1, 2].

150.2 NFS Upper Implementation

150.2.1 Mount Installation Protocol and NFS Remote Procedure

Implementation of NFS is separated into two separate programs to achieve, namely Mount installation agreement and NFS remote procedure calls. Mount Installation agreement is the beginning of file access. Its main function is to get a different file system structure on the remote machine and return the handle to the file system root to be accessed, as the later on the root of the file system operations.

150.2.2 Specific Processes of Access Files

In NFS, the system data structure called "file handle" achieves the targets file manipulation on a remote machine at a time as shown in Fig. 150.1. First of all, resolve the file name in the local. This process is similar to traditional UNIX file names in the resolution process. Analyze a part of a full path name. It begins from the root and the path of the hierarchy, and repeatedly removes from the path of the next section, and finds the one with the name of a file or subdirectory.

In the NFS, one destination file handle is not a step gained , but more steps have to be achieved. First, on the NFS server, the hierarchical file structure information is gained by the Mount installation Protocol, and the file system root handle obtained. After obtaining a handle to a remote file system's root, combined with the results of the local file name resolution, it can call the NFS remote procedure. In the current remote file system root handle, it checks out file handles in various subdirectories to return, checks for a file handle, and gets the last file to be accessed by handles. File handles are shown in the flowchart below. It shows the exchange of information between the client and the server when the client is able to find a path—/a/b/c file in the hierarchy of the server [3].

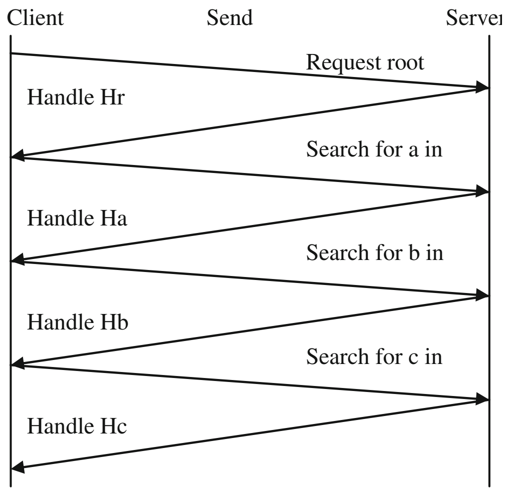


Fig. 150.1 The process of file access

150.3 The Lower of Network File System: Remote Procedure Call (RPC)

Remote procedure call model is based mainly in the procedure call mechanism in a traditional programming language. Procedure call provides a powerful abstraction and it allows the programmer to divide a program into smaller fragments, manageable, easy to understand. It can be performed to give the program's conceptual model of simplicity of implementation. Figure 150.2 shows how the remote procedure call model divides a program into two pieces, each executing on a separate computer.

RPC is a powerful technique for constructing distributed, client–server based applications. It is based on extending the notion of conventional or local procedure calling, so that the called procedure need not exist in the same address space as the calling procedure. The two processes may be on the same system, or they may be on different systems with a network connecting them. By using RPC, programmers of distributed applications avoid the details of the interface with the network. The transport independence of RPC isolates the application from the physical and logical elements of the data communications mechanism and allows the application to use a variety of transports. RPC makes the client/server model of computing more powerful and easier to program. When combined with the protocol compiler clients transparently make remote calls through a local procedure interface.

According to the process model, a single lead flows through the entire process. Computer begins execution from a main program and continues until a procedure call. The call is changed to a specified procedure code until a return statement is encountered. In RPC, a remote procedure call passed control to the calling process,

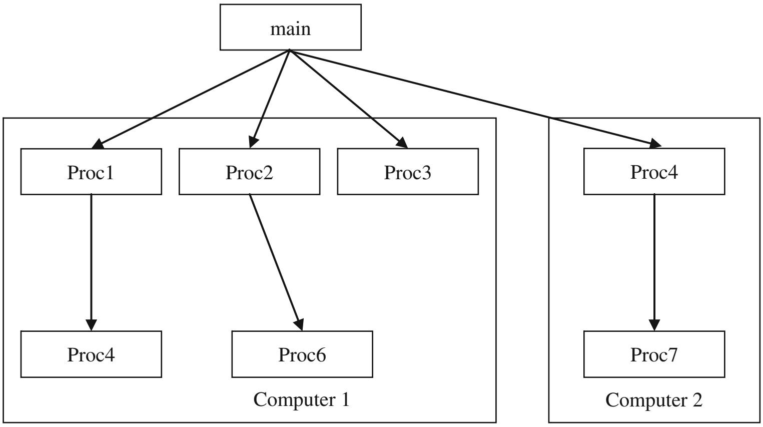


Fig. 150.2 Calling mode of RPC

pending execution of the called procedure. Remote servers for this process, when finished, gave the client a response, which corresponded to the return process model. On return of the control to the caller, the called procedure execution stopped. This process can be nested [4].

NFS is implemented by clients and servers together: on the customer side, calls to use the remote file system through a number of core functions; on the server side, provided by the NFS server to listen on the process to file data. There are two main listening processes: moutd and nfsd. Installation of the moutd is to listen for client requests, and send the appropriate response, such as client and server addresses, etc; and nfsd processes are to listen for client requests read and write files and return the corresponding file data. File access is completely transparent to the customer, and NFS can span a variety of servers and hosting platforms. Other file systems on the ground floor is through disk access, while NFS on the ground floor is through the RPC protocol for file access [5].

NFS's main advantage is that it can consume a large amount of disk space or user data. That is to say, shared on an NFS server only simply by NFS install transparent access to a local directory. Transparent access refers that the user interface is consistent between the general access to these files and local files, without the need for additional commands as shown in Fig. 150.3.

150.4 NFS Installation and Configuration

At present almost all Linux distributions have the NFS service, and it is installed by default such as Red Hat Enterprise Linux, because the NFS service needs the support both nfs-utils and portmap package, so before installing the NFS server determine whether in the system these two packages are already installed.

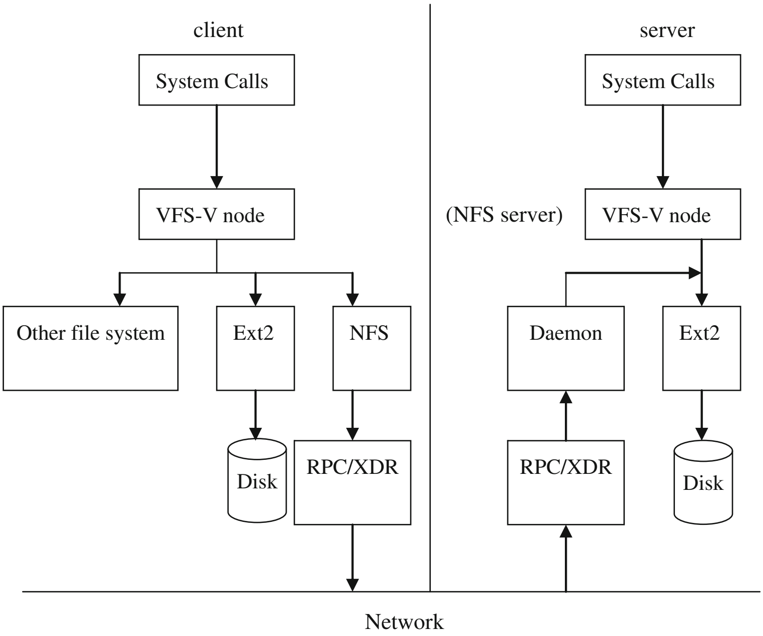


Fig. 150.3 Structure of NFS

General login as root user execute the command: rpm—q nfs-utils portmap.

System current NFS and portmap services are already installed. If these two services are not already installed on the system, look for Red Hat Enterprise Linux 4 x 2nd in the RPM installation package file portmap-4.0-63.i386.rpm, and nfs-utils-1.0.6-46.i386.rpm and install.

NFS service configuration method is relatively simple, simply set NFS in the main configuration file/etc./exports, and then start NFS services.

/mnt/share *(ro)

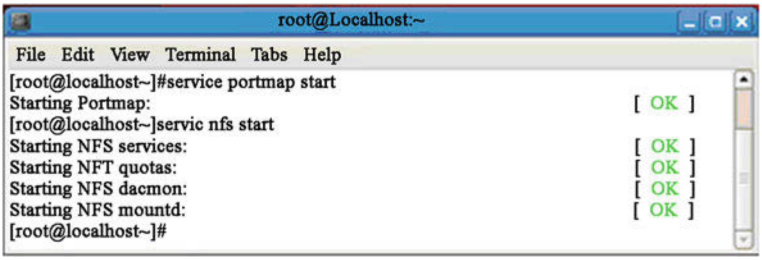
/mnt/share the output directory, available to any user connections to this computer in the network to read the read-only directory.

/mnt/cdrom 222.17.242.150(rw)

/mnt/cdrom the output directory, only for the IP address 222.17.242.150 clients to read and write operations.

150.5 NFS Server Running and Testing

To enable NFS servers work correctly, it needs to start portmap and NFS, and portmap must precede NFS's start. The result of command executing is as shown in Fig. 150.4.



```
[root@localhost~]#service portmap start
Starting Portmap:                                    [ OK ]
[root@localhost~]servic nfs start
Starting NFS services:                               [ OK ]
Starting NFT quotas:                                 [ OK ]
Starting NFS dacmon:                                 [ OK ]
Starting NFS mountd:                                 [ OK ]
[root@localhost~]#
```

Fig. 150.4 Start portmap and NFS

Use the Showmount command to test the output directory of the NFS server status, IP address of the NFS server to 222.17.242.34. The NFS Server configuration is completed. Clients can use the Showmount command learned when it can share a resource on a remote NFS server. The NFS server using the mount command (222.17.242.34)/mnt/shares shared directories mounted on to a directory on this computer for use. Before using the Mount command, first determine whether a local mount directory exists or not, if not create/mnt/nfs directory.

150.6 Summary

The server and client in the network system are relative concepts and a machine can be the server or client. Rational allocation of system resources using NFS can be cross-hosting. Shared file access across operating systems and file system format conversion between systems, and keeping the owner of the file group permissions in UNIX operating systems, the networked storage is a relatively new subject in the present international, which also occupies an important position in the network storage access methods. When an NFS server exports a subdirectory of a local file system, it leaves the rest un-exported. The NFS server must check whether each NFS request is against a file residing in the area that is exported. This check is called the subtree check.

To perform this check, the server includes information about the parent directory of each file in NFS file handles that are handed out to NFS clients. If the file is renamed to a different directory, for example, this changes the file handle, even though the file itself is still the same file. This breaks NFS protocol-compliance, often causing misbehavior on clients such as ESTALE errors, inappropriate access to renamed or deleted files, broken hard links, and so on.

Acknowledgments This work is supported by the Foundation of office of Education of Hainan Province under Grant Hjkj2010-55 (Colleges Scientific Research Projects of Hainan Province: Based on J2EE Project-teaching Interactive System of “2+1” Training Mode) and Hjsk2011-92 (Research on system maintenance talent training mode of ability-oriented).

References

1. Bi J, Xie P, Liu Y (2001) Network file system NFS [J]. Autom technol appl 02:131–135
2. Chen Y, Feng S (2003) NFS network file system and security [J]. Tuha oil gas 02:122–126
3. Wang J (1997) The comparison and analysis of several kinds of UNIX network file system [J]. Comput Commun 05:110–113
4. Liu Z, Qian W, Li S (2008) Research on data consistency of the network file system [J]. Nav Electron Eng 07:012–015
5. Choudhury BR (2008) Understand the new trends of NFS version 4 [J]. Financial Comput South China 09:0333–0339

Chapter 151
Ontology Relation Based Construction Algorithm of Characteristics Level

Gao Xiaobo, Fang Xianmei and Zhang Yanwang

Abstract In the Chinese opinion mining, the relevant scholars will focus on how to accurately receive the semantic emotion of opinion word as their breakthrough points, but the accurate access to features and characteristics of the relationship between the relatives were few studied. Correlation level analysis of characteristics will play an important role in the following semantic emotion analysis and understanding of the entire review. This paper describes the different concepts and definitions of ontology and characteristics level, and analyzes the existing construction algorithm of characteristics level. Finally, in the comments on the past different Chinese corpus, the word-level feature extraction algorithm proposed an improved method. After the analysis of specific grammatical structure in Chinese, the algorithm finds whether there are different characteristics of hierarchical relationships between the words with specific grammatical structures and Chinese internet commercial searching engine results.

Keywords Ontology · Data mining · Characteristics level · Algorithm

151.1 Introduction

With the wide use of Web 2.0 technology on the internet, people's online life has also undergone a major change [1]. Online shopping, online distance education, network news, teleconferencing, etc., and a variety of network applications

G. Xiaobo (✉) · F. Xianmei · Z. Yanwang
Department of Computer and Information Science, Hechi University,
Yizhou 546300, China
e-mail: aspone@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_151,

are springing up in people’s lives. With the vigorous development of various applications, the information people got also increased exponentially. However we can quickly and accurately get the information we need in such an explosion space of information, which is bound to be pursued by every modern man.

Facing this challenging problem, WEB data mining has been raised [2]. The data on the Web are not only the text with dispersion, but also arbitrary. It also includes all available information resources such as background transaction database and network user behavior. Therefore, Web mining technology is an integrated technology involving Web technology, data mining, computational linguistics, informatics and other fields. Text mining is in a dominant position here. Text mining is a research branch of data mining [3]. The technology applies to both text segmentation and feature extraction technologies; it could transfer the text data into structured text data to be described, and then uses the subclustering, association analysis and data mining technology to form structured text, in order to explore the concept and the concept of correlation described in the text.

The same as history tells us, the revolution of demand will drive another revolution of the technology. How to efficiently use this subjective information has become the new topic of Web data mining. Which comments are really needed? What are the comments made about? Faced with such a number of existing problems which can not be solved by search engines, opinion mining officially gets on the stage for Web data as a subject.

151.2 Ontology Concepts

Ontology, also known as entity theory, derived from the philosophy branch of metaphysics, the theory could date back to ancient Greek philosopher Aristotle (384–322 BC) [4]. In the philosophy, ontology is defined as “the system description of objective existences over the world, namely, ontology”, which is to study the abstract nature of objective existences, to study and decompose the objective things of world and to find each basic parts of the things. To a certain extent, ontology is a explanation or description about objective system.

The concept of this philosophy was first introduced to the study of natural sciences—artificial intelligence theory; Neches et al. first defined the ontology: “to give the basic terms and relationships which constitute related vocabulary, and use these terms and relationships vocabulary to definite the extension of these rules.” Since then, concepts and roles of ontology attract more and more attention in the field of natural sciences, and gradually become a research hotspot in many disciplines.

In essence, ontology is a set make up by concepts of certain field and relations between these concepts, relations reflect the constraints and contact between concepts. These relations can be regarded as a special concept; new relationship can exist between the relationships. It is knowledge organization approach the same way as traditional classification, thematic approach. The largest difference is that the theory can do better in organizing knowledge. Because of the characteristics of knowledge itself, the traditional methods generally only reflect the knowledge in one-dimensional tree structure; however, ontology can be represented with the network structure, which is a more complete and effective organization of knowledge.

The relationship between the concepts to be used in ontology construction:

Is-a relationship. In building the field concepts, the field concepts are divided into basic concepts, role concept and role owner. The field concept is "concept definition set" in general sense, the words or terms used in expressing the concept represent the common language used in the exchange of views between representatives. Therefore, in building ontology, field classification should be made based on "IS-A" relationship.

Part-of relationship. It constitutes the overall—part relationship between part concept and overall concept. The main semantics of "part-of" relationship is that when "overall concept" created, all its "part concept" instance is created accordingly. However, in field analysis, it should distinguish the basic concepts, role concept and role owners.

151.3 Construction Algorithm of Characteristics Level

The existing characteristics level extraction contains a lot of specific information extracted from regular text corpus and irregular text corpus. Irregular text is evaluation text edited from the web site. Therefore, the products evaluations in such text are completed complying with the specifications of the product, the terms in object description are not very strong randomness. Therefore a construction algorithm of characteristics level was specially proposed.

As follows:

Step 1: According to the writing feature of regular text, extract feature words and feature descriptors in the text.

Step 2: Process mutual information calculation on the result of word segmentation, merge them into the hierarchy based on the relationship degree between words, the common parts between words and the appearance location between words.

Step 3: Process word segmentation on the irregular text, using Bootstrapping algorithm for extraction of the characteristics of irregular text.

Step 4: According to the features of segmentation theme that feature word occurred, the different feature word should be lower theme. Similarity calculation should be processed between the extracted feature word and each level formed in Step 2; the new features of the product or characteristic attribute values as the highest similarity of the next level characteristics.

This algorithm can get the relationship between the characteristics levels, but in dealing with irregular corpus, there are often several comments object to a text description, as the irregular corpus cannot control the described content.

For example, the comments corpus on the screen, sometimes features word that has nothing to do with the screen, such comments appeared on the screen may concern about the music, shape, keyboard and other feature word. In this case, these characteristics are often classified as the lower screen features by the algorithm. In order to overcome this problem, we often need to manually delete the non-feature words. We need to avoid the classification algorithm of defects in corpus theme, so as to complete the process with more intelligent features hierarchically.

151.4 Grammatical Structure with the Concept Relationship of Ontology

Ontology collection is a collection of concepts and concept relationships. There are upper and lower levels between concept relations, or the overall and partial relations and so on. Overall and partial relations embody the relationships of entity concept and its parts. Any kind of entity is made up of many parts, then the ontology will have to reflect this collection of relationships, such as: computer is consisted by the monitor, case, keyboard, mouse, etc., and each concept can be composed of other concepts, for example: the case is a combination of the motherboard, hard drives, chips, shells and other case parts.

The upper and lower level is a fundamental and critical issue in Knowledge Acquisition from Text. Their acquisition methods can be divided into two categories: one is based on the model, which is mainly uses linguistics and natural language processing technology, obtain the upper and lower level mode through lexical analysis and syntax analysis, and then obtain upper and lower level relationship with pattern matching; another one is based on statistical methods, which is mainly based on corpus base and statistical language model, obtain the upper and lower level mode through cluster calculations. The model-based method extracts particular model of grammar structure appeared in the corpus text, and analyzes the structure of the phrase that appears in the concept relation according to the semantic model. In English, upper and lower level relationship between ontology concepts is mainly reflected by the sentence structure Is-a, etc. Such as:

NP_1\{“,”\} “Such as” NP $List_2$
NP_1\{“,”\} “and other” NP_2
NP_1\{“,”\} “including” NP $List_2$
NP_1 “is a” NP_2
NP_1 “is the” NP_2 “of” NP_3
“the” NP_1 “of” NP_2 “is” NP_3

It considers that the above phrase contains not only the NP phrase, but also the NP phrase relationship between the upper and lower level. For example: NP $List_2$ is the lower concept of NP_1 in NP_1\{“,”\} “Such as” NP $List_2$. In fact, the grammatical structure has another kind of relationship between ontology concepts. That is the relationship between part and overall. For example, NP_2 “of” NP_3 can be interpreted as part of Computer. Therefore, we could obtain the relationship between these two concepts through a specific syntactic structure. Then extract the particular mode using Know it all network information extraction system, finally determine which feature category the feature words in, according to the probability assessed value of PMI.

In Chinese, the concept of relationship is also in progress. In the article [4, 5], author makes the following modes as basis of words extraction:

Mode 1. < ? C1 > {< ? | or | otherwise | or | and | also |between > <?C2} < is | was | indicate |that is > .
Mode 2. < ? C1 > {< ? | or | otherwise | or | and | also |between > <?C1} < each | of | this > <varies | type |some | catalog |class > < of > .
Mode 3. < like| as | include | divide into | contain | embrace | cover | have | is | indicate |that is > <?C1 {< ? | or | otherwise | or | and | also |between > <?C2}.

Then form upper and lower relationship between a word in set C1 or C2 and the existing word. However, because there is ambiguity and words are affected by context, therefore, C1 or C2 in the collection of other words were made synonyms and LSA analysis, and finally form different relationships between the upper and lower through the concept of clustering.

151.5 Web-Based Word Mutual Information Algorithm

Mutual information is a measure of useful information; it refers to the correlation between the two sets of events. The mutual information between two events *X* and *Y* is defined as:

$$I(X,Y) = H(X) + H(Y) - H(XY) \tag{151.1}$$

where *H*(*X*,*Y*) is the joint entropy, which is defined as:

$$H(X,Y)= -\sum_{x,y} p(x,y)\log p(x,y) \tag{151.2}$$

However, in normal circumstances, we use the concept of pointwise mutual information, that is, the mutual information between two specific samples in calculation distribution.

The point mutual information is usually used to calculate the correlation degree between the two elements. Mutual information used very widely in natural language processing.

$$\mathrm{PMI}(\mathrm{word}_1, \mathrm{word}_2) = \log \frac{p(\mathrm{word}_1 \& \mathrm{word}_2)}{p(\mathrm{word}_1)p(\mathrm{word}_2)} \tag{151.3}$$

$\mathrm{PMI}(\mathrm{word}_1, \mathrm{word}_2)$ is the concurrence frequency of word_1 and word_2 in the corpus, $p(\mathrm{word}_1)$ and $p(\mathrm{word}_2)$is the frequency appeared alone in the corpus.

According to the consecutive appearance frequency of the two words in a large number of samples, we analyze the relations between single random variables (the appearance frequency of word_1,word_2) in the statistical sample. In order to determine whether the two words word_1 and word_2 is a word match.

Later, Oren Etzioni made the following improvements to PMI in his thesis:

$$\mathrm{PMI}(I,D)=\frac{|\mathrm{Hits}(D+I)|}{|\mathrm{Hits}(I)|} \tag{151.4}$$

$\mathrm{PMI}(D+I)$ is the occurrences times of phrase D and word I, $\mathrm{Hits}(I)$ is the occurrences times appeared alone of the word I. This improvement in his study is made for researching the relationship between phrases and words. Select a specific phrase containing the semantics. For example, LA is a city make X is a city as the phrase, words to be analyzed as: NY, London, etc. Use Web search engine results to search the frequency of NY is a city (London is a city) and NY (London), calculate the relationship of NY, London by PMI to determine whether it is a city. In the experiments, he did success. Combined with the context, we research the phrase and product feature words, and do the following improvements when using PMI algorithm:

$$\mathrm{PMI}(I,D)=\frac{|\mathrm{Hits}(D+I)|}{|\mathrm{Hits}(I)|} \tag{151.5}$$

word_1 and word_2 to be analyzed in two words, D is the conjunction of word_1 + "of" + word_2 and word_1 + "of" + word_2 in Chinese often equal to word_1 + word_2, for example: the size of the screen is often can be expressed as the screen size. White body can be expressed as white body. Therefore, use the network correlation engine to research terms; this must be taken into consideration. The expression "size of screen" may be not found at any web page, but "screen size" can be used. So the two words are considered as a transformation of the word "of" under the semantic template. It is a count of PMI ('screen', 'size') in Hits (screen + size).

151.6 Construction Algorithm of New Characteristics Level

Through the description about existing algorithm of characteristics level, although there are algorithms that have been proven can be very good product characteristics level in the analysis of the text corpus. But characteristics analysis of irregular algorithm with Bootstrapping, the algorithm itself, as candidate features is not well focused by the level of relationship. Therefore, the algorithm results in the Bootstrapping adds network-based PMI algorithm to generate the characteristics of the candidate collection characteristics level classification.

Algorithm steps is described as follows:

Step 1: Using the description in the regular text to analyze the characteristics level and characteristics hierarchy of corpus. In the calculation of the relationship between words when using the PMI algorithm based on network computing, it is no longer used PMI with irregular corpus, and then get a FC collection of characteristics level.

Step 2: Bootstrapping algorithm uses the feature of irregular text corpus to make extraction. Feature words are the candidate set *D*.

Step 3: Use specific rules of grammar corpus of non-performing operations. There is inherent hierarchy between the extracted features words. We could obtain the regular set *R*.

Step 4: Use the calculation relationship between characteristics of network PMI on the D word and each level of each intermediate node in the FC. In accordance with the principles of lower priority, feature words and the characteristics level merge.

Step 5: Bottom-up method for features distinguishing will be in different sublayers of the same conceptual distinction between feature words.

151.7 Conclusion

In the Chinese opinion mining, the relevant scholars will focus on how to accurately receive the semantic emotion of opinion word as their breakthrough points, but the accurate access to features and characteristics of the relationship between the relative was few studied. Correlation level analysis of characteristics will play an important role in the following semantic emotion analysis and understanding of the entire review. This paper describes the different concepts and definitions of ontology and characteristics level, and analyzes the existing construction algorithm of characteristics level. Finally, in the comments on the past different Chinese corpus, the word-level feature extraction algorithm proposed an improved method. After the analysis of specific grammatical structure in Chinese, the algorithm finds whether there are different characteristics of hierarchical relationships between the words with specific grammatical structures and Chinese Internet commercial searching engine results.

Acknowledgments This paper is supported by Research Project of Education Department in Guangxi (201010LX455).

References

1. Maoshu N (2007) Mining and studying on semantic understanding based opinions. Excell Master Eng 10:25–28
2. Kim S, Hovy E (2004) Determining the sentiment of opinion. In: Proceedings of the international conference on computational linguistics

3. Tianfang Y, Xiwen C et al (2008) A survey of opinion mining for texts. J Chin Inf Process 22(3):71–80
4. Turney P (2009) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the association for computational linguistics. USA, pp 417–424
5. Morinaga S, Yamanishi K, Tateishi K et al (2009) Mining product reputations on the web. In: Proceedings of knowledge discovery and data mining, Edmonton, pp 341–349

Chapter 152
Adaptive Variance Scaling in Clayton Copula EDA

Ru Yang, Li-Fang Wang and Jian-Chao Zeng

Abstract Estimation of Distribution Algorithms based on copula (cEDA) divide the estimated probabilistic model about the promising population into two parts, the marginal distribution of each variable and a copula function. The copula function combines the marginal distribution of each variable together as the joint distribution. The selection of marginal distribution can affect the optimization results. In the research on Clayton cEDA, empirical distribution and normal distribution are used as the marginal distribution respectively, the results are compared with each other, then the lack of using normal distribution is simply analyzed, an adaptive variance scaling strategy introduced to the algorithms to improve the optimization efficiency, and the experiments are used to illustrate the results.

Keywords Estimation of Distribution Algorithms (EDAs) · Copula theory · Clayton copula function · Marginal distribution · Empirical distribution · Normal distribution

R. Yang (✉) · J.-C. Zeng
Complex System and Computational Intelligence Laboratory,
Taiyuan University of Science and Technology,
Taiyuan 030024, China
e-mail: yr19831029@163.com

J.-C. Zeng
e-mail: zengjianchao@263.net

L.-F. Wang
Colloge of Electrical and Information Engineering,
Lanzhou University of Technology, Lanzhou 730050, China
e-mail: wlf1001@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_152,

152.1 Introduction

The estimation of distribution algorithms (EDA) is based on Genetic Algorithms. It does not use crossover and mutation, establishes probability distribution model by individuals of better fitness in the current generation, and then samples to get new individuals by the model. EDA uses the distribution feature of individuals to reflect the correlation of individuals, so it is important to establish a probability.

Copula theory is the research topic in economics and statistics field nowadays. It has great superiority in terms of establishing a model with large amounts of data. Copula theory and EDA have common property, that is, to establish a model from a sample and to sample by the model, so they have high consistency. Therefore, introducing copula theory into EDA (cEDA) will make the algorithm simple, and can establish accuracy.

Copula theory studies the marginal distribution of each variable and the correlation of variables respectively, the correlation between variables is expressed by a copula function. There is not much research about cEDA, applying binary copula into EDA [1–3] is earlier. The literatures [1, 2] select different copula functions as research subjects, and [3] apply copula function into MIMIc algorithm to replace the algorithm of condition normal. The latter is stablishes the distribution based on Clayton copula and empirical margins [4]. It is different from [1, 2] not only on copula function but also on it is high dimension function. However, all of these research mainly solve how to apply copula theory into EDA and how to sample. The research of the influence of margin distributions about effect of algorithms is shortage. We research the affect of margin distributions based on Clayton cEDA [4].

Nowadays, in continuous EDA, many scholars use Normal distribution as probabilities model, for example EMNA [5], EGNA [6], BOA [7] and IDEA [8]. These algorithms are highly complex in the term of times and spaces, and many scholars find that some of these algorithms have premature phenomenon, so the algorithm [9–14] which can improve the efficiency.

We compare margins using empirical distribution with normal distribution first, and then analyze the lack of normal distribution adopting the method in [15], and finally apply the method in [10] into cEDA,

152.2 The Estimation of Distribution Algorithm Based on the Copula

The two important operations in cEDA are estimating probability model with elite population and sampling by the model.

First, estimate probability mode; if C denotes D dimension copula function with variable $(u_1, u_2, \ldots u_D)$, $F_1(x_1), F_2(x_2), \ldots, F_D(x_D)$ represent margin distributions of each variable, then according to the copula theory, $H(x_1, x_2, \ldots, x_D) =$

$C(F_1(x_1), F_2(x_2)\ldots F_D(x_D))$ is the joint distribution of the D dimension variable, so in the cEDA, the work we need to do is selecting copula function and each margin distribution.

Second, sample by the model, the sampling process of cEDA [4] can be described simply as follows:

select S individuals composed of elite population by certain selection strategy based on the fitness in the current population, as $(x_1, x_2, \ldots, x_D)$;

establish the margin distribution model $F_i(x_i)$ of each variable by elite population;

sample from the copula function C. Generate m vectors $(u_1, u_2, \ldots u_D)$, which obey the joint distribution C;

get m new individual using the inverse of margin distribution $x_i = F_i^{-1}(u_i)$;

get new generation of population by the new individual replacing the old one.

The sample from Clayton copula can be seen in [4].

152.3 The Selection of Margin Distribution

The margin distribution can be the kinds of distribution functions in the statistic, such as empirical distribution, normal distribution, index distribution, χ^2 distribution and so on.

When the margin distribution of each variable adopts normal distribution, the establishment of normal model is similar to optimization in continuous domains by learning and simulation of Gaussian networks [5], so we can use the analysis method in [15] applying to cEDA, and calculating the mean and variance in the t generation. The process of analysis is as follows:

Because each variable x_i obeys to normal distribution, and the value of x_i is required in a range when optimizing function, so x_i meet to the property of truncated normal distribution.

Definition 1 [17]: if $x \sim N(\mu, \sigma^2)$, x_1, x_2 are two known real numbers, and $x_1 \leq x_2$, then on the condition of $x_1 \leq x \leq x_2$, the distribution which x obeyed is called truncated normal distribution, denoted $(\mu, \sigma^2; x_1, x_2)$.

Theorem 1 [17]: if $x \sim N(\mu, \sigma^2; x_1, x_2)$, then

$$E(x) = \mu + \sigma \cdot \frac{\phi(y_1) - \phi(y_2)}{\Phi(y_2) - \Phi(y_1)} = \mu + \sigma \cdot d \tag{152.2}$$

$$D(x) = \sigma^2 \cdot \left\{ 1 + \frac{y_1\phi(y_1) - y_2\phi(y_2)}{\Phi(y_2) - \Phi(y_1)} - d^2 \right\} = \sigma^2 \cdot a \tag{152.3}$$

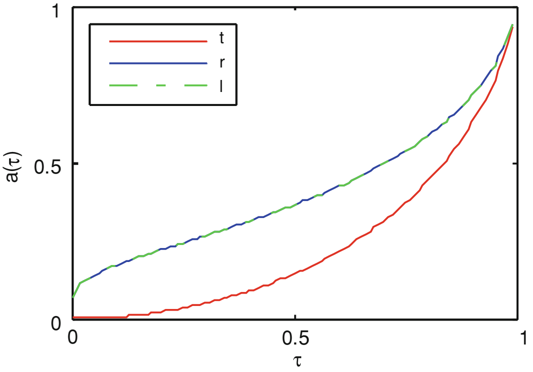


Fig. 152.1 The *curves* of $a(\tau)$

In that $y_1 = \frac{x_1 - \mu}{\sigma}$, $y_2 = \frac{x_2 - \mu}{\sigma}$, ϕ and Φ respectively denoted probability density function and distribution function of standard normal distribution. When $x_1 \to -\infty$, called right truncated normal distribution, when $x_2 \to \infty$,called left truncated normal distribution. τ denotes the population selection rate, then $0 < \tau < 1$, it is property is similar to the quantile of the normal distribution, we can get the formula as follows:

$$\Phi(y_2) - \Phi(y_1) = \begin{cases} \tau & x \le y_2 \quad \text{and} \quad y_1 \to -\infty \\ \tau & y_1 \le x \le y_2 \\ 1 - \tau & x \ge y_1 \quad \text{and} \quad y_2 \to \infty \end{cases} \tag{152.4}$$

Using inverse function can get $y_2 = \Phi^{-1}(\tau)$ on the condition of $x \le y_2 \, \text{and} \, y_1 \to -\infty$; $y_1 = \Phi^{-1}\left(\frac{1-\tau}{2}\right)$, $y_2 = \Phi^{-1}\left(\frac{1+\tau}{2}\right)$ on the condition of $y_1 \le x \le y_2$; $y_1 = \Phi^{-1}(1 - \tau)$ on the condition of $x \ge y_1 \, \text{and} \, y_2 \to \infty$. If substituting the value of y_1 and y_2 into Eqs. (152.2) and (152.3) according to the three situations, we can get the function of d and a about τ, denoted $a(\tau)$, $d(\tau)$, the graphic of $d(\tau)$ and $a(\tau)$ shown as Figs. 152.1 and 152.2. *t* denotes both ends of the truncated normal distribution, *r* denotes right truncated normal distribution and l denotes left truncated normal distribution in the image.

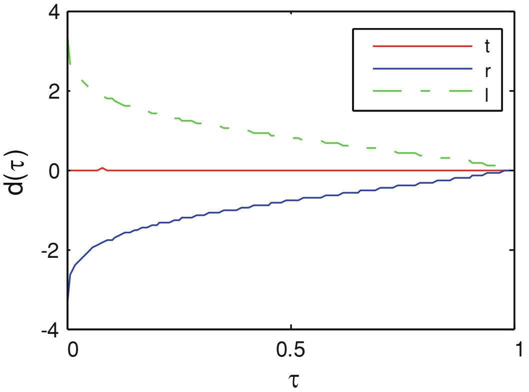
Because margin distribution only represents the distribution of variable itself, we can get the mean $\mu^{t+1} = \mu^t + \sigma^t \cdot d$ and variance σ^t by applying Definition 1 and Theorem 1:

$$\mu^{t+1} = \mu^t + \sigma^t \cdot d(\tau) = \mu^0 + \sigma^0 \cdot \sum_{k=1}^{t} \left(\sqrt{a(\tau)}\right)^k \cdot d(\tau) \tag{152.5}$$

$$\sigma^{t+1} = \sigma^t \cdot \sqrt{a(\tau)} = \sigma^0 \cdot \left(\sqrt{a(\tau)}\right)^{t+1} \tag{152.6}$$

From the image we know that $0 \le a(\tau) \le 1$, so Eqs. (152.5) and (152.6) can further calculate:

Fig. 152.2 The *curves* of $d(\tau)$



$$\lim_{t\to\infty}\mu^t = \mu^0 + \sigma^0 \cdot d(\tau)\cdot \lim_{t\to\infty}\sum_{k=1}^{t}\left(\sqrt{a(\tau)}\right)^k = \mu^0 + \sigma^0 \cdot d(\tau)\cdot \frac{1}{\sqrt{1-a(\tau)}} \tag{152.7}$$

$$\lim_{t\to\infty}\sigma^t = \sigma^0 \cdot \lim_{t\to\infty}\left(\sqrt{a(\tau)}\right)^t = 0 \tag{152.8}$$

Above all, we can get the result that the changing of mean μ^t is restricted when the variance σ^t convergence.

In order to solve premature problem, through analysis we found that the smaller of the variance, the smaller the search detection area of EDA. Therefore, the best way of expanding the search area and improving the probability distribution of the slope stage at the same time is increasing variance. In the adjustment of the variance of algorithm, the adaptive variance model put forward by Bosman et al. [10–13] did better, its main idea is that if a generation of fitness has obviously improved, then it showed the algorithm was on the slope stage, and the variance should be raised in order to expand the detection area, however, if there was no better fitness, then it showed that the algorithm has evolved into the near optimal value. We should reduce variance and improve convergence speed, at the same time, continuing to search, avoid variance too large and increase constantly the search scope which caused the missing of the optimal solution. Its main operation is that use $c \cdot \sigma$ substitute σ, the value of c can be adjusted dynamically, that is to say, when the variance needs to increase, then $c = c/\eta$,$\eta \in (0,1)$, while the variance needs to decrease, then $c = c \cdot \eta$, and set initial value $c = 1$, $\eta = 0.94$.

In the study of the margin distribution of cEDA using normal distribution, in order to improve the efficiency of the algorithm, the paper applied the adaptive variance model in [10] when studying on modeling the margin distribution of each variable, but because the sample vector of cEDA $X = (x_1, x_2, \ldots, x_n)$. was obtained by inverse function of normal distribution, in order to avoid the value of x_i beyond the scope of optimization function required because the variance was too large,

Table 152.1 The optimization results of using different margin distributions about cEDA and UMDAc, MIMICc

Function	Algorithms	Mean	Var	Min	Max
F1 Sphere function	UMDAc	2.5374e−23	7.4692e−28	1.1927e−23	4.4637e−23
	MIMICc	6.8774e−24	1.9882e−24	3.2909e−24	1.0672e−23
	cEDA(empirical)	2.1242e−10	2.4378e−10	8.9423e−12	9.6979e−10
	cEDA(normal)	1.1690e−14	2.7950e−14	1.5654e−18	1.2009e−13
	cEDA(adaptive)	8.0085e−13	1.0786e−12	6.9405e−15	4.0617e−12
F2 Parabolic ridge function	UMDAc	−7.2621e + 8	2.3795e + 8	−1.4863e + 9	−3.2238e + 8
	MIMICc	−6.1508e + 8	3.3864e + 8	−1.8133e + 9	−2.7532e + 8
	cEDA(empirical)	−4.9804e + 9	5.4297e + 9	−2.6224e + 10	−5.9220e + 8
	cEDA(normal)	−9.0823e + 8	4.8793e + 8	−2.9511e + 9	−3.7188e + 8
	cEDA(adaptive)	−1.3168e + 13	2.6264e + 13	−1.2869e + 14	−6.5157e + 11
F3 Rosenbrock function	UMDAc	7.5599	0.1658	7.2054	7.9313
	MIMICc	6.5379	0.14447	6.2789	6.8858
	cEDA(empirical)	6.7760	0.46909	6.4392	8.7013
	cEDA(normal)	6.0937	0.049055	5.9690	6.2247
	cEDA(adaptive)	4.2395	0.4709	3.3562	5.1241
F4 Sumcan function	UMDAc	−6.2358e + 4	20041	−9.9800e + 4	−2.3749e4
	MIMICc	−6.2464e + 4	3.2110e + 4	−9.9571e + 4	−8.8995e3
	cEDA(empirical)	−8.6949e + 4	9.7341e + 3	−9.6942e + 4	−5.5958e4
	cEDA(normal)	−8.5018e + 4	8.4343e + 3	−9.6480e + 4	−5.6654e4
	cEDA(adaptive)	−9.9032e + 4	770.92	−9.9857e + 4	−9.6868e + 4
F5 Schwefel function	UMDAc	0.018147	0.04983	9.2561e−4	0.20863
	MIMICc	1.8805e−3	8.9680e−4	4.6147e−4	3.8026e−3
	cEDA(empirical)	1.1432e−4	2.2415e−4	3.7966e−6	9.4935e−4
	cEDA(normal)	1.1439e−7	3.6351e−8	5.5179e−8	2.1551e−7
	cEDA(adaptive)	4.3953e−7	2.9783e−7	7.4447e−8	1.1885e−6
F6 Griewank function	UMDAc	4.2575e−3	2.3319e−2	0	0.12773
	MIMICc	1.9006e−2	5.4173e−3	9.1612e−3	3.2211e−2
	cEDA(empirical)	3.4067e−2	2.5470e−2	1.6369e−10	8.7543e−2
	cEDA(normal)	1.6930e−10	4.4906e−10	1.0103e−14	2.3256e−9
	cEDA(adaptive)	4.2505e−13	9.3518e−13	9.9920e−16	401425e−12

set $1 < c < 2.5$, if $c < 1 or c > 2.5$, then the value of c was reset to 1, and the algorithm was continuing.

152.4 Experiments

In order to test performance of the algorithm, and comparison with other algorithms, we conduct the experiment by six functions. The six test functions are all minimization problems. The variable dimension is 10, the top five functions corresponding to population size are 2000, the last of the population scale for 750, truncation selection operators is used to select elite population, the selection rate is 0.2, the mutation operator is also adopted in the algorithm, the mutation rate is 0.05, the experiment result is shown in Table 152.1.

Acknowledgements This work is supported by the Youth Research Fund of ShanXi Province (No. 2010021017-2), the Science Fund of Shanxi Institution of Higher Learning (No. 2010015) and the Nature Science Fund of ShanXi Province (No. 2009011017-3).

References

1. Wang L, Zeng J (2010) Estimation of distribution algorithms based on copula theory. In: Chen YP (ed) Exploitation of linkage learning in evolutionary algorithms, pp 137–160 Springer, Heidelberg. ISBN: 3642128335, 265 p
2. Wang L, Zeng J, Hong Y (2009) Estimation of distribution algorithms based on archimedean copula. In GEC 2009: proceedings of the first ACM/SIGEVO summit on genetic and evolutionary computation. ACM, New York, pp 993–996
3. Salinas-Gutierrez R, Hernandez-Aguirre A, Villa-Diharce ER (2009) Using copulas in estimation of distribution algorithms. In: Hernandez Aguirre A et al (ed.) LNAI 5845, MICAI 2009, pp 658–668
4. Wang L, Wang Y, Zeng J, Hong Y (2010) An estimation of distribution algorithm based on clayton copula and empirical margin. In: 2010 international conference on life system modeling and simulation & 2010 international conference on intelligent computing for sustainable energy and environment (LSMS & ICSEE 2010), Wuxi, China, pp 17–20
5. Larranaga P, Etxeberria R, Lozano JA, Pena JM (2000) Optimization in continuous domains by learning and simulation of Gaussian networks. In: Pelikan M et al (eds) Proceedings of the optimization by building and using probabilistic models OBUPM workshop at the genetic and evolutionary computation conference GECCO-2000. Morgan Kaufmann, San Francisco, pp 201–204
6. Larranaga P, Lozano JA, Bengoetxea E (2001) Estimation of distribution algorithms based on multivariate normal and gaussian networks. Technical report KZZA-IK-1-01. Department of Computer Science and Artificial Intelligence University of the Basque Country, Madrid
7. Ahn CW, Ramakrishna RS, Goldberg DE (2004) Real-coded bayesian optimization algorithm: bringing the strength of BOA into the continuous world. In: Deb K (ed) Proceedings of the genetic and evolutionary computation conference GECCO 2004, Springer, Berlin, pp 840–851
8. Bosman PAN, Thierens D (2000) Expanding from discrete to continuous estimation of distribution algorithms: the IDEA. In: Proceedings of the 6th international conference on parallel problem solving from nature-PPSN VI, Springer, pp 767–776

9. Duque TS, Goldberg DE, Sastry K (2008) Enhancing the efficiency of the ECGA. In: Rudolph G et al(eds) PPSN 2008.LNCS, Springer, Heidelberg, 5199:165–174
10. Grahl J, Bosman PAN, Rothlauf F (2004) The correlation-triggered adaptive variance scaling IDEA(CT-AVS-IDEA). In: Yao X (ed) Proceedings of the 8th annual conference on genetic and evolutionary computation conference. Gecco 2006, ACM Press, New York, USA, 2006, pp 397–404 PPSNVIII, Springer, Berlin, pp 352–361
11. Bosman PAN, Grahl J, Rothlauf F(2007) SDR: a better trigger for adaptive variance scaling in normal EDAs. In: GECCO'07: Proceedings of the 9th annual conference on genetic and evolutionary computation, ACM Press, New York, USA, pp 492–499
12. Peter Bosman AN, Grahl J (2008) Matching inductive search bias and problemstructure in continuous Estimation-of-Distribution Algorithms. Eur J Oper Res 185:1246–1264
13. Ocenasek J, Kern S, Hanse N, Koumoutsakos P (2004) A mixed bayesian optimization algorithm with variance adaptation. In: Yao X (ed) Parallel Problem Solving from Nature-PPSNVIII. Springer, Berlin, pp 352–361
14. Yuan B, Gallagher M (2005) On the importance of diversity maintenance in estimation of distribution algorithms, In: Beyer HG, O' Reilly UM (eds) Proceedings of the genetic and evolutionary computation conference GECCO-2005, vol 1. ACM Press, New York,USA, pp.719–726
15. Grahl J, Minner S, Rothlauf F (2005) Behaviour of UMDAc with truncation selection on monotonous functions. In: The 2005 IEEE congress on evolutionary computation IEEE CEC 2005
16. Marshall AW, Olkin I (1988) Families of multivariate distributions. J Am Stat Assoc 83:834–841
17. Dong'en Z (1998) Estimation of parameter for truncated normal distribution via the EM algorithm. J Beijing Inst Light Ind 16(2):123–129

Chapter 153
The Analysis on the Influencing Factors of Hospital Costs for Cerebral Infarction Patients

Xiaohong Wang, Guoli Wang, Jianhui Wu and Xinlei Guo

Abstract To control the irrational increase of medical costs, and improve the rational use of health resources. 2,218 medical records of cerebral infraction patients between 2007 and 2008 was collected from a tertiary hospital in Tangshan city, and univariate and multiple linear regression analysis was conducted for risk factors of hospital charges. The average hospital cost of cerebral patients was 8,371.22 yuan, univariate analysis results show that: gender, rescue, payment methods, hospital stay and treatment outcomes were important factors that has impact on hospital cost. Multiple linear regression results show that: hospital stay, rescue, payment methods, gender, treatment outcomes and admissions were the affect factors of hospital cost. It was found that hospital stay was a major factor of hospital cost, and comprehensive measures should be taken to shorten the hospital costs.

Keywords Cerebral infarction · Hospital costs · Influencing factor

153.1 Introduction

In recent years, with medical costs rising rapidly, “see a doctor is difficult and expensive” is a key issue to ordinary people. “The fourth national heath services survey results”, announced by the ministry of health in early 2009, show that:

X. Wang (✉)
Tangshan Centers for disease control and prevention,
Hebei United University, Tang Shan 063000 China
e-mail: wgl3726393@yahoo.com.cn

G. Wang · J. Wu · X. Guo
Hebei United University, Tang Shan 063000, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_153,

in 2008, the average of hospital costs of city were 8,958 yuan, while the rural were 3,685 yuan [1]. High medical costs, brings especially to ordinary people a serious economic burden. Controlling high medical costs, optimizing the allocation of health resources, and establishing multi-level health insurance have become a hot social concern in recent years [2]. In this chapter, take cerebral infarction patients for example, univariate and multiple linear regression analysis was conducted for risk factors of hospital charges, in order to provide theoretical basis for improving the using efficiency of health resources.

153.2 Material and Method

153.2.1 Research Subjects

Medical records of cerebral infraction patients between 2007 and 2008 were collected from a tertiary hospital in Tangshan city, excluding missing cases and non-logical cases. At last, 2,218 effective cases were included.

153.2.2 Research Methods

For the raw data extracted from HIS system, summary analysis was conducted by Excel 2003 database. Patients' medical records information including: total medical costs, gender, age, martial status, admissions, rescue, payment methods, hospital stay, treatment outcomes etc were taken into consideration.

153.2.3 Statistic Methods

For the total medical costs a logarithmic transformation to normal distribution was conducted, and SAS 8.2 statistical package was applied for data processing. ① t-test was conducted to compare the mean between two groups, analysis of variance was conducted to compare the mean among groups, SNK-test was conducted to compare the mean between each two groups; ② with the logarithm of the total costs as the dependent variable, and gender, age, martial status, admissions, rescue, payment methods, hospital stay, treatment outcomes as independent variables, inclusion and exclusion criteria were 0.05. Multiple regression was conducted to screen the influencing factor of hospital costs.

153.3 Results

153.3.1 Medical Records Report

In the total of 2,218 patients, with male 1,254, female 964, mean age (65.65 ± 11.29) years and mean hospital stay (17.05 ± 9.22)d, 2,199 were married, 19 were unmarried, divorced and widowed. 1,568 were initial admission, 650 were non-initial admission; 199 were rescued and 2,019 were not rescued; 1,995 were health insurance patients, 223 were non-health-insurance patients; 1,628 were cured, 561 were improved, 11 were not cured and 18 were dead. The hospital costs are shown in Table 153.1.

153.3.2 Univariate Analysis for Influencing Factors of Hospital Costs

The coefficient of skewness of hospital costs showed that, hospital costs was positively skewed distribution, and was not a normal distribution after data changed; therefore, rank sum test was conducted to compare hospital costs between different gender, age, martial status, admissions, rescue, payment methods, hospital stay and treatment outcomes. The results are shown in Table 153.2.

Table 153.1 showed that, gender, rescue, payment methods and treatment outcomes to certain extent have a impact on hospital costs. The hospital costs of male patients were higher than female; patients who were rescued higher than not rescued; medical insurance patients higher than no medical insurance patients; the hospital stay was longer, the hospital costs was higher; there were no significant differences between cured, improved and dead patients, but these were higher than non-cured patients.

153.3.3 Multiple Linear Regression Analysis for Influencing Factors of Hospital Costs

The logarithm of the total costs as the dependent variable, and gender, age, martial status, admissions, rescue, payment methods, hospital stay, treatment outcomes as independent variables, multiple regression was conducted to establish multiple linear regression model. Assignment methods for each variable are shown in Table 153.3, model results are shown in Table 153.4.

Multiple linear regression results showed that, gender, admissions, rescue, payment methods, hospital stay and treatment outcomes were the influencing factors of hospital costs.

Table 153.1 Descriptive analysis for hospital costs of cerebral infraction patients

Mean	Median	Std. deviation	Skewness	Kurtosis
8,371.22	7,071.13	5,454.47	2.39	8.38
Min	Max	P25	P75	
1,035.71	4,1791.22	5,045.84	9,897.85	

Table 153.2 Univariate analysis results of hospital costs

Variable	Rating	n	Median	Quartile range	Z	P
Gender	Male	1,254	7,428.58	5,129.40	5.35	0.00
	Female	964	6,661.97	4,578.56		
Marital status	Married	2,199	7,069.61	4,858.44	0.03	0.98
	Other	19	7,964.98	3,626.86		
Number of admissions	1	1,568	7,059.15	4,985.72	1.14	0.25
	≥2	650	7,086.36	4,306.57		
Rescue	Yes	199	7,964.98	5,733.16	3.96	0.00
	No	2,019	6,991.67	4,813.13		
Payment	Medicare	1,995	7,268.30	4,839.56	7.60	0.00
	Non-medicare	223	5,528.18	4,420.48		
Variable	Rating	n	Median	Quartile range	Chi-square	P
Age (years)	<50	181	7,114.19	5,361.75	6.98	0.14
	50∼	538	7,081.30	4,464.57		
	60∼	571	7,306.76	4,630.65		
	70∼	703	6,991.67	4,765.65		
	80 and over	225	6,444.39	5,843.58		
Days of hospital	<5	81	2,078.22	1,138.28	1,097.49	0.00
	5∼	315	4,348.42	2,526.03		
	10∼	713	6,248.03	3,006.51		
	15∼	491	7,670.01	3,661.77		
	20∼	321	9,107.07	5,044.85		
	25 and over	297	1,3204.14	7,512.87		
Treatment outcome	Cure	1,628	7,257.25	4,377.81	23.16	0.00
	Improve	561	6,458.20	5,968.21		
	Healed	11	6,413.52	5,609.64		
	Death	18	5,650.19	9,568.44		

153.4 Discussion

Stroke is one of the three fatal diseases, for which hospital costs accounts for 2–4% of each national healthcare costs. With the world population aging, the proportion will further increase [3]. In china, the overall incidence of cerebral infraction has not been reported. According to a regional survey, the incidence of each year was about 166.07–199.96 per 100,000 [4]. If the costs of each patient with cerebral infraction were 8,400 yuan, the national, society and family will spend 19.52–

Table 153.3 The assignment methods for influencing factors of cerebral infraction patients

Factors	Code	Quantitative methods
Gender	x1	Male = 1, female = 2
Marital status	x2	Married = 1, other = 2
Age	x3	Years
Number of admissions	x4	1 time = 1, ≥2 times = 2
Rescue	x5	No = 0, yes = 1
Payment	x6	Medicare = 1, non-medicare = 2
Days of hospital	x7	Days
Treatment outcome	x8	Cure = 1, improve = 2, healed = 3, death = 4

Table 153.4 Multiple analysis regression results for influencing factors of hospital costs

Variable	Regression coefficient	Standardized partial regression coefficient	t	P
Constant	3.7226		145.36	<0.0001
Gender	−0.0325	−0.0644	4.21	<0.0001
Number of admissions	−0.0253	0.0460	2.99	0.0029
Rescue	0.0898	0.1026	6.67	<0.0001
Payment	−0.0689	0.0829	5.33	<0.0001
Days of hospital	0.0182	0.6698	43.56	<0.0001
Treatment outcome	−0.0271	0.0557	3.59	0.0003

23.52 billion. As a development country with large population, it is a huge medical expenses.

In this Chapter, univariate and multiple linear regression results showed that, rescue, payment method, hospital stay and treatment outcomes were the factors that affected the hospital costs. The results was consistent with other reports[5, 6].

It was different from previous research, this study showed that, the hospital cost was different for different gender, male was higher than female. Medical research showed that, female is stronger than male for the overall resistance to the body, and more male are in subhealth state. According to the survey results of PLA 114 hospital, there are 15 diseases which are main causes of death. Male is higher than female, and most of these diseases are chronic diseases. At present, the medical costs is high which made the idea to focus on the severe but light prevention for chronic diseases. Therefore, enhancing prevention and heath care in community, especially making a prevented plan according to the specific situation of male patients is necessary to reduce morbidity and mortality of stroke, improve physical and mental suffering of patients and control rapid rising of medical costs.

In this Chapter, univariate analysis showed that admissions did not affect the hospital costs, and multiple linear regression showed that admissions affected the hospital costs. Studies have shown that there are complex relations between hospital costs and its influencing factors, and that there may exist collinearity between factors. In this study, stepwise regression was conducted to screen variables when multiple linear regression model was established, so that effectively controlled the collinearity between factors, so the results were more reliable.

Multiple linear regression results showed the hospital stay has the greatest impact on the hospital costs, the longer hospital stay, the higher hospital costs. Hospital stay not only reflected the severity of stroke patients, but also had close relations to medical standards, management levels and logistics supports. The severity is an important measure for the medical services demand strength of patients. Critically ill patients often face a greater death and the risk of loss physiological functions, and need more effective medical interventions, more medical resources and a longer hospital stay, so naturally the more hospital costs occurred.

According the above analysis, to control the hospital costs of cerebral infraction patients, it is important to shorten the hospital stay. Reducing hospital stay has become an important means to effectively control the growth of medical costs. On the one hand, improving the medical profession level to shorten the hospital stay by early accurate diagnosis, making a timely treatment plan, developing a reasonable treatment; on the other hand, improving the hospital management level to make medical departments, administrative functions departments, logistics and other departments take a serious measures to meet the needs of clinical departments, such as speed up laboratory tests results and shorten the appointment time. Effectively reducing the length of hospital stay, objectively reduced the hospital costs, and improved the hospital's efficiency and management level, and also improved the use efficient limited health resources, reduced the disease burden of patients and increased the social benefits of medical services. In other countries, by reviewing and restricting the hospital budget, strengthening the hospital management, improving the hospital control mechanism, developing home care and other non-hospital treatment the medical costs were controlled and unnecessary medical act were reduced [7]. These methods are worth learning and a reference for us.

References

1. Ministry of health of the People's Republic of China. http://www.moh.gov.cn/publicfiles/business/htmlfiles/mohbgt/s3582/200902/39201.htm
2. Xing H, Ren G (2007) A influential factors analyzing of hospitalization fees for patients with acute appendicitis. Mod Prev Med 34(20):3983–3985
3. Wentworth DA, Atkinson RP (1996) Implement of an acute stroke program decreases hospitalization costs and length of stay. Stroke 27(6):1040–1043
4. Huang M, Huang Z, Zeng J, et al (2001) Dynamic analysis of incidence and mortality of stroke and the risk factors in the communities in Shanghai, in the 1990s. Chin J Epidemiol 22(3):198–201
5. Lin L, Le C (2008) Analysis of medical charge and influencing factors of inpatients with stroke. Chin Health Econ 27(9):79–80

6. Song X, Xu L , Wang X et al (2007) Comparative study on the cerebral infarction inpatients under different payment. Chin Health Econ 26(10):55–58
7. Chu Z, Li G (1997) A study of foreign medical insurance cost. Foreign Med Sci Health Econ (2):53–56

Part XIV
Social and Economical Systems

Chapter 154
Analysis of the Internal Force Within Rural Financial Ecosystem

Fei Bao Lu, Wu Yi Zhang and Yi Yang

Abstract The financial ecosystem is essential for rural development. This paper argues that there is an internal force existing within the rural financial ecosystem. The quality of the entire system should achieve a truly integrated learned from law of universal gravitation and Coulomb's law, constructed out of the internal force of rural financial ecosystem model, and carried out the formula detailed analysis of various elements and then find the key to rural finance of ecological systems.

Keywords Internal force · Rural financial ceosystem · Elements

154.1 Introduction

Outbreak of the current global financial crisis and a series of financial risks in China has revealed that the world's financial system is facing many problems. People have to think and explore the reasons for its formation [1, 2]. After some of

F. B. Lu (✉)
Faculty of Business Administration, Chongqing University
of Science and Technology, Chongqing, China
e-mail: lfb19752001@163.com

F. B. Lu · W. Y. Zhang
Faculty of Management and Economics, Kunming University
of Science and Technology, Kunming, China
e-mail: wuyi20000@yahoo.com.cn

Y. Yang
Department of Mathematics and Physics, Chongqing University
of Science and Technology, Chongqing, China
e-mail: yangyi-2001@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_154,

the survey and subsequent research, the people have recognized the complexity of financial risks. We can learn from ecological principles to solve these problems. All causes of the financial risk can be attributed to an imbalance in the financial eco-system. When the financial imbalance reaches a certain level of ecosystem it is easy to trigger the financial crisis. The concept of financial ecology is draw on the British ecologist Tansley (1935) ecosystems (Ecosystem) [3].

In 2004, China's central bank governor Zhou Xiaochuan made the first official "financial ecology" concept [4, 5]. It includes the laws of social credit system, accounting and auditing standards, intermediary service system, the progress of enterprise reform and bank-enterprise relationships. The elements of interdependence and interaction jointly promote the development of financial industry.

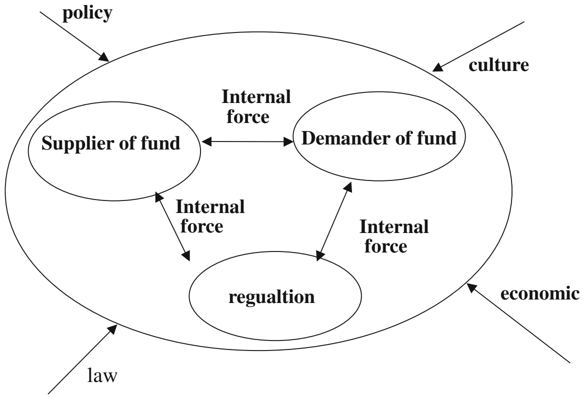
Rural financial ecology is a branch of the financial ecology in the rural context of this particular space [6]. In the rural context of this particular space, living environment and internal financial organizations formed a dynamic balance between the interaction of the system. Strengthening the financial ecological environment is the fundamental requirement for building a harmonious countryside. If we optimize the rural financial environment, we can optimize resources, and promote development of the rural economy. Financial ecological environment can improve the quality of bank assets to ensure national, regional, economic and financial security in rural areas. Rural financial environment determines the financial resources of the rural appeal of the economy.

154.2 Description of Rural Finance Ecological System

Rural financial ecosystem includes: Rural financial products and services, producers or suppliers (Rural credit cooperatives, rural commercial banks, rural cooperative insurance and other financial institutions in rural areas), demander of rural financial products and services (Farmers, village, town organizations, rural economic development, rule of law in rural areas, rural culture, customs, institutions and traditional environment) and regulators (Rural financial decision-making bodies and regulatory bodies) [6–8].

Development and changes of things must be considered from a system as a whole. Because of the existence of force. Elements can gather together. The role of this force can be roughly divided into two. One is external force (Government intervention force, the legal binding force, rural customs, as well as the overall macroeconomic situation). External features are mandatory. The role of external factors is to reduce the distance between the elements of the financial ecosystem. The second is the internal force (Economic conditions in rural areas, infrastructure conditions, credit culture, the quality of the peasants, and the character, experience, values, etc.). The role of internal forces of elements of the system is to realize the integration of quality. External force provides the environment of the system and the internal force provides the dynamic stability. Internal forces and external forces work together to form integrated system.

Fig. 154.1 Rural financial ecosystem



Evidences show that there are reluctant to be involved in. As the week of farmers credits within the rural economy, many financial institutions fear the presence of high risk. Banks are willing to give large-scale credit loan to urban enterprises, agriculture for many years has become a prominent problem in rural economic. So, the various elements of the financial ecosystem must be integrated. (Fig. 154.1).

154.3 Internal Force Model of Financial Ecosystem

The main elements of internal force consist of three parts: pyschological distance, quality and time (Table 154.1).

Learned from the law of universal gravitation and Coulomb's law [9], we believe that rural financial ecosystem relates to the psychological distance, quality and time factors. Model is as follows:

$$F(t) = \frac{Q1(t)Q2(t)\ldots\ldots Qn(t)}{d(t)^2} \tag{154.1}$$

F (t)	refers to the ecosystem of the internal forces of rural finance
Q1 (t) Q2 (t)…… Qn (t)	is the system the quality of the different elements. They are a function of time
D (t)	is the psychological distance between the elements, it is also a function of time

Table 154.1 Elements of the internal force

Internal force within rural financial ecosystem	Psychological distance	Advertising, television, news
		Emotional exchange (entertainment, communication)
		Credit, personal credit
	Quality	Capacity (cooperation, business, deal with complex issues
		Values (consumption concept, development concept)
		Education (training, learning)
		Experience (on crops, natural disasters)
		Knowledge (scientific and technological knowledge, life knowledge)
		Character
	Time	Function

154.3.1 Quality

We believe that in the rural finance system, the quality of the individual lies in the overall quality of the farmers, including farmer's education, experience, knowledge, ability, character, will and values. As we all know, in everyday life, if farmers improve their overall qualities, it will enhance the credit awareness of rural society, the ratio of escaping and wasting financial debt will go down. In the process of collecting loans, the illegal phenomenon will be reduced a lot. Improved financial awareness can enhance efficiency of financial instruments and the state monetary policy in less developed rural areas. The transmission of monetary policy also has primary carrier. Transmission will be relatively smooth, potentially affecting the financial sustainability and enhance the security of financial environment. Farmers' overall qualities will determine the entire financial system development. If we want to improve the overall quality of the farmers, it is necessary to educate farmers. Through the attractive form, we can strengthen moral training of farmers and improve moral quality of farmers; through regular legal education, we make farmers know law, understand law and become law-abiding, and correctly exercise their rights and obligation. Through collective education, we can enhance the farmers' overall concept and abandon selfishness and narrow-minded "small-scale peasant consciousness". Through the training of scientific and cultural knowledge make farmers abandon the feudal ideology, and make them become rich through scientific and technological study. Only when the quality of individuals enhance, the efficiency of financial institutions at all levels will improve. The situation of power of government intervention in the rural areas will be reduced. As a result, the entire financial ecosystem could be raised to a large extent.

154.3.2 Psychological Distance

In the financial eco-system, when the quality of individual farmers and organizations have raised, the increase of quality will enlarge the gravity among the elements, thus increasing the possibility of integration [10]. The essential elements of the financial ecosystem is human, the psychological distance among each element has an impact on the internal forces and integration effect of system.

First of all, we ought to create an atmosphere of rural credit culture. Lack of credit environment is the direct cause which leads to financial environmental degradation in rural areas. Currently, the construction of the rural credit system is seriously lagging behind, social credit services is not complete and intermediary services do not meet regulation. Due to lack of effective personal credit system, the risk of personal loans cannot be monitored timely. So these factors affect to a large extent the internal forces of various elements in the system, and cannot form a huge gravitation. Meanwhile the monetary policy cannot be properly implemented. We should further enhance the sharing level of credit data, improve business and personal credit and credit evaluation system, to effectively solve the information asymmetry problem existing between the financial institutions and farmers.

Secondly, we should increase publicity using various media, such as television, press, radio and other communication to fully carry out the ideology, in order to reach a consensus on the issue. Only by reducing the risks of rural financial system, we can achieve the purpose of shortening the Psychological distance. We should strengthen communication in several aspects, such as among people, between people and organizations, and among organizations. The smaller psychological distance is, the greater the internal integration strength will be. In addition, we are supposed to narrow the psychological distance by using a variety of incentives ways and means,and give full play to individual farmers' energy to realize goals effectively.

154.3.3 Time

No matter quality or psychological distance, we need time. The entire ecological system of rural finance will continue to change as time goes by. No matter the improvement of both the farmers' quality, or creation of credit cultural environment, and measures and methods of improving ecological environment, we need time. Such things cannot be achieved overnight, especially for the improvement of quality, we may need to do long-term planning. When we consider the running of system, we must take the time factor into our consideration.

All in all, if we want to improve the overall efficiency of the system, we must increase the internal forces. If we want to implement the internal forces, we must improve the system's overall function,and narrow the psychological distance among elements. Therefore the final analysis is to improve the overall quality of the farmers, and create a credit culture in rural areas, carry out positive publicity, and enhance exchange of feelings. In this way, the internal forces of the whole system can be enhanced, to create the harmonious financial environment which consists of policy environment, economic environment, legal environment, and financial services level. In addition, effective coordination mechanisms should be established among main departments which is related to coordination mechanism of financial supervision, including information sharing, discussion on major issues, joint conference, etc., to avoid duplication of regulatory and supervision vacuum, so as to improve regulatory efficiency and encourage financial innovation, and maintain financial operation and stability of financial markets. In this way, we can fundamentally improve the economic environment in rural areas, improve rural financial infrastructure, and thus promote the healthy development of the whole system.

References

1. Sai T (2009) Rural financial theory of optimization of the ecological environment 1:047–053
2. Qiuming W (2004) Integrated management. Economy Science Press, Beijing
3. Harrington J (1979) Computer integrated manufacturing [M] Reprint New York. Robert E. Krieger Publishing Co, Melbourne
4. Xiaochuan Z (2004) Some on rural financial reform ideas [J]. Econ Perspect 9:27–48
5. Xu N (2005) Financial ecological environment in China [J]. Financ Res (11):21–27
6. Jing W (2010) Mutation analysis of the financial ecosystem. Financ Theory Pract 8:58–63
7. Liu H (2006) County financial ecology and the development of new socialis countryside in China. Ecol Econ 2(4):412–423, (2006/11)
8. Jiang M (2009) Rural finance the construction of new countryside Empirical, Zhong Agricultural University (Social Science) (1):101–107
9. Zhen Z (2010) Decision-making system based on ecology model and analysis [J]. Jishou University (Natural Science) 3:213–216
10. Qiang Z (2009) Based on the "functional paradigm" of the rural financial system optimization choice. Hebei University of Economics and Technology in March 2009

Chapter 155
The Study on the Early Warning Mechanism of Agricultural Management Risk

Guofu Zhang

Abstract The intertwined basic characteristics of natural reproduction of agricultural production and economical reproduction determine the agricultural high-risk and serious threat. After joining WTO, it faces the serious challenges of agriculture in developed countries. The risks that Chinese agriculture faces can not only be eliminated, but will also increase with the increase of its interleaving degree, making huge losses to agricultural economy, and affecting the sustainable development of agriculture. Therefore, it is necessary to establish a perfect agricultural risk prevention mechanism. The content of agricultural business risks is analyzed in this paper, and many risks are proposed in this paper, such as natural risk, technical risk, economic risk, etc. From risk identification, early warning of natural disasters, market forecasting and monitoring and futures market, some feasible approaches and specific measures are proposed for how to build a mechanism to prevent the agricultural management risk.

Keywords Prevention · Agricultural business risks · Mechanism

155.1 Introduction

Agricultural management risk refers to the uncertainty of internal and external environment of agriculture, the complexity of agricultural production and management and the limited capacity of agricultural enterprise in the process of

G. Zhang (✉)
College of Economics and Management,
Heilongjiang Bayi Agricultural University, Daqing, China
e-mail: byndzgf@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_155,

agricultural production and management, leading to the deviation of actual revenue and expected revenue of agriculture, and even to the failure of agricultural production and management.

For the agricultural management risks and early warning mechanism, many scholars studied the existing risks, characteristics, causes and risk prevention measures of Chinese agriculture [1, 2]. Some scholars made a further study on the characteristics, development status and the various existing contradictions of Chinese agriculture; the implementation ideas of agricultural risk protection system with Chinese characteristics are proposed based on this. That is, to establish: an agricultural price protection system; an agricultural buffer stocks and risk fund system; an agricultural policy loans system; a stable growth mechanism of government on agricultural inputs; an agricultural insurance system; and an agricultural futures market. However, the study of the generation mechanism of agricultural risks is not enough. It has been proposed that the risks that agriculture faces is mainly agricultural crop insurance, disaster relief and other natural risks and price protection, futures market and other market risks, and the risk management measures of crop insurance, disaster relief, price protection and futures markets, but there is little study on the early warning mechanism of natural risks and market risks. The characteristics of the agricultural risk in transition period were also analyzed, which shows that the Chinese agricultural market risk management should be a multi-composite structure management mode of government, market, enterprise, farmers self-management, that is, the government should be responsible for system selection and policy selection in agricultural risk management. Meanwhile the agricultural futures markets should be established for agricultural organization management, and then to build a new agricultural insurance system. In addition, the foreign agricultural risk management experience is introduced by some scholars, and a technique risk that Chinese agriculture faces is analyzed.

According to the studies of scholars at home and abroad it is found that there are more results of partial study for agricultural risk prevention; there is little comprehensive and systematic study on agricultural risk prevention, and only the basic levels of risk identification, risk evaluation and risk management have been analyzed before. For the risk categories and sources that China's agriculture faces after joining WTO, there is no thorough analysis, and the study of the harm on China's agriculture risk is almost non-existent, thus, the feasibility and operability of agricultural risk management measures are yet to be studied. There is no study on the system of prevention, control and resolving of agricultural risks. All this requires in depth study and discussion.

155.2 Analysis of China's Agricultural Management Risk

Agricultural management risks mainly include natural risks, agricultural economic risks and agricultural technical risks. Agricultural natural risks refer to losses from irregular changes in the natural forces to agriculture, which mainly include agrometeorological disaster risk, agricultural biotechnology disaster risk, agricultural biotechnology disaster risk and agricultural biological disaster risk. Economic risks of agricultural systems are mainly from poor management; market forecast errors, price volatility, changes of consumer demand, inflation, exchange rate changes, etc. in the economic activity. Market risk mainly includes price risk, competitive risk, information risk and credit risk. Agricultural technical risk is the possibility of the deviation of the actual revenue and expected revenue by using the agricultural technology. In agriculture, the development of modern science and technology not only widens the types of traditional agricultural products, but also reduces the dependency of agriculture on natural resources, which has greatly improved agricultural productivity. However, there also exists a huge risk.

155.3 Analysis of Agricultural Management Risk

Analysis of the causes of agricultural risks is mainly to analyze the risk factors of agricultural systems. Risk factors include physical risk factors and human risk factors [3, 4]. While specific to different risks, the risk factors will have a different focus. Even the same risk can be different at different times. In general, the risk factors of agricultural natural risks are mainly physical risk factors, technical risks, economic risks, etc. and risk factors of other risks that are mainly human risk factors. The real risk factors that produce agricultural natural risks are climatic factors, geologic and morphologic factors and biological factors. The characterization of agriculture is the inherent factor of various agricultural risks. Market mechanism is the formation factor of the market risk of various agricultural risks. Market failure and government failure are the economic reasons to produce agricultural risk. Human factors are the main risk factors to produce a variety of agricultural risks. Economic benefits are the fundamental factors that lead to various risk factors. Economic moral fall is the ethical risk factor of agricultural risk.

155.4 The Building of the Early Warning Mechanism of Agricultural Management Risk

Agriculture is a multi-risk industry, under the influence of the natural environment and the market economy. There exist uncertainty of agricultural production and management, which often brings unexpected losses and risks for agricultural production and management. Recognizing and identifying risk, and then taking effective measures to prevent them is the precondition to guard against agricultural management risk. The building of agricultural risk prevention mechanism will be represented from risk identification, natural disaster early warning mechanism, market forecasting and monitoring mechanism, futures market system construction, etc.

155.5 The Building of Agricultural Risk Identification Mechanism

Risk identification is the process of the systematic and continuous collection of risk sources identification, risk harm and risks loss exposure before or when the risk appears.

155.5.1 The Identification of the Influence Factors of Risk Source

Risk source is the source of factors and hazards that will lead to negative or positive consequences. For example, when building an agricultural high-tech garden, the level of the quality of local labor is an important factor. It mainly includes physical environment, social environment, political environment, legal environment, operating environment, economic environment, environmental awareness, etc.

155.5.2 On the Analysis Method of Agricultural Risk Identification

Risk is not a one-dimensional concept; no single method can achieve all the purposes at the same tim. The common risk identification methods include environmental analysis, expert investigation, insurance investigation, decomposition analysis and screen scene analysis.

155.6 Construction of Natural Disaster Early Warning Mechanism

In recent years, there have been frequent occurrences of natural disasters in China's agriculture, leading to larger and larger destruction on agricultural production and management, and making agricultural management in great risk, bringing obstacles to the sustainable development of agriculture. In order to reduce agricultural management risk and enhance agricultural competitiveness, it is necessary and urgent to establish an agricultural natural disaster early warning system that fits the national condition of China.

155.6.1 The Prediction Method and Model of Agricultural Natural Disasters

In prediction methods analysis of agricultural natural disasters the first is disaster site prediction (spatial prediction). The best way is to draw up various prognostic maps. The second is time prediction. Long-term observation records to establish an accurate probability model are used as the basis for time prediction. Nowadays, this method is used for time prediction of the weather disasters of floods, wind, and drought and geological disasters of earthquakes and landslides, and its accuracy depends on the reliability and time scale of the observation value. The third is precursor event. Many disasters have precursory events. Some disaster roles of natural events can be predicted accurately according to the precursor events.

In quantitative analysis of agricultural natural disaster prediction, the level of risk of natural disasters mainly depends on three factors:first, the extent of catastrophic events, including the intensity (scale) and frequency (probability) of the disaster, and under normal circumstances, the higher the level of catastrophe, the greater the disaster risk; second, the value of the affected property and the vulnerability of the affected property and disaster prevention capabilities, and under normal circumstances, higher the density of the affected property, the worse the ability of the disaster prevention; the higher the vulnerability, the greater the disaster risk.

The third is risk dispersion degree, which is the dispersion degree of disaster losses with random changes. Under normal circumstances, the smaller the dispersion degree of disasters, the lower the degree of disaster risks; the greater the dispersion degree of disasters, the larger the degree of disaster risks. Based on the above knowledge, the risk index f that represents natural disaster risk degree can be expressed as:

$$f = Q * G * Y(1 - Z_X) * [1 + f(b)] \tag{155.1}$$

$$f(b) = \sqrt{\frac{\sum\limits_{i=1} (fi - f)^2}{n - 1}} \tag{155.2}$$

In the equation: f is risk index; G is occurrence probability of disaster; Y is the vulnerability index of the affected body; Z_X is disaster prevention; $f(b)$ is variation coefficient. In which

$$f(b) = \sqrt{\frac{\sum\limits_{i=1} (fi - f)^2}{n - 1}} \tag{155.3}$$

In the equation: fi is the risk index of the i year; f is the average risk index of many years; n is the number of years of statistics times.

In the establishment of an agricultural natural disaster early warning mechanism aAccording to China's national conditions, the main construction contents of the mechanism mainly include the integration of resources and technical equipment; the formation of expert teams; the full use of the advantages of modern science and technology, especially information technology, and then to achieve the scientific and systematic agricultural natural disasters early warning mechanism; it learns from the basic theoretical research results, comprehensive technological capabilities and advanced equipments from developed countries, to make a comprehensive monitoring and early warning for agricultural natural disasters; the establishment of an effective natural disasters risk emergency system can minimize the loss caused by agricultural natural disaster risks.

155.7 The Building of Market Early Warning Mechanism

In the market economy, the agricultural market risk is the biggest risk in agricultural production and management. Therefore, it is necessary to study the early warning management for agricultural market risk.

155.7.1 Quantitative Analysis of Market Early Warning Model

Multi-level fuzzy comprehensive evaluation method can deal better with multiple factors, ambiguity, and subjective judgments, etc. Specific steps are as follows.

Determine the fuzzy evaluation index set U. Set early warning indicator system of market risk as U, divided into two levels of factors. The first layer factor is U = (U1, U2, U3, U4, U5, U6) = (marketing risk, market demand risk, market competition risk, market price risk, technological innovation risk, and market overall risk), the second layer factor is the specific indicators that reflect the first layer factor.

Determine the evaluation set and its assignment. Generally, the evaluation levels of various factors in determining model are divided into five levels, namely V = (Vl, V2, V3, V4, V5) (very good, good, general, poor and very poor). Assign to the evaluation set as very good l, good 0.8, general 0.6, poor 0.4, very poor 0.2, which are used to indicate the assignment set, and then V = (l, 0.8, 0.6, 0.4, 0.2).

Determine the single factor evaluation matrix (membership matrix) R. There exists membership degree between elements of fuzzy set and itself. In market risk assessment, the membership degree of the relationship between index and risk in index system can be represented by the membership degree. The membership relation can be recorded as R = l, otherwise R = 0, and the taking value interval of membership degree.

Determine index weight set W. According to the importance of each index factor, it assigns the corresponding weights to various indicators, and its influence degree should be consistent with the degree that affects the last layer, thus forming the weight set of evaluation indicator factors. It is recorded as Wk = (Wkl, Wk2,…, Wkm). Assessment weights of agricultural market risk can be determined by the obtained first-hand information, expert evaluation and experience accumulated by evaluators over the years based on questionnaires. Agriculture should be combined with its own focus and weakness during the management process, and pay more attention to the risk indicators that is easy to appear.

Calculate the fuzzy comprehensive membership degree value set B. Set the single-factor evaluation matrix of n indicators in a large indicator as Rk, the specific indicators score Bk = Rk*VT can be obtained according to the single factor evaluation matrix and evaluation set V.

Comprehensive evaluation. The evaluation result Sk, Sk = Wk*VT of this indicator can be obtained from Bk and index weight Wk. Repeat the above steps, the 6 factors subset Uk in U can be seen as a single factor of U, distribute the weight W, W = (W1, W2,…, W6) of the role of Uk in U; the comprehensive evaluation of U S, S = W, B = (W1, W2,…, W6) (W1, W2,…, W6) T can be obtained from the evaluation results Sk of each U.

155.7.2 Market Risk Early Warning Method

The determination of market risk early warning line. In the fuzzy forecast of market risk, the level of risk warning line can be set. The warning line should be set according to the characteristics of agriculture and its own specific circumstances. The operating status of agriculture can be divided into four levels, which are W1 prosperity state, W2 quasi-adversity state, W3 stress state and W4 crisis state. Set three warning lines for the four states in order, the warning values are respectively 0.25, 0.5 and 0.75. The four state values are respectively W1 = (0.75, 1], W2 = (0.5, 0.75], W3 = (0.25, 0.5] and W4 = (0, 0.25).

Alarm type of market risk early warning. According to the different corresponding risk state of the comprehensive evaluation results of fuzzy matrix, different levels of warning n can be sent in the way of “Light Card”. When the risk status is in normal, “green card” will be shown; when the risk status is in a low risk state, the “yellow card” will be shown; when the risk status is in low risk state, “red card” will be shown; when the risk status is in high risk status, the “double red card” will be shown. The purpose of light card is the early warning risk management decision-makers to adjust policies timely, thus reducing the possibility of loss or crisis, and then increasing the possibility of revenue.

155.8 The Establishment of Agricultural Futures Market System

Futures future market is the place for futures trading and the sum of futures trading. Agricultural future is the earliest future type in the world, and the futures market was first produced in the agricultural market. Its construction contents mainly include three aspects. The first is to rationalize and strengthen the regulatory system of futures market and improve the trading system. The second is to improve the laws and regulations of the futures market, and create a favorable policy environment. The third is to foster investment management entities and to carry out investment business. The fourth is to enhance the development and management of futures and strive to cultivate mature products.

155.9 Conclusions

Risk identification is the prerequisite for the prevention of agricultural management risk, and the construction of agricultural natural and market risk early warning mechanism is the basis to prevent agricultural management risk, and the establishment and perfect futures market system is the strong guarantee to prevent agricultural business risk. According to the construction of risk identification mechanism, natural disaster early warning mechanism and market early warning mechanism, it forms a sound risk prevention system, and the risk is nipped in the bud and before the bud stage to achieve the purpose of preventing agricultural management risk.

Acknowledgments This research is supported by: Grant From Heilongjiang Province Office of Education Early Warning Mechanism of Agricultural Management Risk of Heilongjiang of province.

References

1. Su L (2007) Reflections on agricultural risk. Northern Econ (14):22–26
2. Wang X (2008) On identification and control of agricultural risk. Modern Agric 3:087–090
3. Zhao L (2009) The initial explore agricultural insurance policy on Heilongjiang province. Mod Agric 4:151–154
4. Li Han C (2009) The study of T industrialized operation of agriculture and risk prevention system build. Agric Econ 12:145–147

Chapter 156
Stock Option Pricing Formula Volatility Estimation

Shujun Wang, Nana Li and Xianrui Meng

Abstract In practical applications, there is the phenomenon of volatility smirk in Classic B–S option pricing formula. In order to eliminate this phenomenon, we add an independent compound Poisson process to geometric Brownian motion, and get the corresponding option price formula. This article has done two tasks: Firstly, we estimate the volatility of classic B–S option pricing formula with stock and warrant market data, and find that volatility smirk is obvious. Secondly, we estimate the volatility of our option pricing formula with stock and warrant market data, and find that our pricing formula eliminates the smirk, and is better than classical B–S formula.

Keywords Options pricing · Jump-diffusion process · Volatility smirk

156.1 Introduction

Fisher Black and Myron Scholes in 1973 through the famous Black–Scholes option pricing model to derive the Black–Scholes European call and European put option pricing formula [1], which makes the theory of option pricing has been a breakthrough of the solution. That

S. Wang (✉) · N. Li
Tangshan College, Tangshan, People's Republic of China
e-mail: shujunwang88@hotmail.com

N. Li
e-mail: lalazinana@126.com

X. Meng
Hebei United University, Tangshan, People's Republic of China
e-mail: xianruimeng@yahoo.com.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_156,

Theorem 1 (*European call option pricing formula*) [2]: *set maturity date T, the Executive price K, the stock price process* $\mathrm{d}S = \mu S\mathrm{d}t + \sigma S\mathrm{d}W$ *for the value of European call option*

$$C(S,t) = SN(d_1) - K\mathrm{e}^{-rT}N(d_2). \tag{156.1}$$

where, $d_1 = \dfrac{\ln\frac{S}{K} + (r + \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}$, $d_2 = \dfrac{\ln\frac{S}{K} + (r - \frac{\sigma^2}{2})T}{\sigma\sqrt{T}}(d_2 = d_1 - \sigma\sqrt{T})$.

Theorem 2 (*European put option pricing formula*) [3]: *set maturity date T, the Executive price K, the stock price process* $\mathrm{d}S = \mu S\ \mathrm{d}t + \sigma S\ \mathrm{d}W$ *for the value of European put option*

$$P(S,t) = -SN(-d_1) + K\mathrm{e}^{-rT}N(-d_2). \tag{156.2}$$

where d_1 and d_2 have the same meaning with above.

In practical applications, there is the phenomenon of volatility smirk in Classic B–S option pricing formula. In order to eliminate this phenomenon, we add an independent compound Poisson process to geometric Brownian motion, and get the corresponding call option price formula [4]:

$$C(K,T) = \mathrm{e}^{-rT}\sum_{n=0}^{+\infty}\frac{(\lambda \mathrm{T})^n\mathrm{e}^{-\lambda T}}{n!}(\mathrm{e}^{a+\frac{b}{2}}N(d_1) - KN(d_2)). \tag{156.3}$$

put option price formula:

$$P(K,T) = \mathrm{e}^{-rT}\sum_{n=0}^{+\infty}\frac{(\lambda T)^n\mathrm{e}^{-\lambda T}}{n!}(KN(-d_2) - \mathrm{e}^{a+\frac{b}{2}}N(-d_1)). \tag{156.4}$$

where $d_1 = \dfrac{a + b - \log K}{\sqrt{b}}$, $d_2 = \dfrac{a - \log K}{\sqrt{b}}$.

156.2 Classic B–S Option Pricing Formula Volatility Estimation

In Classic B–S option pricing formula, closer examination in the Black–Scholes equation coefficients [5], we find that growth stocks not among them. However, we need to use and the risk-free interest rate volatility. In a regular basis of the maturity dates, the corresponding option exercise price and the stock price, these factors will be used to option pricing. In turn, fixed maturity date and exercise price of each, the option that corresponds to the stock price can also use the Black–

Table 1 BS formula stock volatility estimation table

Option price $C(K, T)$	The corresponding stock price options S_0	Maturity date T (Units:year)	Implementation of price K	Stock volatility σ
4.72	20.01	0.312	26.91	1.5101
5.00	19.62	0.310		1.6270
4.99	20.00	0.307		1.3772
4.77	19.66	0.304		1.5230
4.31	20.15	0.323		1.5851
4.64	19.67	0.316		1.5847
4.72	19.98	0.314		1.5087
4.33	20.37	0.321		1.3612
4.21	19.08	0.326		1.4710
3.78	19.37	0.334		1.3233
4.50	19.51	0.318		1.5054
4.01	20.12	0.325		1.3107
3.68	18.71	0.329		1.3840
4.01	22.87	0.342		0.9994
4.12	25.09	0.359		1.2716
3.46	18.99	0.340		1.2803
3.72	18.97	0.331		1.3600
3.97	20.67	0.336		1.2218
4.09	22.67	0.345		1.0292
3.46	21.08	0.347		1.0569

Scholes model inverse solution lies in the Black–Scholes formula, volatility. This volatility is called implied volatility. If the Black–Scholes model is completely correct, then calculate the implied volatility should be the only (or the volatility of rates would be calculated base value equal to or in the vicinity of a certain volatility), not as to different maturity date or exercise price varies, it is because the volatility of volatility changes described in assets, has nothing to do with their derivatives on. But in fact not the case.

As shown by Shanghai Automotive, for example, collected at different time of the option price, and the corresponding stock price, and the strike price and expiration date; in addition, we assume that: risk-free interest rate (to take one year time deposit interest). These data in Classic B–S option pricing formula to calculate the value of the volatility, to see for the different data sets, the value of the stock volatility is the same. Specific data is shown in Table 156.1.

Through the above data shows that prices for different options, exercise price, expiration date, the corresponding stock price and options, Classic B–S option pricing formula is not the only volatility, but, with the option price, strike price, maturity corresponding to the stock price and options vary. The implied volatility of the shape of a curve form, known as volatility skew or volatility smile. The skewness of securities derivatives markets, exchange rates and interest rate derivatives market derivatives are present. Curve for the formation of such volatility is not constant but there are many reasons to explain. Different explanations may apply in different markets.

In short, volatility skews a challenge to the financial model. An ideal model should be possible to eliminate all volatility skew.

156.3 Come Before this Option Pricing Formula of Stock Volatility in the Estimation Denote

The authors have come before the price of call option

$$C(K,T) = \mathrm{e}^{-rT}\sum_{n=0}^{+\infty}\frac{(\lambda T)^n \mathrm{e}^{-\lambda T}}{n!}\left(\mathrm{e}^{a+\frac{b}{2}}N(d_1) - KN(d_2)\right) \tag{156.5}$$

where $a = \log S_0 + (\mu - \frac{1}{2}\sigma^2)T + nE(X)$, $b = \sigma^2 + n\mathrm{Var}(X)$, $d_1 = \dfrac{a + b - \log K}{\sqrt{b}}$, $d_2 = \dfrac{a - \log K}{\sqrt{b}}$.

The formula used in the stock growth rate μ, volatility σ, risk-free interest rate r and the jump intensity, expectation, variance. In a regular basis of the maturity dates and the corresponding option exercise price and the stock price, these factors will be used to option pricing. Fixed maturity date of each turn, and the exercise price, the option that corresponds to the stock price, you can also use the above formula (5) Anti-solve in the volatility of the formula.

The formula (5) gives the call option price for an item and infinite type (i.e. a number), because the series is convergent, so here just take the first three and to calculate the price of an option. In addition, in order to better calculate the value of the stock volatility, we assume: in the option pricing formula with jumps, the jump in intensity$\lambda = 7$; risk-free interest rate $r = 0.036$ (to take one year time deposit interest);

$E(X) = 0$, $V(X) = 1$, which X obey the expectations of 0 and variance 1 of the standard normal distribution.

Expectations of the subject 0, variance 1 standard normal distribution, that is, the probability density function $f(x) = \frac{1}{2\pi}\mathrm{e}^{-\frac{x^2}{2}}$, there $E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\mathrm{d}x = 0$, $\int\limits_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1$.

Therefore,

$$\begin{aligned} E(\mathrm{e}^X) &= \int\limits_{-\infty}^{+\infty} \mathrm{e}^x \cdot f(x)\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \mathrm{e}^x \cdot \frac{1}{2\pi \mathrm{e}^{-\frac{x^2}{2}}}\mathrm{d}x = \int\limits_{-\infty}^{+\infty} \frac{1}{2\pi}\mathrm{e}^{-\frac{x^2-2x}{2}}\mathrm{d}x \\ &= \mathrm{e}^{\frac{1}{2}} \int\limits_{-\infty}^{+\infty} \frac{1}{2\pi}\mathrm{e}^{-\frac{(x-1)^2}{2}}\mathrm{d}x = \sqrt{\mathrm{e}}. \end{aligned}$$

The stock growth rate $\mu = r + \lambda - \lambda E(\mathrm{e}^X) = r + \lambda - \lambda\sqrt{\mathrm{e}}$.

Table 2 Parameter estimation of the stock volatility table

Option price $C(K, T)$	The corresponding stock price options S_0	Maturity Date T (Units:year)	Implementation of price K	Stock volatility σ
4.72	20.01	0.312	26.91	1.3006
5.00	19.62	0.310		1.3346
4.99	20.00	0.307		1.2998
4.77	19.66	0.304		1.2693
4.31	20.15	0.323		1.3160
4.64	19.67	0.316		1.3299
4.72	19.98	0.314		1.3145
4.33	20.37	0.321		1.2972
4.21	19.08	0.326		1.3642
3.78	19.37	0.334		1.3483
4.50	19.51	0.318		1.3322
4.01	20.12	0.325		1.2907
3.68	18.71	0.329		1.3260
4.01	22.87	0.342		1.3028
4.12	25.09	0.359		1.3614
3.46	18.99	0.340		1.3529
3.72	18.97	0.331		1.3342
3.97	20.67	0.336		1.3362
4.09	22.67	0.345		1.3410
3.46	21.08	0.347		1.3189

Here as shown by Shanghai Automotive, for example, to collect the option price, and the corresponding stock price, and the strike price and expiration date. These data into the option pricing formula (3.1), calculate the value of the parameter to see the stock for the different data sets, changes in volatility. Specific data are shown in Table 156.2.

Through the data in Table 156.2 shows that prices for different options, exercise price, expiration date and the corresponding stock price options, this option pricing formula derived from the volatility of the change is relatively stable, the basic changes are around 1.3. The difference between the maximum and minimum is 0.0949. and then classic B–S option pricing formula, the accuracy increased by 6.6 times. Therefore this shows that the diffusion process of the jump is calculated based on option pricing formula to a large extent eliminate the volatility skew, is a relatively ideal model.

References

1. Hull J (2000) Options, futures and other derivatives [M], 4th edn. Prentice-Hall, Englewood Cliffs, pp 156–230
2. Sun J (2007) Financial derivatives pricing models—introduction to mathematical finance. Economic Publishing House, Beijing, China

3. Chen DF (2007) Preliminary mathematical finance. Machinery Industry Press, Beijing
4. Ye ZH, Wang GL, Lin JZ (2007) Financial mathematics introduction to derivative pricing. Posts & Telecom Press, Beijing
5. Sun HX (eds) (2008) Random process. Machinery Industry Press, Beijing

Chapter 157
Analysis of the Impact on Bullwhip Effect Based on Different Treatment of Information Needs

Sanyou Ji and Yong Wang

Abstract Through the quantitative calculation of bullwhip effect, the effect that the centralized information processing can reduce bullwhip effect more significantly than the scattered information processing while using the moving average forecast method in demand forecasting is proved. Thus custom demand can be determined more accurately. This analysis provides a scientific and effective method for business managers, and it is very convenient for retailers, distributors, and manufacturers to forecast demand reasonably.

Keywords Centralized information · Bullwhip effect · Forecasting · Analysis

157.1 Introduction

In the study of market demand for a product, it is interesting that the change of orders always has bigger than the retail volume even though the retail volume of that product is fairly stable. To address this situation, the moving average forecast method is used for forecasting. The analysis of the impact on bullwhip effect is given based on different treatment of information needs.

S. Ji (✉) · Y. Wang
College of Logistics Engineering, Wuhan University of Technology,
Wuhan, Hubei, China
e-mail: jisanyou@126.com

Y. Wang
e-mail: wangy118@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_157,

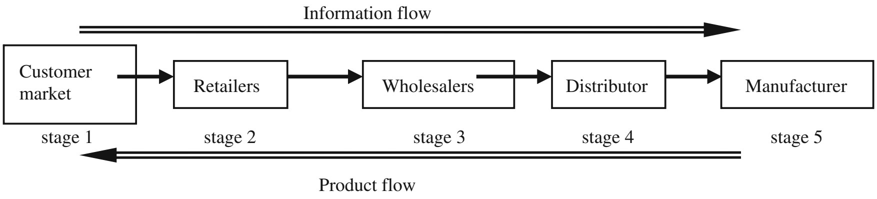


Fig. 157.1 Supply chain

157.2 The Quantitative Description of Moving Average Forecast Method's Affection to Bullwhip Effect

157.2.1 Structure of Supply Chain

Structure of supply chain is the essential reason for the formation of bullwhip effect supply chain structure. To simplify the problem, a structure of supply chain can be defined in Fig. 157.1. It can be assumed that there were only one product, and its demands in the given period follows a normal distribution [1].

A simple four-stage supply chain is enumerated in Fig. 157.1. It includes customer market, retailers, wholesalers, distributors, and manufacturers. Retailers observed customer demand, and then order goods from wholesalers, wholesalers order goods from distributors, and distributors order goods from manufacturers.

157.2.2 Quantization of Bullwhip Effect

The bullwhip effect means the demand order variability in the supply chain will be amplified as they moved up the supply chain as shown in Fig. 157.1. So far, there are three general methods of measurement or the bullwhip effect. The first is variance and ratio method. It is clear, simple and direct, applies especially to quantify and deal the practical and complex supply.

The second and third methods are the maximum amplitude method of frequency response curve and the noise bandwidth method. The latter two methods can be proved strictly but just valid in simple and linear production inventory system. When the supply chain is a complex system, the quantization of bullwhip effect will be very difficult by using the second and third methods.

Generally speaking, the quantization of bullwhip effect is described by the proportional relation between the variance of order fluctuations and customer demand in supply chain system. In the quantization research of bullwhip effect, Simchi-levi and some other scientists has arrived two widely recognized inequalities which are used to quantify the leading time of bullwhip effect. In this

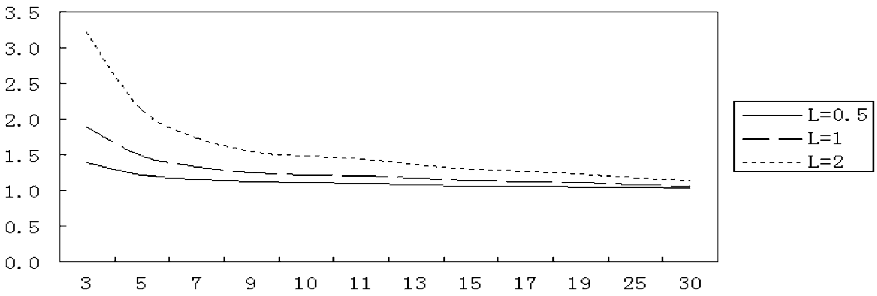


Fig. 157.2 The function graph demand variations in given leading times

model, the sample observations of customer demand observations are independent and identically distributed are independent and identically distributed variables.

So we can make quantitative calculations over bullwhip effect. That is to say, the demand variability to manufactures can be calculated and can be compared with that to retailers. Set Var(D) as the customer demand variance which is observed by retailers, Var(Q) as the order demand variance from retailers to wholesalers. Their proportional relation can be described as follows.

$\frac{\mathrm{Var(Q)}}{\mathrm{Var}(D)} \geq 1 + \frac{2L}{p} + \frac{2L^2}{p^2}$ (where L is the leading time, p is parameter of demand variations)

As shown in Fig. 157.2, according to the lower limit of demand variability in different leading time L, the demand variations $f(p) = 1 + \frac{2L}{\mathrm{p}} + \frac{2L^2}{P^2}$ is function of P, while p is very large and L is very short, the bullwhip effect caused by prediction error can be neglected. Bullwhip effect will magnify with the increase of p and the decrease of the leading time.

157.2.3 The Affection of Moving Average Forecast Method's to Bullwhip Effect

Reference [2] proves that, the fluctuations of market demand information forecasted by using simple moving average method (where $p > 5$) is smaller than the actual fluctuations. This prediction method is used in this chapter.

157.2.4 Moving Average Method

Maximum inventory levels are calculated as follows: $L \times \mathrm{AGV} + z \times \mathrm{STD} \times \sqrt{L}$

where AGV is the average value of customer's day (or week) requirement, and STD is the standard deviation of that. Constant value z is the safety factor which can be found out from the statistical table. The function of z is to guarantee the probability that shortages do not occur during the leading time equal to a given service level. In order to implement this inventory policy, retailers have to estimate the average value and the standard deviation in accordance with the customer demand data. Therefore, the maximum inventory level will respond to changes of the average value and the standard deviation.

Specifically, the maximum inventory level y_t during t is estimated according to the observed demand: $y_t = u_t L + z\sqrt{\mathrm{L}}\mathrm{S}_t$ where $u_t = \frac{\sum_{i=t-p}^{t-1} D_i}{p}$, $S_t^2 = \frac{\sum_{i=t-p}^{t-1} (D_i - u_t)^2}{p-1}$. In each terms, retailers should calculate new average value and standard deviation according to the recent p demand observations.

157.3 The Impact of Different Information Treatment on Prediction

Considering the value of sharing customer demand information within the supply chain, the supply chain described as in Fig. 157.1 is used in this chapter. For determining the impact of centralized demand information on bullwhip effect, two types of supply chain should be distinguished, one uses the centralized mode to process demand information, and the other one uses the scattered mode.

157.3.1 Centralized Mode to Process Demand Information

In this way, retailers of the first stage in the supply chain observe customers' requirements; forecast the average demand by using moving average P ($p > 5$) demand observations, which is projected to identify target inventory levels. Then orders are issued to wholesalers, the second phase of the supply chain. When the wholesalers received orders and the average forecast demand data, they use the predicted values to determine their target inventory levels, and then issued orders to distributors, the third stage of the supply chain. Similarly, the distributors receive the orders and the average forecast demand information, the use of the predictive value to determine the target inventory level and place orders to the fourth stage of the supply chain—manufacturers.

In this mode, each stage of the supply chain receives the average demand information from the retailers, and then determines the min–max inventory policy according to it. Therefore, demand information, forecasting methods and inventory strategies are put together in this chapter.

Based on the above analysis, it can be elicited that the relationship between $\mathrm{Var}\left(Q^k\right)$ (the variance of order placing in each stage) and $\mathrm{Var}(D)$ (the variance of customer demand) is as follows.

Stage 1:

$$\frac{\mathrm{Var}(Q^1)}{\mathrm{Var}(D)} \geq 1 + \frac{2L_1}{p} + \frac{2L_1^2}{p^2}$$

;

Stage 2:

$$\frac{\mathrm{Var}(Q^2)}{\mathrm{Var}(D)} \geq 1 + \frac{2(L_1 + L_2)}{p} + \frac{2(L_1 + L_2)^2}{p^2}$$

Stage 3:

$$\frac{\mathrm{Var}(Q^3)}{\mathrm{Var}(D)} \geq 1 + \frac{2(L_1 + L_2 + L_3)}{p} + \frac{2(L_1 + L_2 + L_3)^2}{p^2} \cdots\cdots$$

Stage K:

$$\frac{\mathrm{Var}\left(q^k\right)}{\mathrm{Var}(D)} \geq 1 + \frac{2\sum_{i=1}^{k-1} L_i}{p} + \frac{2\left(\sum_{i=1}^{k-1} L_i\right)}{p^2}$$

where D is the custom demand (received from the retailers);

K is the difference in supply chain including retailers, wholesalers, distributors, manufacturers, and so on;
q^k is the order amount placed from stage $k - 1$ to stage k;
L_i is the leading time between stage i and stage $i + 1$;
P is the capacity of the demand sample: the order observations;
Var(D) is the variance of customer demand received from the retailers;
Var(q^k) is the variance of demand in stage K, or it can also be described as the variance or demand order amount placed from stage $k - 1$ to stage k.

157.3.2 Decentralized Mode to Process Demand Information

In this way, wholesalers predict the average demand based on moving average method according to $p(p > 5)$ observations (that is to say the recent p orders from retailers), determine the target inventory levels, and then place orders to each distributors. Distributors make moving average according to the p orders placed by wholesalers, establish the average value and the standard deviation, determine the target inventory levels and then place orders to manufacturers. Based on the above analysis, the relationship between variance of orders placed in each stage $\mathrm{Var}\left(Q^k\right)$

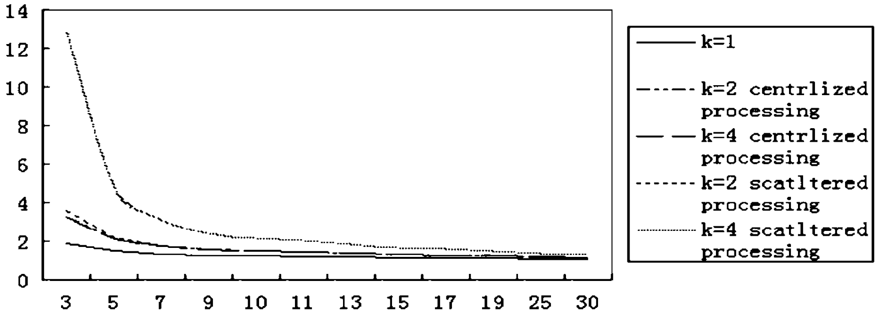


Fig. 157.3 Degree of changes under centered processing mode and scattered processing mode

the variance of orders placed in each stage) and Var(D) (the variance of customer demands) can be deduced as follows.

Stage 1: Stage 2:

$$\frac{\mathrm{Var}(Q^2)}{\mathrm{Var}(D)} \geq \left(1+\frac{2L_1}{p}+\frac{2L_1^2}{p^2}\right)\left(1+\frac{2L_2}{p}+\frac{2L_2^2}{p^2}\right)$$

Stage 3:

$$\frac{\mathrm{Var}(Q^2)}{\mathrm{Var}(D)} \geq \left(1+\frac{2L_1}{p}+\frac{2L_1^2}{p^2}\right)\left(1+\frac{2L_2}{p}+\frac{2L_2^2}{p^2}\right)\left(1+\frac{2L_3}{p}+\frac{2L_3^2}{p^2}\right)\ldots\ldots$$

Stage K:

$$\frac{\mathrm{Var}\left(q^k\right)}{\mathrm{Var}(D)} \geq \prod_{i=1}^{k-1}\left(1+\frac{2L_i}{p}+\frac{2L_i}{p^2}\right)$$

157.3.3 Comparison of the Two Treatments

All references [2–4] described that, no matter which prediction method it chose, variance of order amount will gradually increase between each stage in supply chain. Thus variability of orders placed by wholesalers is bigger than it placed by retailers. The difference between the two demand information treatments is that the demand increment from one stage to another is different.

Figure 157.3 shows the comparison of the two treatments. By using the centralized mode, order variance is calculated in total leading times and increasing in a simple superimposing way [5, 6]. By using the scattered mode, order variance redoubled. Therefore, if the demand information is attainable in each stage of the supply chain, bullwhip effect can be reduced significantly.

157.4 Conclusion

Through the above analysis and discussion, the method proposed in this chapter can control bullwhip effect effectively. It is conducive for setting the safety stock. For retail enterprises, it can help to determine customer demand reasonably while guaranteeing the service levels remain still. Thereby, the variability of orders to manufacturers will be reduced significantly, the cost will be saved, and competitiveness of enterprises can be enhanced.

References

1. Simchi-Levi D, Kaminsky P (2004) Designing and managing the supply chain. China Financial and Economic Publish House, China
2. Wan J, Li MQ, Kou JS (2003) Analysis and control on the bullwhip effect in the course of forecasting and processing of demand information. J Ind Eng Eng Manag 4:370–373
3. Xie H (2009) Comparison of the influence of different demand forecasting methods upon the bullwhip effect. Logist Technol 12:207–210
4. Liu H, Wang P (2007) Bullwhip effect analysis in supply chain for demand forecasting technology. Syst Eng Theory Pract 7:27–33
5. Lee H, So KC, Tang C (2000) The value of information sharing in a two-level supply chain. Manag Sci 46:626–643
6. Li G, Wang S, Yan H, Yu G (2005) Information transformation in a supply chain: a simulation study. Comput Oper Res 32:707–725

Chapter 158
An Epidemiological and Analysis of Sleep Problems of Children Aged 0–7 in Fengrun District, Tangshan

Ling Xue, Shulan Pang, Weijun Guan, Ruigeng Liu and Yang Zhao

Abstract To understand the status, the related factors of sleep problems of children aged 0–7, and the influence on their health, 831 children aged 0–7 years old were sampled from Fengrun district of Tangshan. Questionnaires responded by the parents showed that the children sleep for 11.22 ± 1.69 h in daytime and 9.03 ± 1.19 h at night. A total of 22.3% of the children reported to have sleep problems in which 23.1% were girls and 21.3% were boys. The three most frequent sleep problems were night sweating (8.7%), early waking (4.3%), and snoring (3.8%). The risky factors that initiated sleep problems includes mothers' anxiety, sleep disorder of fathers, children's catching cold, snore of fathers, and the difficult temperament of children. The survey showed that there was certain sleep problems among children aged 0–7 years in Fengrun of Tangshan. The possible negative influence on the growth and development of children deserves more attention from both the parents and those who work in medical industry.

Keywords Children · Sleep problems · Epidemiology · Survey

158.1 Introduction

Man spends 1/3 of his life in sleep as it is an important physiological process for body's recovery and adjustment. Enough sleep can help children to develop both intellectually and physically. With the development of modern medicine,

L. Xue (✉) · S. Pang · W. Guan · R. Liu · Y. Zhao
Hebei Province Key Laboratory of Occupational Health
and Safety for Coal Industry, Division of Maternal,
Child and Adolescent Health, Hebei United University,
Tang Shan 063000, China
e-mail: xuel2009@gmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_158,

infectious and malnutrition diseases that once seriously affected children's health had decreased significantly, which makes the children's sleep problems relatively more prominent. Sleep problems referred to a variety of psychological disorder during sleep. Symptoms related to sleep problems are snoring, waking by choke, mouth breathing, sleep apnea, squirm, sweating, body twitching, enuresis, etc. Chronic sleep problems can lead to neurohumoral changes. Slight symptoms are inattention, mental retardation, growth retardation, and growth difficulties. Critical disorder can cause sudden death during sleep [1]. In this study, epidemiological survey was adopted to know about the general situation of sleep problems in Fengrun district. The results would provide basis for prevention of sleep disorders of children, improvement of the sleep quality and the promotion of children's physical and mental development.

158.2 Study Objectives and Methods

158.2.1 Study Population and Data Collection

This was a cross-sectional survey of samples selected randomly from Fengrun district in Tangshan city in May, 2008. Eight hundren and thirty one community children aged 0–7 joined the survey. The children surveyed must live with their parents or at least live with one parent in the past year, and have no serious congenital or infectious diseases. In this study, 896 questionnaires were distributed, and 831 questionnaires were valid, which made the response rate 92.7%. The gender proportion of male and female were 53.1 and 46.9%. Age was categorized as <1, 1-years, 2-years, 3-years, 4-years, 5-years, and 6–7 years. All the respondents acknowledged the assessments and the study.

158.2.2 Study Questionnaire

The questionnaire made reference to the revised version of clinical questionnaire of the Sleep Center at Children's Hospital, University of Sydney, Australia [2]. On this basis, the questionnaire was revised according to the situation in Tangshan and included seven parts, a total of 61 points. The questionnaire covered age, sex, nationality, growth and development history, disease history, birth and feeding conditions, sleeping conditions in the past month, general family conditions, living condition, parents' sleep condition, etc. The questionnaires were completed by the parents.

158.2.3 Diagnostic Criteria for Sleep Problems

In this study, symptoms of sleep problems include the following, snore, throat choke, apnea, squirm, mouth breathing, sweating, body twitching, odontoprisis, somniloquy, sleepwalking, enuresis, awake by choke, difficulty in falling asleep, early awakening, nasal congestion, nightmares, night terrors, limb pain, early sleep, and nocturnal waking. No children have other organic diseases than respiratory diseases. And enuresis is for children over 5 years only. The criteria of sleep problems were that any of the above symptoms occur once or more than once a week.

158.2.4 Statistics Analysis Method

Statistical analysis was carried out by using SPSS 12.0 software. The statistical description of the socio-demographic and general sleeping conditions variables was performed by using frequencies, percentages and chi-square test. Non-conditional logistic regression analysis was used in multivariate statistical analysis of sleep problems.

158.3 Results

158.3.1 Demographic Characteristics of the Participants

Totally 896 questionnaires were distributed, and 831 were eligible for data analysis, which means the valid rate was 92.3%. In this 441 (53.1%) participants were males, 390 (46.9%) were females. The proportions of the participants of different age were 155 (18.7%) for 0 year old, 21 (2.5%) for 1 year old, 20 (2.4%) for 2, 140 (16.8%) for 3, 175 (21.1%) for 4, 238 (28.6%) for 5, 76 (9.1%) for 6 and 6 (0.7%) for 7, respectively. Most of participants (68.9%) were 2–6 years old.

158.3.2 General Situation of Sleeping

This study assesses general situation of children' sleep by studying the sleep time. The results are shown in Table 158.1. The comparison between all day sleep time and night sleep time among different age group was of statistical significance.

Table 158.1 Frequency distribution of the participants' sleep situation ($n = 831$)

Ages group	Cases	Total sleep time (h)	Night sleep time (h)
0-	155	12.50 ± 2.66	8.70 ± 1.92
1-	21	11.88 ± 1.60	9.00 ± 1.52
2-	20	11.08 ± 1.20	9.05 ± 0.95
3-	140	11.26 ± 1.01	9.34 ± 0.91
4-	175	10.92 ± 1.03	9.15 ± 0.82
5-	238	10.67 ± 1.23	8.96 ± 0.96
6-7	82	10.54 ± 1.10	9.09 ± 0.87
Total	831	11.22 ± 1.69	9.03 ± 1.19

Table 158.2 Comparison between children of different ages regarding the incidence of sleep problems ($n = 831$)

Ages group	Headcount	Sleep problems		χ^2	P
		Cases	Incidence (%)		
0-	155	54	34.8	24.076	0.001
1-	21	7	33.3		
2-	20	6	30.0		
3-	140	18	12.9		
4-	175	35	20.0		
5-	238	48	20.2		
6–7	82	17	20.7		
Total	831	185	22.3		

158.3.3 Status of Sleep Problem

The incidence of sleep problems among the 831 children was 22.3% (185/831), and that of boys' sleep problems was 23.1% (102/441) while girls' was 21.3% (83/390). The difference of the incidence between boys and girls is not of statistical significance ($\chi^2 = 0.408$, $P = 0.523$), but there were significant differences among different ages ($\chi^2 = 24.076$, $P < 0.001$). Children of younger age generally had higher incidence of sleep problems, as shown in Table 158.2.

158.3.4 Related Symptoms of Sleep Problems of Children

Related symptoms of sleep problems are presented in Table 158.3. In this study, the top three high incidence of sleep problems were sweating at night (8.7%), early waking (4.2%), and snoring (3.7%).

158.3.5 Analysis of the Course of Sleep Problems of Children

In this study, the single factor analysis was processed first. Sleep problems were not regarded as the dependent variable. The independent variable were including

Table 158.3 The incidence of related symptoms of sleep problems

Symptoms	Male ($n = 441$)		Female ($n = 390$)		Total ($n = 831$)	
	Cases	Incidence (%)	Cases	Incidence (%)	Cases	Incidence (%)
Throat choke	1	0.2	0	0.0	1	0.1
Apnea	1	0.2	0	0.0	1	0.1
Sleepwalking	0	0.0	1	0.3	1	0.1
Enuresis	1	0.2	2	0.5	3	0.4
Night terrors	1	0.2	3	0.8	4	0.5
Awake by choke	1	0.2	4	1.0	5	0.6
Nightmare	2	0.5	3	0.8	5	0.6
Limb pain	2	0.5	3	0.8	5	0.6
Sleep early	4	0.9	4	1.0	8	1.0
Nasal congestion	6	1.4	3	0.8	9	1.1
Awake at night	3	0.7	7	1.8	10	1.2
Somniloquy	5	1.1	5	1.3	10	1.2
Body twitching	8	1.8	2	0.5	10	1.2
Mouth breathing	11	2.5	9	2.3	20	2.4
Odontoprisis	18	4.1	4	1.0	22	2.6
Squirm	9	2.0	13	3.3	22	2.7
Difficulty in falling asleep	8	1.8	15	3.8	23	2.8
Snore	19	4.3	12	3.1	31	3.7
Early awakening	20	4.5	15	3.8	35	4.2
Sweating	40	9.1	32	8.2	72	8.7

Table 158.4 The unconditional logistic regression analysis of sleep problems

Variable	$\widehat{\beta}$	Wald χ^2	P	OR	OR 95% C.I.
Difficult temperament	0.857	3.949	0.047	2.357	1.012 ~ 5.488
Snore of father	1.035	7.626	0.006	2.815	1.350 ~ 5.869
Cold of children	1.282	11.173	0.001	3.602	1.699 ~ 7.637
Sleep disorder of father	1.782	4.806	0.028	5.942	1.208 ~ 29.235
Anxiety of mother	1.837	4.903	0.027	6.278	1.235 ~ 31.914

seven parts, a total of 61 issues, just as situation of mother during pregnancy, conditions of birth and feeding, situations suffering from respiratory diseases, general situation of family, living condition, parents sleep condition, children behavior. An unconditional logistic regression analysis was applied in this study. Inclusion criteria was $\alpha = 0.05$. The dependent variable was having sleep problems or not, the independent variables were 24 statistically significant variables in the single factor analysis. The results observed were that risk factors of sleep problems including anxiety emotion of mother, sleep disorder of father, cold of children, snore of father, and the difficult temperament of children. (Table 158.4).

158.4 Discussions

Infant sleep quality of children directly affected children's brain development and hormone levels The study about infant sleep problems was significant for children's growth and development. In this study, 831 children of 0–7 years old accepted the survey. The mean of total sleep time was (11.22 ± 1.69) h, which was decreasing with the increasing age. The result was higher than the data Shaoping Yang got in the analysis on sleep situation of 3–6 years old children in Wuhan (10.04 ± 1.82) h. The incidence of sleep problems was 22.3%, which is consistent with the most study results abroad (20–25%) [3, 4], but significantly lower than the reported results in Shanghai of 1–23 months children and in Sweden [2, 5].

The study results revealed that primary risk factors of sleep problems include anxiety of mother, sleep disorder of father, cold of children, snore of father, and the difficult temperament of children. The children whose mother had anxiety was 6.278 times more possible to have sleep problems than those whose mother had no anxiety. The mother being down in spirits, anxiety, or depression during pregnancy period or in long-term, the risk of sleep disorders in children would increase, consistently with Armstrong's results [6]. If the parent suffered from anxiety or depression for long-term, the negative emotions would inevitably affect the neuropsychological development of children and these changes would be reflected during sleep directly or indirectly. However, the conclusion does not exclude that sleep behaviors of children was affected by negative emotions of parents through inheritage [7]. Those children, whose fathers have sleep disorders, had higher risk of sleep problems. In this survey, compared with parents without sleep disorders, the incidence of sleep problems of children whose parents had sleep disorders had increased by 4.942 times. The results also showed that father's snoring affected children's sleep. In China, infants often sleep with their parents in a bed or a room. Children's sleep environment and quality may be affected by fathers snoring. In addition, respiratory diseases had the highlighted impact on children's sleep. Cold was a risky factor to sleep problems in this study. The occurrence of parasomnias might be related to the normal development of children. Due to the growth, the airway of respiratory system became arrow due to teeth and pharyngeal lymphoid tissue (tonsil and adenoid) in the physiological development peak, so prone to somniloquy, odontoprisis, and snore [8]. If having a cold at this time, symptom of airway obstruction would gradually aggravate by tonsil and adenoid hypertrophy, caused by chronic inflammation or allergic lesions of mouth, pharynx and nose, and it may result in sleep disorder.

In abroad study in infant, children of difficult temperament were prone to wake up significantly more frequently than those easily raising children at night [9]. The results also showed in this study, children of difficult temperament were more prone to sleep problems. Psychological and behavioral problems tended to occur to children of difficult temperament when not adapted to the environment. Children in difficult temperament were sensitive to living and sleep environment, and difficult to adapt to strange surroundings, so the sleep would be affected.

For children, sleep is an important guarantee for their growth and development. The impeccable structure of sleep had a very important role to development and maturation of the central nervous system. Sleep quality affected children's body and mind development directly. Therefore, it was of great significance to be more concerned with children's sleep problems so as to prevent potential unhealthy factors that may influence the physical and mental health of children and create a good sleep environment.

References

1. Morrison DN, McGee R, Stanton WR (1992) Sleep problems in adolescence. J Am Acad Child Psychiatry 31:94–99
2. ThunstrÖm M (1999) Severe sleep problems among infants in a normal population in Sweden: prevalence, severity and correlates. Acta Paediatr 88(12):1356–1363
3. Qian W, Wei Y (2004) An epidemiological study on sleep problems in children aged 2 to 12 years old in Gangzhou. J Appl Clin Pediatr 19(12):1078–1080
4. Mindell JA, Owens JA, Carskadon MA (1999) Developmental features of sleep. Child Dolesc Psychiatr Clin N Am 8(4):695–725
5. Fan J, Chonghuai Y, Shenghu W et al (2003) An epidemiological study on sleep problems in children aged 1 to 23 months in Shanghai. Chin J Prev Med 37(6):435–438 November
6. Armstrong KL, O'Donnell H, McCallum R et al (1998) Childhood sleep problems: association with prenatal factors and maternal distress/depression. Pediatric Child Health 34(3):263–266
7. Shaoping Y, Anna P, Bin Z et al (2007) Analysis on sleep characteristics and its risk factors of preschool children in Wuhan city. Chin J Public Health 23(2):164–165
8. Liu Y, Liu H, Yang H et al (2007) Survey on 1007 dyssomnia cases of preschool children. Med Soc 20(3):39–41
9. Wolke D, Söhne B, Riegel K et al (1998) An epidemiologic longitudinal study of sleeping problems and feeding experience of preterm and term children in southern Finland: comparison with a southern German population sample. J Pediatr 133:224–231

Chapter 159
Study on Medical Scale of Reproductive Women in Rural Areas and Its Influencing Factors Analysis

Xiao Dong Xie, Hui Sun and Yong Hong Xiao

Abstract To know the medical scale of child-bearing age women in rural areas and its influencing factors and to provide a reasonable basis for the allocation of medical resources for the government, 903 women of child-bearing age in rural areas of Hebei Province were chosen with multistage cluster sampling method and their medical scales were investigated by the self-scale questionnaire, then data were analyzed with χ^2 test and logistic regression method. The results showed that the medical scale below county grade was 76.0% for child-bearing age women in rural areas. Occupation, education level, number of pregnancies, induced abortion times were the main factors influencing the medical scale of child-bearing age women. So the government should strengthen country health services and expand the scope of services to provide appropriate health service environment. With the national medical health system reform, the government has reinforced the medical facilities to provide the health service for rural residents, especially for child-bearing age women, to make the health service near, convenient and inexpensive. But the related study on rural women of child-bearing age medical scale and its influencing factors are still in the blank fields, therefore, the paper studied the medical scale of reproductive women in rural areas and its influencing factors analysis to make government plan health service and reasonably distribute the medical resources.

Keywords Women of child-bearing age · Influencing factors · Medical scale

X. D. Xie · H. Sun · Y. H. Xiao (✉)
Department of Prevent Medicine, North China Coal Medical College,
Tangshan, Hebei, China
e-mail: kycxyh@tom.com

X. D. Xie
e-mail: xexodg@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_159,

159.1 Subjects and Methods

159.1.1 Survey Object

In Hebei province, random sample of ten countries were chosen using a multi-stage cluster sampling method, each country was randomly selected as one township, then each village was collected in the township, in which all child-bearing age women from 20 to 59 years old, the total survey 903 cases were regarded as survey objects.

159.1.2 Methods

Using self-designed questionnaire, all objects were investigated by face-to-face interview survey. Content of questionnaire included: (1) General situation includes seven factors: sex, age, ethnicity, education level, occupation, marital status and income. (2) Reproductive health: including whether the patient suffered gynecological diseases, pregnancy or artificial abortion times. (3) Medical scale: ① county and township level and below, ② county-level, ③ above the county-level.

159.1.3 Statistical Analysis

Using SPSS13.0 statistics soft, univariate analysis was carried with χ^2 test, multivariate logistic regression was used to analyze the influencing factors.

159.2 Results

159.2.1 Information Overview

In total 903 cases of child-bearing age women from 20 to 59 years old, the average age was 34.28 years old, cultural degree in high school and above was 44.3%, elementary school 39.3 and 16.4% junior high school, marital status in first-marriage was 74%, other 26%.

Table 159.1 Univariate analysis of general situation influencing medical scale

Variable	Influencing factors	Group below	Below county	County	Above county	χ^2	P
*X*1	Age (years)	20～	78	224	72	1.492	0.474
		30～	68	209	90		
		40～	50	139	55		
*X*2	Education	Primary school	53	77	28	23.639	0.000
		Junior high school	80	235	79		
		High school and above	63	260	110		
*X*3	Professional	Farmers	108	234	96	10.254	0.006
		Workers	42	154	46		
		Intellectuals	21	98	50		
		Individual	25	88	25		
*X*4	Annual household income (million)	<1	57	202	75	0.904	0.636
		1～	92	229	87		
		2～	47	141	55		
*X*5	Marital status	First-marriage	178	522	75	15.135	0.001
		Other	Cases	38	50		

159.2.2 Univariate Analysis Influencing Medical Scale

159.2.2.1 General Situation

There were significant differences in the medical scale among different marital status, education level and occupation ($P < 0.05$) (Table 159.1).

159.2.3 Reproductive Health

The effects of gynecological diseases, pregnancies, abortion on the medical scale were significant ($P < 0.05$), (Table 159.2).

159.2.4 Multivariate Analysis

Choosing the medical scale as a dependent variable, significant factors by single factor analysis were analyzed with logistic regression model. These factors were assigned in Tables 159.3 and 159.4.

Discuss: The influence could be seen from the table, affecting women of childbearing age to medical scale are various factors [1], Logistic regression results

Table 159.2 Analysis of the reproductive health influencing medical scale

Variable	Influencing factors	Group below	Below county	County	Above county	χ^2	*P*
*X*6	Gynecological diseases	Had not contracted	136	235	104	9.255	0.010
		Suffered	89	216	121		
*X*7	Number of pregnancies (time)	1	61	210	82	10.072	0.007
		2	60	189	64	23.639	0.000
		3 or more	50	72	39		
*X*8	Abortion (times)	0 not done	109	314	109	7.880	0.019
		1	31	115	67		
		2 or more	40	72	29		

Table 159.3 The variables and their assignment

Factors	Variable name	Assignment		
Education	*X*2	Primary school = 1 Junior high school = 2 High school = 3		
Professional	*X*3(1) farmers	1	0	0
	*X*3(2) workers	0	1	0
	*X*3(3) intellectuals	0	0	1
	*X*3(4) individual	0	0	0
Marital status	*X*5	1 = first-marriage	0 = other cases	
Gynecological diseases	*X*6	1 = suffered	0 = had not contracted	
Number of pregnancies (time)	*X*7(1) 1	1	0	
	*X*7(2) 2	0	1	
	*X*7(3) 3 or more	0	0	
Abortion (times)	*X*8(1) 1	1	0	
	*X*8(2) 2 or more	0	1	
	*X*8(3) Not done	0	1	
Medical scale	*Y*	1 = county-level and below 2 = county-level 3 = county-level and above		

showed: (1) when county-level and below compared with above the county-level in the medical scale choice, from occupation, farmers and workers choosing the county and township level were more than individual ($P < 0.05$), because county-level and below may be a simple procedure, need not spend a very long time and medical expenses. There was no significant difference in medical scale choice between career 3 (Intellectuals) and 4 (Individual) ($P > 0.05$), while intellectuals accounted for lower in the composition of the sample. The medical scale choosing county-level and below in primary school culture was more than high school degree and above ($P < 0.05$), but there was no significant difference in the medical scale choice between middle schools and high schools ($P > 0.05$), and there was

Table 159.4 Logistic regression results

Medical scale		B	Std. error		χ^2	df	95% confidence interval for exp (B)	
P	OR						Lower bound	Upper bound
1 Intercept	0.554	0.241	5.269	1	0.022	–	–	–
[Professional = 1]	1.044	0.345	9.172	1	0.002	2.841	1.455	5.584
[Professional = 2]	0.704	0.303	5.411	1	0.020	1.495	0.273	0.895
[Professional = 3]	−0.647	0.441	2.148	1	0.143	0.524	0.221	1.224
[Professional = 4]	0b	.	.	0	.	.	.	.
[Education = 1]	0.973	0.352	7.642	1	0.006	2.645	1.327	5.270
[Education = 2]	0.453	0.293	2.397	1	0.122	1.573	0.886	2.793
[Education = 3]	0b	.	.	0	.	.	.	.
[Pregnancies = 1]	−0.526	0.299	3.102	1	0.078	0.591	0.329	1.061
[Pregnancies = 2]	−1.083	0.329	10.824	1	0.001	0.339	0.178	0.645
[Pregnancies = 3]	0b	.	.	0	.	.	.	.
[Abortion = 1]	−0.397	0.288	1.903	1	0.168	0.672	0.383	1.182
[Abortion = 2]	−0.579	0.294	3.884	1	0.049	0.560	0.315	0.997
[Abortion = 3]	0b	.	.	0	.	.	.	.
2 Intercept	1.342	0.349	14.757	1	0.000	–	–	–
[Professional = 1]	−0.540	0.298	3.291	1	0.070	0.583	0.325	1.044
[Professional = 2]	−0.069	0.308	.050	1	0.823	0.934	0.511	1.706
[Professional = 3]	−0.628	0.282	4.967	1	0.026	0.533	0.307	0.927
[Professional = 4]	0b	.	.	0	.	.	.	.
[Education = 1]	0.185	0.304	.370	1	0.543	1.203	0.663	2.184
[Education = 2]	0.717	0.241	8.852	1	0.003	2.049	1.277	3.287
[Education = 3]	0b	.	.	0	.	.	.	.
[Pregnancies = 1]	−0.017	0.259	.005	1	0.946	0.983	0.592	1.632
[Pregnancies = 2]	−0.526	0.274	3.681	1	0.055	0.591	0.345	1.011
[Pregnancies = 3]	0b	.	.	0	.	.	.	.
[Abortion = 1]	−0.024	0.245	0.010	1	0.922	0.976	0.605	1.577
[Abortion = 2]	−0.702	0.281	6.228	1	0.013	0.496	0.286	0.860
[Abortion = 3]	0b	.	.	0	.	.	.	.

no significant difference between artificial abortion 1–3, pregnancy induced abortion number 1–3 times and pregnancy ($P > 0.05$), while compared with pregnancy 3 times, pregnancy 2 times more inclined to choose at or above the county-level on the scale ($P < 0.05$). As the pregnancy and abortion increased, the number selecting the higher scale increased, there medical conditions may be a good major hospitals, doctors may be on a high standard [2] (2) Town and at or above the county-level, when compared between town and at or above the county-level, there were no significance difference in the medical scale between career 1 and 4, career 2 and 4 ($P > 0.05$), women of career 3 and 4 more likely chose the county and above scale of a medical institution ($p < 0.05$), intellectual economy condition is good, their health sense is strong. Cultural degree 1 and 3 in choosing medical scale was no difference ($P > 0.05$), there was no significant difference in choosing the medical scale between cultural degree two and three ($P < 0.05$). The reproductive health understanding and demand in high education degree women are higher and can actively take own of rights [3]. There was no significant difference in choosing the medical scale between abortion 1 and 3, artificial abortion 2 and 3, the number of pregnancies 1 and 3 ($P > 0.05$), women of pregnancy 2–3 times more likely chose above the county-level hospitals ($P < 0.05$). We can find out the medical scale choice of child-bearing age women pay attention to saving time, simple procedures, low healthcare costs and so on [4] therefore government should improve and perfect the county hospital as the center, regarding village hospital at or above the county-level with the village for hub, rural health act foundation three-level medical and health service system, forming ailment out township, a serious illness into hospital medical services in the new pattern [5].

159.3 Conclusion

It was necessary for the related department to increase the financial investment for town and village medical institution, to improve the medical condition of village clinics and improve the new rural cooperative medical system, so as to ensure the reproductive health of rural child-bearing age women.

References

1. Zhang Y, Zhang L, Chong JL, Yang O (2009) Seeking behavior of rural populations factors. J Hosp Adm, 05
2. Changfu MA, Xuehui S, Xiuwei C (2009) Residents choose a different medical institutions factors affecting the analysis of the Chinese hospital management 05
3. Yu kai DU, Wei min F, Cheng-Liang X (2004) and so informed choice of contraceptive women of childbearing age influencing factors of public health, vol 20 (2)
4. Dong Fu Q, Ai tian Y, Qingyue M et al (2007) Changes in the flow of rural patients with medical trend analysis. Chin Health Serv Manag 23(12):845–847
5. Zhao Yu Y, Na Z (2007) Rural residents seeking behavior and its influencing factors—based on a survey in Northern Town X [J]. Nanjing Agric Univ (Soc Sci) 7(3):12–17

Chapter 160
The Exploratory Research Using BP Neural Network to Analyze the Influencing Factors of Hospitalization Expenses in Acute Appendicitis

Jianhui Wu, Jie Tang, Guoli Wang and Sufeng Yin

Abstract It has no requisition to the distribution of hospitalization expenses using BP neural network, and the network can fit the complex relations between input and output variables. The article establishes the model of BP neural network, and analyzes influencing factors of hospitalization expenses in acute appendicitis. The results display: the first factor influencing hospitalization expenses is the days in hospital. In order to control expenses of acute appendicitis, hospitals should improve the levels of diagnosing and treating, and decurtate days in hospital.

Keywords Neural network · Hospitalization expenses · Acute appendicitis · Influencing factors · Days in hospital

160.1 Introduction

In recent years, the problem of expensive hospitalization expenses in our country has been concerned. The research to analyze the influencing factors of hospitalization expenses and make medical resources reasonable has become a focus. The researches before [1] concerning the influencing factors of hospitalization expenses in acute appendicitis mostly use multiple linear regression method. However, expenses are generally skewed distribution and influenced by many complex factors; multiple linear regression method is limited in the application.

J. Wu (✉) · J. Tang · G. Wang · S. Yin
Hebei Province Key Laboratory of Occupational Health
and Safety for Coal Industry, Division of Epidemiology
and Health Statistics, School of Public Health,
Hebei United University,
Tang Shan 063000, China
e-mail: wujianhui555@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_160,

This article analyzes influencing factors of hospitalization expenses in acute appendicitis by the establishment of a BP neural network model.

160.2 Principle of BP Neural Network

Back propagation (BP) network is a multi-layer mapping network that can back propagate errors and modify the internal structure. It can solve the smallest error of mean square between actual output and expected output. When the parameters appropriate, the network can conve a smaller mean square and is widely used in many aspects.

BP network includes one input layer, one output layer and one or more hidden layers, and neurons of different layers are connected by weights. The information propagates from input layer processed by hidden layer to output layer, if not the expected output, then turn back propagation and recursive calculate the error. Gradually adjust the weights in each layer when error propagates and ultimately make the error within the permitted extent [2]. BP neural network has no requisition to the distribution of hospitalization expenses, and can fit the complex relations between input and output variables.

160.3 Principle of Sensitivity Analysis

The sensitivity analysis based on the model of BP neural network is to change part of the input and observe the change of output, and to determine the degree of importance of each input variable. The categorical variables use normalized values, and continuous variables use four equal portions of the split point, that is 0, 0.25, 0.5, 0.75 and 1. Changing the inputs, recording the differences between the maximum and minimum outputs, then calculating the proportions of maximum output that the differences account for. Ultimately the sensitivity is the mean of all the recorded proportions [3].

160.4 Neural Network Model and Sensitivity Analysis

160.4.1 Case Study

Collect information of acute appendicitis patients from 2007 to 2009 in some hospital in Tangshan City, and delete cases with missing values, ultimately the effective cases are 1,083. Select variables of gender, marital status, whether the first admission, the cost categories, treatment outcome, admission condition, age,

Table 160.1 Factors and code

Code	Factor	Specific coding
$X1$	Gender	1 = male, 2 = female
$X2$	Marital status	1 = married, 2 = others
$X3$	Whether the first admission	1 = the first time, 2 = not the first time
$X4$	Cost categories	1 = average patient, 2 = health insurance, 3 = cooperative medical care, 4 = retired cadres, 5 = other insurance
$X5$	Treatment outcome	1 = heal, 2 = improve, 3 = healed
$X6$	Admission condition	1 = general, 2 = acute, 3 = disease risk
$X7$	Age	(years)
$X8$	Days in hospital	(days)
$X9$	Whether surgery	0 = not surgery, 1 = surgery
$X10$	Whether other diagnoses	0 = none, 1 = with other diagnoses
Y	Total expenses	(dollars)

days in hospital, whether surgery, whether other diagnoses and expenses, then encode in excel (Table 160.1).

160.4.2 Methods

Establish BP neural network and sensitivity analysis by Matlab7.1.0. Take expenses as output variable, and the other 10 variables as inputs. Programing and selecting different number of neurons in hidden layer, different training algorithm, repeatedly train the data for 100 times, and set parameters of the network. Finally establish one model that fitting ability and simulation capability of the network reach an appropriate level. Then by sensitivity analysis, observe the main factors that have impaction on hospitalization expenses.

160.4.3 Results of Neural Network Modeling

Normalize the data to [0,1], and set transfer-function between hidden and output layers as logsig function. Set indicator of error performance of network as SSE, and its target as 0.05. The maximum training step is set to 1000. Through repeated training of different numbers of neurons in hidden layer and different training algorithms, the eventual model and main parameters of the model are established as shown in Table 160.2.

Table 160.2 Main parameters of the network model

Parameters of neurons	Parameters of training	Simulating results of test set	Fitting results of training set
Hidden layer:1	Training algorithm:trainlm	R = 0.81625	R = 0.86211
Neurons in hidden layer:15	Training times when stopped: 12	$R2$ = 0.66626	$R2$ = 0.74324
Neurons in input layer:10	Learning speed: 0.01	$R_{\text{adj}}^2 = 0.66315$	$R_{\text{adj}}^2 = 0.\ 74085$
Neurons in output layer:1	Indicator of error:*SSE*	SSE = 8.4022e + 008	SSE = 2.6585e + 009
	SSE [a]when stop training:1.29072	MSE = 7.8379e + 005	MSE = 2.4799e + 006
		RMSE = 2236.4	RMSE = 1751.1

[a] SSE is to the data normalized

Table 160.3 Results of sensitivity analysis

Order	Factor	Sensitivity
1	Days in hospital	0.6370
2	Whether surgery	0.4884
3	Age	0.3644
4	Treatment outcome	0.2589
5	Whether other diagnoses	0.2234
6	Cost categories	0.2146
7	Admission condition	0.1977
8	Whether the first admission	0.1955
9	Marital status	0.1047
10	Gender	0.0851

160.4.4 Results of Sensitivity Analysis

Results show that the top three factors are days in hospital, whether surgery and age. Marital status and gender have little influence on expenses. The order of factors is in Table 160.3.

160.5 Conclusion

Althouth the neural network has the advantages such as widely used and high fault tolerance, there are no clear standards in setting parameters of model. It needs to training network of different parameters repeatedly and eventually establish a relatively appropriate model.

In this article, the BP neural network model is established and then conducts sensitivity analysis that is based on information of hospitalization expenses in acute appendicitis. The results show that days in hospital, whether surgery and age are the top three factors influencing expenses of acute appendicitis. This is consistent with the results of some previous studies [4, 5]. Focus on improving the level and efficiency of hospital, and shorter days in hospital on the basis of ensuring proper and high-quality treatment, medical resources can be taken good advantage of and expenses of patients can be reduced [6]. This is one effective way to control hospitalization expenses in acute appendicitis and optimize allocation of health resources.

References

1. Zhang H, Tan P (2009) Analysis of hospitalization fee and its influencing factors of 10296 cases with acute appendicitis. Chin Health Econ Mag 23(3):66
2. Wang J, Chen J (2010) Principle and design tips of BP neural network. China J Health Stat 25(5):547–549
3. Zhang W, Zhu L, Wang J (2011) BP neural network based analysis of factors influencing hospitalization expenses in TCM hospitals. Chin J Hosp Adm 21(3):161–165
4. Liao S (2006) Analysis of hospitalization fee and its influencing factors of 1063 cases with acute appendicitis. J Hosp Stat 10(4):223–224
5. Ye X, Lv J, Tan S (2008) Analysis of influencing factors of hospitalization fee of 660 cases with acute appendicitis. J Hosp Stat 14(1):32–33
6. Yang H (2010) Discussion about shorten the average hospitalization days. Mod Hosp 9(2):100–101

Chapter 161
The Study on Toxicokinetics and Distribution of CdSe Quantum Dots in Rats

**Houjun Xu, Qingzhao Li, Yu Su, Yulan Hao,
Licheng Yan and Haijuan An**

Abstract We aimed to investigate the difference in toxicokinetics and tissue distribution of different sizes and soluble cadmium selenide (Quantum dot) QD in rats. Male SD rats were divided randomly into five groups by size and dissolubility. The tail vein of rats was exposed to each CdSe QD. After the exposure the blood samples were taken from rats in the following time: 1st, 12th, 24th, 36th, 48th, 60th h. The liver, kidney, brain and testes were removed the next day and the concentration of cadmium in the blood and the organs of each group using ICP-AES was monitored. The content of cadmium in the blood had a trend of decreasing gradually after CdSe QDs exposure. The bigger the particle diameter, the sooner the concentration of cadmium in blood and the distribution. The metabolic rate of dissolved CdSe is faster than other indissolvable ones. The metabolism of CdSe QD was a first order two-compartment model. These results reveal the potential risks of their future applications in medicine and in semiconductors.

Keywords Nanoparticles · Cadmium selenide · Quantum dots (QDs) · Toxicokinetics · Distribution

161.1 Background

Nanotechnology offers many benefits in various fields, such as drug delivery, imaging, water decontamination, information and communication technologies, as well as for the production of stronger, lighter materials [1]. Synthesis of nanomaterials

H. Xu (✉) · Q. Li · Y. Su · Y. Hao · L. Yan · H. An
Hebei Province Key Laboratory of Occupational Health and Safety for Coal Industry,
Division of Epidemiology and Health Statistics, School of Public Health,
Hebei United University, Tang Shan 063000, China
e-mail: houjunxu@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_161,

has become increasingly more common since the early 1980s. Various kinds of nanomaterials, such as (QDs), carbon nanotubes and fullerenes, have been synthesized, and quite a few have been commercialized (e.g. CdSe QDs). The nanotechnology market is predicted to be valued at $1 trillion by 2012, so the likelihood of exposure to synthesized nanomaterials will exponentially increase. Thus, there is an immediate need for research to address uncertainties about the health and environmental effects of nanoparticles. The interactions of nanoparticles with cells and tissues are poorly understood in general, but certain diseases have been proven to be associated with uptake of nanoparticles.

161.2 Introduction

Common cadmium selenide is beige or red crystal. It is mainly applied in the field of electron emission and spectral analysis. The median hethel dose (LD50) is 2.425 mg/kg in mice for oral use, so it has the adverse effect for taking and breathing into body. Quantum dots (QDs) are colloidal nanocrystalline semiconductors with unique optical and electrical properties, because of their large specific surface and high chemical activity [2]. As a new type of inorganic fluorophore, it is gaining widespread recognition and is rapidly applied to fluorescent labeling of cellular proteins [3], cell tracking, and even imaging in vivo [4]. Although some reports have evaluated the cytotoxicity of a few of QDs in different cell lines in different circumstances [5–8], little is known about its effect on the health and environment of cadmium selenide QD, and also its general toxicity in vivo and invitro. In fact, many QDs may seem harmless, but they can be destabilized because of their sequestration in tissues and long-term exposure to the bioenvironment. Cell functions and structures could be damaged when cells are exposed to unstable, poorly coated QDs. Even if the QDs are well modified, the potential risks still exist.

161.3 Materials and Methods

SD rats, ♂, clean, 10-weeks old, weight 200−250 g, were purchased from: Beijing Academy of Medical Sciences, Institute of Laboratory Animal Permit No.: SCXK (Beijing) 2005-0013.

Main reagents and instruments CdSeI: size 2−3 nm; CdSeII: size 5−6 nm; CdSeIII: size 7−8 nm; CdSeIV: the surface with a diameter of 2−3 nm wrapped mercaptoacetic acid, soluble; purity of 99.5% specific surface area of 60m^2/g (by the Institute of Environmental Science and Engineering, Nankai University to provide.) Digestion: nitric acid, hydrogen peroxide. Electronic balance, ETHOSA Microwave Digestion System (U.S. MILESTONE products), IR-1000 inductively coupled plasma atomic emission spectrometer (ICP-AES) (Thermo American products) and so on.

Groups of 30 SD rats were randomly divided into 5 groups, namely CdSe I, CdSe II, CdSe III, CdSe IV and the control group, $n = 6$, free drinking water consumption.

Preparation of drug solvent CdSe nano-ionized water was added to 10 ml, 3000 r/min centrifugal 10 min and supernatant, repeated three times. And then adding the concentration of 10 mg/ml saline configured the turbid fluid, so that after full the liquid was evenly ultrasonic. The CdSe I was coated by 2–3 nm mercaptoethanol surface to form soluble cadmium selenide.

Each group of animals was exposed to, respectively, four one-off tail vein injection of cadmium selenide nanoparticles of about 0.1 ml (about 40 mg/kg dose), and control group injected with saline.

Toxicokinetics experimental detection index: in rats after intravenous injection of 1, 15, 45 min exposure and 3, 10, 24 h, respectively, from the tail vein blood 1 ml, at 4°C after adding anticoagulants. By ICP-AES instrument for quantitative analyse of cadmium selenide. Tissue distribution experiment: All the animals in the first 2 days after exposure were killed after ether anesthesia, whichever is the liver, kidney, testis, after digestion with ICP-AES instrument for quantitative analysis of cadmium selenide.

Statistical analysis of data $\bar{x} \pm s$, dealing compartment model with 3p97 software processing and calculation of pharmacokinetic parameters T1/2 (half time), K (elimination constant), V(L) (apparent volume of distribution), AUC (area under the curve), CL (clearance), then analysis of variance with SPSS13.0 software.

161.4 Result of Study Methodology

The method of the experiment specific, high sensitivity, detection limit is 0.15 μg/L. 2d consecutive day precision detection method is the relative standard deviation <5%, by the standard curve r2 $\geq$ 0.99.

161.5 Result of Toxicokinetics

Twenty four hours after exposure in each group, was the time point blood concentrations of cadmium in blood samples, see Table 161.1. The blood sampling time as the abscissa, the concentration of cadmium in blood samples for the longitudinal coordinates, make the element cadmium in plasma concentration–time curve as shown in Table 161.1.

Increasing the time after exposure gradually decreased blood in the cadmium content. In the rat tail vein 24 h after exposure, in each experimental group the

Table 161.1 Various time points of the blood concentrations of cadmium after caudal vein was exposed ($\bar{x} \pm s$, n = 6, μg/g)

Type of CdSe nanoparticles					
Time	CdSeI	CdSeII	CdSeIII	CdSeIV	Control
1 min	37.36 ± 5.66	21.42 ± 4.09	15.44 ± 2.19	1.94 ± 0.09	–
15 min	13.47 ± 4.95	9.45 ± 1.77	6.49 ± 1.25	1.54 ± 0.03	–
45 min	5.99 ± 1.21	5.42 ± 0.92	4.68 ± 0.75	1.37 ± 0.10	–
3 h	4.00 ± 0.93	2.90 ± 0.50	2.33 ± 0.59	1.01 ± 0.04	–
10 h	1.93 ± 0.07	1.75 ± 0.38	1.57 ± 0.07	–	–
24 h	0.88 ± 0.05	0.77 ± 0.08	–	–	–

Note "–"indicates not detected

Table 161.2 Parameters after rat tail vein administration of drug

Parameters of drug action	CdSeI	CdSeII	CdSeIII	CdSeIV
K	0.0015 ± 0.0002	0.00143 ± 0.0002	0.00172 ± 0.0001	0.00249 ± 0.0004**
V(L)	3.23 ± 0.621	4.000 ± 0.728	3.641 ± .0842	12.13 ± 3.132**
T1/2 (min)	456.77 ± 52.630	450.21 ± 47.335	295.66 ± 31.650*	236.88 ± 19.772**
AUC	4102.72 ± 199.622	3490.60 ± 202.450*	2435.27 ± 164.750*	567.76 ± 80.21**
CL (L/min)	0.00447 ± 0.0003	0.00573 ± 0.0004	0.00884 ± 0.0007*	0.0375 ± 0.0011**

* Compared with the CdSe I, $P < 0.001$ **Compared with the CdSeI,IIand III, $P < 0.001$

concentration of cadmium in blood samples had the following relationship: CdSeI > CdSeII > CdSeIII > CdSeIV > control group. In the first 24 h after exposure to blood, only CdSeI and CdSeII group can contain trace amounts of cadmium detected which in the other group were not detected.

Table 161.2 shows CdSeIV of the elimination rate constant K and the apparent volume of distribution V(L), where soluble cadmium selenide (CdSeIV) in the body eliminates at speed higher than the other three groups ($P < 0.001$), while the CdSeIII and CdSeIV of semi-reduced Period (T1/2), area under the curve (AUC) and clearance (CL) were lower than the first two groups ($P < 0.001$), CdSeI and CdSeII with basically the same parameters.

161.6 Result of Tissue Distribution

The experimental study of CdSe nanoparticles in each group 24 h after the exposure in the rat liver, kidney, testis and the distribution of the results are given in Table 161.3.

Table 161.3 after intravenous exposure to liver, kidney and testis concentrations of cadmium element ($\bar{x} \pm s$, n = 6, μg/g)

Type of CdSe nanoparticles	Organization		
	Liver	Kidney	Testis
CdSeI	349.77 ± 25.56**	54.05 ± 5.45*	28.43 ± 3.48**
CdSeII	70.63 ± 9.21*	51.80 ± 5.63*	22.47 ± 2.92*
CdSeIII	65.36 ± 7.41*	40.47 ± 5.30	18.98 ± 4.03*
CdSeIV	44.83 ± 5.66	38.00 ± 7.59	10.43 ± 2.36
F	122.771	30.654	15.262
P	<0.001	<0.01	<0.01

Note: *Compared with CdSeIV, $P < 0.01$ **Compared with CdSeII, CdSeIII and CdSeIV, $P < 0.01$

161.7 Conclusion

The metabolism of nano-CdSe was a first order two-compartment model. The smaller the size and the less soluble the nano-CdSe, the slower it was eliminated from the blood and organs.

References

1. Colvin VL (2004) Sustainability for nanotechnology: making smaller safer and changing the way industry thinks in the process. Scientist 18:26–27
2. Bruchez M Jr, Moronne M, Gin P, Weiss S, Alivisatos AP (1998) Semi-conductor nanocrystals as fluorescent biological labels. Science 281(5385):2013–2016
3. Sukhanova A, Devy J, Venteo L, Kaplan H, Artemyev M et al (2004) Biocompatible fluorescent nanocrystals for immunolabeling of membrane proteins and cells. Anal Biochem 324(1):60–67
4. Gao X, Cui Y, Levenson RM, Chung LW et al (2004) In vivo cancer targeting and imaging with semiconductor quantum dots. Nat Biotechnol 22(8):896–976
5. Chan WH, Shiao NH, Lu PZ (2006) CdSe quantum dots induce apoptosis in human neuroblastoma cells via mitochondrial dependent pathways and inhibition of survival signals. Toxicol Lett 167(3):191–200
6. Kirchner C, Liedl T, Kudera S et al (2005) Cytotoxicity of colloidal CdSe and CdSe/ZnS nanoparticles. Nano Lett 5(2):331–338
7. Lovric J, Chao SJ, Winnik FM et al (2005) Unmodified cadmium telluride quantum dots induce reactive oxygen species formation leading to multiple organelle damage and cell death. Chem Biol 12(11):1227–1234
8. Zhang T, Stiwell JL, Gerion D, Ding L, Elboudwarej O et al (2006) Cellular effect of high doses of silica-coated quantum dot profile with high throughout gene expression analysis and high content cellomics measurements. Nano Lett 6(4):800–808

Chapter 162
Study on Hydrological Simulation of Gan River Based on SWAT-X Model

Yujie Fang, Wenbin Zhou and Dinggui Luo

Abstract Hydrological simulation is the basis of water resources management and utilization. In this study, soil and water assessment tool (SWAT) model was applied to Gan River Basin for hydrological simulation on ArcView3.3 platform. The basic database of Gan River Basin was built using ArcGis9.2. Based on the LH-OAT parameter sensitivity analysis, the sensitive parameters of runoff were identified, including CN2, Gwqmn, rchrg_dp, ESCO, sol_z, GW_revap, SOL_AWC, sol_k, canmx and Alpha_BF, and then model parameters related to runoff were calibrated and validated using data observed in 20 hydrological stations during 2001–2008. The simulation showed that the simulated values were reasonably comparable to the observed data ($\mathrm{Re} < 20\%$, $R2 > 0.7$ and Nash-suttcliffe > 0.7), suggesting the validity of SWAT model in Gan River Basin.

Keywords SWAT model · Distributed hydrological model · Hydrological simulation · Gan River · Parameter calibration

Supported by State Water Pollution Control and Treatment Project (2008ZX07526-008) and National Science and Technology Support Program (2007BAC23B02) in China.

Y. Fang · W. Zhou (✉)
Key Lab of Poyang Lake Environment and Resource Utilization
Ministry of Education, Nanchang University, Nanchang, China
e-mail: wbzhou@ncu.edu.cn

Y. Fang
e-mail: 289293905@qq.com

D. Luo
School of Environmental Science and Engineering,
Guangzhou University, Guangzhou, China
e-mail: ldggq@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_162,

162.1 Introduction

Variation of water resource depends on many aspects of environment, economy and society; it has the potential to severely impact upon environmental quality, economic development and social well-being. Hydrological simulations have been performed for hundreds of years all over the world. From a modeling perspective, hydrologic models can be divided into two categories: lumped models and distributed models.

As one of the distributed models, the soil and water assessment tool (SWAT) allows a number of different physical processes to be simulated in a watershed and is a public domain and open source integration model that allows the users to infer modifications for tailor-made applications. In the incipient stage of this model, a number of investigators evaluated the applicability of this model in different regions [1]. Based on the databases (soils, land use, and topography) for the conterminous U.S. at 1:250,000 scale, Arnold et al. [1] simulated the hydrologic balance for each soil association polygon (78,863 nationwide) without calibration for 20 years using dominant soil and land use properties and validated the model by comparing simulated average annual runoff with long-term average annual runoff from USGS stream gage records, and their results showed that the large-scale hydrologic balance could be realistically simulated using a continuous water balance model. Hernandez et al. [2] described a procedure for evaluating the effects of land cover change and rainfall spatial variability on watershed responses, and found that the model was able to characterize the runoff responses of the watershed due to changes of land cover. Subsequently, some scholars have evaluated its sensitivity to different parameters (including soil data resolutions, watershed subdivisions and topography) of runoff yield and sediment yield [3]. In recent years, SWAT has been widely applied for a number of studies in catchments of quite different sizes, such as the responses of water resources to land use changes and to climate variability [4, 5], sediment yield [6], agricultural management [7] and pollution modeling [8].

Based on AVSWAT-X to simulate hydrology cycle in Gan River basin, our objectives are: (1) to investigate the sensitivity analysis of parameters, and calibrations and validation of this model; (2) to evaluate the validity of model in Gan River for water resources management and utilization.

162.2 Materials and Methods

162.2.1 Swat Model Description

The SWAT model is a distributed hydrological model developed by the United States Department of Agriculture-Agriculture Research Service (USDA-ARS), and it incorporates of several ARS models and is a direct outgrowth of the Simulation

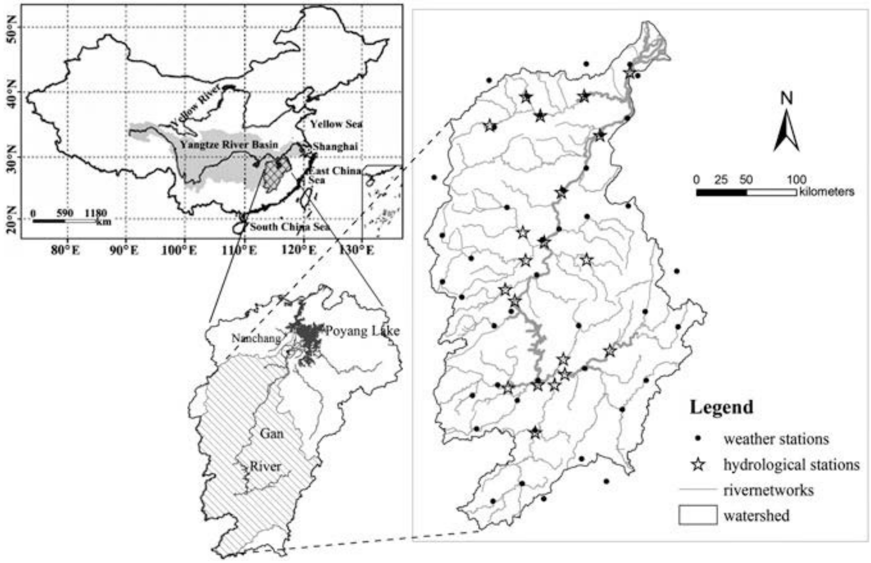


Fig. 162.1 Location of study area

for Water Resources in Rural Basins (SWRRB) [9]. The most significant improvements of the model between releases include: SWAT94.2, SWAT96.2, SWAT98.1, SWAT99.2, SWAT2000, SWAT2005, and SWAT2009.

Water balance, which is the basic drive of SWAT model, includes rainfall, runoff, seepage, evapotranspiration, base flow, interflow and so on. To predict stream generation, SWAT uses a modified version of soil conservation service (SCS) curve number (CN) method.

Hydrological Process changes with different land cover and land use. Based on a digital elevation model (DEM) and stream networks, SWAT delineates watersheds into subbasins, which are further subdivided into hydrologic response units (HRUs) with homogeneous land use, soil type and management practices. With subbasins set-up, flow from each HRU in a subbasin is summed and then routed through channels, ponds and reservoirs to the watershed outlet. SWAT calculates surface runoff at daily time steps.

162.2.2 Study Area Description

Gan River [between 24°29′–29°5′N and 113°46′–116°38′E (Fig. 162.1)] locates in the south bank of Yangtze River, and belongs to Poyang Lake Basin river system. Gan River is the seventh biggest tributary of Yangtze River and the biggest river in Jiangxi province. From south to north, Gan River flows its way across the city of Ganzhou, Jian, Fengcheng, Zhangshu, Nanchang and pours into the Poyang Lake.

With a basin area of about 81,500 sq km, it belongs to subtropics moist monsoon climatic region, with moderate climate, abundant rainfall (average annual long-term rainfall H in the basin is 1400–1800 mm) and adequate lighting. Red soil covers a large part of the basin, accounting for approximately 58%. With high forest cover, the forest land accounts for about 65%.

During recent years, with economic growth, environmental deterioration and the enhancement of the public environmental awareness, the study on issues of water pollution in Gan River basin has already become more in-depth and systemized.

162.2.3 Input Data Prepared for the SWAT Model Setup

Data required in this study include the DEM of the Gan River basin, soil properties, land use, weather and climate and observed basin discharge. These data were obtained and are detailed below.

DEM. The DEM of the basin was derived from topographical data at the resolution of 1:250,000. The data were obtained from Jiangxi Academy of Environmental Science. The resolution for the basin DEM is 90 × 90 m.

Soils. Soil data at the resolution of 1:250,000 were obtained from a soil survey completed by Bureau of Land Management of Jiangxi Province in 1990. Four major soil types and their percentage distributions in the basin are: Red soil, which covers 58.03% of the basin area, hydragric paddy soil (20.51%), grayed paddy soil (3.50%) and yellow–red soil (3.09%).

Land use. According to the survey completed by the Department of Soil Survey of Jiangxi Province in 2000, the land use in the Gan River basin can be categorized into forested land. The resolution of these land cover is 1:100,000.

Meteorological data. In the SWAT model, the required meteorological inputs for daily calculations of hydrological processes are daily precipitation, maximum/minimum temperatures, net radiation (determined from observed solar and terrestrial radiation), near- surface wind and relative humidity of the air.

Runoff observations. Runoff observations were used for comparisons against the modeled surface flow in calibration and validation. Daily streamflow data from the 20 hydrological stations were collected for these comparisons.

162.3 Results and Discussion

162.3.1 Construction of SWAT-X Model

Firstly, DEM and river networks were imported to the model for generalization [creations of digital streams, and divisions of sub-basins and hydrologic response unit (HRU)] of Gan River basin based on the platform of ArcView3.3. This study

Table 162.1 Final value of the parameter after calibration

Parameter	Range of typical value	Final value of the parameter
CN2	35–98	50–69
Gwqmn	0–5000	500–800
rchrg_dp	0–0.2	0.2–0.3
ESCO	0–1	0.5–0.65
GW_revap	0–0.2	0.02–0.15
Slope	0–1	0.146–0.195

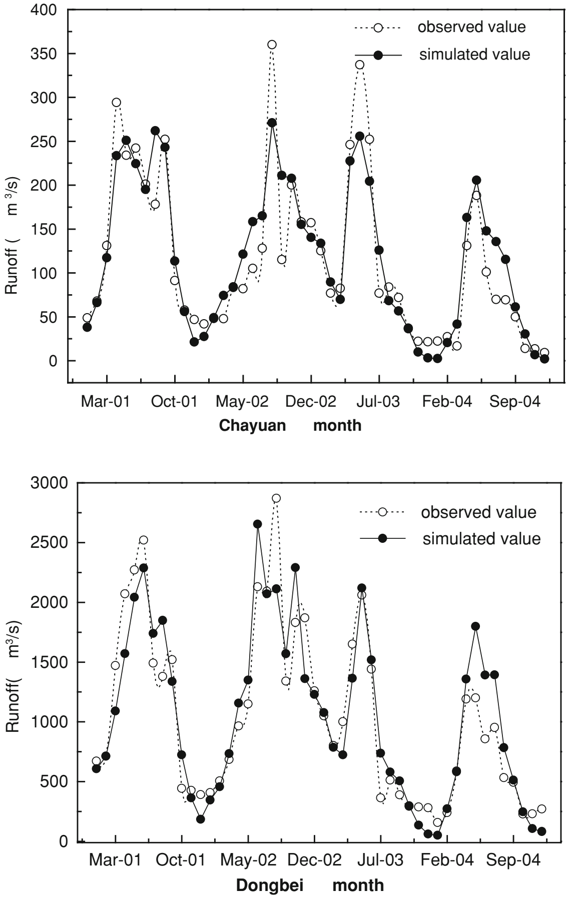
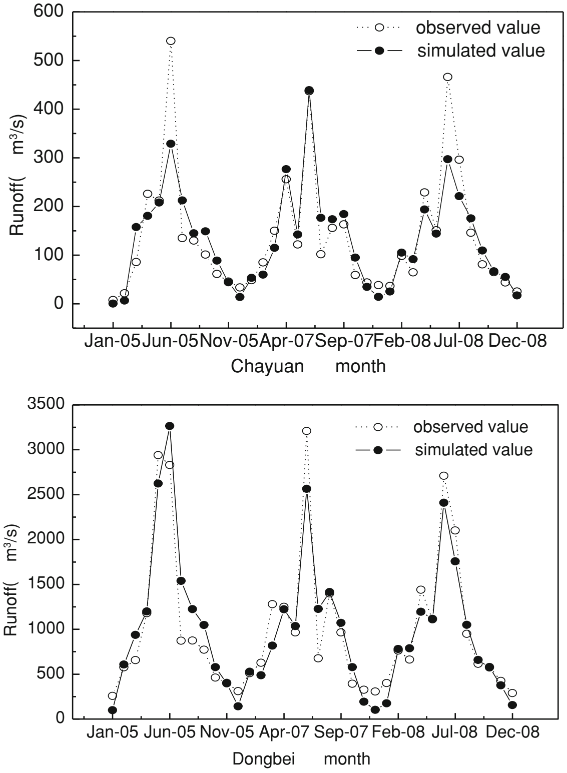


Fig. 162.2 Runoff in calibration period of months (Chayuan and Dongbei stations as example)

basin was divided into 81 subbasins (Fig. 162.1), through the thresholds of channels (40 000 ha), land use (5%) and soil types (5%). Secondly, the meteorological input files including rainfall data, maximum/minimum temperature, wind speed and relative humidity were imported. In the model, the method of SCS runoff curve, Penman–Monteith method and Variable Storage was selected for simulations of runoff and potential evaporation, and calculations of channels, respectively.

Fig. 162.3 Runoff in validation period of months (Chayuan and Dongbei stations as example)



162.3.2 Sensitivity Analysis

In this study, we adopted the method of LH-OAT sensitivity analysis proposed by Morris in 1991 and is embedded into the sensitivity analysis module of the SWAT-X version [10]. Using this method, we can effectively select the influential factors of model simulation, which affect results and usability of the model obviously.

We made a sensitivity analysis with observations at Bashang hydrological station. LH Sampling interval is set to 10, OAT changing parameters is set to 0.05 and Stochastic Population is set to 2003 with 280 runs. The sequence of sensitive parameters is CN2, Gwqmn, rchrg_dp, ESCO, sol_z, GW_revap, SOL_AWC, sol_k, canmx, Alpha_BF. Other parameters do not have obvious performance.

162.3.3 Calibration and Validation of SWAT Model

According to the sensitivity analysis, 8 parameters are selected as sensitive parameters, including CN2, Gwqmn, rchrg_dp, ESCO, sol_z, GW_revap,

Table 162.2 Summary of streamflow calibration and validation

Stations	Re(%)		R2		Ens	
	Calibration period	Validation period	Calibration period	Validation period	Calibration period	Validation period
Chayuan	5.881	0.085	0.8412	0.8192	0.839	0.803
Julongtan	2.546	15.07	0.8171	0.9265	0.853	0.903
Bashang	2.240	3.015	0.7932	0.8324	0.840	0.823
Fenkeng	2.332	6.833	0.8963	0.8536	0.852	0.826
Xiashan	10.510	4.394	0.8696	0.8757	0.911	0.916
Hanlinqiao	−7.668	5.886	0.8791	0.8272	0.909	0.905
Dongbei	0.342	1.226	0.8455	0.8872	0.941	0.949
Linkeng	11.586	19.587	0.7832	0.7726	0.971	0.987
Shangshalan	12.482	−4.37	0.843	0.8894	0.827	0.883
Saitang	6.036	8.531	0.8719	0.8428	0.903	0.912
Baisha	−1.466	19.892	0.8468	0.8278	0.940	0.957
Ji'an	3.178	0.664	0.8755	0.895	0.953	0.953
Xintian	−5.363	4.29	0.8847	0.8991	0.872	0.940
Xiajiang(2)	0.717	7.778	0.8716	0.9081	0.954	0.959
Zhangshu	6.453	9.757	0.858	0.9048	0.945	0.953
Weifang	3.854	12.342	0.8068	0.728	0.990	0.993
Yifeng	1.392	−9.166	0.7705	0.7594	0.996	0.999
Shanggao	10.476	8.109	0.8162	0.8112	0.789	0.842
Gaoan	9.775	5.222	0.8469	0.7933	0.791	0.775
Waizhou	5.022	2.550	0.8474	0.9048	0.944	0.962

SOL_AWC, sol_k, canmx, Alpha_BF and slope. The monthly observed data at 20 hydrological stations during 2001–2003 were used for the monthly runoff calibration. The final calibration parameters are detailed in Table 162.1.

In this study, the observed data in 1999–2000 were used for practice, and data measured during 2001–2003 for calibration and data from 2004 to 2008 (not including 2006) for validation. The daily observed data in 20 hydrological stations during 2001 and 2008 (not including 2006) were used during the daily runoff calibration and validation.

Figures 162.2 and 162.3 are the simulated and observed runoff of the process of comparison in calibration and validation period.

The result was evaluated using Re (average relative error), R2 (correlation coefficient) and Ens (Nash-Sutclife coefficient) [11] and showed good agreement between observation and simulation (Table 162.2).

Seen from Figs. 162.2 and 162.3 and Table 162.2, the simulated monthly Runoff was compared with observed at the 20 hydrological stations. Re was less than 20%, $R2 > 0.7$ and $Ens > 0.7$. These indicated that the simulation of streamflow was reasonable, and SWAT model was applicable to be used in Gan River Basin. Most Re with positive value indicated that the simulation result was comparatively larger than observed data. The potential reasons were that: (a) Simulation result of each year may dramatically influence upon the whole modeling, as the calibration and validation period was not long; (b) In this study, the amount of agricultural and industrial water intake was not considered because of difficulty in collecting materials.

162.4 Conclusions

As the latest version of SWAT model, SWAT-X further improved the pragmatism of model and reliability of result. The databases were established using ArcGis9.2 and were used to simulate runoff in Gan River Basin with SWAT-X. Based on the LH-OAT parameter sensitivity analysis, the sensitive parameters of SWAT were identified, including CN2, Gwqmn, rchrg_dp, ESCO, sol_z, GW_revap, SOL_AWC, sol_k, canmx and Alpha_BF. Then, the model parameters related to runoff were calibrated and validated using the daily observed flow at the 20 hydrological stations distributed throughout the whole river basin during 2001–2008. The simulation showed that the simulated values were reasonab compared to the observations ($Re < 20\%$, $R2 > 0.7$ and $Ens \geq 0.7$), suggesting the validity of SWAT model in Gan River Basin. This study supported an important foundation for further water resources management and utilization and point and non-point source pollution simulation and control.

References

1. Arnold JG, Srinivasan R (1999) Continental scale simulation of the hydrologic balance[J]. J AWRA 35(5):1037–1052
2. Hernandez M, Miller SN, Goodrioh DC, Goff BF, Kepner WG, Edmonds CM, Jones KB (2000) Modeling runoff response to land cover and rainfall spatial variability in semi-arid watersheds[J]. Environ Monit Assess 64:285–298
3. Geza M, McCray JE (2008) Effects of soil data resolution on SWAT model streamflow, water quality predictions[J]. J Environ Manag 88:393–406
4. Guo H, Qi H, Jiang T (2008) Annual and seasonal streamflow responses to climate and land-cover changes in the Poyang Lake basin, China[J]. J Hydrol 355(1–4):106–122
5. Van Liew MW, Garbrecht J (2003) Hydrologic simulation of the little washita river experimental watershed using SWAT[J]. J Am Water Resour Assoc 39(2):413–426
6. Huang Z, Xue B, Pang Y (2009) Simulation on streamflow, nutrient loadings in gucheng lake low yangtze river basin based on SWAT model. Quat Int 208:109–115
7. Tuppad P, Kannan N, Srinivasan R, Rossi CG, Arnold JG (2010) Simulation of agricultural management alternatives for watershed protection[J]. Water Resour Manag 24:3115–3144
8. Coffey R, Cummins E, O'Flaherty V, Cormican M (2010) Analysis of the soil and water assessment tool (SWAT) to model Cryptosporidium in surface water sources[J]. Biosystems Eng 106:303–314
9. Williams JR, Nicks AD, Arnold JG (1985) Simulator for water resources in rural basins[J]. J Hydraul Eng 111(6):970–986
10. Griensven V (2007) Sensitivity auto-calibration uncertainty and model evaluation in SWAT2005[Z]. Grassland, soil and water research service, Temple, TX
11. Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models, Part-1: a discussion of principles[J]. J Hydrol 10(3):282–290

Chapter 163
A Study on the Change Trends of Regional Water Resource Carrying Capacity

Hongquan Liu and Hongru Liu

Abstract Based on the systematic optimization and the theory of sustainable development, the paper established the model for calculating water resources carrying capacity and took the "population carrying capacity" as its integration index. The water resources carrying capacity model was tested by taking Zhangjiakou as a practical example and analyzed the situation of WRCC in Zhangjiakou.

Keywords Regional water resources carrying capacity · Population carrying capacity · Standard of living

163.1 Introduction

China is a country that is famous for the shortage of water resources and the per capita water resource is 2200 m^3, one third of the world average. For Zhangjiakou that lies to northern China in Hebei province, the per-capita water resources are 400 m^3, only one fifth of the national average and one fifteenth of the world average. Zhangjiakou is the most water-deficient area in the world [1]. Water resources carrying capacity (WRCC) is the study on the supporting ability of water resources system to society. The specializing in WRCC is seldom abroad, and for

H. Liu (✉)
College of Urban and Rural Construction, Agricultural University of Hebei,
Baoding 071001, People's Republic of China
e-mail: lhqlhqlhq_2001@163.com

H. Liu
Academic Affairs Office, Hebei University of Technology,
Tianjin 300130, People's Republic of China
e-mail: liuhongruhero@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_163,

the shortage of water resources, the studies of WRCC are centralized in China [2–4]. By now WRCC is an immature theory, the method and conception are not well-researched [4–8]. Based on the optimization and the theory of sustainable development, the author studied the WRCC of Zhangjiakou.

163.2 The Model of Water Resources Carrying Capacity

The WRCC is not the static state, but enhanced. Current research results of WRCC are mostly of static research. This research took 2000 as the status quo and studied the WRCC of Zhangjiakou in different target years (2000, 2010, 2020 and 2030). At last the article analyzed the change trend of WRCC in Zhangjiakou.

163.3 Determining Person's Demand Under Certain Living Level

A person's demands are many-sided. In this paper, a vector for a person's demand is given.

$$\overrightarrow{R} = (r_1, r_2, \ldots, r_m) \tag{163.1}$$

The formula: ri ($i = 1, 2\ldots m$) represents average per person demand for i aspect; the vector $\overrightarrow{R}$ represents the average per person demand. A person's demand in any aspect can be changed to the demand for water resources. During certain period, the Ui represents the water efficiency coefficient of people's demand for i aspect and it is the unit water's efficiency for i aspect. For agricultural, Ui is the value of agriculture value for unit water resources, yuan/m^3; for industrial, it is the value of industry for unit water resources, yuan/m^3; Specially $Ui = 1$, for living water in the paper. Human's demand is various and water resources must meet all aspect of a person's demand at the same time.

$$\overrightarrow{R}\ \text{water} = (r_1/U_1, r_2/U_2, \ldots, r_m/U_m) = (r_{\mathrm{w}1}, r_{w2}\ldots r_{wm}). \tag{163.2}$$

In the formula: the vector rwi ($i = 1, 2\ldots m$) represents the amount of water resources that is needed for meeting the average per person demand for the i aspect.

163.4 The Amount of Water Resources Supplied to the Society During a Given Period

In this paper Wj ($j = 1, 2\ldots n$) represents the amount of water resources for kind of j that can be supplied to the society. Bji represents the distribution coefficient of the kind of j water resources to meet people's demand for i aspect.

$$Wji = Wj \bullet Bji, \tag{163.3}$$

where *Wji* represents the amount of water resources to meet people's demand of *i* aspect can be get from the kind of *j* water resources.

$$Wi = \sum_{j=1}^{n} Wj \bullet Bji, \tag{163.4}$$

where *wi* represents the amount of water resources needed to meet people's demand of *i* aspect. Human's demand is various and water resources must meet all aspect of person's demand at the same time.

$$\vec{W} = \{w1, w2, \ldots wm\}, \tag{163.5}$$

where $\vec{W}$ represents the amount of water resources for every water-use sector.

163.5 Determining the Maximum Population (POP') That Water Resources Can Carry Under Certain Water Allocation Plan

When water resources distribution coefficient *Bji* is fixed, the maximum population POP′ that water resources can carry under certain living level can be obtained.

$$\mathrm{POP}'_i = \left(\sum_{j=1}^{n} Wj \bullet B_{ji}\right)/r_{wi} = (W_1, W_2, \ldots W_n) \bullet \begin{matrix} B_{1i} \\ B_{2i} \\ \vdots \\ B_{ni} \end{matrix} /r_{wi} = wi/r_{wi}. \tag{163.6}$$

In the formula: POP′i represents the population that can be carried by water resources to meet a person's demand for i aspect. Human's demand is various and water resources must meet all aspect of a person's demand at the same time.

$$\mathrm{POP}' = \min\{\mathrm{POP}'_i\}, \tag{163.7}$$

where POP′ is the maximum population that the water resources can be carried under certain water supply scheme. POP′ is the water resources carrying capacity under certain *Bij*. For different *Bij*, the value of WRCC will be changed.

163.6 Determining the Maximum Population (WRCC) That Water Resources Can Carry Under Certain Living Level

Different water resources allocation plan, the WRCC is different. What we want to do is to determine the maximum population that the regional water resources can carry.

$$\begin{aligned}\mathrm{WRCC} = \max\{\mathrm{POP}'\} &= \max\left\{\min\left\{\left(\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{Wj} \bullet \mathrm{Bij}\right)/r_{wi}, j = 1, 2, \ldots m\right\}\right\} \\ &= \max\{\min\{wi/r_{wi}, j = 1, 2, \ldots m\}\}\end{aligned} \tag{163.8}$$

163.7 Water Resources Carrying Capacity in Zhangjiakou

In this paper, we took Zhangjiakou of China as an example. We studied the WRCC of Zhangjiakou in different years and under different living level. In view of the development of the social economy, for 2000 and 2010, we studied the WRCC under three living levels (clothing and food, moderate prosperity and affluence); for 2020, we will study study the WRCC under two living level(moderate prosperity and affluence); for 2030, we will study the WRCC only under affluence living level.

163.8 Standards of Living

Based on the reality of Zhangjiakou, there are three standards of living as Table 163.1.

Comprehensive quota of water for life:

$$R = Rt \times st + Rc \times sc \tag{163.9}$$

In the formula: R represents the comprehensive quota of water for life, Rt represents the quota of water for urban life, st represents the rate of urban people to population, Rc represents the quota of water for rural life, sc represents the rate of rural people to population.

163.9 Available Water Resources Amount

In this paper, the author did not consider water consumes of ecosystem and only studied WRCC in normal flow year (50%) see Table 163.2.

163.10 Computed Result

The predicting population of 2,000 is actual numerical value. But for other years, is predicting population. Based on the above situation, bringing the parameters of Zhangjiakou into the WRCC model, we can get the results of WRCC in Zhangjiakou as follow (Fig. 163.1). Unit: 10,000.

Table 163.1 Status of water utility level in Zhangjiakou

Items	Food and clothing	Moderate prosperity	Affluence
Comprehensive quota of water for life (m3/person·day)	0.08	0.12	0.15
Per capita GDP (yuan/person)	5,000	20,000	40,000

Table 163.2 Available water resources in Zhangjiakou [10^8 m^3]

Surface water	Shallow ground water	Deep ground water	Other water	Total
5.23	1.38	6.42	0.08	13.10

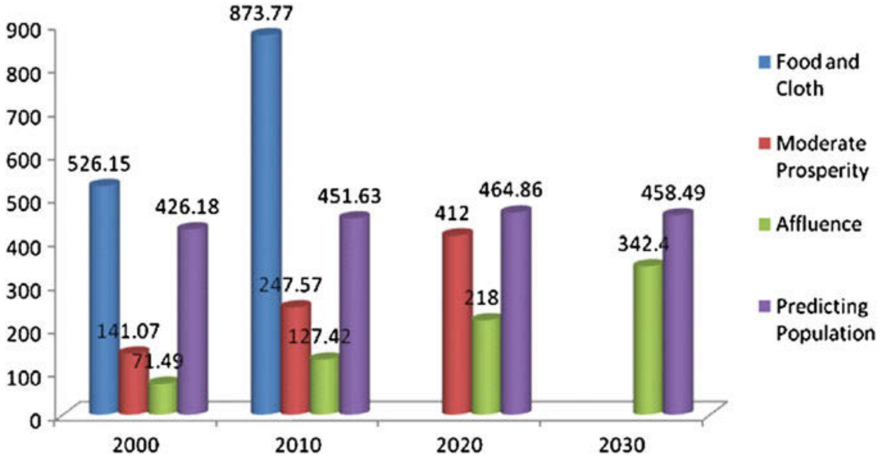


Fig. 163.1 WRCC of Zhangjiakou

163.11 Conclusions

The water resources system of Zhangjiakou only can carry the predicting population under food and cloth living level in 2000 and 2010. Obviously, the WRCC of Zhangjiakou is very fragile. Transferring water from other valley, preventing water pollution and saving water resource are very necessary.

Acknowledgement This article is funded by the project of Hebei Education Department (Z2010150): "Study on Regional Water Resource Carrying Capacity Based on Climate Change" and the International Cooperation Project between China and Germany (2007DFA21160):"Sustainable Water and Agricultural Land Use in the Guanting Watershed under Limited Water Resources". At last, I wish to thank the team of Hebei water resources planning for providing the base data of the study.

References

1. Bai Y (2010) Think about the water resources of Zhangjiakou. J Hebei North Univ (Soc Sci Ed) 12:53–57
2. Harris JM et al (1999) Carrying capacity in agriculture: globe and regional issue. Ecol Ecnom 129:443–461
3. Rijisberman et al (2000) Different approaches to assessment of design and management of sustainable urban water system. Env Impact Assess Rev 129:333–345
4. Falkenmark M, Lundqvist J (1998) Toward water security: political determination and human adaptation crucial. Nat Resour Forum 21:37–51
5. Xiang F, Ji G (1999) Comprehensive assessment of regional water resources bearing capacity. Resour Env Yangtze Basin 2:31–37
6. Xu Y (1993) Comprehensive assessment of the water resource carrying capacity in arid region. J Nat Resour 8:229–237
7. Li L, Guo H (2000) The study on water resources carrying capacity in chaidamu basin. Environ Sci 3:20–30
8. Yaolong F, Wenxiu H (2003) The study on water resources carrying capacity. Adv Water Sci 14:109–113

Chapter 164
Investigation on Life Quality of Cataract Patients in Different Gender and Age, Treated by the Light Engineering Half Year Later in Tangshan

Qi Ren, Wei Cui, Ruiqi Zhang and Yun Li

Abstract Cataract is the first reason that leads to blindness in the world. And this will make the patients' life quality significantly declined. Cataract patients generally need the surgery to treat this disease, which is a heavy burden to the poor patients. In order to deal with this problem, Tang Shang government carries out a light engineering to help the poor patients for free. This research investigated the patients treated by this project half year later aimed to understand their life quality. And then provide the reference opinions to improve the operation effect and life quality of the patients treated for cataract . The investigation random selected 212 patients treated from workers hospital half year later, and divided them into different groups by gender and age. The results showed that male patients' life quality is better than that of female patients'; physical function is related with gender and age.

Keywords Cataract · Quality of life · Influencing factors · SQOL-DVI scale

164.1 Introduction

Cataract is a kind of disease where crystalline lens becomes opaque and hampering the light into intraocular, consequently affecting the vision. Crystalline lens opacification could affect vision gradually serious with the illness development. Especially in advanced stages, it can obviously affect vision even leading to

Q. Ren (✉) · W. Cui · R. Zhang · Y. Li
Hebei Province Key Laboratory of Occupational
Health and safety for Coal Industry, Division of Social Medicine,
School of Public Health, Hebei United University,
Tang Shan 063000, China
e-mail: Jessie_az@hotmail.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_164,

Table 164.1 Distribution of subjects divided into different groups

	Groups	*n*	Percentages (%)
Gender	Male	96	45.3
	Female	116	54.7
Age (years)	≤54	26	12.3
	55～	42	19.8
	65～	74	34.9
	75–97	70	33.0

blindness. Study found, cataract disease seriously influences not only patients' vision, but also their psychology, economy, daily work and life [1]. At the same time, it can make a negative change in patient's survival quality. At present, the incidence of cataract has been increasing globally. Statistics show that China's blind patients occupied about 18% of total number in the world. And cataract patients are about one-fifth of the total number in the world [2]. Along with the ageing society coming, there will be 50 million new cases annually in the next 50 years [3]. In China, cataracts occupied the major position in spectrum of disease in the elder people no matter in city or countryside [4].

The Light Engineering is organized by Tangshan disabled persons' federation and co-organized by the Workers Hospital. The poor cataract patients registered the disabled persons' federation can enjoy the free medical examination and appropriate surgical treatments according to their conditions. This welfare project provides timely treatments to poor cataract patients, which could certainly improve their life quality. Therefore, the investigation of life quality of cataract patients treated by Light Engineering could provide the basis and evidences to improve the operation effect and life quality.

164.2 Subjects and Methods

164.2.1 Research Subjects

The 212 subjects are randomly selected from cataract patients, registered in the disabled persons' federation and treated by Light Engineering half year later.

164.2.2 Research Methods

The Scale for Quality of Life-Damaged Vision Illness (SQOL-DVI) was used to investigate the subjects. The SQOL-DVI includes four contents that reflected life quality, which are symptoms and visual function, physical function, social activities and mental health.

Table 164.2 Comparison of the subjects' results in different gender

Life Quality	Male ($n = 96$)	Female ($n = 116$)	t	P
Symptoms and visual function	58.26 ± 15.60	55.50 ± 15.16	1.302	0.194
Physical function	25.09 ± 6.58	22.47 ± 7.27	2.726	0.007
Social activities	31.93 ± 8.68	31.92 ± 8.75	0.004	0.997
Mental health	30.00 ± 8.99	29.27 ± 9.74	0.564	0.570

Table 164.3 Comparison of the symptoms and visual function results in different age groups

Age	n	$\bar{x} \pm s$	F	P
≤54	26	55.85 ± 13.35	2.478	0.062
55～	42	61.14 ± 12.15		
65～	74	55.99 ± 16.32		
75–97	70	53.77 ± 16.28		
Total	212	56.75 ± 15.39		

Table 164.4 Comparison of the physical function results in different age groups

Age	n	$\bar{x} \pm s$	F	P
≤54	26	27.27 ± 4.41	9.145	< 0.001
55～	42	25.50 ± 7.36		
65～	74	24.36 ± 6.71		
75–97	70	20.47 ± 6.94		
Total	212	23.60 ± 7.07		

164.3 Results

The distribution of investigated cataract patients was shown in Table 164.1. It indicates the number and percentages of different groups, which were divided by gender and ages.

Comparison of the subjects' results in different gender can be seen from Table 164.2. The scores of male are higher than that of female in four aspects. The difference of results between physical function and the other three aspects are statistically significant.

Comparison of the symptoms and visual function results in different age groups can be seen from Table 164.3. The results indicate that the patients in 55～ age group have the best symptoms and visual function. But the 75–97 age group show the worst situation.

Comparison of the physical function results in different age groups was indicated by Table 164.4. The scores of physical function are declined with the increase in age. The patients in ≤54 age group get the highest score in physical function, which is 27.27 ± 4.41. And who are in 75–97 age group have the lowest

Table 164.5 Comparison of the social activities results in different age groups

Age	n	$\bar{x} \pm s$	F	P
≤54	26	32.15 ± 6.99	0.550	0.648
55 ~	42	32.62 ± 8.34		
65 ~	74	32.47 ± 8.84		
75–97	70	30.84 ± 9.38		
Total	212	30.84 ± 9.38		

Table 164.6 Comparison of the mental health results in different age groups

Age	n	$\bar{x} \pm s$	F	P
≤54	26	29.54 ± 8.91	0.327	0.806
55 ~	42	30.74 ± 8.41		
65 ~	74	29.62 ± 9.99		
75–97	70	28.91 ± 9.60		
Total	212	29.60 ± 9.40		

score, 20.47 ± 6.94. The difference of results between 75 and 97 age group and the other three groups are statistically significant.

Comparison of the social activities results in different age groups was indicated in Table 164.5. The highest score in 55 ~ age group is (32.62 ± 8.34) and the lowest score in 75–97 age group is (30.84 ± 9.38).

Comparison of the mental health results in different age groups was indicated by Table 164.6. The highest score is in 55 ~ age group (30.74 ± 8.41) and the lowest score is in 75–97 age group (28.91 ± 9.60).

164.4 Conclusions

The comparison results of life quality in different gender shows that male is better than female in four relative aspects, which are symptoms and visual function, physical function, social activities and mental health. Especially, in physical function, the score of male is much higher than that of female. In the daily life, male's physical function is usually better than female's. Even if they are in disease condition, they are still stronger than female because the difference of physiological structure. However, there are no significant differences in other three aspects between the two gender groups. This might be caused by the similar age and state of cataract condition of patients. So their results of symptoms and visual function are similar. And also, the patients are all from poor families, and most of them are peasants. This means they have similar life style and daily activities. Thus, their results of social activities and mental health level are not significantly different.

The results comparison of life quality in different age groups indicate that 55 ~ age group have the highest scores in symptoms and visual function, social activities and mental health. But in physical function aspect, ≤54 age group shows the highest score in four age groups. And in this aspect, there is a significant character that the scores declined according with the age increasing. The age group of ≤54 is the youngest among all the investigated cataract patients. They certainly have the best physical function. According with the increasing age, old patients would suffer not only cataract but also many other diseases, which could obviously influence their life quality. Therefore, older patients express worse physical function. For symptoms and visual function, social activities and mental health, the results of 55 ~ age group are better than the other three groups. This is because patients in this age group do not have the expectation as high as that of ≤54 age group. At the same time, they do not need to suffer other cataract complications as older patients. So they express better results. But for ≤54 age group patients, they still need to pay more attention to their daily life, such as reading and house works. So they still expect high operation effects. This is the reason for them to express worse satisfaction with the treatments and life quality [5]. However, no matter in which aspects of life quality, 75–97 age group shows the lowest scores among the four age groups because their comprehensive health conditions are much worse than the other age groups. Especially, they also suffer from other physical diseases and decrease of life enthusiasm, which could negatively influence patients' life quality.

References

1. Jayamanne DG, Allen ED, Wood CM et al (1999) Correlation between early, measurable improvement in quality of life and speed of visual rehabilitation after phacoemulsification. J Cataract Refract Surg 25:1135–1139
2. Zhang S (2005) Cataract and medical treatment. J China Pharm 16(18):1439–1440
3. Wang J (2004) A study on cataract epidemiological research and pathogenesis. J Chin Rural Physician 20(11):9–10
4. Xia W, Zhao B, Zhao H et al (2008) Analysis of life quality of cataract patients after surgery. Shanghai Nurs 8(1):42–43
5. Tian J, Dang Y (2009) Study on the pertinence of living quality in patients of different ages after cataract surgery. China J Chin Ophthalmol 19(3):180–181

Chapter 165
Detection of Serum Vascular Endothelial Growth Factor (VEGF) and Endothelin, in Type 2 Diabetic Nephropathy Patients and Its Clinical Significance

Yongqiang Zheng, Jianfen Wei, Xiaojun Li, Ling Xue and Guoyu Qiao

Abstract To investigate the relationship between the content of serum factor (VEGF) and endothelin with complication of type 2 diabetic mellitus. One-hundred patients of diabetic nephropathy, fifty diabetic mellitus without renal dysfunction and fifty healthy volunteers were enrolled. The serum levels of vascular endothelial growth factor (VEGF) were measured with enzyme end method and endothelin were measured with radioimmunoassay method. The results show that vascular endothelial growth factor (VEGF) and endothelin were significantly higher in diabetic nephropathy patients than the other patients. Vascular endothelial growth factor (VEGF) and endothelin can be a reliable indicator for early impairment of renal function, and will be helpful for diagnosis of type 2 diabetic nephropathies.

Keywords Type 2 diabetic mellitus · Type 2 diabetic nephropathy · VEGF · ET

165.1 Introduction

The vascular endothelial growth factor (VEGF) is a highly specific mitogen vascular endothelial factor, and plays an important role in angiogenesis [1]. In recent years studies have shown that VEGF is closely linked with the occurrence

Y. Zheng (✉) · J. Wei · X. Li
Hospital of Hebei United University, Tangshan 063000, China
e-mail: zhengy_q@163.com

L. Xue
Public Health of Hebei United University, Tangshan 063000, China

G. Qiao
Clinical laboratory of Tangshan Union Hospital, Tangshan 063000, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_165,

and development of diabetic microvascular. Endothelin (ET) is secretion of vascular endothelial cell active substance, and widely distributed in vascular smooth muscle, heart, brain, nerves and other tissues. It is one of the strong vasoconstrictors, and involved in regulation of nerve and blood vessel function directly or indirectly as a neurotransmitter. Ischemia, hypoxia, ischemia and reperfusion can significantly promote substantial release of ET. Diabetic Nephropathy (DN) is one of an important microvascular complications of type 2 diabetes, a major cause of diabetes' disability and death. We learn the relationship between ET and DN by testing patients' levels of VEGF and ET. They are reported like than.

165.2 Materials and Methods

165.2.1 Study

Taking 150 patients with type 2 treated in North China Coal Medical College Hospital from November 2008 to September 2009, all patients in 1997 were found in the American Diabetes Association (ADA) diagnostic criteria, while excluding diabetes associated with acute chronic infectious diseases, cardiovascular disease, renal insufficiency, and other serious systemic disease and DM acute metabolic complications. They were divided into 3 groups according to urine albumin excretion rate (UAER): normal albuminuria group (UAER < 20 μg/min), microalbuminuria group (UAER $\geq$ 20 μg/min but $\leq$ 200 μg/min) and clinical proteinuria (UAER $\geq$ 200 μg/min). Number of patients in each group was 50. There are 76 males and 74 female cases, and they are 50 to 85 years old. The average age is 54.06 ± 8.18 years. The Control group includes 50 patients from the hospital health examination center, including 36 males and 24 females, aged 48–81 years, mean 53.09 ± 7.28 years.

165.2.2 Research Methods

For all patientsblood is taken in the morning without having breakfast. After after centrifugating at 3000r/min for 15 and 30 min, take the upper serum, and keep it at −20°C in the refrigerator. All patients were measured by testing blood sugar (glucose oxidase), blood lipids series (measuring total cholesterol by cholesterol oxidase method, measuring triglycerides by triglyceride phosphate oxidase and measuring low density lipoprotein by polyethylene sulfate precipitation method), VEGF (horseradish peroxidase enzyme immunoassay method) and ET (radioimmunoassay) [2].

165.2.3 Statistical Methods

Building a database of all data using excel, showing it by $\bar{x} \pm s$, analysis using SPSS11.5 statistical software, counting material comparison with $\chi 2$ test,

Table 165.1 VEGF levels in each group and the ET ($\bar{x} \pm s$) of the comparison (pg/ml)

Group	Sample size	VEGF	ET
Control group	50	67.84 ± 28.69	70.26 ± 28.56
Type 2 diabetes	150	117.36 ± 26.19★	128.96 ± 23.26★

Note: ★ compared with control group P < 0.01

Table 165.2 VEGF and ET levels in each group ($\bar{x} \pm s$) of the comparison (pg/ml)

Group	Sample size	VEGF	ET
Control group	50	67.84 ± 28.69	70.26 ± 28.56
Normal albuminuria group	50	86.34 ± 22.31★	107.46 ± 17.54★
Microalbuminuria group	50	116.59 ± 23.36★▲	128.36 ± 28.31★▲
Clinical albuminuria group	50	155.84 ± 25.96★▲◆	145.57 ± 16.35★▲◆

Note: ★ compared with control group P < 0.01, ▲ and normal albuminuria group P < 0.01, ◆ and microalbuminuria group *P* < 0.01

Table 165.3 The correlation analysis among VEGF, ET levels and fasting blood glucose, glycated hemoglobin, high sensitivity-C reactive protein and homocysteine

Index	Fasting blood glucose		Glycated hemoglobin		Hs–C reactive protein		Homocysteine	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*
VEGF	0.56	<0.01	0.78	<0.01	0.86	<0.01	–	–
ET	0.66	<0.01	0.67	<0.01	–	–	0.76	<0.01

measurement of data using t test, variance analysis or analysis of covariance and factor analysis of correlation analysis using linear correlation.

165.3 Results

The levels of VEGF and ET comparison between Type 2 diabetes and normal control serum (Table 165.1) shows that patients with type 2 diabetes have significantly higher levels of VEGF and ET ($P < 0.01$).

The serum levels of VEGF and ET. A comparative study showed that for paients with diabetes mellitus complicated with proteinuria group (clinical proteinuria and microalbuminuria group) serum levels of VEGF and ET was significantly higher than diabetic patients without proteinuria group and control group ($P < 0.01$), Table 165.2:

The correlation analysis among VEGF, ET levels and fasting blood glucose, glycated hemoglobin, high sensitivity-C reactive protein and homocysteine (Table 165.3) shows that VEGF levels and fasting blood glucose, glycated hemoglobin, high sensitivity-C reactive protein were positively correlated; ET Levels and fasting plasma glucose, glycosylated hemoglobin, homocysteine was positively correlated, and $P < 0.01$.

Table 165.4 Serum VEGF and ET and diabetic nephropathy analysis of covariance (adjusted mean ± standard deviation correction) (pg/ml)

Covariate	Groups	VEGF	ET
Age	Normal albuminuria	88.39 ± 15.05	108.98 ± 4.56
	Microalbuminuria	118.16 ± 17.10*	130.86 ± 5.13*
	Clinical albuminuria	158.07 ± 17.22*▲	150.69 ± 6.12*▲
Disease process	Normal albuminuria	88.53 ± 15.18	118.65 ± 4.89
	Microalbuminuria	116.58 ± 16.06*	136.25 ± 4.39*
	Clinical albuminuria	159.99 ± 16.28*▲	157.34 ± 5.16*▲
Body mass index	Normal albuminuria	82.03 ± 15.48	105.23 ± 5.02
	Microalbuminuria	115.29 ± 18.02*	125.36 ± 7.01*
	Clinical albuminuria	154.29 ± 19.17*▲	146.26 ± 6.12*▲

Note: age, duration, adjusted mean body mass index were 57.36 (years), 7.55 (years), 26.38 compared with normal albuminuria group difference was significant ▲ and microalbuminuria group was significantly different

The relationship between VEGF, ET and diabetic nephropathy and age, disease duration and body mass index is related to blood vessel diseases. We analyzed by age, disease duration and body mass index of covariance variables, and observed the relationship between VEGF, ET and DN after controlling covariance variable interference. Age adjusted mean = 57.36 (years), duration = 7.55 (years), adjusted mean body mass index = 26.38, proteinuria group showed that diabetes mellitus (clinical proteinuria and microalbuminuria group), and serum VEGF ET levels were significantly higher than that of diabetic patients without proteinuria group ($P < 0.01$), while clinical albuminuria group was higher than the microalbuminuria group ($P < 0.01$), and results are shown in Table 165.4.

165.4 Discussion

The Specificity acts on VEGF in vascular endothelial cells with an increase venule permeability, and promote angiogenesis and the maintenance of vascular function and so on. VEGF is the reaction of vascular endothelial permeability and proliferation of indicator [3]. In this study, by analysis of covariance after controlling age, body mass index, duration of confounding factors, VEGF synthesis and secretion of patients with diabetic nephropathy continues to increase. Diabetes produces the enzyme saccharification making the structure and function of RBC abnormal, microvascular basement membrane thickening, hemosclerosis, erythroid shrinks, the volume and surface narrowing, all affect oxygen exchange, leading to widespread anoxia, stimulating increased secretion of VEGF [4]. DN, the glomerular epithelial cell foot processes to increase VEGF expression and secretion of VEGF receptors may be increased through the glomerular basement membrane, and endothelial cells of VEGF receptor binding changes in the structure and function of endothelial cells to increase glomerular capillary permeability, promoting mesengial cell synthesis of extracellular matrix and the mechanism of renal hypertrophy in the occurrence and development of DN [5]. Glomerular endothelial cells, VEGF is an important regulator of glomerular endothelial cells that produce VEGF, which can mediate the secondary capillary endothelial cell proliferation and repair. Endothelial cells stimulated by VEGF and collagenase increased glomerular overexpression of VEGF can cause proteinuria, which may be through the decomposition of protein and destruction of the glomerular basement membrane (GBM) cause. Proteinuria and glomerular caused by a variety of ways, and progressive renal interstitial fibrosis increased the progress of DN [6]. This study found that serum VEGF levels of diabetic patients was significantly higher than the normal control, microalbuminuria group, clinical albuminuria. Albuminuria was significantly higher than the normal group.

The ET is endothelial cells secretion of active substances, involved in regulation of nerve and blood vessel function directly or indirectly as a neurotransmitter. Ischemia, hypoxia, ischemia and reperfusion can significantly promote the massive release of ET. Increased ET has relationship with the following factors: (1) hypoxic endothelial cells to stimulate synthesis and release of ET; (2) damaged endothelial cells directly disclose ET; (3) decreased blood perfusion led to decreased endothelial cell shear stress suffered by increased ET [7]. DN ET-1 levels in patients have a positive correlation between glomerular sclerosis. Liu [8], who detected in the renal cortex of DN ET content found that in 12-week DM rats, though still in high filtration state, renal cortex homogenate ET-1 levels were significantly higher than the normal group ($P < 0.01$), and confirmed endothelial cell injury and vary with the disease. This study used in the analysis age, disease duration and BMI. The covariate analysis of covariance, excluding the role of these factors, displayed that the interference level of the control group was significantly higher than normal. Microalbuminuria group and clinical albuminuria were significantly higher than normal albuminuria group. As reported in the literature consistently, we can see that, VEGF and ET development and progression of diabetic nephropathy plays an important role, through a series of complex mechanisms involved in the progress of DN and to some extent, reflects the severity of the disease.

References

1. Hui L (2002) Vascular endothelial growth factor and diabetic nephropathy. Foreign Med Sci Endocrine Sect 22(3):186–187
2. Cha DR, Kim NH, Yoon JW et al (2000) Role of vascular endothelial growth factor in diabetic nephropathy. Kidney Int 58(77):104–112
3. Yinna W, Rongfen L, Huilan L (2004) Vascular endothelial growth factor and diabetes and kidney complications. Foreign med Sci urinary Sect 24(6):822–824
4. Jianhua Y, Chenghong M, Bin Z et al (2006) Vascular endothelial growth factor in early diabetes capillary pathological changes of change and significance. Acad J Guangdong Colg Pharm 22(2):197–198

5. Peng Y, Deming D (2007) Vascular endothelial growth factor and diabetes capillary pathological changes. J Yangtze Univ (Natural Science Edition) Med V 4:321–322
6. Lei X, Wei L (2008) The changes of serum vascular endothelial growth factor and C-reactive protein levels in type 2 diabetic patients. J Harbin Med Univ 42:70–71
7. Shengying Q, Xinghua C, Ge Y (2004) Study on the mechanism of diabetic peripheral neuropathy (DPN) possible role of microvascular damage and dysfunction. J Radioimmunol 17(3):171–174
8. Bing L, Zhangsuo L (2003) Measurement of nitric oxide and endothelin-1 in renal cortex of diabetic nephropathy rats. J Henan Med Univ 38(3):390–392

Chapter 166
The New NURBS Interpolation Algorithm Based on Constant Feed Rate and Deceleration in STEP-NC Machining

Lizhi Gu, Hao Wu and Qi Hong

Abstract Interpolation technology is very important in STEP-NC machining, and affects machining accuracy and efficiency directly. Based on the analysis of major interpolation methods, is put forward an important interpolation algorithm for Non-Uniform Rational B-Spline (NURBS) curve that combines the constant feed rate interpolation for regular area of the curve and S-shaped curve deceleration interpolation for special miniature region or the cusp place. Two featured curves were initially discriminated according to the chord error. By the given interpolation, a universal NURBS curve was designed and visualized. The relevant algorithm for the curve was simulated in VC++6.0. Results have shown that the algorithm is feasible and effective.

Keywords STEP-NC · NURBS · S-shaped interpolation · Constant feed rate interpolation

L. Gu (✉) · H. Wu · Q. Hong
College of Mechanical Engineering Automation,
Huaqiao University, Quanzhou 362021, China
e-mail: gulizhi888@163.com

H. Wu
e-mail: wuhao19860405@163.com

Q. Hong
e-mail: lusia_361005.student@sina.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_166,

166.1 Introduction

Nowadays, the processing of curve (surface) is used more and more in products for spaceflight, aviation, mould, and so on. But traditional computer numerical control (CNC) system machines are programmed in G & M codes (formalized by ISO6983), which only has the insert function with line and curve, causing some other shortcomings in application [1, 2].

The CNC used polygonal line instead of the curve, and the polygonal line is a discontinuous first e. So the surface is not smooth.

In the processing of curve (surface), if done with machine tool, it easily leads to overshoot and cannot guarantee the quality and precision, otherwise it becomes low efficiency and poor quality.

Sometimes, processing parts with complex surface shape needs to store large numbers of segments, the CNC system memory capacity is very small compared to the capacity of this requirement, and therefore needs to store in segment. It will not only reduce the reliability of the system, but also reduce processing efficiency.

In order to overcome the shortcomings of ISO6983, International Organization for Standardization ISO developed CAM and CNC on the basis of STEP. The new data standards for CNC, called STEP-NC (ISO14649) replace the traditional data standard ISO6983, promoting CNC Systems intelligence, integration and network development.

STEP-NC standard not only contains a detailed geometrical description but also the detailed technology description. Furthermore, it supports spline data. In STEP-NC the standard of geometrical description free curve and surface are described with uniform NURBS.

Therefore, the chapter will take the processing of the NURBS surface characteristics for the object to deeply study the technology of generating the free curve tool path automatically and the corresponding interpolation methods in STEP-CNC system [3].

166.2 New Algorithm Background

Spline has a wide range use in curve surface, the main curve (surface) models are: B-Spline, Bezier and Coons. One of the most widely used is NURBS, and a NURBS curve is generally expressed as follows [4, 5]:

$$p(u) = \frac{\sum_{i=0}^{n} \omega_i d_i N_i^k(u)}{\sum_{j=0}^{n} \omega_j d_j N_j^k(u)} = \sum_{i=0}^{n} d_i R_j^k(u) \tag{166.1}$$

$$R_i^k(u) = \frac{\omega_i N_i^k(u)}{\sum_{j=0}^{n} \omega_j N_j^k(u)}_{u \in [0,\, 1]} \tag{166.2}$$

where u is a scalar parameter which varies from 0 to 1 and it is a list of weigths-data. Given a knot list $U = [\mathrm{u}_0, \mathrm{u}_1, \ldots, \mathrm{u}_\mathrm{n}, \ldots, \mathrm{u}_{\mathrm{n+k}}]$, $k \geq 1$ and $n \geq 0$, the associated normalized B-splines, $N_j^k(u)$ with the k-th power in Eq. 166.1, are defined by

$$N_i^k(u) = \frac{u - u_i}{u_{i+k-1} - u_i} N_i^{k-1}(u) + \frac{u_{i+k} - u}{u_{i+k} - u_{i+1}} N_{i+1}^{k-1}(u)\,(i = 0, 1, \ldots, n, k > 0) \tag{166.3}$$

The Definition of NURBS In EXPRESS. The description of NURBS curves reference standard Part 42 of STEP, which is specifically expressed for geometry and topology. It is described as follows:

ENTITY rational_b_spline_curve

SUBTYPE OF (b_ spline_curve);

weightses data: LIST [2:?] of REAL;

DERIVE

weights: ARRAY[O:upper_index_on_control_points] of REAL

:= list_to_ array (weights_data,0,upper_index_on_control_points);

WHERE

WRI:SIZEOF (weights_data) = SIZEOF (SELF\b_spline_curve.control_points list);

WR2:curve_weights_positive (SELF);

END_ENTITY;

Based on the concept and principle of NURBS above, and analysis of the definition of NURBS in STEP-NC, a new interpolation algorithm is gonging to be put forward dependent on STEP-NC standard.

166.3 Constant Feed Rate Interpolation of NURBS Curve

The STEP-NC procedure of NURBS surface not only contains a detailed geometrical description but also the detailed technology description, its geometrical description is harmonized with STEP data format. Therefore it mainly relies on the geometry information about the surface model to design the cutting tool way [6].

The Interpolation in Constant Feed Rate. In order to avoid the perturbation of velocity in the processing, the article proposes the interpolation technology which feeds back in constant rate. Namely in each interpolation cycle, the interpolation curve's arc length is equal, but parameter u is inhomogeneous division.

$$p(u) = (x(u), y(u), z(u)) \tag{166.4}$$

where $p(u)$ is spline curve, time function u is curve parameter.

$$u(t_i) = u_i\, u(t_{i+1}) = u_{i+1} \tag{166.5}$$

We can expand the first-order Taylor in the parameter of the time *t*, and obtain the corresponding approximate algorithm:

$$u_{i+1} = u_i + \frac{du}{dt_i}(t_{i+1} - t_i) + H.O.T \tag{166.6}$$

where the curve parameter u$_i$ is the No.$_i$ interpolation points, the time parameters t$_i$ is the corresponding No.$_i$ interpolation points, u$_i$ + 1 is the No.$_i$ + 1 interpolation points, the time parameters t$_i$ + 1 is the corresponding No.$_i$ + 1 interpolation points. *H.O.T* is defined as the high trace. And interpolation speed $v(u_i)$ can be defined as follows:

$$v(u_i) = |\frac{dp(u)}{dt}|_{u=u_i} = |\frac{dp(u)}{du}|_{u=u_i}\frac{du}{dt}|_{t=t_i} \tag{166.7}$$

Curve interpolation cycle time *T* is the difference between the adjacent time. So we can get the function from the above:

$$\Delta u_{i+1} = u_{i+1} - u_i = \frac{v(u_i)T}{|\frac{dp(u)}{u}|_{u=u_i}} = \frac{F.T}{\sqrt{x^2 + y^2 + z^2}} \tag{166.8}$$

where Δu_i + 1 is the incremental of interpolation parameter. *F* is feed rate which remains unchanged in the interpolation process. Then we would determine the interpolation step Δs, and *L* is the arc length of spline curve, *M* is the step number which required for the whole spline interpolation.

$$M = \text{round}(\frac{L}{F.Ts})\Delta s = \frac{L}{M} \tag{166.9}$$

Then we can determine the next interpolation point and the incremental value of corresponding coordinates Δx, Δy, Δz:

$$\Delta s = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} \tag{166.10}$$

So obtained from the above kinds of formula:

$$F(u) = a_6u^6 + a_5u^5 + a_4u^4 + a_3u^3 + a_2u^2 + a_1u + a_0 \tag{166.11}$$

where the parameter value *u* can be obtained by Newton iteration method, where the coefficients a_6, a_5, a_4, a_3, a_2, a_1, a_0 are real numbers and must have the only real solution. Assumption $u_{i,k}$ is the parameter values of corresponding points (*xi*, *k*, *iy*, *k*, *zi*, and *k*) and $u_{i,k-1}$ is the corresponding points of the parameter values $(x_{i,k-1}, y_{i,k-1}, z_{i,k-1})$. In order to improve the solution efficiency and find a similar solution quickly, set the initial iteration $u_0 = u_{i,k} + (u_{i,k} - u_{i,k-1})$, the iterative error is set to: $\varepsilon = |u_{i+1} - u_{_i}| \leq 10^{-3}$, the above formula can be interpolated calculated. But to achieve in the interpolation function numerical control system, also need to address the accuracy, speed, efficiency, and many other key technologies.

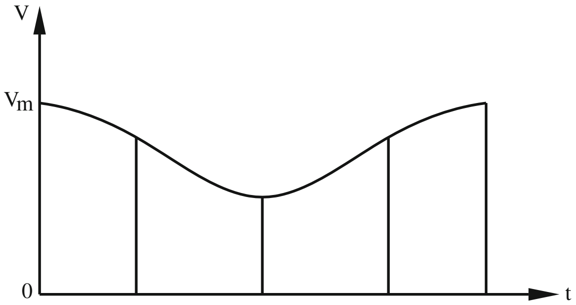


Fig. 166.1 The S-shape ADC/DEC

S-shaped Deceleration Processing. The interpolation technology with the constant feed rate is used to the NC processing, when the radius of curvature becomes quite small region or meets the cusp place, it is easy to occur the impact or the vibration, if the processing speed does not change. Therefore the article proposes the perspective method of S-shape ADC/DEC which distributes in symmetry and is shown in Fig. 166.1. This article mainly takes the S deceleration as the example to study, establishes the S curve deceleration model, plans the S deceleration path, then according to the characteristic of the NURBS, seeks for the speed change sensitivity point, carries on the deceleration processing ahead of time, makes the entire moderating process to have high flexibility and continuity.

Figure 166.2 expressed the moderating process of S-shape, the maximum acceleration is $\alpha_{\max}$, the maximum speed is v_m, and the biggest jerk is $J_{\max}$. The entire processing is controlled in the definition scope of $J_{\max}$. In order to avoid the impact which comes from the machine tool of the oversized acceleration, the changes of S-shape speed assume the smooth transition. In ordinary, the deceleration of S-shape divides into three stages. However, the presence of uniform deceleration segment may be to the point of initial speed (v_s), maximum speed (v_m), maximum acceleration $(\alpha_{\max})$ and the biggest jerk $(J_{\max})$. (1) If there is no uniform deceleration phase: $v_m - v_s \leq a^2_{\max}/J_{\max}$, and there are three segments in the phase: $t_1 = a_{\max}/J_{\max}$, $t_2 = 0$, $t_3 = t_1$. (2) If it has uniform deceleration phase: $v_m - v_s \leq a^2_{\max}/J_{\max}$, and there are three segments in the phase: $t_1 = a_{\max}/J_{\max}$, $t_2 = (v_m - v_s)$, $a_{\max} - t_1$, $t_3 = t_1$. And the corresponding interpolation cycle functions are: $n_1 = \text{ent}\,(t_1/T)$, $n_2 = \text{ent}\,(t_2/T)$, $n_3 = \text{ent}(t_3/T)$. Therefore, we can create a mathematical model of S-shaped curve from the function above.

In order to determine the location of acceleration and deceleration, first we can determine the interpolation points in each curve from the interpolation algorithm, so there is no chord length error when the real-time interpolation NURBS accumulated, but there is chord error (contour error). Something must be introduced to control the chord high error. When the chord error accumulated to the critical value, the location at this time is the deceleration start position, and is shown in Fig. 166.3. At the time of real-time interpolation, the system of interpolation cycle is very small, and each feed step is very short, so it can be approximated as standard circular arc, and the chord error can be calculated as follows:

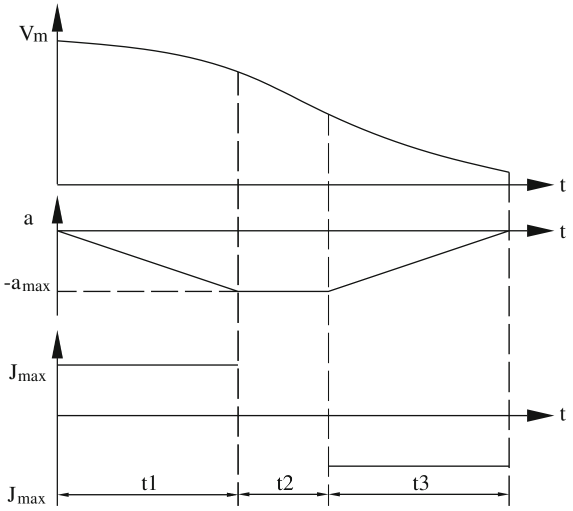


Fig. 166.2 The moderating processing of S-shape

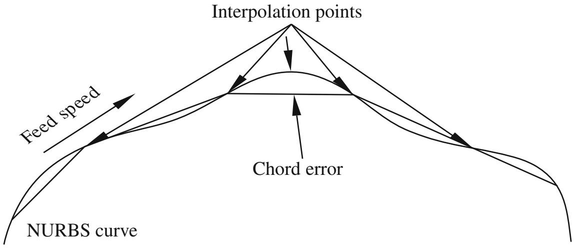


Fig. 166.3 The deceleration start position

$$h = \frac{(FT)^2}{8r} \tag{166.12}$$

where F is the feed rate, T the interpolation cycle, r the arc radius of curvature. We should monitor the size of interpolation error in the real-time interpolation. When the actual error is within the allowable range, the interpolation step can still be calculated according to the given feed rate. Otherwise, it is necessary to decrease processing speed on the basis of the actual S-shaped curve, so that the actual feed rate can be changed to meet the actual precision.

166.4 Simulation Results

In order to verify the validity of the new NURBS interpolation algorithm, an instance programmed by Visual C++ and Open GL was demonstrated through simulation. The attribute data of NURBS surface instance and the corresponding

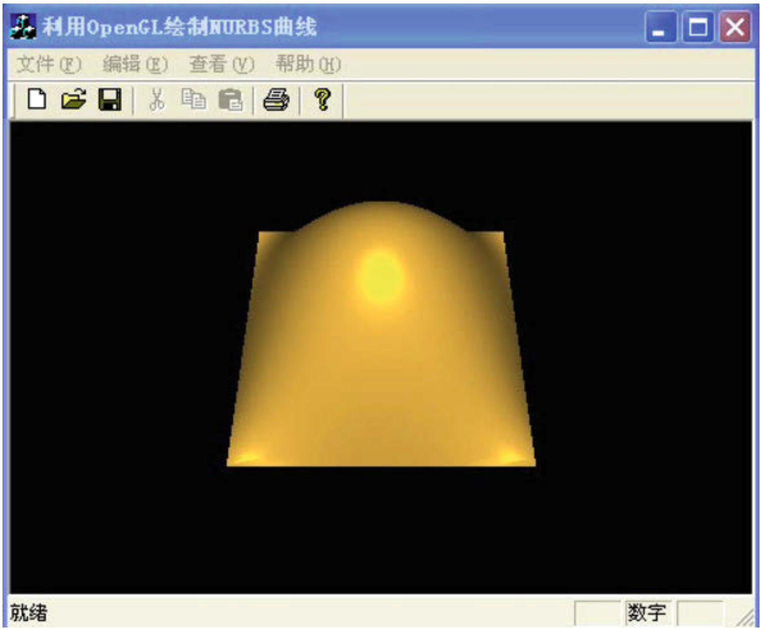


Fig. 166.4 NURBS curve

manufacturing data were first extracted from the new NC program by STEP-NC. A NURBS curve in Fig. 166.4 is generated by the new interpolation algorithm provided by the author.

166.5 Summary

The paper has presented a new NURBS interpolation technology based on constant arc length increment and developed a data pre-processing module. By an instance test, the interpolation algorithm has proved to be valid and reliable, and higher machining precision and quality of surface can be obtained.

Acknowledgment This study is supported by Fujian Natural Science Fund (2010J01295), by Xiamen Science and Technology Project Fund (3502Z20093031), Quanzhou Science and Technology Project Fund (2008G5), and also by Heilongjiang Natural Science Fund (E201043).

References

1. Xu R, X Le, Congxin LI et al (2008) Int J Adv Mfg Technol 36:343–354
2. Chuan S, Tong Z, Peiqing YE et al (2007) Chin Mech Eng 18(12):1421–1425
3. Liming S, Jun LI, Zhenping F et al (2005) J Aerospace Power 22(9):1499–1504
4. Liu X, Yamazaki K (2005) Int J Mach Tools Mfg 45:433-444
5. Yeh SS, Hsu PL (2002) Comput-Aided Des 34(4):229−237
6. Information on http://www.step-nc.org

Chapter 167
The Model of the Shanghai World Expo Influence City Infrastructure to Build Based on Multiplier Effect

Liang Yuxin

Abstract This paper focuses on the economy and establishment of investment multiplier model for the Shanghai world expo. The AHP gray theory was established in forecasting model GM (1, 1). Consider power construction, transportation, post and telecommunications construction, the construction of public facilities for the expo in three aspects as positive roles. Through comparison of 2009 Shanghai's investment of actual urban infrastructure construction, and Shanghai urban investment of infrastructure from 2003 to 2009 under the world exposition conditions successfully, and then based on the GM (1, 1) time response equation get the differences in urban advanced development.

Keywords Investment multiplier effect · Hierarchical analysis · GM (1, 1) forecast model · After · Long-term benefits expo

167.1 Introduction

Multiplier Effect: it is a kind of Macroeconomic effect. It is a variable in economic activity caused by the decrease of the total economic output change in chain reaction degree. In economics, the multiplier effect is a more complete spending/ income multiplier effect,as a concept of macroeconomics, is a kind of macroeconomics control measure, and also the total economic changes in spending its disproportionate changes in demand.

L. Yuxin (✉)
Hebei United University, Tangshan 063000, Hebei Province, China
e-mail: 43698059@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_167,

In the multiplier effect, investment multiplier is contrast causality refers to investment the spontaneous demand with certain incentives and marginal propensity consume in general, as consumption conditions for the results of the national income.

Multiplier effect model

$$\boxed{\Delta Y = K \times \Delta I}$$

Among them

$$\boxed{K = 1/(1 - \mathrm{MPC})}$$

When investment amount changes, the inevitable causes changes in the national economy which are not only directional and multiply sensex, namely the investment effect multiplier. General spontaneous demand incentives refer to all possible changes' aggregation, the resulting changes all put together as a result of national income. Shanghai world expo is mainly composed of government investment, so, government investment in Shanghai by economic impact, can use investment multiplier model for analysis. Investment multiplier K is the effect of Shanghai GDP expo contribution.

167.2 Model Assumption

Assuming only the initial investment demand is present and it causes the consumer needs to be fulfilled;

Investment multiplier play does not consider "bottleneck" department restriction;
Various factors on tourism contribution consistently every year;
Electric power construction, transportation, post and telecommunications and public facilities for Shanghai infrastructure built has certain representativeness.

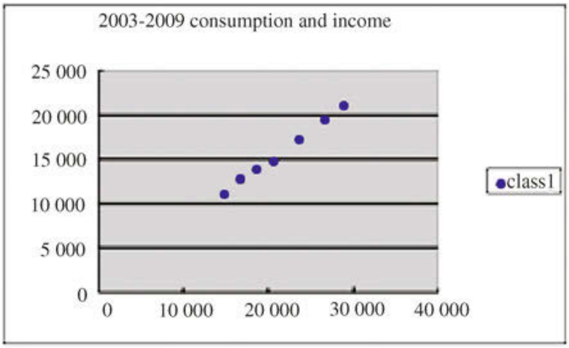
167.3 Symbols Explain

ΔY, National income incremental; K, Investment multiplier; $\Delta\alpha$, Fixed capital increment; $\Delta\beta$, Yield increment; ΔI, Investment incremental, MPC, Marginal propensity to consume; ΔC, Consumption changes; ΔP, Changes in income; C, Per capita consumption; P, Per capita income; b, Marginal propensity to consume MPC; a_{ij}, The ratio of the influence of tourism to A_i and A_j.

Table 167.1 2003–2009 per capita consumption and income in Shanghai

Year	Per capita income	Per capita consumption
2003	14,867	11,040
2004	16,683	12,631
2005	18,645	13,773
2006	20,668	14,762
2007	23,623	17,255
2008	26,675	19,398
2009	28,838	20,992

Fig. 167.1 2003–2009 per capita consumption and income fitting map



167.4 Modeling and Solving

167.4.1 Multiplier Effect of Shanghai World Expo

167.4.1.1 Determination of MPC

The world expo in Shanghai for investment spending and the influence of GDP, depends on the multiplier K size, and for this value during the expo according to the MPC will decide

MPC $= \Delta C/\Delta Y =$ Each year the changes of consumption/changes in gross domestic product

To calculate the benefit of the GDP, you need to first calculate the value of the MPC. Again according to the relationship between consumer functions and, calculating from 2003–2009 after the Shanghai world expo awarding the obtained GDP growth within a few years.

Refer to the relevant data [1] to get 2009–2003 per capita consumption and income data of Shanghai, see Table 167.1.

Per capita income as independent variables and the per capita consumption for function, using the least squares fitting, and carry on the regression analysis, see Fig. 167.1.

Table 167.2 Before world expo organizing, From 1995 to 2002 national actual infrastructure investment in Shanghai (unit: billion yuan)

Project Year	Electric power construction	Transporting postal	Utilities
In 1995	57.33	79.36	137.09
In 1996	77.61	147.21	153.96
In 1997	80.24	146.10	186.51
In 1998	89.58	181.46	260.34
In 1999	83.05	166.16	252.18
In 2000	64.61	117.52	267.77
In 2001	72.22	168.42	270.14

Note transportation post includes transportation and post and telecommunications

get $C = CO + bP = 0.7028P + 634.02C$

By type (1) (2) get

$$\Delta Y = K \times \Delta I = 1/(1 - \mathrm{MPC})\Delta I$$

Here in order to explain the following [2]: ΔY is 2003–2010 increment of national income;

ΔI is 2003–2010 the world expo 2010 government investment;

MPC From 2003, after success in its bid by linear regression model MPC = 0.7028 is taken to determine its value.

167.4.1.2 The World Expo in Shanghai Investment Contribution

According to the statistics, Shanghai world expo total investment of 30–40 billion yuan, multiplier investment model is used in type (3) to get

$$\begin{aligned}\Delta Y &= 1/(1 - \mathrm{MPC})\Delta I = 1/(1 - 0.7028) \times (3000{-}4000)\\ &= 10094.21{-}13458.95\mathrm{RMB}\end{aligned}$$

This is because of the government investment in the increase of 10,094.21 produced about 13,458.95 billion yuan

This is the multiplier effect contribution.

167.4.2 Shanghai World Expo on City Infrastructure Investment Impacts

For Shanghai world expo before organizing the late, extract urban infrastructure construction, transportation [3–5], post and telecommunications power utilities three aspects need to be analysed.

Table 167.3 Model deviation analysis

Project Year	Electric power construction	Transporting postal	Utilities
In 1995	57.33	79.36	137.09
In 1996	77.61	147.21	153.96
In 1997	80.24	146.10	186.51
In 1998	89.58	181.46	260.34
In 1999	83.05	166.16	252.18
In 2000	64.61	117.52	267.77
In 2001	72.22	168.42	270.14

For searching data,countries of urban infrastructure investment get in Shanghai in the recent years, see Table 167.2.

First, from power construction GM$(1,1)$ [6] model is used to carry out the forecast analysis, according to the data $x^{(0)} = \left\{x^{(0)}_{(1)}, x^{(0)}_{(2)}, \ldots\ldots\ldots, x^{(0)}_{(n)}\right\} = \{57.33, 77.61, \ldots\ldots, 72.22\}$

The first time of original data accumulation

$$x^{(1)}_{(K)} = \sum_{i=1}^{K} x^{(0)}_{(n)} = (57.33, 134.94, 215.18, 304.76, 387.81, 452.42, 524.64)$$

Establish equation X meet first-order univariate differential equation [7], there is $\frac{\mathrm{d}x^{(1)}}{\mathrm{d}t} + ax^{(1)} = u$

Among them a constant coefficient, u of system often set for input, differential equation

$$\hat{x}^{(1)}_{(K+1)} = \left(x^{(0)}_{(1)} - \frac{u}{a}\right)e^{-ak} + \frac{u}{a} \tag{167.1}$$

Type of $k = 1, 2, \ldots, n-1$, when $k \geq n$ we can calculate $\hat{x}^{(1)}_{(K+1)}$ fitted values.

Do the least-square estimation computation, get $\hat{U}$ value

$$\hat{U} = \begin{bmatrix} \hat{a} \\ \hat{u} \end{bmatrix} = (B^T B)^{-1} B^T Y$$

MATLAB solving $\hat{a}$ and $\hat{u}$, will receive generation value in type (4) finally to get the power construction time response equation

$$x^{(1)}_{(k+1)} = 3059.9 - 3006.7e^{-0.0280k} \tag{167.2}$$

Similarly for transport the time response equation in the post

$$x^{(1)}_{(k+1)} = 171179e^{0.000912k} - 171100 \tag{167.3}$$

Table 167.4 Prediction during the world expo in Shanghai in 2003–2009 national infrastructure investment (unit: billion yuan)

Project Year	Electric power construction	Transporting postal	Utilities
In 2003	68.63	155.10	350.10
In 2004	66.72	155.30	384.70
In 2005	64.87	155.40	422.60
In 2006	63.04	155.50	464.30
In 2007	61.32	155.70	510.20
In 2008	59.61	155.80	560.50
In 2009	57.90	156.00	615.70

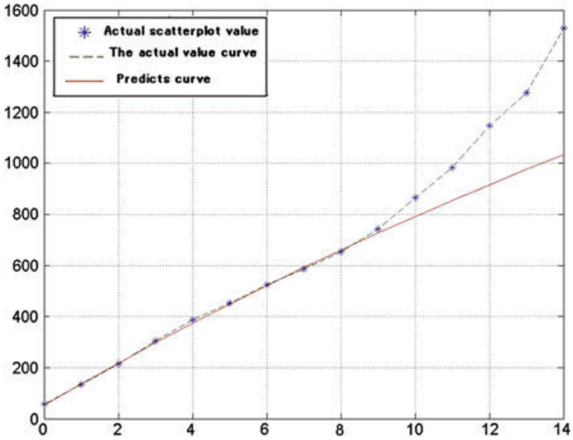


Fig. 167.2 1995–2009 power construction forecasting results and actual value of contrast

Similarly for public facilities for the time corresponding equation

$$x^{(1)}_{(k+1)} = 1836.39e^{0.0941k} - 1699.3 \tag{167.4}$$

167.4.2.1 Using the Data in 2002 Test the Model (Unit: Billion)

$$e = \frac{|\hat{x}^{(0)}(k) - x^{(0)}(k)|}{x^{(0)}(k)} = \frac{\text{Both poor value}}{\text{Practical value}}$$

Model deviation can be seen above, for GM (1, 1) model,as the deviation degree is small,the model is feasible. Using this model, after the success of bid for 7 years, power construction, transportation, post and telecommunications and public facilities,and the three aspects of investment to carry on the forecast, get the data refered in the following Tables 167.3 and 167.4.

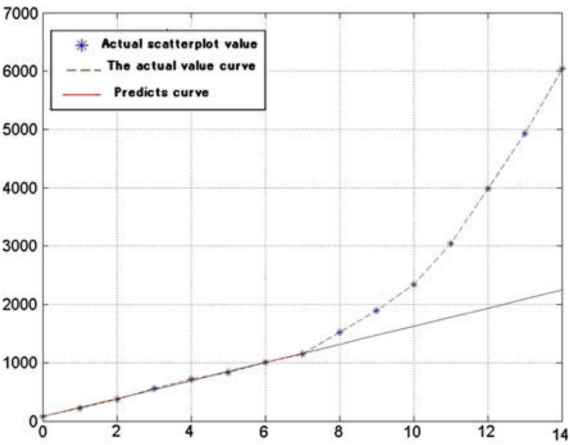


Fig. 167.3 1995–2009 predictive value and practical values of transport post and the contrast

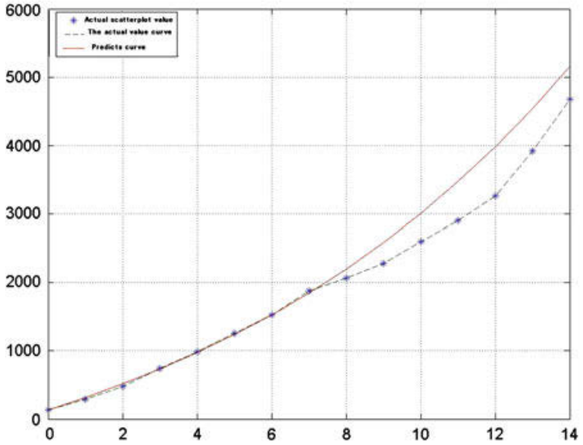


Fig. 167.4 1995–2009 predictive value and practical values of public facilities of contrast

According to relevant data, using MATLAB get power construction, transportation, post and telecommunications, public facilities in 1995–2009 with the actual values of the prediction from the contrast, see Figs. 167.2, 167.3 and 167.4.

Our model is based on the world expo effect,for the analysis of the general data, which were analyzed and predicted its assumptions of Shanghai's economic development role, that makes the problem a simple mathematical description of the intuitive. In fact, the impact on the economy of the world expo obtained by many factors can work. It is considered with the economic role in boosting employment, the per capita consumption level changes on the influence of the consumer market economy, etc. These are under the influence of the world expo in research on the economic development of question that needs careful consideration. As time was limited, we only discussed a relatively simple case; this is the direction which needs to be improved.

References

1. Shanghai bureau of statistics (2010). http://www.stats-sh.gov.cn/2004shtj/tjnj/tjnj2009.html, 2010-9-10
2. Shanghai tourism nets, information disclosure reports (2010). http://lyw.sh.gov.cn/, 2010-9-10
3. National bureau of statistics, statistical yearb (2010). http://219.235.129.58/welcome.do, 2010-9-10
4. Baidu statistical data of Shanghai, the number of domestic tourists to come over and percapita consumption expenditure, Shanghai statistics 1999–2003 (2010). http://tjsj.baidu.com/pages/jxyd/27/14/52a101f77e9a594a8cfe14939600d0a5_0.html, 2010-9-10
5. Jiang Qi Yuan (2003) Mathematical model. Higher education press, Beijing
6. Wang Geng (2008) Modern method of mathematic modeling. Science press, Beijing
7. Samson Y (2009) Long-term effects, Shanghai world expo in Shanghai, The new company vol 7, pp 33–35

Part XV
Web Service

Chapter 168
A Model for Distributed Web Service Discovery

Li Chen, Zi-lin Song and Zhuang Miao

Abstract In traditional architecture of Web services, the description of Web services is stored in centralized registry. The disadvantages in that case are that the existence of registries influences the overall performance very negatively and that any failure on the function of the registry would lead to a system failure. In this paper, we proposed a model for distributed Web service discovery called JWSD (JXTA-based Web Service Discovery) in order to overcome these problems. The JWSD model can mainly be represented through two viewpoints: system viewpoint and technology viewpoint. The detail of these two viewpoints is discussed in this paper. Our model could provide a scalable Web service discovery in large scale.

Keywords SOC · Web service discovery · JXTA · System viewpoint · Technology viewpoint

168.1 Introduction

Service-oriented computing (SOC) [1] is emerging as a new, promising computing paradigm that centers on the notion of service as the fundamental element for developing software applications. According to Papazoglou and Georgakopoulos [1], services are self-describing components that should support a rapid and low-cost composition of distributed applications. Services are offered by service

L. Chen (✉) · Z. Song · Z. Miao
Institute of Command Automation, PLA University of Science and Technology,
Nanjing, 210007 China
e-mail: ivan_chenli@tom.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_168,

providers, which procure service implementations and maintenance, as well as supply service descriptions. Service descriptions are used to advertise service capabilities, behavior, and quality, and should provide the basis for the discovery and binding of services.

The Web service model includes three roles, namely requesters, providers, and registries, where providers advertise their services to registries, and requesters query registries to discover services. The current Web services infrastructure relies on Web services description language (WSDL) [2], simple object access protocol (SOAP) [3], and universal description and discovery interface (UDDI) [4]. WSDL is an XML-based language for describing what a service does and how to invoke it. SOAP is a standard protocol for exchanging messages over HTTP between applications. UDDI allows for the definition of global registries where information about services is published.

Currently, UDDI is the universally accepted standard for Web service discovery. In traditional architecture of Web services, the description of Web services is stored in UDDI which is a centralized registry. The disadvantages in that case are that the existence of registries influences the overall performance very negatively and that any failure on the function of the registry would lead to a system failure.

In order to handle the above drawbacks a new solution for a directory service has to incorporate various technologies. One promising solution proposed here is using Peer-to-Peer (P2P) technology to introduce advantages of P2P networks to enhance the scalability and robustness. Additionally, P2P technology is used to create decentralized registries minimizing problems like performance bottlenecks and eliminating a single point of failure in the service-oriented architecture.

Toward synergizing P2P networks and Web services, there have many proposed methods in different research communities [5–15]. Castro et al. [9] proposed building a universal ring in a DHT P2P network to support service advertisement, service discovery, and code binding. These functions are based on three operations, namely the persistent store, the application-level multicast, and the distributed search. Every exposed service (a piece of code) has a code certificate to identify the correctness of the service (code binding). Since there may be many Web services with the same functionality and name globally, the code certificate will help the user find the correct service to invoke. It is essential for finding the correct service that the information used to generate the service key for discovery should be the same as the one used to generate the service key for advertisement. This is the problem of lack of semantic information and only supporting exact searching in structured P2P networks. If the name of service request is not equal to the name of service advertisement, such that the result of hashing service request cannot equal to the result of hashing service advertisement, the registry will return failure because none of service advertisements can match the service request.

Benatallah et al. [10] proposed a declarative dynamic composition and execution framework for Web services in P2P networks. But their work did not consider the Web service publishing and discovery in P2P networks, which is the primary aspect to overcome the scalability issues.

Banerjee et al. [11] proposed Distributed UDDI Deployment Engine (DUDE) to address the scalability issues with UDDI. DUDE proposed the leveraging of structured Distributed hash table (DHT), a P2P system that forms a structured overly, allowing more efficient routing than the underlying network, as a rendezvous mechanism between different registries. In their approach service description message dispersion to several distributed UDDI registries promote scalability and replication. But such an approach cannot cope with dynamism, mobility and scenarios where inter-communication between entities is not possible in a point-to-point fashion.

Similarly Wang et al. [12] proposed the Web services architecture based on a P2P network. They proposed a classification of peers on the basis of computation power and memory into Service-peers and Super-peers, with Super-peers responsible for publishing of Web services, query routing and formation of peer-groups for Service-peers. Such architecture assumes homogenous communication capabilities of peers, which cannot be considered in heterogeneous environments as robot swarms. Moreover we do not consider a full coverage, of an entity, of rest of the network and mobility of entities can restrict the communication coverage of an entity.

Du et al. [13] proposed the notion of an active and distributed service registry (ad-UDDI), an active monitoring mechanism enabled UDDI to maintain periodic service information. They considered a layered approach for distributed UDDI registry into a management root layer, a business layer comprising of domain specific ad-UDDI registries and a service layer.

Liu et al. [14] proposed a kind of Web service discovery model based on unstructured P2P network. When users have a service request, the model firstly searches in the neighbor nodes and then broadcast the service request to the whole network with find flooding. Every node compares the service request to all the registered services and returns the result to the request node. The disadvantage of this method is that every service request will be broadcasted to almost every node in the model. Obviously every user's request will possess the whole bandwidth and computing resource of the network, the efficiency is suspicious.

Banaei-Kashani et al. [15] also suggests a P2P service discovery method using the Gnutella protocol and semantic technologies. Although these methods are intuitive and simple, they can cause high network consumption and lower the possibility of finding a service.

Corresponding to the shortage of methods proposed above, in this paper we introduce the model named JWSD (JXTA-based Web Service Discovery). Any service provider who wishes to make a service available on a JXTA network needs to create an advertisement of the service and then publish the advertisement to the registry node of its community in JXTA. Publishing an advertisement allows service requesters in the same community to find it. When a service requester wants to search for a desirable service, it first sends the service request to the registry node in its community. If the registry node does not store any relevant service, it will send the service request to other registry nodes. When the registry node finds an advertisement, it usually puts it in its local cache. Other service requesters can then retrieve it from there as well as retrieving the advertisement from the actual service provider. This mechanism provides additional redundancy and scalability of JXTA networks.

The rest of the paper is organized as follows. In Sect. 168.2 we introduce the related technology of our work. In Sect. 168.3, we discuss our JWSD model in detail. The last section presents the conclusions and future work.

168.2 Related Technology

The goals of the peer-to-peer infrastructure JXTA [16] are interoperability, platform independence and ubiquity. JXTA defines a common set of protocols for building Peer-to-Peer applications. These applications avoid the problem of existing peer-to-peer systems of creating incompatible protocols. Therefore JXTA offers the means to develop a hybrid overlay network and to orchestrate the deployed applications and services.

Figure 168.1 shows an example of a possible network topology. In order to distinguish "normal" peers from rendezvous peers, these are typically called edge peers, because of the position at the edge of the overlay network. The rendezvous peers (nodes in dark gray) are located at the center of the network acting as superpeers, while the normal edge peers (nodes in light gray) are connected to their corresponding rendezvous peers. Peers in the JXTA environment organize themselves in peergroups, which represent a set of peers sharing a common interest and have agreed upon a common set of policies, e.g. a membership policy. Based on the Resolver Service Protocol, which provides resolution operations such as resolving a peer name into an IP address, defined in the JXTA specifications, the JXTA overlay network provides a default resolver service based on rendezvous peers. Rendezvous peers are peers that index advertisements to facilitate the discovery of resources in a peergroup. Rendezvous peers are defined in the scope of peergroups to reduce the communication complexity. Any peer can potentially become a rendezvous peer, except prohibited due to security considerations. Rendezvous peers maintain an index of published advertisements using the Shared Resource Distributed Index (SRDI) service. Peers use SRDI to push advertisement indices to their rendezvous peers. The rendezvous/edge peer hierarchy allows resolver queries to be propagated between rendezvous only, which reduces the amount of peers involved in a search operation significantly.

168.3 JWSD Model

In this section, we mainly consider the construction of JWSD model. The JWSD model can be represented through two viewpoints: system viewpoint and technology viewpoint. System viewpoint defines the function of nodes and the

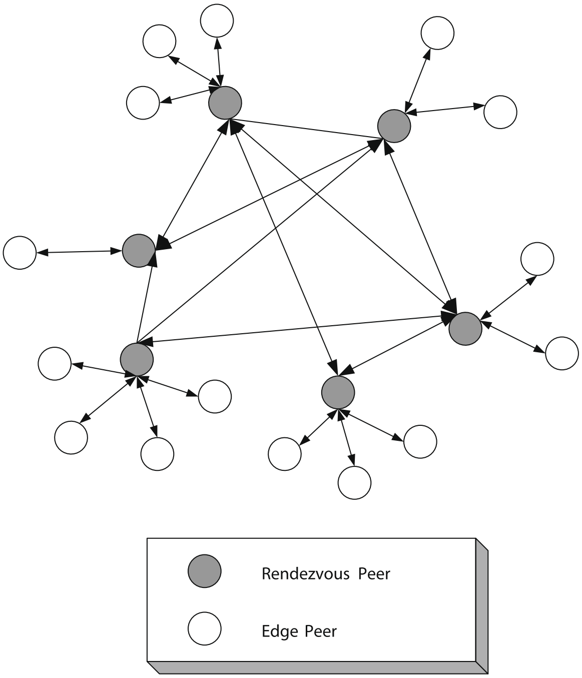


Fig. 168.1 The topology of JXTA

relationship between nodes. Since our model is based on JXTA, the system viewpoint is similar to the topology of JXTA. There are two kinds of actors in our model. The edge of model is service providers and service requesters. The service providers and service requesters connect to a local registry node and are organized to a registry community. The registry communities connect to each other based on JXTA (Fig. 168.2).

Technology viewpoint defines the configuration document in the model and technical realization of node's function. Technology viewpoint is shown in Fig. 168.3.

Any provider who wishes to make a service available on a JXTA network needs to create an advertisement of the service. An advertisement is a small piece of XML data that announces the existence, and some properties of a service. (In this paper, the advertisement is usually a WSDL document.) The provider then needs to publish the advertisement. Publishing an advertisement allows other service requesters in the same community to find it, using a standardized search mechanism, until the expiration time of the advertisement has passed. At that time, the service provider should publish a new advertisement, if it still wishes to provide the service. When a service requester wants to search for a desirable service, it first sends the service request to registry node in its community. If the registry node does not store any relevant service, it will send the service request to other registry nodes. When the registry node finds an advertisement, it usually puts it in its local

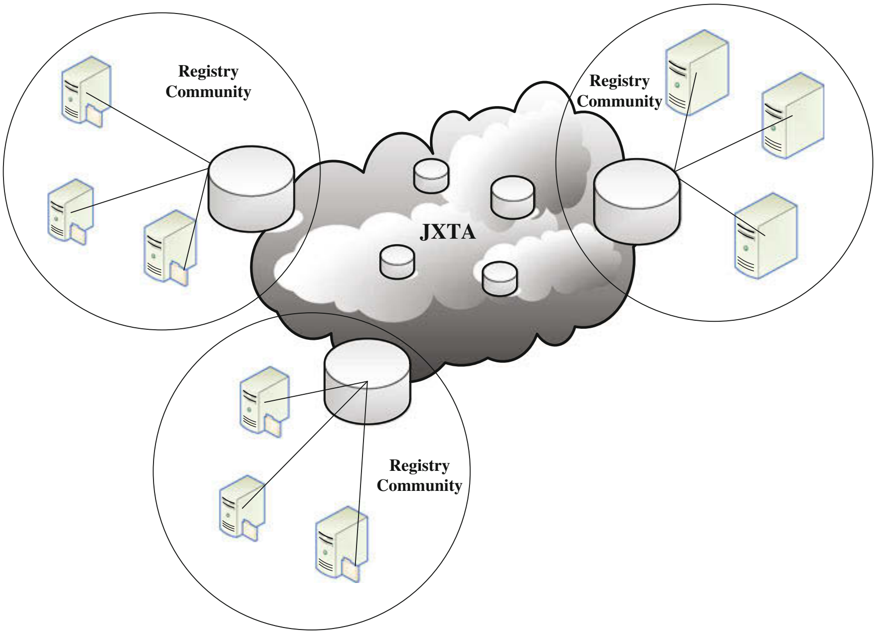


Fig. 168.2 System viewpoint of JWSD model

cache. Other requesters can then retrieve it from there as well as retrieving the advertisement from the actual service provider. This mechanism provides additional redundancy and scalability of JXTA networks.

168.4 Conclusions and Future Work

With the development of Web service, the number of available Web service is growing day by day. How to improve the efficiency of service discovery continues to be an important issue. In this paper, we first introduced the peer-to-peer infrastructure JXTA and proposed the JWSD model in order to overcome the disadvantage of traditional architecture of Web services. The JWSD model can be represented through two viewpoints: system viewpoint and technology viewpoint. The detail of these two viewpoints is discussed in this paper. Our model could provide a scalable Web service discovery in large scale.

In practice, Web services discovery may not be sufficient for the user's demand because of complicated application. How to compose such services in order to satisfy the high-level requirement especially in the distributed environment will be our future works.

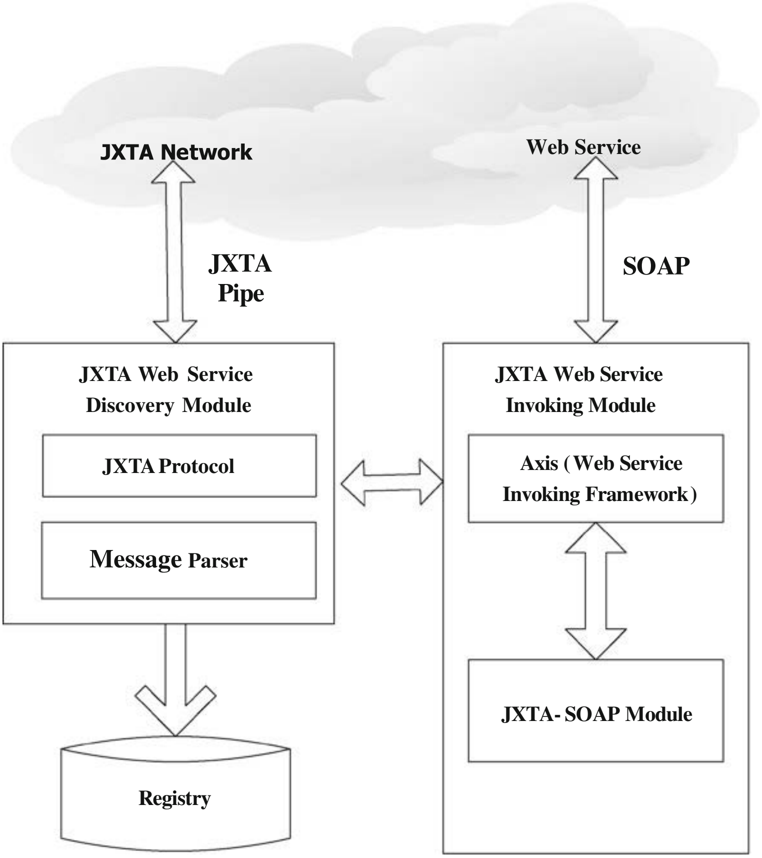

JXTA Network
Web Service
JXTA Pipe
SOAP
JXTA Web Service Discovery Module
JXTA Protocol
Message Parser
JXTA Web Service Invoking Module
Axis (Web Service Invoking Framework)
JXTA- SOAP Module
Registry

Fig. 168.3 Technology viewpoint of JWSD model

References

1. Papazoglou MP, Georgakopoulos D (2003) Service-oriented computing. Commun ACM 46:25–28
2. Chinici R, Moreau J, Ryman A, Weerawarana S (2007) Web services description language (WSDL) version 2.0 Part 1: core language. W3C Recommendation
3. W3C (2001): simple object access protocol (SOAP) 1.2 (W3C working draft 17)
4. Universal description discovery and integration (UDDI) technical white paper (2007), Accessed 8 June 2007
5. Anadiotis G, Kotoulas S, Lausen H, Siebes R (2009) Massively scalable web service discovery. In: Proceedings of 2009 international conference on advanced information networking and applications. IEEE, Bradford, pp. 394–402
6. Emekci F, Sahin OD, Agrawal D, Abbadi AE (2004) A peer-to-peer framework for web service discovery with ranking. In: Proceedings of IEEE international conference on web services (ICWS'04). IEEE, San Diego

7. Paolucci M, Sycara KP, Nishimura T, Srinivasan N (2003) Using DAML-S for P2P discovery. In: Proceedings of 2003 international conference on web services (ICWS'03). IEEE
8. Gagnes T, Plagemann T, Munthe-Kaas E (2006) A conceptual service discovery architecture for semantic web services in dynamic environments. In: Proceedings of the 22nd international conference on data engineering workshops (ICDEW'06). IEEE
9. Castro M, Druschel P, Kermarrec A, Rowstron A (2002) One ring to rule them all: service discovery and binding in structured peer-to-peer overlay networks. In: Proceedings of the SIGOPS European workshop, France
10. Benatallah B, Sheng QZ, Ngu AHH, Dumas M (2002) Declarative composition and peer-to-peer provisioning of dynamic web services. In: Proceedings of 18th international conference on data engineering. IEEE, pp 297–308
11. Banerjee S (2005) Scalable grid service discovery based on UDDI. In: Proceedings of 3rd international workshop on Middleware for grid computing. ACM, pp 1–6
12. Wang ZQ, Hu YY (2007) A P2P network based architecture for web service. In: Proceedings of international conference on wireless communications, networking and mobile computing. IEEE, Shanghai, pp 3446–3449
13. Du ZX, Huai JP, Liu YH (2005) Ad-UDDI: an active and distributed service registry. In: Proceedings of 6th international workshop on technologies for E-services. Springer, Trondheim, pp 58–71
14. Liu ZZ, Wang HM, Zhou B (2007) A two layered P2P model for semantic service discovery. Chin J Softw 18:1922–1932
15. Banaei-Kashani F, Chen C, Shahabi C (2004) WSPDS: web services peer-to-peer discovery service. In: Proceedings of the international conference on internet computing. CSREA Press, Las Vegas, pp 733–743
16. Oaks S, Traversat B, Gong L (2002) JXTA in a Nutshell. O'Reilly, Sebastopol

Chapter 169
The Role of Website Quality in the E-Banking Development: The Chinese Perspective

Zhengwei Ma and Jinkun Zhao

Abstract *Purpose*. The purpose of this research is to analyze factors of website quality that may influence e-banking quality in the Chinese commercial banking sector. Moreover, the paper also aims to analyze the relationship between website quality and customer satisfaction, and to find some key variables for keeping high-level online banking customer satisfaction. *Design/methodology/approach*. The paper describes the positive effect of website quality about e-banking customer satisfaction in China. After the validation of measurement scales, the hypothesis is contrasted through structural modeling. Finally, the authors validate the hypothesis and a measurement model. *Findings*. The data showed that website quality have direct and significant effect on e-banking quality in China. Besides this, the authors found that website quality is positively related to e-banking customer satisfaction. Finally, it is observed that efficiency, interactivity, security, information, ease of use and content are major factors to affect customer satisfaction in the e-banking service. *Originality/value*. This study proposes a model for analyzing empirically the link between website quality and customer satisfaction in the e-banking sector.

Keywords Website quality · Customer satisfaction · E-banking

Z. Ma (✉)
School of Business Administration, China University of Petroleum-Beijing,
Beijing, China
e-mail: mzw8632425@yahoo.com.cn

J. Zhao
Teaching Supervision Department, Harbin University, Harbin, China
e-mail: zjk1998@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_169,

169.1 Introduction

Given the fact that banks invest billions in the Internet infrastructure, customer satisfaction and customer retention are increasingly developing into key success factors in e-banking. But low customer satisfaction is a major encumbrance to depress development of online-banking service in China. Patricio et al. [1] focus-group study found that customers with different patterns of use (e.g., frequency of use and type of operations performed) for an e-banking service tend to value different website attributes, several of which are related to website quality.

Therefore, website quality has significant relationship with customer satisfaction. Website quality is a key factor to affect customer satisfaction.

In the present study, the creation of the measurement items went as follows. Initially, lists of items from existing instruments were compiled that would capture the five broad dimensions of website quality identified in the following areas: privacy/security, information quality, ease of use, graphic style and fulfilment. Where theory is less than well developed, it is beneficial to use both academic and practical perspectives [2, p. 348]. Therefore, using this base list of items, several iterative focus-group discussions were conducted with managers from several banks' online-banking departments in an attempt to choose one item to adequately represent each of the main quality dimensions in e-banking. During these discussions, it was considered important, for the topic of website quality, to break the notion of quality into security, interactivity, efficiency, information, ease of use, content, accuracy, technology and design.

Considering the previous considerations, the paper is structured as follows: firstly, authors carry out a deep review of the relevant literature concerning the variables included in the study; secondly, authors formalize the hypotheses; thirdly, authors explain the processes of data collection and measures validation; fourthly, authors present the results and conclusions of the study. Finally, authors mentioned potential future research.

169.2 Literature Review

In this section, authors review the relevant literature and the focus-group discussions with banks' managers, authors summarize the variables included in the study: website quality (security, interactivity, efficiency, information, ease of use, content, accuracy, technology and design) and customer satisfaction (number of complaints and overall service quality).

Security is the freedom from danger, risks or doubts. It involves physical safety, financial security and confidentiality. It is one important dimension that may affect users' intention to adopt online-banking services. Encryption technology is the most common feature at all bank sites to secure information privacy, supplemented by a combination of various unique identifiers, such as a password, mother's maiden name, a memorable date; in certain cases, a few minutes of inactivity will automatically log a user off the account. Besides, the "secure socket layer," a widely used protocol for online credit card payments, is designed to provide a private and reliable channel between two communicating entities [3]. So, security is a factor that should be taken account of when measuring website quality.

Interactivity is concerned with how an online-banking website interacts with its visitors. It is defined as the facility for users and online-banking websites to communicate directly with one another. In this sense, interactivity denotes any action a user or a website takes while a user is visiting a website. Lowry et al. [4] asserted that the more interactive a website is, the more likely a user is to experience satisfaction. Two-way communication and active control of users are considered critical dimensions of interactivity, and they both play a key role in determining user satisfaction, which underlined the importance of interactivity, which also determines website quality.

Furthermore, Zeithaml [5] reported a significant correlation between efficiency and e-banking quality. Downloading speeds and response time, in consumers' perception, are two crucial facts of e-banking efficiency. Downloading speed depends on the nature of the site from which information is downloaded, the computing hardware and method of connection used to download information [6]. In addition, Kwon and Chidambaram [7], studying consumer perceptions of quality of e-banking services, concluded that efficiency is one of five dimensions sufficiently representative of customers' perceived e-banking quality. Thus, efficiency should play a decisive role in measuring website quality.

Website quality is important because it enhances customer loyalty [8], a key to the success of e-services [9]. Zeithaml [5] and some researchers provide excellent summaries of most of these studies, in which information quality is identified as one the main dimensions of website quality. Among other results, Novak et al. [10] found, through structural equation modeling, that information quality had a significant impact on website quality. In view of these findings, information should figure in when a researcher measures website quality.

Ease of use has been studied extensively in the context of IT adoption and diffusion [11], and it is one of the important measures for user satisfaction, system adoption or IS success [12]. In some studies, system quality has been represented by ease of use, which is defined as the degree to which a system is "user-friendly" [13]. In the context of online banking, consumers may access the websites based on how easy they are to use and how effective they are in helping them to accomplish their tasks [5].

Pikkarainen et al. [14], and Jayawardhena and Foley [6], claim that the content on an online-banking website will affect the website's acceptance among users. The term "content" denotes the design of the service. It creates value if such designs fit customers' needs and if it is clearly understood and updated. Content, needless to say, is a factor that will affect website quality in online-banking services.

If online-banking information is accurate, up-to-date, objective and authentic, it will be considered reliable. Madu and Madu [15] found that accuracy of information played an important role in the formation of satisfaction. Accurate information is information that can be used effectively for a given purpose. In other words, accuracy gives website users the ability to use the information for their purposes. Therefore, many researchers (e.g., [16] believed that it is essential that accurate information be introduced into online-banking services and that accuracy is an important benchmark by which online-banking users judge the website quality.

Zeithaml [5] states that e-service is a web service delivered through the Internet. In e-services, customers interact with or contact service providers through technologies. Customers have to rely entirely on information technologies in an e-service encounter [5]. An online-banking service's deeds, effort or performance are mediated by information technologies. Zeithaml [5] stated that some dimensions of SERVQUAL may be applied to e-service quality, but in e-service there are additional dimensions, many of which are specifically related to technologies. So, technologies are a factor that affects website quality.

In the virtual environment of e-service, for customers, a website is a main access to online banking and to a successful online process. Thus, a deficiency in website design can result in a negative impression of the website quality, resulting in customers exiting an online-banking transaction. A website is a starting point for customers to gain confidence with the business. Website design can influence customers' perceived image of a company. With good navigation and useful information on its website, a company may easily attract customers to its online-banking services. Websites proprietors, thus, should provide appropriate information and multiple functions for their customers. So, clearly, design is an important aspect of website quality.

Some researchers found that customer complaints had a direct effect on customer satisfaction. They reported that as one-dimensional attributes increased, the level of overall customer satisfaction also increased. Researchers discovered that major gains in customer satisfaction were likely to come from an alleviation of complaints. These researchers, overall, concur that the number of complains is an index of customer satisfaction. This is why, in the present study, the number of complaints were used to measure customer satisfaction.

Service quality is defined as a long-term cognitive judgment regarding an organization's "excellence or superiority". Two main streams of research into the dimensions of service quality exist: the Nordic school, which tends to incorporate the process and outcome dimensions, and the North American school, which draws on SERVQUAL [17]. A customer-oriented quality strategy is critical to service firms as it drives customers' behavioral intention with, for instance, highly perceived service quality leading to repeat patronage and customer loyalty [5]. Accordingly, substandard service quality will lead to negative word-of-mouth, which may result in a loss of sales and profits as the customers migrate to competitors [5]. These factors stress the importance of delivering

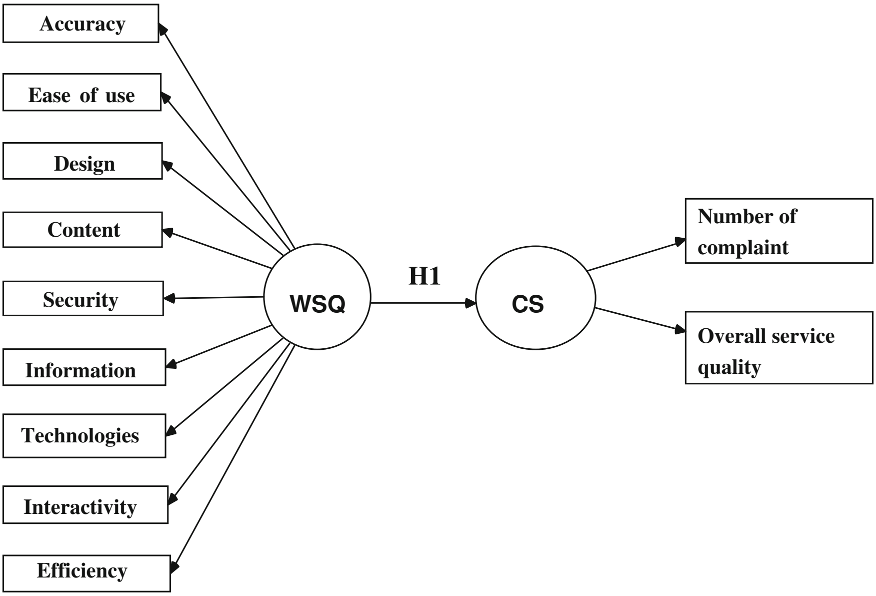


Fig. 169.1 Research factors and hypotheses in the present study

high-level services, especially within an electronic environment, where customers can readily compare service firms and where switching costs are low.

169.3 Constructs for the Present Study and Hypothesis

According to the possible connection between website quality and customer satisfaction in handling private data, a direct relationship might be established between the two concepts [18–20]. And follow the prior study; one construct is addressed in the present study: website quality and customer satisfaction, all of which are elaborated in prior paragraphs. The relationships between these constructs, as embedded in the hypothesis, are now illustrated in Fig. 169.1.

Taking into account the previous considerations, the relationship between website quality and customer satisfaction is evident in personal data handling and should be examined in greater detail. With the aim of testing this connection in the online-banking customer satisfaction, the following hypothesis is proposed:

H1: there will be a positive relationship between website quality and customer satisfaction.

169.4 Data Collection

The generation of the initial questionnaire was ascertained by experts and managers interviews at banks as well as through in-depth discussions with online-banking users. Pre-tests of the initial 24-item questionnaire were carried out with 30 online users to improve the questionnaire. The resulting modified 11-item pool was presented to Chinese users of online banking in drop in survey. Respondents were asked to refer to their own online-banking service (the one they use regularly) when answering the questionnaire. Non-random method of collecting the data (volunteer sampling) generated 198 fully usable questionnaires. The questionnaires of collection are non-random samples. So authors compared some of the survey results with available information about the population. The results are very similar and as a consequence, authors may conclude that our sample represents the profile of the average Chinese online-banking users.

169.5 Results

Authors developed a structural equations model, which the objective of testing is the proposed hypothesis (Fig. 169.2). Authors observed that the hypothesis was supported at the 0.01 level and, in a similar way. Model fit was acceptable [Chi-square = 103.13, 43 df, $p < 0.001$; goodness-of-fit index (GFI) = 0.908; adjusted goodness-of-fit index (AGFI) = 0.859; Comparative Fit Index (CFI) = 0.968; Bollen (IFI) Fit Index = 0.968; Bentler–Bonett Normed Fit Index (NFI) = 0.946; the root mean square error of approximation (RMSEA) = 0.084; normed Chi-Square = 2.398].

It was also notable that this model has allowed authors to explain at a very high level the website quality in customer satisfaction of online-banking service. Besides, according to the standardized estimates, authors may say that customer satisfaction is clearly and positively influenced by website quality in handling personal data ($\beta = 0.73$). And authors found that efficiency, interactivity, securities, information, ease of use and content have larger effect than other three factors for customer satisfaction ($\beta > 0.85$).

169.6 Discussion

Review the results, online-banking customer satisfaction cannot be described as one fact construct. Instead, it represents a multi-factor construct that is composed of website quality judgments with regard to the service categories. This study provides validated measurement scales for each factor [21–22]. The empirical results strongly support the understanding of website quality as integral solutions.

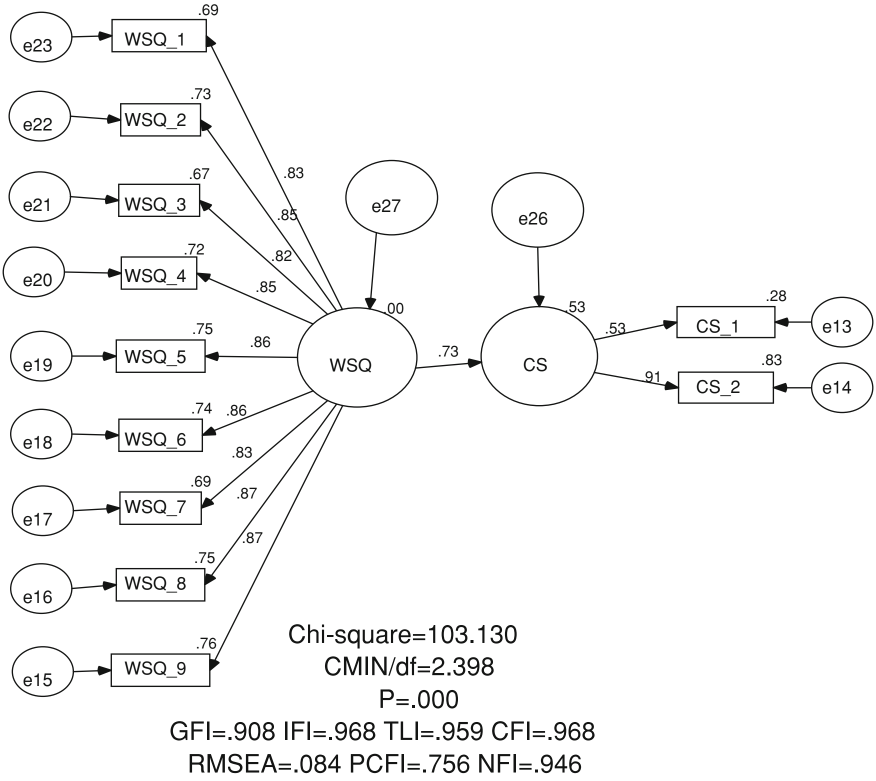


Fig. 169.2 The structural equation model

Based on our findings, management can establish a sequential priority to improve website quality in online-banking service. The sequential priority depends on the influence to customer satisfaction. For example, when limited resources become the barrier to improve all of nine factors, banks can improve the first efficiency, interactivity, securities, information, ease of use and content to be first step. And the banks could put accuracy, technology and design to be second step.

169.7 Limitations

There are several limitations to the present study. First, the sample was China-focused, with all of the respondents residing in China. The participants in this survey may have possessed attributes and behaviors that differed from those in other parts of the world. Second, the sample was restricted to the consumers of banks and may have possessed attributes and behaviors that differ from those of consumers in other business sectors. Next, as mentioned earlier, in the data collection section, since it was impossible to send follow-up surveys, no attempts were made to ascertain the existence of non-response bias by comparing responses to the first-wave surveys with those to a second wave.

References

1. Patricio L, Fisk R, Cunha J (2003) Improving satisfaction with bank service offerings: measuring the contribution of each delivery channel. Manag Serv Qual 13(6):471–482
2. Hensley R (1999) A review of operations management studies using scale development techniques. J Oper Manag 17(3):343–358
3. Hutchinson D, Warren M (2003) Security for internet banking: a framework. Logist Inform Manag 16(1):64–73
4. Lowry PB, Spaulding T, Wells T, Moody G, Moffit K, Madariaga S (2006) A theoretical model and empirical results linking website interactivity and usability satisfaction. Paper presented at the 39th Hawaii international conference on system sciences, Hawaii, Jan 2006
5. Zeithaml VA (2000) Service quality, profitability and the economic worth of customers: what we know and what we need to learn. J Acad Mark Sci 28(1):67–85
6. Jayawardhena C, Foley P (2000) Changes in the banking sector: the case of internet banking in the UK. Internet Res Electron Netw Appl Policy 10(1):13–30
7. Kwon HS, Chidambaram L (2000) A test of the technology acceptance model: the case of cellular telephone adoption. Paper presented at the third Hawaiian international conference on system sciences, Jan 2000
8. Boyer K, Hult G. (2005) Extending the supply chain: integrating operations and marketing in the online grocery industry. J Oper Manag 23(6):642–661
9. Reichheld F, Schefter P (2000) E-loyalty: your secret weapon on the web. Harv Bus Rev 78(4):105–114
10. Novak T, Hoffman D, Yung Y (2000) Measuring the customer experience in online environments: a structural modeling approach. Mark Sci 19(1):22–42
11. Davis RD (1989) Perceived usefulness, perceived ease of use and user acceptance of information technology. MIS Q 13(1):319–339
12. Moore GC, Benbasat I (1991) Development of an instrument to measure the perceptions of adopting an information technology innovation. Inform Syst Res 2(3):192–222
13. Doll WJ, Torkzadeh G (1988) The measurement of end-user computing satisfaction. MIS Q 12(2):259–274
14. Pikkarainen T, Pikkarainen K, Karjaluoto H, Pahnila S (2004) Consumer acceptance of online banking: an extension of the technology acceptance model. Internet Res 14(3): 224–235
15. Madu CN, Madu AA (2002) Dimensions of e-quality. Int J Qual Reliab Manag 19(3): 246–258
16. Waite K, Harrison T (2002) Consumer expectations of online information provided by bank websites. J Financial Serv Mark 6(4):09–322
17. Parasuraman A, Berry LL, Zeithaml VA (1991) Refinement and reassessment of the SERVQUAL scale. J Retail 67(4):420–450
18. Ahmed I, Gul S, Hayat U, Qasim M (2001) Service quality, service features and customer complaint handling as the major determinants of customer satisfaction in banking sector: a case study of National Bank of Pakistan. www.wbiconpro.com/5%5B1%5D.ISHFA.pdf. Accessed date 14 Jan 2001
19. Bauer HH, Hammerschmidt M (2002) Financial portals in the internet. Paper presented at the WSEAS conference on E-commerce, Janeiro, Oct 2002

20. Hair JF Jr, Anderson RE, Tatham RL, Black WC (1998) Multivariate data analysis. Prentice-Hall, Englewood Cliffs
21. Li H, Suomi R (2007) Evaluating electronic service quality: a transaction process based evaluation model. Paper presented at the European conference on information management and evaluation, Sept 2007
22. Parasuraman A, Zinkhan G (2002) Marketing to and serving customers through the internet: an overview and research agenda. J Acad Mark Sci 30(4):286–295

Chapter 170
Web Service Composition Method Using Hierarchical Reinforcement Learning

Hao Tang, Wenjing Liu and Lei Zhou

Abstract Through web services composition, distributed applications and enterprise business processes can be integrated by individual service components developed independently. In this chapter, we concentrate on the optimization problems of dynamic web service composition, and our goal is to find an optimal composite policy. We introduce a hierarchical reinforcement learning technique, i.e., a continuous-time unified MAXQ algorithm, to solve large-scale web service composition problems in the context of continuous-time semi-Markov decision process (SMDP) model under either average- or discounted-cost criteria. Finally, this proposed algorithm is tested in a simulation, and the experimental results show that it has better optimization performance than the flat Q-learning.

Keywords Web service composition · Hierarchical reinforcement learning · Semi-Markov decision process (SMDP) · MAXQ

H. Tang (✉) · W. Liu · L. Zhou
School of Computer and Information, Hefei University of Technology,
Tunxi Road No.193, Hefei, 230009 Anhui, People's Republic of China
e-mail: htang@hfut.edu.cn

W. Liu
e-mail: winnykris@163.com

L. Zhou
e-mail: zhouleizhl@hotmail.com

H. Tang
Engineering Research Center of Safety Critical Industry Measure and Control
Technology, Ministry of Education, Hefei, People's Republic of China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_170,

170.1 Introduction

A single web service usually cannot fulfill the requirements of a user, while web service compositions provide a way to combine a set of simple web services into more powerful services or new value-added services that can satisfy the user's needs. So, in the modern high-tech world, web service composition has played more and more important roles in many domains [1, 2], typically in e-commerce and enterprise application integration, and has attracted many researchers' attention.

So far, a number of methods have been proposed to achieve dynamic web service composition. For example, in Wang et al. [3], a new algorithm is applied for dynamic service composition based on reinforcement learning (RL) and logic of preference. However, in large-scale web service composition problems, the current tabular RL methods suffer from the "curse of dimensionality", which is the exponential growth of computational and memory requirements with the number of system state variables.

Wang et al. has adapted MAXQ algorithm to dynamic service composition, which is based on discrete-time semi-Markov decision process (DT-SMDP) [4]. But in real-word applications, web service compositions are always related to run time. Therefore, it is more practical to use continuous-time hierarchical reinforcement learning (HRL) algorithms to solve web service composition problems. In this article, we will propose a dynamic web service composition method based on MAXQ algorithm, in the context of continuous-time semi-Markov decision process (CT-SMDP). It is effective in dealing with the "curse of dimensionality" and the "curse of modeling" for practical large-scale service composition.

170.2 Web Service Composition Model

The framework of the dynamic web service composition model is shown in Fig. 170.1 [1].

First, the service providers advertise their atomic services at a global market place. Once a service requester submits his requests and the task acceptor accepts the information, the composed service engine will try to solve the requirement by composing the atomic services advertised by the service providers. Then the execution engine will accept the corresponding flowchart, and send the service specification to the service matchmaking, which will find the most appropriate atomic web services and return the information to the execution engine. Finally, the execution engine invokes and executes each atomic service, and the result will be then sent back to the service requester.

The dynamic web service composition problem can be modeled as an SMDP [5–7], which is a more general model than MDP. When a composition process evolves to a task node, the composed service engine should decide to select a

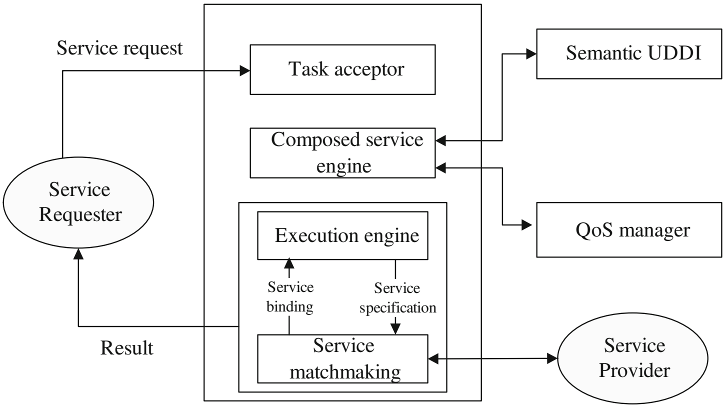


Fig. 170.1 Web service composition model

concrete web service, and the moment of making decision is called decision epoch. Here, we use T_n to denote the nth decision epoch with $T_0 = 0$. If there are l task nodes to be bound in a web service composition problem, the system state s is then defined as a conjunction of status of each task node, i.e., $s = <s_1, \ldots, s_k, \ldots, s_l>$, where s_k corresponds to the kth task node of this service composition. $s_k = 1$ represents that this node is active and has been bound to a concrete web service, while $s_k = 0$ means that this node is not active. Let Φ be the system state space, i.e. $\Phi = \{1, 2, \ldots, N\}$, and $s \in \Phi$. At time T_n and state s_n, the action a_n is selected from the set of all possible candidate web services $A(s_n)$, i.e., $a_n \in A(s_n)$, and write $A = \bigcup A(s_n)$. Then, a stationary policy π represents a mapping from states to actions, i.e. $\pi : \Phi \to A$. Here $p(s_{n+1}|s_n, a_n)$ is the transition probability, that is, under a concrete action a_n, the system transits from current state s_n to next state s_n with probability $p(s_{n+1}|s_n, a_n)$ or still stays at state s_n with probability $1 - p(s_{n+1}|s_n, a_n)$. Let $\tau_{ss'}$ be the interval time between T_n and T_{n+1}, and it can also be called the sojourn-time of state s_n under action a_n, which follows a random distribution. Then, the web service composition problem can be modeled as an SMDP. Suppose that the expected cost the system pays every unit time at state s_n under action a_n and before transiting to next state s_{n+1} is denoted by $f(s_n, a_n, s_{n+1})$. Then, we take the following infinite-horizon expected cost criteria [7]

$$\eta_\alpha^\pi(s) = E\left[\sum_{n=0}^{\infty} \int_{t_n}^{t_{n+1}} \alpha e^{-\alpha t} f(s_n, a_n, s_{n+1}) \mathrm{d}t | s_0 = s\right] \qquad \forall s \in \Phi \qquad (170.1)$$

Here, α denotes a discount factor, $0 < \alpha < 1$, and when $\alpha > 0, \eta_\alpha^\pi(s)$ represents the long-run expected total discounted cost under policy π. As a special case, if $\alpha \to 0$, the limitation $\eta_0^\pi(s)$ represents the following infinite-horizon expected average cost

$$\eta^{\pi} = \lim_{N \to \infty} \frac{1}{t_N} E\left[\sum_{n=1}^{N-1} \int_{t_n}^{t_{n+1}} f(s_n, a_n, s_{n+1})\mathrm{d}t\right] \tag{170.2}$$

170.3 Continuous-Time MAXQ Algorithm for Service Composition

To construct MAXQ decomposition for the web service composition problem, we should first identify a set of individual subtasks that we believe important for solving the overall task. More formally, the MAXQ method decomposes the target task M into a set of subtasks $\{M_0, M_1, \ldots, M_m\}$ and decomposes a hierarchical policy π into a set of policy $\{\pi_0, \pi_1, \ldots, \pi_m\}$ with π_i corresponding to subtask M_i [8, 9]. In this case, let us define four tasks as follows:

- *Root*. This is the whole web service composition task.
- *Input*. In this subtask, the goal is to get the request information so as to invoke a concrete service.
- *Output*. In this subtask, the goal is to obtain the service output data.
- *Comp*. In this subtask, the goal is to select an appropriate composite model during the composition process. In other words, it is to move the service process from its current state to target state.

The decomposition for the $V^{\pi}(i,s)$ is also shown by Fig. 170.2. Each circle is a state of the composition process, and suppose subtask M_i is initiated at state s_I and terminated at state s_T. If i is a primitive action, then $s = s_I$, $s' = s_T$. The interval time $\tau_{ss'}$ between state s and its next state s' is the sojourn-time of state s, which is assumed to be exponential distribution during our experiments in the next section. If i is a composite action, $\tau_{ss'}$ is the cumulative time that can be separated into several interval times, and is no longer exponential distribution. Then, the value function $V^{\pi}(i,s)$ of state s for M_i in the MAXQ algorithm is broken into two parts: the value of the subtask M_j that is independent of the parent task M_i, and the value for completing M_i after executing subtask M_j that of course depends on the parent task M_i. So, we have

$$V^{\pi}(i,s) = \begin{cases} f'(s,i,s') & \text{if } i \text{ is primitive} \\ Q^{\pi}(i,s,\pi_i(s)) & \text{if } i \text{ is composite} \end{cases} \tag{170.3}$$

where

$$f'(s,i,s') = k_1(s,i) + \int_0^{\tau_{ss'}} k_2(s,i,s\text{'}) \times e^{-\alpha t}\mathrm{d}t - T_{\alpha}(\tau_{ss'}) \times \tilde{\eta} \tag{170.4}$$

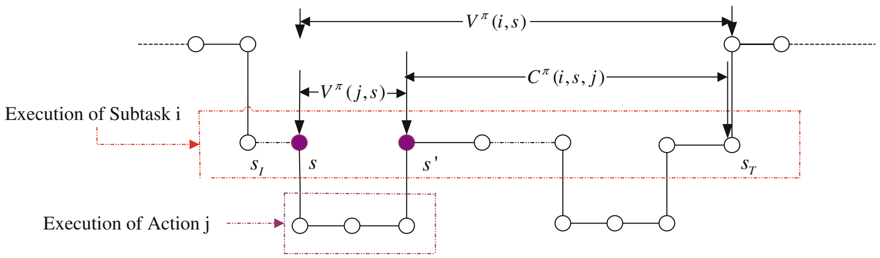


Fig. 170.2 Value function decomposition for MAXQ algorithm of continuous-time SMDP

$$Q^{\pi}(i,s,j) = V^{\pi}(j,s) + C^{\pi}(i,s,j) \tag{170.5}$$

Here, $k_1(s,i)$ represents the immediate cost of executing primitive action i at state s; $k_2(s,i,s')$ represents the time-related cost that the system pays every unit time, such as response time. In addition, $\tilde{\eta}$ represents the estimate of average cost, satisfying

$$\tilde{\eta} := \frac{S_f}{S_w} \tag{170.6}$$

with s_f and s_w being learned, respectively, as follows:

$$S_f := S_f + \beta_n\left(f'_{\alpha=0}(s,i,s') - S_f\right) \tag{170.7}$$

$$S_w := S_w + \beta_n(\tau_{ss'} - S_w) \tag{170.8}$$

Here, β_n is a stepsize, and $f'_{\alpha=0}(s,i,s')$ is calculated by (170.4) with $\alpha = 0$, which denotes accumulated cost from s to s' under primitive action i.

The projected value function for the root is then decomposed into the ones for the individual subtasks and the individual completion functions recursively by equations (170.3–170.5). Furthermore, $V^{\pi}(i,s)$ and $C^{\pi}(i,s,j)$ are updated, respectively, as follows:

$$V^{\pi}(i,s) := (1-\gamma)V^{\pi}(i,s) + \gamma \times f'(s,i,s') \tag{170.9}$$

$$C^{\pi}(i,s,j) := (1-\gamma)C^{\pi}(i,s,j) + \gamma\left(e^{-\alpha T}(C^{\pi}(i,s',j*) + V^{\pi}(j*,s')) - T_{\alpha}(T) \times \tilde{\eta}\right) \tag{170.10}$$

$$j* = \arg\min_{j \in A_i(s')}(V^{\pi}(j,s') + C^{\pi}(i,s',j)) \tag{170.11}$$

Here, $T_{\alpha}(\tau) = \int_0^{\tau} e^{-\alpha t}\mathrm{d}t = \frac{1-e^{-\alpha\tau}}{\alpha}$ for any discount factor $\alpha > 0$ and time $\tau > 0$, and let $T_0(\tau) = \lim\limits_{\alpha \to 0} T_{\alpha}(\tau) = \tau$. In addition, γ is a stepsize, and T is the current total cumulative time for executing subtask M_i.

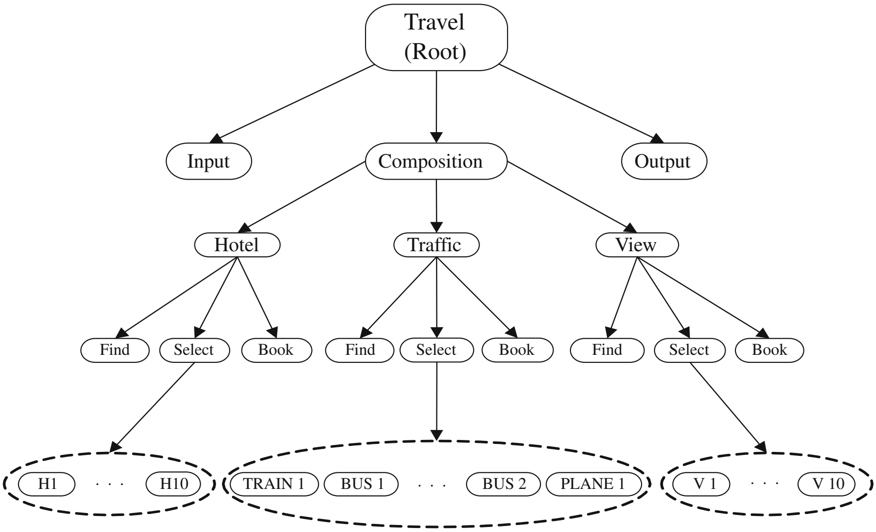


Fig. 170.3 Travel reservation model

Due to the introduction of the term $T_{\alpha}(\cdot) \times \tilde{\eta}$ in Eqs. (170.4) and (170.10) according to the idea of performance potential for SMDP [7, 10], the unified MAXQ algorithm can then be established for both discounted and average criteria, which is the difference between our algorithm and other MAXQ algorithms.

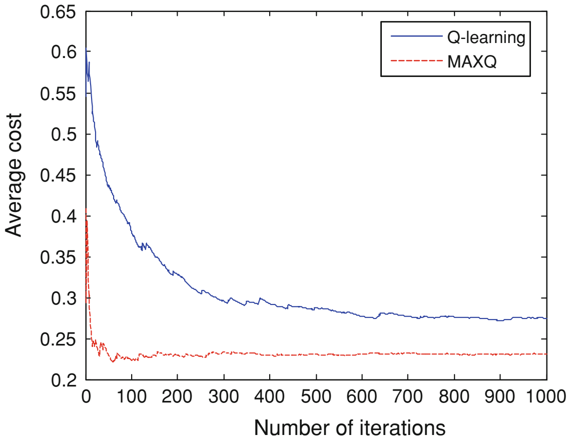
170.4 Experiments

170.4.1 Simulation Model

Let us assume that a customer wants to travel to some place. He/she will first talk to the travel agent who notes the customer's requests and generates a corresponding *trip request* document that may contain several requirements *plane/train/bus tickets hotels, car rental, excursions,* etc. The travel agent performs all bookings and when he is done, he puts the *trip request* either into the *canceled requests* or into the *completed requests* data base. A completed document is sent to the customer as an answer to his request. If the booking fails, the customer is contacted again and the whole process reiterates.

As Fig. 170.3 shows, the travel reservation problem is decomposed into four levels. The highest level is the task of *input*, *composition* and *output*. Furthermore, the *composition* task decomposes into *hotel*, *traffic* and *viewpoint*, which constitute the second level. Then they all decompose into three subtasks-*find, select, book* respectively. In the last level, there are primitive actions which are all candidate

Fig. 170.4 Average cost of the two algorithms



web services that can be bound to the corresponding parent task node. The goal is to find an optimal composite policy.

170.4.2 Experimental Results

As a case study, we suppose there are three web service classes, such as hotel, traffic and viewpoint, and the number of candidate services for each web service class is ten, as shown in Fig. 170.3. For comparison, we used Q-learning to simulate this problem at first. In Q-learning, ε—*greedy* actions are necessary for exploration, especially at the beginning of the optimization. So we use $\varepsilon = 0.3$, learning step $\gamma = 1/\left(8 \times N(s,i)^{0.2}\right)$, here $N(s,i)$ is the number of state-action pairs (s,i) that has been visited. $k_1(s,i)$ is different for every state-action pairs (s, i), and $k_2(s,i,s') = 0.8$

First, we consider the average case, i.e. $\alpha = 0$. Two optimization plots are provided respectively in Fig. 170.4, where each y-axis denotes the average cost of each algorithm in 1,000 episodes. In this problem, a simulation episode is staring from any state to the termination state. Figure 170.5 shows the results of Q-learning and MAXQ algorithm for discounted cases with the discount factor $\alpha = 0.01$. We observed that, compared with the Q-learning, the performance of the MAXQ method is more efficient whether in average case or in discounted cases. This is because the MAXQ method accelerated the learning speed and also improved the optimizing precision via hierarchy and action abstractions. Due to the utilization of discount factorα, the curve for discounted case is better than the average case.

On the other hand, we extended the number of tasks to test the other performance values, such as the success rate and the computation cost. Figure 170.6

Fig. 170.5 Discounted costs of two algorithms

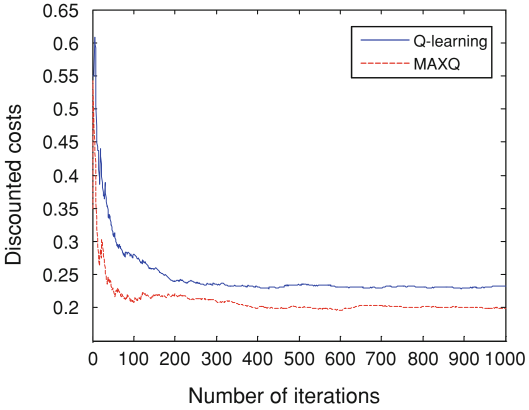
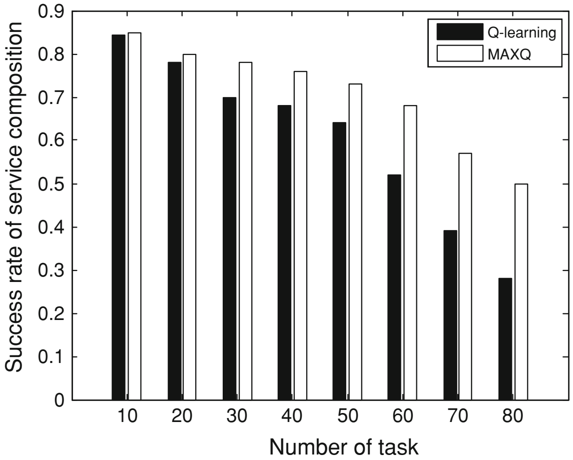


Fig. 170.6 The web service composition success rates of both algorithm with different of tasks



shows the success rates of Q-learning and MAXQ algorithm with different number of tasks, respectively. We can see that when the number of tasks is ten, the success rates of these two algorithms are almost the same. But as the number of tasks increases, the success rates of Q-learning reduce faster than success rates of MAXQ algorithm. Besides, the computation costs of both algorithms are shown by Fig. 170.7. Obviously, when the number of tasks becomes large, the computation cost of Q-learning grows faster than that of MAXQ algorithm. These two figures fully illustrate that the MAXQ algorithm is more effective in large-scale services composition problems.

The differences for the two algorithms are also listed in Table 170.1. The average cost of MAXQ algorithm is almost 15.8% less than that of Q-learning. It is because the policies learned in subproblems can be reused for multiple parent tasks, so the accuracy of optimization has been increased. On the other hand, the value functions learning in subproblems can be shared, so the learning speed is

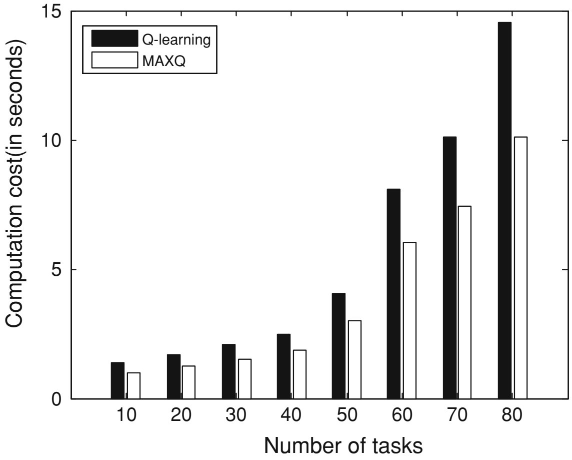


Fig. 170.7 The computation cost of both algorithms with different number of tasks

Table 170.1 Results for different algorithms

Algorithms	Q-learning	MAXQ
Average cost	0.2746	0.2311
Time for 1,000 steps	1.6688 s	1.7661 s
Time for getting good results	1.3678 s	1.0067 s

accelerated. When the curves tend to smooth, we consider that the algorithm gets a good result. So we can see that the time spent for getting good results by MAXQ algorithm is about 26.4% less than that by Q-learning, as shown in Table 170.1.

170.5 Conclusions

The web service composition problems, under either average- or discounted-cost criteria, are solved effectively by using continuous-time unified MAXQ algorithm. Compared with Q-learning, the proposed algorithm tends to be more suitable for solving the "curse of dimensionality" in large-scale web service composition problems. The simulation results also demonstrated that the MAXQ algorithm has the advantages of high effectiveness and high learning speed. In addition, as part of our ongoing work, we will extend our algorithm to support multi-agent web service composition problems.

References

1. Rao J, Su X (2004) A survey of automated web service composition methods. In: Proceedings of the first international workshop on semantic web services and web process composition, pp 43–54
2. Zhao H, Doshi P (2006) Composing nested web processes using hierarchical semi-Markov decision process. AAAI workshop on AI-driven technologies for services-oriented computing, pp 75–84

3. Wang HB, Tang PP, Hung P (2008) RLPLA: a reinforcement learning algorithm of web service composition with preference consideration. IEEE congress on services Part II, pp 163–170
4. Wang HB, Guo XH (2009) Preference-aware web service composition using hierarchical reinforcement learning, Web Intelligence/IAT Workshops, pp 315–318
5. Bradtke SJ, Du MO (1995) Reinforcement learning methods for continuous-time Markov decision problems, in advances in neural information processing systems 7. MIT Press, Cambridge, pp 393–400
6. Mahadevan S, Marchalleck N, Das T, Gosavi A (1997) Self-improving factory simulation using continuous-time average-reward reinforcement learning. In: Proceedings of the 14th international conference on machine learning, pp 202–210
7. Cao XR (2003) Semi-Markov decision problems and performance sensitivity analysis. IEEE Trans Autom Control 48(5):758–769
8. Dietterich TG (2000) Hierarchical reinforcement learning with the MAXQ value function decomposition. J Artif Intell Res 13:227–303
9. Dietterich TG (1998) The MAXQ method for hierarchical reinforcement learning. In: Proceedings of the fifteenth international conference on machine learning, pp 118–126
10. Tang H, Arai T (2009) Look-ahead control of conveyor-serviced production station by using potential-based online policy iteration. Int J Control 82(10):1916–1928

Chapter 171
Research on Agent for Mail Classification

Ying-hui Sun, Ying-juan Sun and Dong-bing Pu

Abstract This chapter researches on agent for mail classification. An Email Intelligent Classification system is designed to classify the e-mails into four kinds by their significance. It has proved the feasibility of e-mail classification by experiment.

Keywords Agent · Mail classification · Intelligence

171.1 Introduction

In the information society, the access, processing and transmission of information are the primary task of social operation. In the future society, it is not what we have that users will use, but we will be able to provide what users will need. Intelligent agent is produced to rather facilitate users and provide help to them [1].

Generally speaking, intelligent agent is a calculating entity, which can operate in a particular environment, adapt to the changes in the environment and take flexible and independent action to meet their design goal [2].

Y. Sun
College of Computer Jilin Normal University, Siping, China
e-mail: sunyh178@163.com

Y. Sun (✉)
College of Computer Science and Technology Changchun Normal University,
Changchun, China
e-mail: syj_pyf@sohu.com

D. Pu (✉)
College of Computer Science and Information Technology
Northeast Normal University, Changchun, China
e-mail: pudb@nenu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_171,

E-mail is currently the most widely used means of communication in a computer, which further promotes the exchanges between people. But the more e-mail, the more frequent coordinating work needs to be done. If the computer can automatically finish coordinating e-mail, it will enable people to avoid many of the tedious work. There are also some current e-mail systems with filtering function, but it is simple interception just under spam addresses, there are not really intelligent services. Some foreign intelligence agent systems (for example: Lydia E-mail Agent, etc.) can be smart to classify e-mail, but it usually requires users to interact with the software constantly, which was very inconvenient.

Based on the above, this paper presents an e-mail classifying Agent algorithm, named ILDS. According to the degree of importance, ILDS algorithm classifies e-mails received into important e-mail, Less important e-mail, Deleted e-mail and Strange e-mail. Learning the address, theme, text and user habits of the received mail enables us to achieve the intelligent mail classification.

171.2 Introduction of e-mail System

A. **E-mail Address**
An e-mail address is known as an electronic mail address. Let us use an e-mail address to analyze its composition. For example, it is sss@sohu.com.cn. @ divides the address into two parts: “sss” on the left of the address stands for the receiver’s account name and the part on the right stands for domain name.

B. **E-mail Composition**
Usually e-mail has five main parts: the addressee’s address, the sender’s address, the mail theme, the mail text and annexes.

C. **E-mail Transfer Process**
Use the mail client software to create a new mail. Enter the mail addressee’s address, theme, text, and add annex if necessary. Send the mail through the mail server to the receiving mail server [3, 4].

171.3 Mail Classification

Every user has a yardstick on the importance of receiving the mail. According to the degree of importance, ILDS algorithm classifies it into important mail, less important mail, deleted mail and strange mail.

Important mail refers to the one with relatively high frequency of communication or of great significance. Less important mail refers to the one with low frequency of communications or of no significance—ordinary mail. Strange mail refers to the one of the first communication. Delete mail refers to the advertisement or other junk mail, which is the most irritating mail and an important source of the virus.

171.4 Intelligent Mail Classification

To better distinguish mail, we give different types of mail different thresholds. The highest threshold is the important mail, followed by less important mail, the lowest for the deleted mail. Without loss of generality, the threshold between 0 and 1 is deleted mail; the one between 1and 2 is less important mail; the one between 2 and 3 is important mail.

A. Classify e-mail According to the e-mail Addresser's Address

(1) Initialization of system
Firstly, establish the set of initial important e-mail address, less important e-mail addresses and deleted e-mail address. The establishment of sets can be done through mail sampling. Through sampling, establish the initial glossary index storehouse of the important e-mail, less important e-mail and deleted e-mail.

(2) The process of e-mail classification

  (a) Extricate the sender's address from the boxes.
  (b) Judgment.

  If the address is in the set of important e-mail addresses, the mail will be classified as important e-mail. If the address is in the set of less important e-mail addresses, the mail will be classified as less important e-mail. If the address is in the set of strange addresses, judge the e-mail through other methods. If there is no @, just delete the e-mail. In addition, if there are several addresses in the address column, and there are two or more addresses with the same prefix, the e-mail is classified as deleted e-mail.

B. Classify e-mail According to the User's Habits
ILDS algorithm does not deal with the e-mail first received that does not belong to the present set of addresses and classifies it as stranger e-mail. When receiving the e-mail address again, the algorithm will deal with it, which can increase accuracy of the intelligent e-mail classification.

(1) The e-mail address belongs to the set of strange mail addresses

  (a) Delete e-mail
  If the user does not read the e-mail and just deletes it after opening the mail, it is classified as deleted mail.
  (b) Reply e-mail within 48 h
  The speed of replying e-mail is not the only important indicator, but it also reflects the importance of a mail. Therefore, the speed of replying e-mail in

accordance with users gives a weight value d_1. $\mathrm{d}_1 = 3 - (\mathrm{t}_2 - \mathrm{t}_1)/24$,where t_2 stands for the reply time and t1 for open box time, whose unit is hour.

(c) Do not reply e-mail within 48 h
This kind of e-mail belongs to the one that needs no reply or junk e-mail, so we are not sure of it. It relies on the content of the mail. The weight value of the e-mail is $\mathrm{d}_1 = -1$.

(2) The e-mail address does not belong to the set of strange mail addresses
This kind of e-mail belongs to the first mail. If the user reads it, the e-mail for the user is useful. It should be classified as strange e-mail. If the users do not read it, the e-mail is junk mail. It should be classified as strange e-mail.

C. Classify e-mail According to its Theme and Text

(1) Get the weight value d21 according to the theme of e-mail

(a) Vocabulary weights
The frequency that vocabulary appears in different types of e-mail theme shows the different weights. Vocabulary weights correspond to e-mail weights: the weights of the themes of important e-mails are between 2 and 3; the weights of the themes of less important e-mails are between 1 and 2; the weights of the themes of deleted e-mails are between 0 and 1(statistics is held in terminology indexing database). Once classify an e-mail, the index of vocabulary will be updated. In addition, the same word may appear in different vocabulary indexing databases with different weights. See specific solution 4.4.Under the theme of the right to obtain items worth challenged.

(b) To get weights d21 according to e-mail theme

① Extract the vocabulary of e-mail theme.

② Get the weighted average ad [i] of the i-word to make the emergence of vocabulary total number ($t = t_1 + t_2 + t_3$): the number of the emergency of vocabulary in important e-mails is t1; the number of the emergency of vocabulary in less important e-mails is t_2; the number of the emergency of vocabulary in deleted e-mails is t_3, and make the weights of vocabulary in important e-mails f_1; the weights of vocabulary in less important e-mails f_2; the weights of vocabulary in deleted e-mails f_3. The representation of each word to e-mails is related to the frequency it appears in the e-mail, so make:

$$\begin{cases} \mathrm{ad}[i] = \frac{f_1 \times t_1}{t} + \frac{f_2 \times t_2}{t} + \frac{f_3 \times t_3}{t} & (t \neq 0) \quad (171.1) \\ \mathrm{ad}[i] = 0 & (t = 0) \quad (171.2) \end{cases}$$

③ Get the weighted average ad of each word in the theme and make the n words in the theme, then

$$\begin{cases} \text{ad} = (\text{ad}[1] + \text{ad}[2] + \ldots + \text{ad}[n])/n & (n \neq 0) \quad (171.3) \\ \text{ad} = 0 & (n = 0) \quad (171.4) \end{cases}$$

④ Get the weights $d_{21} = \text{ad}$.

(2) To get weights d_{22} according to e-mail text

(a) Vocabulary weights
The statistics of vocabulary in the text is held in index database. The calculating method of weights is the same as that of the theme vocabulary.

(b) To get weights d22 of e-mail according to e-mail text
According to the text classified e-mails, make main statistics of the weighted average of relatively high frequency of vocabulary. Take the weighted average of vocabulary from the former n words (*n* from 5 in Test System). Statistical method is to identify the vocabulary of high frequency, then use the same method as vocabulary to theme the same values to calculate the value of the theme vocabulary to get d_{22}.

① Extract n words from the text of the mail.

② How to get the weighted average ad[i] to make the emergence of vocabulary total number ($t = t_1 + t_2 + t_3$): the number of the emergency of vocabulary in important e-mails is t1; the number of the emergency of vocabulary in less important e-mails is t_2; the number of the emergency of vocabulary in deleted e-mails is t_3, and make the weights of vocabulary in important e-mails f1; the weights of vocabulary in less important e-mails f_2; the weights of vocabulary in deleted e-mails f_3.

$$\begin{cases} \text{ad}[i] = \frac{f_1 \times t_1}{t} + \frac{f_2 \times t_2}{t} + \frac{f_3 \times t_3}{t} & (t \neq 0) \quad (171.5) \\ \text{ad}[i] = 0 & (t = 0) \quad (171.6) \end{cases}$$

③ Get the weighted average ad of each word in the text

$$\text{ad} = (\text{ad}[1] + \text{ad}[2] + \cdots + \text{ad}[n])/n \quad (171.7)$$

④ Get the weights $d_{22} = \text{ad}$.

(3) To get e-mail weights d_2 according to the theme and text

$$\begin{cases} d_2 = (d_{21} + d_{22})/2 & (d_{21} \neq 0) \quad (171.8) \\ d_2 = d_{22} & (d_{21} = 0) \quad (171.9) \end{cases}$$

(4) To classify e-mail according to d_1 and d_2

$$\begin{cases} d = d_1/2 + d_2/2 & (d_1 \neq -1) \quad (171.10) \\ d = d_2 & (d_1 = -1) \quad (171.11) \end{cases}$$

If $0 < d \leq 1$, the e-mail is the Deleted e-mail. If $1 < d \leq 2$, the e-mail is the Less important email. If $2 < d \leq 3$,the e-mail is the Important e-mail.

D. The Work After e-mail Classification

(1) Glossary index storehouse
In order to better classify e-mail according to e-mail theme, establish three glossary index storehouses: theme glossary index storehouse of important e-mail; theme glossary index storehouse of less important e-mail; theme glossary index storehouse of deleted e-mail. Theme glossary index storehouse is used to make the statistics for receiving the theme vocabulary mail messages. Establish three text glossary index storehouses: text glossary index storehouse of important e-mail; text glossary index storehouse of less important e-mail; text glossary index storehouse of deleted e-mail.

(a) Glossary index storehouse of important e-mail
The vocabulary is stored in the storehouse according to the frequency of e-mail's appearance ranking. Each word has different weights in the storehouse. If there are n words, the weight of word "i" is $d[i] = 3 - i/n$. The weight is 0 if the word is not in the storehouse

(b) Glossary index storehouse of less important e-mail
The vocabulary is stored in the storehouse according to the frequency of e-mail's appearance ranking. Each word has different weights d in the storehouse. If there are m words, the weight of word "j" is $d[j] = 2 - j/m$. The weight is 0 if the word is not in the storehouse.

(c) Glossary index storehouse of deleted e-mail
Because for those deleted e-mails, the higher frequency of occurrence of words, the better the attributes of the deleted e-mail can be expressed and the smaller its weight should be. Therefore, the vocabulary of deleted e-mail is stored in the storehouse according to the frequency ranking. Each word has different weights d. If there are l words in the storehouse, the weight of the k word is $d[k] = 1 - k/l$. The weight is 0 if the word is not in the storehouse

(2) The work after e-mail classification

(a) Important e-mail
Extract the vocabulary of theme, and update important e-mail the theme index storehouse. Extract the vocabulary of theme, and update the theme index storehouse of important e-mail. Extract the former n words of high frequency in the text, and update the text index storehouse of important e-mail. Download the e-mail from the Internet to important mail folder. If the e-mail address originally belongs to the unfamiliar e-mail address set, e-mail addresses of senders will be moved to the important e-mail address sets and the sender's e-mail will be moved to the unfamiliar mail folder.

(b) Less important e-mail

Extract the vocabulary of theme, and update the theme index storehouse of less important e-mail. Extract the former n words of high frequency in the text, and update the text index storehouse of less important e-mail. Download the e-mail from the Internet to less important mail folder. If the e-mail address originally belongs to the unfamiliar e-mail address set, e-mail addresses of senders will be moved to the less important e-mail address sets and move the sender's e-mail to the unfamiliar mail folder.

(c) Strange e-mail

Put the e-mail address of senders in the strange e-mail address set. Download the e-mail from the Internet to strange mail folder.

(d) Deleted e-mail

Add the e-mail address to the address set of the deleted address. Extract the vocabulary of the theme, and update it and update the glossary index storehouse of deleted e-mail theme. Extract the former n words in the text, and update the glossary index storehouse of deleted e-mail text. Delete the e-mail from the Internet. If the e-mail address originally belongs to the unfamiliar e-mail address set, then delete e-mail addresses from unfamiliar addresses sets, and delete the e-mail address from unfamiliar mail folder.

171.5 Testing system

A. System Testing

(1) Establishing initial knowledgebase

First, add the Address of the initial important e-mail, less important e-mail, and deleted mail to database. Input five important e-mails, the less important e-mails, deleted e-mails to get the knowledgebase. Add some vocabulary of deleted e-mail.

(2) Test e-mail

Send an e-mail from user2006@sina.com.cn. The address is not in the known storehouse. The theme is: greet Dongbing. The text is: How are you? I've safely arrived.

The statistical result of the first letter: a strange e-mail.

The statistical result of the second letter: an important e-mail.

B. Testing Conclusions

Through the test, we verify the feasibility of the algorithm to achieve the purpose of an intelligent mail classification. The correct rate of e-mail classification needs be tested further in the future.

171.6 Conclusions

Through the study of e-mail classification Intelligent Agent System, we made a new method of intelligent classified e-mail—ILDS. Testing system certified the correctness of the algorithm, and the system has some practical value. In this paper, there are still many shortcomings that need constant improving. The choice of initial e-mail is not enough, and the representation is not strong. Users should adjust the e-mails classified. The e-mail classification Intelligent Agent System makes users learn to operate and update the corresponding information.

Acknowledgment This chapter is supported by (1)Project of Research on Science and Technology of Jilin Education Ministry of China under Grant No.2007-172 and 2010-383 (2) Science-technology Development Project of Jilin Province of China under Grant No. 20115056. (3) the master and doctor launched project of Jilin Normal University.(4) the Natural Science Foundation of Changchun Normal University.

References

1. Yixiong W, Hao W, Hongfei G (2006) A mobile agent communication and coordination model. Comput Eng 32(4):122–127
2. Xianfu M, Shumin H, Gongping T (2006) Distributed data service model based on the mobile agent. Comput Eng 32(1):273–275
3. Feras AO, Nabil B, Juan AC, Prabhat M (2010) Differential evolution for learning the classification method PROAFTN. Knowl-Based Syst 23:418–426
4. Chang PC, Fan CY, Dzan WY (2010) A CBR-based fuzzy decision tree approach for database classification. Expert Syst Appl 37:214–225

Chapter 172
Study on the Development of Domestic and International Internet Medicine Market

Feng Chen, Bin Dun, Ying Lu and Jian Zhou

Abstract With the popularity of the Internet, e-commerce market has a positive effect in promoting the development of the pharmaceutical medicine distribution system in China. By studying the Internet medicine market situation and the development trend in the United States, Europe, Canada and Japan, this paper summarizes the development of Chinese e-commerce features, makes comparative analysis of the situation in China, then presents useful recommendations of the pharmaceutical market measures. Internet medicine market in developed countries in terms of online pharmacy and information quality control and supervision are rather perfect, which is what China should learn from. The successful experience of developed countries' e-commerce of pharmacy will be a profound enlightenment to the Internet medicine market in China.

Keywords Internet medicine market · Medicine e-commerce · Online pharmacies · Situation at home and abroad

This work is supported by the Key Projects in the National Science & Technology. Pillar Program during the Eleventh Five-year Plan Period under Grant No. 2009BAH48B01.

F. Chen · B. Dun · Y. Lu
Information Center of State Food and Medicine Administration,
Beijing, China

J. Zhou (✉)
Beijing ITOWNET Cyber Technology Ltd, Beijing, China
e-mail: zhoujian@itownet.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_172,

172.1 Introduction

The rapid development of the Internet, whether for using the Internet to obtain information about medicine, or the medicine trade are no longer strangers, which provides development opportunities for pharmaceutical e-commerce [1]. Online information is everywhere, all-inclusive, including information about medicines in this particular article. Medical e-commerce as a modern commodity circulation model, becomes a set of information, automation, standardization as a whole and greatly enhances the efficiency of medicine distribution, while also greatly reducing the flow of medicines in circulation level and the trading process in order to reduce the flow of medicines costs and distribution costs, thereby reducing medicine prices. For the pharmaceutical industry it is of great significance. At present, China's development of pharmaceutical e-commerce is still very slow. Compared with China, foreign countries have carried out pharmaceutical services on the Internet earlier, from whose experience China can learn, and not only accelerate the development of China's pharmaceutical e-commerce, but also develop China's pharmaceutical industry.

172.2 Paper Preparation

Foreign pharmaceutical development of electronic commerce has the following three characteristics: Developed modern pharmaceutical logistics system; Specialized market intermediary system more complete; Attaches great importance to the whole process of medicine distribution security.

172.2.1 Status of U.S. Internet Medicine Market

The full sense of the medical e-commerce development is accompanied by the rise of the Internet. In the 1990s, the Internet first gained popularity in the United States. Openness of the Internet and low access costs made the pharmaceutical enterprises carry out a more effective way of e-commerce, and so it made pharmaceutical companies, wholesale businesses, retail outlets for a wide range of data exchange and information integration possible. After 10 years of development, with deepening of the U.S.'s improved pharmaceutical e-commerce, there came the formation of a business to business (B2B), online retail (B2C), third-party e-medicine (public trading platform) and other forms coexist [2]. Buying medicines has become the sixth main reason for personal surfing on the Internet [3]. The Internet has brought to American life dramatic changes and through the Internet after finding medical information, people can better participate in the process of treatment.

Drugstore.com, one of America's largest online medicine sales, provides customers with products from pharmaceuticals to fitness equipment to books and magazines [4]. To help customers choose their own non-prescription medicines, Drugstore.com not only gives consumers the information about pharmaceuticals in addition to price, but also the use of medicines, usage, side effects, warnings, cross-reactivity, storage and measures to be taken when overdose or missing a dose. In recent years, the rise of centralized medicine procurement organizations (Group Purchasing Organization, GPOs) and welfare organizations to purchase medicines (Pharmacy Benefit Management, PBMs) and other for-profit service agencies in U.S., provide medicine purchase of specialized intermediaries services for medical institutions and health insurance companies.

172.2.2 Status of the Internet Medicine Market in Europe

In many European countries, the control of online pharmacies is more stringent. According to the information from the Council of Europe, Switzerland, Norway, Spain, Portugal, Austria, Italy, Finland and other countries do not allow sale of medical products through the network. In European countries, Sweden has the fastest growing Internet medicine service, its capital Stockholm is known as the Internet capital of Europe. The online sale of medicines in Sweden is allowed; all pharmacies in Sweden are state-owned by the company Apoteket AB operation, which does not allow any other person to sell medicines to the public. The French Government allows online medicine transaction, but strictly prohibits online sales of prescription medicines. Germany approves online medicine transaction, but enterprises must have the relevant qualification requirements and delivery should be within 48 h.

In Germany, whether general or online pharmacy pharmacies, practitioners must have a pharmacist accreditation, and added to the German Association of Pharmacists under the Pharmaceutical Association. Almost 90% of pharmacies under the Pharmacists Association carry out online booking services, so that patients can buy medicines from home [5]. The German government has very strict rules on the Internet sale of medicines, whether online or offline where only the pharmacies can operate medicines. The German government will release on the B2B and B2C loose harsh to B2C e-commerce [6]. The Royal Society has just issued a series of online pharmacy industry standards. The government has also released a white paper designed to support e-commerce, and actively support the development of electronic commerce. It is clear that European countries will produce polarization in e-commerce: medicine e-commerce in some countries may be developed, such as the Netherlands, Denmark, Sweden, Switzerland and the United Kingdom. It may also have Poland in Central Europe and Belgium. Less developed countries in medicine e-commerce may include Germany, France, Italy, Spain, Portugal and Austria [7].

172.2.3 Canadian Internet Medicine Market Status

For Canadians, the Internet has become their main source of access to knowledge [8]. Since the invention of the Internet 10 years ago, it has changed the daily lives of Canadians, has changed the way Canadians access to health information and their relationship with the doctor [9]. In addition, over the past decade, the provision of health and nutrition information through commercial web sites undertook intense pressure from the global media, the public's growing desire for access to health information and from stakeholders such as doctors and pharmaceutical information providers division of the strong pressure, led to the emergence of new media companies in North America specializing in health and nutrition information. They are often involved in providing support to health services subsidiary, the apparent provision of health care practitioners or company information and the provision of health information with the media is a new trend.

172.2.4 Status of the Internet Medicine Market in Japan

In Japan, being one of the more developed countries in Asia, the use of the Internet by doctors and patients is very popular. Medical information on the Internet not only makes the people understand the information about the disease, but also helps in improving the quality of life of patients and their families. The Japan Internet Medical Union study found in January 2000 that 34% of patients or family use the online medical information, 34% of 1021 medical institutions use the Internet web site to publish job information or treatment. In addition, about 50% of physicians used the Internet, 65% of which medical information through the network to query and another 56% of access to relevant information [10]. However, the quality of information online is not reliable, sometimes in advertising, and sometimes the publisher information is fake, or content of information is distorted and so on. The lack of an effective medical information system to evaluate and ensure that the site provides users information which is accurate, resulted in a site providing hydrocyanic acid, a chemical substance, to people who tried to commit suicide. The network brings another very serious problem, the disclosure of private information, such as the risk of leaking of personal information of sexually transmitted diseases and AIDS patients, making patients suffer from various social pressures.

172.3 Status of the Internet Medicine Market in China

At present, only more than a dozen retail businesses have "Internet medicine information service qualification certificate" and "Internet and medicine transaction services credentials". Because the vast majority of Chinese Internet users are young, and this group precisely lacks the demand for purchase of medicines, they are not the main consumer groups, which is the elderly. But they have less time for surfing on the Internet. They usually do not get online pharmacies' promotional information through an online pharmacy to buy medicines.

According to the Internet medicine trade, the majority of medicine sales have been from hospitals and pharmacies in China at present and it is a small share of online pharmacies. The industry is concerned that the online pharmacy market is about 10 million RMB per year, while actual sales are far below this figure. China has also set up the necessary online pharmacies chain and distribution, etc. where more than 50 qualified, from the present stage, China's online pharmacies can only say that just started.

Although China's pharmaceutical e-commerce in recent years has seen rapid development, there are still some problems, mainly in the following areas: there are some flaws in medical e-commerce laws and regulations. China's pharmaceutical industry and the low level of development of information varies: the true meaning of e-commerce is a medical logistics, information flow and capital flow as a medicine distribution model, which requires participation in e-commerce firms with high information technology level, so as to ensure the establishment of different bodies of medical e-commerce direct, effective and electronic communication. However, due to the short time of the construction of China's pharmaceutical industry information, many domestic pharmaceutical companies only partially implement the information, it is difficult to achieve a unified management of the computer. While some large pharmaceutical companies, such as Tong RenTang, Guangzhou Pharmaceutical, have implemented enterprise resource planning (ERP), the computerization of most enterprises is still very low, they lack the combined talents of inadequate infrastructure, business, medical and network technology.

172.4 Analysis on the Development of the Internet Medicine Market in Developed Countries

172.4.1 Regulatory Mechanism

For the online pharmacy, the industry protection agencies such as U.S. National Association of Boards of Pharmacy (NABP) issues the Verified Internet Pharmacy Practices Sites (VIPPS) certification, recognizes its pharmaceutical business networks business qualification, and then the certified pharmacy certification mark can be marked on the pharmacy's web page. In addition, U.S. federal agencies and the coordination of various state governments will both regulate online pharmacies. Federal agencies are taking a series of measures to prevent the illegal sale of medicines, the U.S. Food and Drug Administration takes the implementation of effective consumer protection measures as a starting point, rather than making too many restrictions on legitimate online pharmacies. FDA makes efforts to promote consumer medicine e-commerce trust. The United States government which attaches great importance to the protection of consumers, specifically formulated the "Internet Pharmacy Consumer Protection Act" to ensure that consumer interests are not violated. Internet sales of each medicine is required to present "health care membership organizations" and proof of insurance companies and guarantee certificate and other party licensing operations.

China adopts a more cautious approach to its online pharmacies. To ensure the quality of online pharmacies, they should obtain certain qualifications which are very strict. Because the threshold is too high, few pharmacies can get through, in a sense, this is one of the reasons that illegal online pharmacies operate more often. The U.S. industry and government together make the decentralized management model worth learning. Industry accreditation protection agencies, given their quality certification to identify and carry out random checks at any time, assisted by government departments, should be the starting point for consumer protection while not imposing too many restrictions on online pharmacies. This requires the development of the pharmaceutical market to the Internet to foster a favorable legal environment. Therefore, there must be a sound legal system. A sound legal system has two aspects, first, it is necessary to improve the existing legal system and make necessary adjustments, and second, to meet the need for the development of new laws and regulations, related to customs, taxation, payment, copyright, patent, network encryption, consumer rights and interests.

172.4.2 Establish Compensation and Multi-Joint Supervision Mechanism

In the U.S., online pharmacies that sell medicines are required to produce "health care membership organizations" and proof of insurance companies and guarantee certificate before operating. In the event of such medicines for sale on online pharmacies having damaged the population situation, people can get relief or compensation. Under German law, consumers can buy medicines from legitimate online pharmacies whose medical insurance should be registered in Germany or in the European Union. The government asks the public to go online to buy medicines with genuine health insurance companies or consumer association consultation online pharmacies, and online pharmacy practitioners to distinguish carefully the qualifications and the quality of government-issued certification marks, so that they can be compromised when claiming for insurance company compensations.

China can learn from this experience. The insurance company makes compensation when injury and supervision occur. It is necessary to establish China's injury compensation system which should be in the event of injury compensation for victims, to first ensure the interests of the people to maximize compensation.

As a result of injury when the cost of compensation borne by the insurance companies, online pharmacies are normal, whether the quality of clearance of the sale of medicines, medicine information is correct and the insurance company to have a close relationship, which will undoubtedly pull out its ranks of supervisors come.

172.4.3 Promote the Standardization Administration of the Pharmaceutical Industry

The State Food and Drug Administration (SFDA) implemented "action plans to improve national medicine standards", a standard two-step upgrading. First, upgrade the local standards to national standards. Second, national standards to be upgraded to the reunification of the next level. It plans to conduct for 3–5 years a comprehensive clean-up of medicine standards. This means that following the GMP certification, the "standard" theme of a new round of knockout competition in the domestic medicine began, improving production standards, pharmaceutical enterprises, certainly to higher production costs and management costs. However, brand companies are considered to improve the market access threshold, the policy of rectifying the market opportunity. A large state-owned Chinese enterprise leader points out that in the past because the production of traditional Chinese medicine or proprietary standards were not strict, companies were free to change their formulas to distinguish it from other products.

172.4.4 Establish Mechanisms for Medicine Quality and Safety Information Query

To help the public better understand medical information, the U.S. Food and Drug Administration established a medicine safety web site information site CDER, where consumers can check the site of the relevant pharmaceutical products, the latest news, including the package insert, and medicine safety risk-related information, research data after the listing of medicines and medicine-related clinical trials data, various forms of medicine safety and medicine-related information as well as regulations and guidance documents.

The Canadian government also attaches great importance to the development of a network for the pharmaceutical industry. This section focuses on the provision of medicines and other health product safety information, to consumers and health professionals to report adverse events more quickly. In China's medical information network, consumers cannot be easily checked. This network should establish a dedicated web site to facilitate consumers to check whether their own medical information obtained is correct and check the latest information on pharmaceutical products to help consumers make better treatment decisions. A total of 900 pharmacists in Germany share a telephone hotline, as long as a chemist's phone call, day or night can get back. China can follow these pages for the establishment of a national medicine information pharmacist line, so that consumers can consult at any time to avoid medicine injury.

172.4.5 Establish the Networking Medical Associations to Help Ordinary Consumers

EU has already established an international project-Medical Attention Alliance whose purpose is to use the network more securely. It uses high technology for online information and health-related certification, to establish the distinction between reassuring signs. Numerable quality certification is done by a trusted third party. The project also enables consumers to identify the Internet as an expert on health information. Medcertain makes the establishment of such a fully functional demonstration system to use a third party in the future so that consumers and experts are able to avoid harmful information to make correct identification and selection of high-quality information.

Medcertain itself does not make a hierarchy for health information, but a systematic and technical standard system allows individuals and organizations to assess the quality of information online. Some media companies in Canada specialize in health and nutrition information, where the apparent provision of health care practitioners in the combined information is a good trend and meets the public's growing desire for access to health information. In addition, mitigation of the front-line information providers such as doctors and pharmacists who could release the strong pressure. It can be beneficial for China to establish such alliances.

172.5 Conclusion

At present, the development of the Internet medicine market in the major developed countries in Europe and America is in full swing. Medical e-commerce development in China started late, with slow development, the reasons being notably the following: China's medical e-commerce laws and regulations have some defects; pharmaceutical industry information in China is low and uneven levels of development; China has not yet formed a national logistics system; China has not yet formed the habit of electronic payments. By comparing the situation of the Internet medicine market in Europe with other developed countries, we can correctly understand the shortcomings that China has in the development of the Internet pharmaceutical service. Learning from the experience of other countries could take China to the future course of development for a fast, stable, healthy development.

References

1. Chen Y, Shen W, Yan Bo H et al (2006) China's medical status and prospects of e-commerce development [J]. J Clin 9(2):127–128
2. Meng L, Liu Z, Shi B et al (2009) The American medical e-commerce development and its enlightenment to China [J]. China Pharm 17(7):551–553
3. Koong KS, Koong LY, Liu LC et al (2005) Examination of selected medicine availability at online pharmacies [J]. Int J Electron Healthc 1(3):292
4. Geng F, Bian Y (2008) Medical e-commerce B to C model of development and China's pharmaceutical retail chain enterprises to develop e-commerce strategy study [J]. China Pharm 4
5. Meng L, Lian GY, Zhou Y (2007) The development of China can learn from German experience online pharmacy [J]. Chinese Med 16(16):11–12
6. Zhao L (2005) Strict German pharmaceutical retail [J]. Med World 2005(7):60–61
7. Cheng L, Meng R, Wang L, Zhang G, Du S (2010) Network pharmacies will go in [J]. Clin Med 10
8. Yang S (2007) Canada: heavily subsidized the development of national internet culture [J]. Netw Commun 6:79
9. Underhill C, McKeown L (2008) Getting a second opinion: health information and the internet [J]. Health Rep 19(1):1–6
10. Haruyuki T (2001) Internet medical usage in Japan: current situation and issues [EB/OL]. http://www.jmir.org/2001/1/e12

Chapter 173
Evolving Model of P2P Content Distribution Network Based on the Prediction of User Requirements

Huang Yongsheng, Du Huamei and Zhang Genglu

Abstract With the increasing scale of the Peer-to-Peer (P2P) network, the contradictions among the scope expanding, the expenditure of time is reducing and the network loading of content search is growing more serious. In this chapter, an evolving model of P2P content distribution network (CDN) based on the prediction of user requirements is proposed in which the organization of the P2P CDN is evolved continuously according to the improvement of the prediction of user requirements. In the model, the peers and the surrogates that have the requiring contents of a designate peer will be assembled at nearby locations to the designated peer, and then more replicas can be found in smaller search scope and shorter time expending. Simulative experiments indicate the evolving model can acquire much smaller mean response time and mean distribution time.

Keywords Evolving model · P2P · Content distribution network · Requirement similarity · Clustering

173.1 Introduction

Content (Viz. file) sharing is one of the widest used web services of the Internet. As content sharing technology, the P2P CDN (Peer-to-Peer Content distribution Network) achieves quick development in recent years [1–4]. In P2P CDN, a peer will send search requests to a surrogate and other peers when it requires a file.

H. Yongsheng (✉) · D. Huamei · Z. Genglu
Tangshan Key Laboratory of Informationization
Technologies and Engineering Control, School of Management,
Hebei United University, Tangshan 063009, China
e-mail: hyosheng@sohu.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_173,

If the surrogate has not searched the file, the surrogate should search and achieve the file from the center server or other surrogates, which will prolong the response time and the distribution time. Due to the constraints of search scope, a great many peers with replicas of the searching file cannot be found. Consequently, making each peer find the requiring files in the surrogate connecting to it and acquire more replicas of the requiring files from other peers with smaller hops as far as possible by properly organizing the peers and the surrogates becomes a research spot in P2P CDN [5–7].

For the sake of reducing the spending of content search, an algorithm is proposed in [8] which guarantees that the requirement will be found with bounded costs by combining with semantic similarity of the queries. Yoshikatsu [9] proposes a novel information delivery network architecture which built over the existing unstructured P2P network by applying percolation theory to diffuse newly defined reverse-query messages.

In order to enhance the performance of the P2P CDN system, plentiful research have been put forward which focused on the P2P CDN topology, such as routing of content search and distribution, repetition placement, loading equilibrium and requirement location, etc. In [10], a locality aware P2P-based content-distribution network, Flower-CDN, has been developed. In the model, peers keep the content they retrieve and later serve it to other peers that are close to them in locality. When a new client requests some content from a website, a locality-aware DHT quickly finds a peer in its neighborhood that has the content available.

Wauters [11] proposes a set of distributed replica placement algorithms (RPAs), based on an Integer Linear Programming (ILP) formulation of the centralized content placement problem. These algorithms further enhance the CDN performance by optimizing the network and server load, reducing network delays and avoiding congestion. In the chapter [12], an architecture was designed which consists of idle ISP servers that can be rented and released dynamically as the load requires. The research looks at providing a globally accessible storage architecture where all content broadcast over a period of time is available for streaming.

The article [13] shows a content delivery network based on grouping surrogates in which a protocol based on the proximity of surrogates has been developed to connect surrogates from the same group and from different groups in order to provide lower content-distribution times. The chapter [14] proposes a topology building protocol to benefit from the small world feature. Based on the group locality of scientific collaboration, an enhanced group clustering cache replacement scheme has been designed.

The traditional organization of P2P CDN is passively conformed to the requirement of file requests and distribution. In this chapter, an evolving model of P2P content-distribution network (CDN) based on the prediction of user requirements is proposed in which the organization of the P2P CDN is evolved continuously according to the improvement of the prediction of user requirements. In the model, the peers and the surrogates that have the requiring contents of a designate peer will be assembled at nearby locations to the designated peer, and then more replicas can be found in smaller search scope and shorter time expending.

173.2 Analysis of User Requirement of P2P

In P2P CDN, if a peer has the file or files that another peer requires, it is deemed that both peers have requirement to the file or files. Although one has acquired the file or files and the other has not, their requirement to the file of files is same or similar. If peers with same or similar file requirement are adjacent to each other, they can help to each other in file distribution because the files some peers will need are more probable than the files that other peers have. Therefore, acquiring requirement similarity of users is significant in the organization of P2P CDN.

In order to analyze the user requirement, the sets $C = \{c_1, c_2, \ldots c_n\}$ and $P = \{p_1, p_2, \ldots p_m\}$ are proposed to denote the content set and the peer set, respectively. $r_{i,j}$ $(0 \le r_{i,j} \le k, 1 \le i \le n, 1 \le j \le m)$ denotes the requirement degree of the peer p_i to the content c_j and the value of $r_{i,j}$ is larger means the requirement degree of the peer $p_i$to the content $c_j$is larger. The requirement similarity between two peers can be calculated as Eq. 173.1.

$$\mathrm{sim}(p_a, p_b) = \frac{\sum_{c \in C_{ab}} r_{a,c}\, r_{b,c}}{\sqrt{\sum_{c \in C_{ab}} (r_{a,c})^2} \sqrt{\sum_{c \in C_{ab}} (r_{b,c})^2}} \tag{173.1}$$

$\mathrm{sim}(p_a, p_b)$ denotes the requirement similarity of p_a and p_b, C_{ab} is the intersection of the sets C_a and C_b.

In reality, the requirements of users are usually complex. For two users, if the requirements of them are similar at a category of contents, their requirements to another category may be different. In other words, the categories of contents should be taken into account in the requirement similarity calculation between users. $T = \{t_1, t_2, \ldots t_s\}$ denotes the set of content categories, and $t_i \cap t_j \ne \varphi$ (φ means not null) is permitted. The requirement similarity based on the content category can be expressed as Eq. 173.2.

$$\mathrm{sim}(p_a, p_b, t_x) = \frac{\sum_{c \in C_{ab,t_x}} r_{a,i} r_{b,c}}{\sqrt{\sum_{c \in C_{ab,t_x}} (r_{a,c})^2} \sqrt{\sum_{c \in C_{ab,t_x}} (r_{b,c})^2}} \tag{173.2}$$

$\mathrm{sim}(p_a, p_b, t_x)$ describes the requirement similarity between p_a *middot*; and p_b at the category $t_x \cdot C_{ab,t_x} = \{c |\, c \in C_{ab}\ and\, c \in Ct_x\}$is the subset of C_{ab}, and Ct_x is the content set which consists of the contents belonging to the category t_x, $\mathrm{sim}(p_a, p_b)$ is called the partial similarity and $\mathrm{sim}(p_a, p_b, t_x)$ is called the entire similarity correspondingly.

For each shown requirement $r_{i,j}$, it is achieving from the activity that the corresponding peer acts to the corresponding file. The number of shown requirements of a peer is increasing accompanying with the number of activities that the peer takes action on files. Additionally, the requirement of a user is changeable. It is not necessary that what a user required previously is what the user requires now or will require in the future. Accordingly, the requirement similarity between users will be changeable as time goes by.

In order to express the effectiveness of shown requirements for a given period of time, the time zone set $TZ = \{tz_1, tz_2, \ldots tz_h, \ldots\}$is defined in which the time zones are successive by order number. Taking the time zone into accounting, the entire similarity and the partial similarity can be expressed as Eqs. 173.3 and 173.4.

$$\mathrm{sim}(p_a, p_b)_{tz_h} = \frac{\sum_{c \in C_{ab}} r_{a,c} f_{tz_h}(tz(r_{a,c})) r_{b,c} f_{tz_h}(tz(r_{b,c}))}{\sqrt{\sum_{c \in C_{ab}} (r_{a,c} f_{tz_h}(tz(r_{a,c})))^2} \sqrt{\sum_{c \in C_{ab}} (r_{b,c} f_{tz_h}(tz(r_{b,c})))^2}} \quad (173.3)$$

$tz(r_{a,c})$ denotes the time zone when the shown requirement $r_{a,c}$ is achieved. If $i < h, f_{tz_h}(tz_i)$ denotes the efficiency weight of the shown requirements in the time zone tz_i to tz_h.

$$\mathrm{sim}(p_a, p_b, t_x)_{tz_h} = \frac{\sum_{c \in C_{ab}, t_x} r_{a,c} f_{tz_h}(tz(r_{a,c})) r_{b,c} f_{tz_h}(tz(r_{b,c}))}{\sqrt{\sum_{c \in C_{ab,t_x}} (r_{a,c} f_{tz_h}(tz(r_{a,c})))^2} \sqrt{\sum_{c \in C_{ab,t_x}} (r_{b,c} f_{tz_h}(tz(r_{b,c})))^2}} \quad (173.4)$$

173.3 Evolving Model of P2P Content-Distribution Network

In P2P CDN, a peer usually has more than one neighbor which is used to transfer the request and response of the file search. The neighbor relationship between two peers is symmetric. In other words, for two peers, if one peer is the neighbor of the other, the other peer is the neighbor of the peer too. The symmetric neighbor method has two weaknesses. Firstly, it will add the burden when peers enter and exit. For two peers with neighbor relationship, one peer should adjust itself to conform to the other peer's enter or exit. Secondly, the symmetric neighbor method is a constraint for peers to select proper neighbors.

173.3.1 The Configuration of CDN Surrogate

In P2P CDN, every peer will connect with a surrogate. As a peer require a file, it sends search requests to other peer, and at the same time send file request to the surrogate it connects to. If the surrogate does not have the request file, it will achieve the file from the CDN central server or other surrogate which will prolong the time cost of the response to the request and constrain the service capacity of CDN when a great many peers send file requests simultaneously. Therefore, to make every surrogate have the files which the peers connecting to the surrogate require as far as possible can enhance the performance of the P2P CDN.

The peers in a P2P subset have larger similarity with each other can assure the peers finding more file replicas they require, and the file scope the surrogate connecting to the peers of the P2P subset should store become smaller. The clustering method is incited in P2P subset organization. A peer is taken as an element and a P2P subset is a clustered cluster of peers. Meanwhile, the entire similarity and partial similarity between two peers are taken as the distance of the two peers correspondingly. Let $\mathrm{sub}(P)_i$ denote a cluster (viz. a P2P subnet), the average distance between peers of a cluster $ad(\mathrm{sub}(P)_i)$and the distance between two cluster $d(\mathrm{sub}(P)_i, \mathrm{sub}(P)_j)$ is significant factor in clustering. $ad(\mathrm{sub}(P)_i)$and $d(\mathrm{sub}(P)_i, \mathrm{sub}(P)_j)$ can be calculated as Eqs. 173.4 and 173.5 respectively.

$$ad(\mathrm{sub}(P)_i) = \frac{2 \times \sum\limits_{p_x \in \mathrm{sub}(P)_i} \sum\limits_{p_y \in \mathrm{sub}(P)_i} p_{xy}}{\left|\mathrm{sub}(P)_i\right| \times \left(\left|\mathrm{sub}(P)_i\right| - 1\right)} \tag{173.5}$$

p_{xy} denotes the entire similarity or the partial similarity between p_x and p_y in a certain time zone. $\left|\mathrm{sub}(P)_i\right|$ denotes the total count of peers in the cluster $d(\mathrm{sub}(P)_i, \mathrm{sub}(P)_j)$.

$$d(\mathrm{sub}(P)_i, \mathrm{sub}(P)_j) = \frac{\sum_{p_x \in \mathrm{sub}(P)_i} \sum_{p_y \in \mathrm{sub}(P)_j} p_{xy}}{\left|\mathrm{sub}(P)_i\right| \times \left|\mathrm{sub}(P)_j\right|} \tag{173.6}$$

173.3.2 The File Storage Strategy of the CDN Surrogate

If a P2P subnet is connected to a surrogate, the surrogates should store the files which the peers of the P2P subnet are most probable to require. By shown requirement of the peers of the P2P subnet, the requirement degree to a file can be predicted as follows.

$$\mathrm{dpredicton}(c) = \sum_{p_i \in \mathrm{sub}(P)} \frac{\sum_{p_j \in P(c)} r_{j,c} \times \mathrm{sim}(p_i, p_j)_{tz_h}}{\sum_{p_j \in P(c)} \mathrm{sim}(p_i, p_j)_{tz_h}} \tag{173.7}$$

$P(c)$denotes the peer set in which every peer has shown their requirement to the file $c.\mathrm{sub}(P)$ is clustered by entire similarity. Accordingly, if $sub(P)$ is clustered by partial similarity, the requirement degree to a file can be predicted as Eq. 173.8.

$$\mathrm{dpredicton}(c) = \sum_{p_i \in sub(P)} \frac{\sum_{p_j \in P(c)} r_{j,c} \times sim(p_i, p_j, t_x)_{tz_h}}{\sum_{p_j \in P(c)} sim(p_i, p_j, t_x)_{tz_h}} \tag{173.8}$$

The file c belongs to the file category t_x. The files which $\mathrm{sub}(P)$ requires constitute a file set$C(sub(P)).C(sub(P))$ can be expressed as $C(sub(P)) = \{c|$ $\mathrm{dpredicton}(c) \geq d_{\min}\}.d_{\min}$ is a value of the requirement degree. When $d_{\min}$

increases, the element count of $C(\mathrm{sub}(P))$ grows larger and the surrogate which the peers of $\mathrm{sub}(P)$ connect to should store more files. A surrogate can connect to more than one P2P subnet. If the P2P subsets connect to a surrogate includes$\mathrm{sub}(P)_1, \mathrm{sub}(P)_2, \ldots \mathrm{sub}(P)_k$, the files the surrogate should store are$\bigcup_{i=1}^{k} C(\mathrm{sub}(P)_i)$.

173.4 Simulative Experiments and Experimental Analysis

The data for experiments comes from MovieLens data set which comprises 100,000 rating scores to 1681 films by 943 users. The value of rating scores range from 1 to 5. To a value of rating score rated by a user to a film, the value is higher means that the user is more interested to the film. Thirty thousand rating scores of 600 users are abstracted from the data set. The abstracted rating scores are arrange by decreasing time order, and the first 5,000 rating score are taken as activities of the initial time zone. When the experiments are proceeding, the reminder rating scores are input by order.

The abstracted rating scores are divided into seven successive time zones ($tz_0 - tz_6$). For the evolving model, tz_0 is the initial time zone in which the first 5,000 rating scores are used to initialize the P2P CDN system. When the value of a rating score which a user rates a film is larger than 3 (Viz. it is 4 or 5), the user is more interested in the film and the rating activity is taken as a request for a film file too in the experiments. A film file will be distributed to the peer (corresponding to a user) after a request to the file is sent. After a rating score larger than 3 has been handled, a film file will be stored to a corresponding peer. From the time zone tz_1 to the time zone tz_6, each rating score is handled by time order as a shown requirement and the rating scores larger than 3 are taken as a request for a film file meanwhile.

Figure 173.1 expresses the variation of mean response time (*MRT*) in the experiments based on the traditional P2P CDN model, the P2P CDN model without evolution and the evolving P2P CDN model, respectively. Contrasting to traditional P2P CDN model, the *MRT* of the P2P CDN model without evolution and the evolving P2P CDN model is much smaller. Figure 173.1 indicates the organization of P2P CDN according to the requirement similarity can make it more probable that peers can find their requiring file in their adjacent peers. The MRT of evolving model is smaller than the P2P CDN model without evolution means the evolving model can more effectively describe the requirement similarity between peers.

Figure 173.2 expresses the variation of mean distribution time (*MDT*) of the three experiments, respectively. Comparing to the traditional P2P CDN model, the

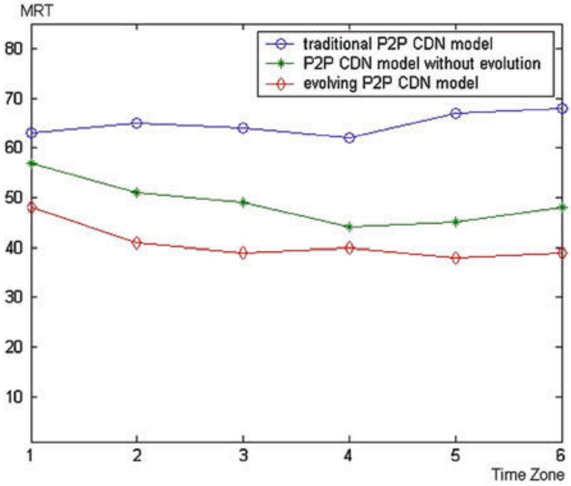


Fig. 173.1 The variation of MRT

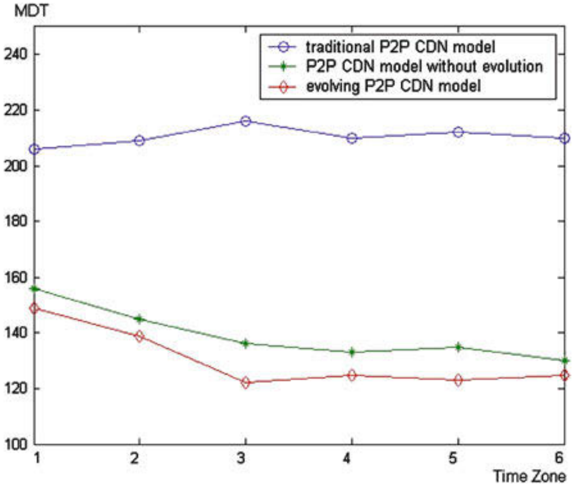


Fig. 173.2 The variation of MDT

evolving P2P CDN model and the P2P CDN model without evolution can acquire much smaller MDT, which means the neighbor selection strategy and the file storage strategy of the CDN surrogate based on the requirement similarity can enhance the distribution performance by quickly finding more replicas of the requiring files from the surrogates and other peers. As the evolving model can more effectively describe the requirement similarity between peers, it can achieve smaller MDT in contrast with the P2P CDN model without evolution.

173.5 Conclusions

The user requirement of P2P CDN can be shown by the activities of the peers. In this paper, context aware technology is applied to achieve the shown requirements and their evolution with time going, and then the requirement similarity between peers is analyzed. Utilizing the entire similarity and partial similarity, the neighbor selection strategy is proposed to organize the peers. By clustering peers with similar requirement, it is more probable for a surrogate to have the files the peers connecting to it require. The organization of the P2P CDN is evolving continuously with varying of the user requirement. Simulative experiments indicate the evolving model can acquire enhanced performance.

References

1. Xuening L, Yin H, Lin C (2009) A Novel and high-quality measurement study of commercial CDN-P2P live streaming [C]. In: Proceedings—2009 WRI international conference on communications and mobile computing, CMC 2009 Institute of Electrical and Electronics Engineers Computer Society, pp 325–329
2. Jiang H, Zhan W, Albert KW, Jun L, Zhongcheng L (2009) A replica placement algorithm for hybrid CDN—P2P architecture [C]. In: Proceedings of the international conference on parallel and distributed systems—ICPADS IEEE Computer Society, pp 758–763
3. Zhang R, Qian W, Zhou A, Zhou M (2009) An efficient peer-to-peer indexing tree structure for multidimensional data. Future Gener Comput Syst [J] 25(1):77–88
4. Giancarlo F, Mastroianni C, Pathan M, Vakali A (2009) Next generation content networks: trends and challenges [C]. In: Proceedings of the 4th Edition of the UPGRADE-CN Workshop on Use of P2P, GRID and Agents for the Development of Content Netw., UPGRADE-CN'09, Co-located Int. Symp. High Perform. Distrib. Comput. Conf., HPDC'09 Association for Computing Machinery, p 49
5. Changyou X, Chen M (2008) Impact of network topology on distance prediction accuracy [C]. In: Proceedings of 2008 IEEE international conference on networking, Sensing and Control, ICNSC Institute of Electrical and Electronics Engineers Computer Society, pp 1425–1429
6. Tian C, Xue L, Hongbo J, Wenyu L, Yi W (2008) Improving bitTorrent Traffic performance by exploiting geographic locality [C]. In: GLOBECOM—IEEE global telecommunications conference. Institute of Electrical and Electronics Engineers, pp 2489–2493
7. Giancarlo F, Mastroianni C (2008) Special section: Enhancing content networks with P2P, Grid and Agent technologies. Future Gener Comput Syst [J] 24(3):177–179
8. Shuang K, Fangchun Y, Sen S (2008) A scalable peer-to-peer overlay for semantic Web services discovery. Chinese J Electron [J] 17(2):361–366
9. Yoshikatsu F, Mori D, Saruwatari Y, Tsuda K (2008) Reverse-query diffusion over unstructured overlay network for content delivery. Int J Comput Appl Technol [J] 33(2–3):131–137
10. Manal ED, Pacitti E, Kemme B. Flower-CDN (2009) A hybrid P2P overlay for efficient query processing in CDN [C]. In: Proceedings of the 12th international conference on extending database technology: advances in database technology, EDBT'09 Association for Computing Machinery, pp 427–438
11. Wauters T, Coppens J, De Turck F, Dhoedt B, Demeester P (2006) Replica placement in ring based content delivery networks. Comput Commun [J] 29(16):3313–3326

12. Trunfio P, Talia D, Papadakis H, Fragopoulou P, Mordacchini M, Pennanen M, Popov K, Vlassov V, Haridi S (2007) Peer-to-peer resource discovery in grids: Models and systems. Future Gener Comput Syst [J] 23(7):864–878
13. Lin SH, Hu JY, Chou CF, Chang IC, Hung CC (2009) A novel social cluster-based P2P framework for integrating VANETs with the internet[C]. In: IEEE wireless communications and networking conference, WCNC Institute of Electrical and Electronics Engineers
14. Turkmen F, Mazzoleni P, Crispo B, Bertino E (2008) P-CDN: Extending access control capabilities of P2P systems to provide CDN services [C]. In: Proceedings—IEEE symposium on computers and communications Institute of Electrical and Electronics Engineers, pp 480–486

Chapter 174
Research on Multi-Agent Automated Negotiation and Dynamic Cooperation Model

Jiang Weijin, Zhong Luo and Wu Xing

Abstract The communication between agents has some special requirements. One of them is asynchronous communication. Communication sequence process is used to describe a model of agents' communication with the shared buffer channel. The essence of this model is very suitable for the multi-agents by multi-agent system communication, so it is the basis for our next step job. Based on the communication model, the distributed tasks dealing method among joint intention agents is explored and with the description of relation between tasks ,we give a figure of agents' organization. Agents communicate with each other in this kind of organization. The semantics of agent communication is another emphasis in this paper. With the detailed description of agents' communication process provided with a general agent automated negotiation protocol based on speech act theory in multi-agent system, we then use communication sequence process to verify this protocol whether it has properties of safety and live ness, in order to prove it is logically right. At last, a frame of this protocol's realization is provided.

Keywords Multi-agent · Automated negotiation · Joint intention · Safety · Dynamic cooperation

J. Weijin (✉) · Z. Luo
School of Computer Science and Technology,
Wuhan University of Technology, Wuhan, China

J. Weijin · W. Xing
School of Computer and Electronic Engineering,
Hunan University of Commerce, Changsha 410205, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_174,

174.1 Introduction

The theory of multi-agent automated negotiation involves extensive applying fields and many kinds of methods, mainly present in argument based automated negotiation [1], game theoretic models and heuristic approaches. In application, it can be divided into two categories [2], agent's negotiation within multi-agent system (MAS) and self-interests between different MAS [3–5]. The theories supporting the interior collaboration of MAS are self-interested, with joint intentions and shared plans and they have been working under the premise of identical intention and target of agent within MAS [6]. This text will discuss the joint intentions in multi-agent automated negotiation of MAS [7].

If multi-agent in MAS interact successfully, there must be three conditions that demand to be satisfied such as : communication structure, that is, how to dispatch and take over the information between agents [8]; communication language, that is, agent is required to understand the significance of the information; Interaction rules, that is, how to organize the conversation between agents [9].

In regard to the research of agent communication structure, we have proposed MAS communication model in the previous parts [10–12]. In the second section, it will be stressed to analyze agent's asynchronous communication mechanism. As for the research of agent communication language, presently there have been many abroad, like KQML, FIPA, ACL, agent talk, etc., hence, the language is not the emphasis in our text. Then, research of interaction rules is the second emphasis in the text. In the third part of the text, the agreement of agent automated negotiation and its validation will be set. In the fourth part, it illustrates and analyzes the complete frame of agent automated negotiations. The fifth is the conclusion of the text.

174.2 MAS Communication Mechanism

Agent is a status course which accomplishes the task automatically with the ability and agreement of communication, for example, P_A represents the course of agent A.

The course of agent makes the agent's ability which can be marked as $Ability\ _{P_A}$and $TASK\ _{P_A}$ means to be able to fulfill the task.

The moving status of the static agent in MAS can be classified as Active, Wait and Run. Agent in the Wait status will be activated after receiving the requests from the other agents and then run. Agent in Run status will negotiate with the other agent or provide services according to the try-best principle. Stateouter stands for the Run status of agent:

$$\mathrm{State}_{\mathrm{outer}} :: \ = \mathrm{Wait}|\mathrm{Active}|\mathrm{Run}$$

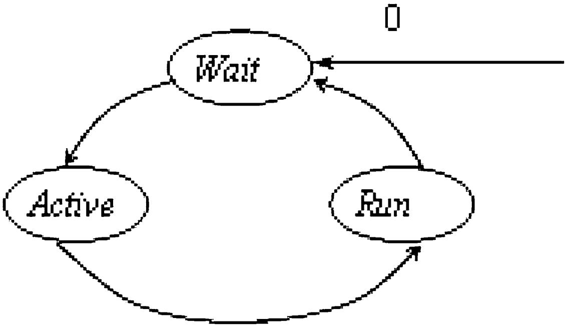


Fig. 174.1 Agent outer state transition figure sprung by events

Agent's collaborating course observed from the outer MAS is the process that Agent runs in the Iouter = Staeouter*.

In an agent's collaborating process with safety and liveness, the circulation of Wait → Active → Run → Wait in Iouter will appear at least once for the agent's launch and acceptance.

Obviously, in the circulation of Wait → Active → Run → Wait, if any one part of the agent cannot fulfill the circulation, it means something has happened unexpectedly causing the deadlock or livelock to the system during the collaborating process,and hence, the theorem will be attested.

Example 1 Agent *A* past passage *C* to transmit one thing dispense with responsive notify *m* to agent *B*. Agent *A* starting tenor is PA and agent *B* starting tenor is PB. The wholly cooperating process for: PA. (*c*) | PB (*c*), with CSP:

$$P_A = \text{Wait} \xrightarrow{o} \text{Active} \rightarrow \text{Outer?}\, x \rightarrow c!\,(m) \rightarrow \text{Wait}$$

$$P_B = \text{Wait} \xrightarrow{\bar{\bar{o}}} \text{Active} \rightarrow c?\,(m) \rightarrow \text{Wait}$$

In Fig. 174.1, outer stands for exterior entity relative to single agent, *o* is the triggered event of outer, *o* and $\bar{o}$ is the coupling event, the said cooperating process possesses activity and security.

More and more application systems ask both the corresponding sides of each other in a position to realize asynchronous communication mode. As a self-contained MAS communication structure, it is not only in a position to realize Agent's synchronous communication, but also able to realize asynchronous communication. Miner'sπ figuring has realized transfer calculations by communication passage, which shows that we can utilize agent's asynchronous communication mode to realize the synchronous. The asynchronous communication's ideal mode shows that both the corresponding sides have each,their own infinite buffer queue. However, it is unpractical to deploy such infinite buffer queue to

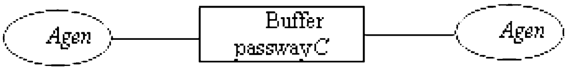


Fig. 174.2 Realize asynchronous communication between agents by using buffer passage

each agent,as sharing the buffer channel may realize agent's transfer between asynchronous communication and synchronous communication better.

Buffer channel C is such an agent which sets independent state switch and message buffer to all its relevant agents and transmits messages for these agents.

Example 2 In Example 1, a buffered passage C is utilized to realize communication process. This example can realize the asynchronous communication between agent A and agent B, and the whole collaborating process is: $P_A\ (C)\ \|\ P_C\ \|\ P_B\ (C)$, as showed below:

Utilizing buffer channel may realize manifold asynchronous communication modes. Introductions of transmit message m through buffer channel as below

$$P_A = \text{Wait} \xrightarrow{o} \text{Active} \rightarrow \text{Outer?}\ x \rightarrow C!\ (m) \rightarrow \text{Wait}$$

$$P_C = \text{Wait} \rightarrow C?\ (m) \rightarrow \text{if}(P_B.\ \text{State}_{\text{outer}} = \ \text{Wait})\ \text{then}\ C!\ (m) \\ \rightarrow \text{Wait else Wait}$$

$$P_B = \text{Wait}\ \xrightarrow{\bar{o}}\ \text{Active} \rightarrow C?\ (m) \cdot \text{Wait}$$

The above process shows that agent can realize asynchronous communication between agents by using buffer passage. P_C stands for buffer channel tenor.

The synchronous communication between agents asks them to be clear about each other's corresponding location. If a MAS system owning N (numerous) agents would like to realize point-to-point communication between Agents, there will be N^2 channels needed to be set up, of which so many would extremely complicate the realization of the agent. Using shared buffer channel can be helpful for realizing channel's transmission between agents (Fig. 174.2).

174.3 MAS Interior Cooperation Mode

When multi-agent in MAS begins cooperation, the reason that there is a conformed joint intention between agents, the process of multi-agent in MAS works according to the principal of "From each according to his ability, abide by the law and behave oneself", that is, each agent is trying its best to cooperate with the other agent.

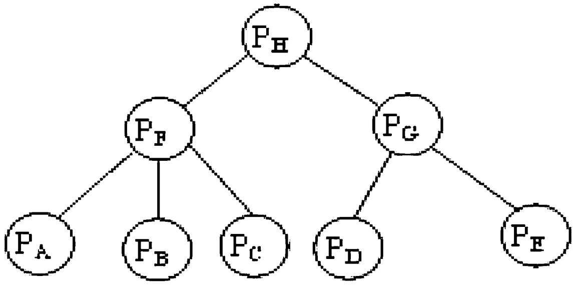


Fig. 174.3 Task tree

The cooperation between agents is aimed at fulfilling certain tasks. As these tasks can be divided into different but related sub tasks, the tasks from the agent's point of view can be described as following: a material task can be regarded as sub-tasks' assembling depending on different abilities of the agent in MAS. Combining divided-task-oriented agents in compliance with sub tasks will position to form a furcation tree of k ($k \geq 2$). Relation between sub tasks is in relation with or to time sequence. Agent's organizing relation is determined by the relation between tasks. Description of sub tasks is given below:

(1) The sequential relationship of the tasks (<), manifests that agent *B*'s task cannot be begin before fulfilling agent *A*'s task. Formalization is described below:

$$\mathrm{TASK}_{P_A} < \mathrm{TASK}_{P_B} |= P_A; P_B$$

Thereinto: TASK $_{P_A}$ and TASK$_{P_B}$ are the start-up tenor PA and PB of agent *A* and agent *B* respectively,which are used to fulfill tasks.

(2) The relation of "AND" between tasks (*V*), indicates that agent *A* and *B* perform simultaneously sub tasks PA and PB. After completing the sub tasks, agent *C* begins its common and subsequential task PC. Formalization is described below:

$$\mathrm{TASK}_{P_A} \lor \mathrm{TASK}_{P_B} |= (P_A \parallel P_B) < \mathrm{TASK}_{P_C}| = (P_A \parallel P_B) < P_C$$

(3) The relation of "OR" between tasks ($\wedge$), indicates that agent *A* and *B* with the relation of "OR" perform simultaneously sub tasks PA and PB, regardless of which one is fulfilled first, agent *C* can begin its subsequential task PC. Formalization is described below:

$$\mathrm{TASK}_{P_A} \wedge \mathrm{TASK}_{P_B} |= (P_A < \mathrm{TASK}_{P_C}) \parallel (P_B < \mathrm{TASK}_{P_C})| = (P_A < P_C) \parallel (P_B < P_C)$$

Example 3 TASKMAS means the task can be fulfilled by MAS, which is divided as the task tree seen in Fig. 174.3.

1) Relations between tasks: $(((P_A \vee P_B \vee P_C) < P_F) \vee ((P_D < P_G) \wedge (P_E < P_G)))\, P_H$
2) With CSP describing $\mathrm{TASK}_{\mathrm{MAS}}$ as follows:

$$\mathrm{TASK}_{\mathrm{MAS}} = (((P_A \parallel P_B \parallel P_C);\ P_F) \parallel (P_D;\ P_C \parallel P_E;\ P_G));\ P_H$$

From the above mentioned: MAS is a task processing distributive system. The agent's ability can be realized by its corresponding tenor. The relations between tasks in MAS have determined that the agent is organized according to its dendriform communication topology which is the precondition for agent's automatic negotiation.

Agent automatic negotiation is the main method for the multi-agent to negotiate, which focuses on three aspects present in negotiation protocol, negotiation object and negotiation policy. Negotiation protocol and negotiation object act as the textual points, but the negotiation policy is clamping on how to look for an Agent each from a negotiation space, best in order to reach consistence, concretion content visible in the literature cited.

Present hypotheses 1 to ensure negotiation agent could each other have partner faith in against due to MAS interior agent according to try-best principle proceed synergic, furthermore MAS possess concurrent combine intent.

Hypothesis 1 Negotiation agents know each other in negotiation policy.

The negotiation is present with the result that the decision of the agent towards internetwork communication negotiatory condition, and agent automatic negotiatory course mission due to specific assignment requires different communication quality guarantee with AND specific network insurance. Text take mission negotiation AND internetwork communication negotiation as agent automatism negotiation is in process in two phases.

MAS interior agent automatic negotiation course includes two phases. The first phase is based on multi-agent automatic negotiation whose negotiation object includes task starting time, task ending time and the relation of the tasks; The second phase is the negotiation of agent's communicating conditions whose negotiation objects include corresponding security policy and network service quality (QOS).

The negotiation procedure of agent *A* with *B* running tasks is given below:

A: Can you run the task T1? (request)
B: No, I cannot. Because I am running task T2 . (reject and state reasons)
A: Can you run the task T1 after completing T2? (conditional request)
B: Yes. (commitment combine confirmation)

In the above dialog, when agent *A* requests agent *B* to run task T1, agent *B* rejects and presents the rejective reasons. Because of the reasons presented by agent *B*, agent *A* puts forth a request again to agent *B*, and then agent *B* replies agent *B* with commitment.

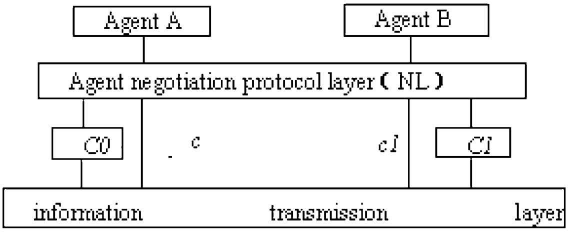

Agent A
Agent B
Agent negotiation protocol layer (NL)
C0
c
cl
Cl
information transmission layer

Fig. 174.4 Agent automatic negotiation protocol model

Agent A, B keep on the above dialog:

A: Can you keep executing task T_2 according constant rate? (conditional request)
B: I can uninterruptedly execute T_2 within 5% at bit rate variation range. (suggestion)
A: I accept your suggestion. (acceptance)

The conditional request set one recommends, at the suggest agent B makes amendment for agent A towards a bit rate claim, as the last agent A take on agent B, negotiation success finishes upon for the upper agent dialog, agent B towards agent A.

According to the top analysis talks about the correlative language with behavior academic theories, we say that the agent automatic negotiation correspondence in the procedure to state row words for certain : request, promise, refuse, advise and counter advise. In view of agreement ,presenting overtime event and agent unsolicited message transmission, so as to increase overtime (timeout) status and inform (inform) state row word. Communication protocol engine of the communication process state follows as of the agent:

$$\begin{aligned}&\mathrm{State_{inner}} ::= \mathrm{Started} \mid \mathrm{Requested} \mid \mathrm{Accepted} \mid \mathrm{Refused} \mid \mathrm{Promised} \mid \\ &\mathrm{Informed} \mid \mathrm{Advised} \mid \mathrm{CAd\text{ - }vised} \mid \mathrm{Timeout} \mid \mathrm{Stopped}\end{aligned}$$

See Fig. 174.4: agent automatic negotiation protocol can be divided into information transmission layer, buffer channel layer and agent negotiation protocol layer from bottom to top, of which buffer channel layer C is one of the needed layers between agents to realize asynchronous communication. If it can realize point-to-point synchronous communication between agents, it can communicate directly through channel C. As to the description of agent automatic negotiation, it mostly focuses on agent negotiation protocol layer, while for the other layers, it only describes their services and running environment in brief. In essence, the function of agent negotiation protocol layer is the description of process.

The agent A description of whole negotiation procedural with agent B, not formal, is as follows: agent A first of all dispatch negotiation beg of agent B received solicit aback, toward request message proceed analyzes, could as per three strain scene dispose to: the first thing, in the event of agent B receivability the solicit of agent A, those agent B to agent A dispatch take send, else dispatch thumb advise, down through upon, the service request block mode, of the such negotiation scene as conventional *C*/*S*. the second thing, provide some agent *B* can provide serve of instruct, but because of the restrict of the resource of system can't very much the serve, so the agent *B* can put forward to agent *A* the serve promises, the agent *A* handles agent *B* the commitment of serve can proceed very much: reject or accept. The third thing, the agent *B* thinks after analyzing the agent *A* request agent *A* some items modification within request empress, can satisfy the agent. A request still, like this agent *B* after proceeding agent *A* some items within request to modification, conduct and actions the suggestion sends out to the agent *A*. Agent *A* for the suggestion of agent *B* can operation proceeding as follows: accept, reject and put forward the counterproposal. Either that of toward agent *A* counterproposal, agent *B* receivability, reject or set own the other one proposal for.

Utilizing process algebra to carry out formalization of communication protocol not only can state logic structure and time sequence nature of the protocol narrowly, but also is favorable to verify the protocol. The nature of the protocol includes liveness and safety. In the protocol system of liveness, its process algebra expression must own the recursion characteristic from initial state to the passing. If protocol stops executing at a certain event and is unable to go on, the system will be deadlocked. If protocol executes some certain events circularly and infinitely but is unable to return to initial state, the system will be livelocked. The system without a deadlock or a livelock will be safe.

Consider protocol $\mathrm{JIAANP} = |||\mathrm{Pi} = (\mathrm{P1}|||\mathrm{P2}|||, \cdots, |||\mathrm{Pn}) = |||(\mathrm{PS}||\mathrm{PR})$, will not carry communication directly between *Pi* and *Pj* $(\mathrm{i} \neq \mathrm{j})$,, and can think that they are separate, namely, can store in and lock or live and lock. So we may prove whether there is deadlock or a livelock appearence between promoter process P_S and acceptor process P_R of negotiation.

Considering three kinds of different conversation scenes, Q1 in agreement JIAANP, Q2, and Q3, among them, the simple message is sent and received in the execution course of Q1 and Q2, not forming circulation in the state changed picture of the agreement, so they will not be formed and extremely locked or lived the lock. There is a proposal and counter proposal circulation in the execution course of Q3, it carries out course and may be formed and locked and lived and locked very much with complexity , because transmit for the agreement overtime proves the difficulty brought for the agreement, so we suppose: the transmission of the network is reliable, the feedback between agent is in time, namely, logic exactness of the agreement that the prerequisite without incident in overtime comes down to prove the agreement.

174.4 Conclusions

This text provides a common and communication-based agent cooperation mode by studying mutual behavior of agent cooperation. The text also uses some effective format ways to depict automatic negotiation protocol of agent process and verify the validity of the protocol's logic. Finally, the text makes an implementation frame for this agreement. While using blackboard mode to realize buffer channel in this implementation frame, it provides an extra deployed agreement stack and at last, presents performance and expandable analyses. In addition to the negotiation between agent in MAS, because the advantage difference of agent group negotiating with agent which has a conform joint intention has great differences on negotiation principle and strategy, the self-interested agent's negotiation agreement between MAS will be our next work for research.

Acknowledgments This work is supported by the Natural Science Foundation of Hunan Province of China No. 10JJ5064, the Science and Technology Plan Project of China under of Hunan Grant No.2009GK2002, and the Technology Innovation Foundation under Science and Technology Ministry Grant No. 09C26214301947.

References

1. Jennings NR, Faratin P, Lomuscio AR et al (2000) Automated negotiation: prospects. Methods and challenges. Pacific Rim international conference on artificial intelligence
2. Grosz B, Sidner C (1990) Plans for discourse[A]. In: Cohen P, Morgan J, Pollack M (eds) Intentions in communication. Bradford Books, MIT Press, Cambridge
3. Rosado IJ, Bernal-Agustin JL (1994) Genetic algorithms in multistage distribution network planning. IEEE Trans Power Sys 9(4):1927–1933
4. Kinny D, Ljungberg AM, Rao E, Tidhar SG, Werner E (1992) Planned team activity. In: 4th European workshop on modeling autonomous agents in a multi-agent world (MAA-MAW)
5. Dillenbourg John A (1992) Self: a computational approach to distributed cognition. Eur J Psychol Educ 7(4):252–373
6. Bredin J, Kotz D, Rus D, Maheswaran RT, Imer C, Basar T (2003) Computational markets to regulate mobile-agent systems. Auton Agents Multi-Agent Sys 6(3):235–263
7. Liu J, Han J, Tang YY (2002) Multi-agent oriented constraint satisfaction. Artif Intell 136(1):101–144
8. Sierra C. Faratin P, Jennings NR (1997) A service-oriented negotiation model between autonomous agents. In: Proceedings of the 8th European workshop on modelling autonomous agents in a multi-agent world (MAAMAW-97), Ronneby, Sweden, pp 17–35
9. Hoare C (1985) RA communicating sequential processes. Prentice Hall International, New Jersey
10. Jiang W, Zhang L, Wang P (2009) Research on grid resource scheduling algorithm based on mas cooperative bidding game. Sci China F 52(8):1302–1320
11. Jiang WJ (2008) Research on distributed solution strategy of complex system based on MAS. J Inf Comput Sci 5:2063–2069
12. Jiang WJ, Wang P (2006) Research on distributed solution and correspond consequence of complex system based on MAS. J Comput Res Develop 43(9):1615–1623

Chapter 175
Ubiquitous Hierarchical Generalized-Sensor Network: Architecture and Application

Zhitong Huang, Jupeng Ding and Yuefeng Ji

Abstract A ubiquitous hierarchical generalized-sensor network (UHGSN) is presented as a typical network model for the future unified communication environment among human society, computer network and the real physical world. Two kinds of novel network elements, the sensor information processing unit and the hierarchical distributed agent server are introduced in this architecture for the effective information communication. The topology and addressing problems of the UHGSN architecture are analyzed in details, and the combined "key word" based characterized searching mechanism is discussed as the basic application in UHGSN, along with the corresponding protocol message definition. Simulation results show the advancements of the presented hierarchical architecture and the job-list based characterized searching mechanism.

Keywords Ubiquitous hierarchical generalized-sensor network (UHGSN) · WSN · Characterized searching · RFID

Z. Huang (✉) · J. Ding · Y. Ji
Key Laboratory of Information Photonics and
Optical Communications (BUPT), Ministry of Education,
Beijing University of Posts and Telecommunications,
Beijing 100876, The People's Republic of China
e-mail: hzt@bupt.edu.cn

J. Ding
e-mail: jupeng7778@163.com

Y. Ji
e-mail: jyf@bupt.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_175,

175.1 Introduction

Due to the rapid developments of mobile communication and the Internet, people can communicate with each other freely regardless of the location or time. However, in this daily lives, people need to get information not only from others but also from the external physical world, so the current human-to-human communication is not satisfying, but a unified communication environment among human society, computer network and the external physical world is required. This objective becomes increasingly realizable since the recent advances in wireless sensor network (WSN) and radio frequency identifier (RFID) have made it possible to acquire huge amount of information from the external physical world. It has given rise to many novel concepts, such as internet of things [1, 2], ubiquitous networks [3], which are concentrating on how to connect the external physical world with the computer network for intelligent control and management.

Current researches on these subjects are mainly focusing on the mechanisms of smart node design, energy saving and event representation. There are still many questions that should be addressed for the complete implementation of such unified communication environment in terms of network architecture, addressing and routing and information aggregation [3–5].

The basic application in such unified communication network is to provide huge amount of valuable information which may be used in economic analysis, academic research and our daily lives. Currently, the key word based context searching is the fundamental technology for information acquisition in the Internet. However, with the growing demands on various digital services, information will be represented not only by text, but increasingly by audio, image, video and even multi-media, which makes the traditional searching mechanisms inadequate. The optimal solution lies on the research of novel information searching technology based on combined "key word" on text, audio, image and video, which is called characterized searching here.

In this paper, a ubiquitous hierarchical generalized-sensor network (UHGSN) is presented as a typical network model for the future unified communication environment. The function of UHGSN for future unified communication is described. The network elements, topology, addressing and functional planes in this architecture are analyzed for its implementation. The characterized searching procedure in the UHGSN is then specifically discussed as the basic application in UHGSN, along with the searching message definition. A UHGSN simulation platform is constructed for performance comparison, and the simulation results show the advantages of the presented hierarchical architecture and the job-list based characterized searching mechanism.

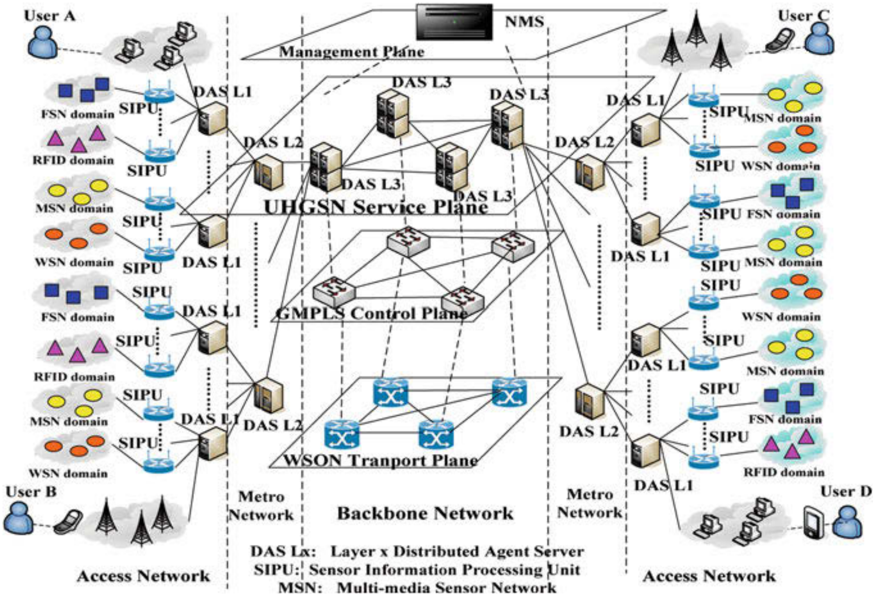


Fig. 175.1 The UHGSN architecture for unified communication

175.2 UHGSN Architecture

Recent advances in sensor technology integrate traditional sensing techniques with the wireless communication to construct different kinds of WSNs. The FSN has also received much attention nowadays as a wired sensor technology due to its sensitivity on pressure and temperature. Besides, the RFID network may be treated as a wireless passive sensor network [6] which is mainly used for object tracking and transportation. Moreover, the camera monitoring network that widely deployed in the traffic and security area can also be considered as a multi-media sensor network (MSN), since it captures and processes data in the forms of image and video. The integration of these sensor networks with current communication network infrastructure builds a ubiquitous generalized-sensor network which can be used to realize the communication among human, computer network and the real physical world.

In such network, for efficient management, not only the sensors should be divided into different domains according to the category and position, but also the captured data and the corresponding information should be managed in different layers based on the location and network scale. Therefore a hierarchical framework UHGSN is constructed, whose architecture is described in Fig. 175.1.

The UHGSN is composed of two kinds of network elements, the sensor information processing unit (SIPU) and the distributed agent server (DAS). Because of the limits on lifetime and node size, the original data captured in each sensor domain are not recommended to be operated by the sensors themselves but by one SIPU which collects the data, translates them into valuable information, and accomplishes further operations on such information including searching and analysis. With the large-scale deployment of WSNs, FSNs, MSNs and other types of generalized-sensor networks, the amount of original data and the corresponding information will be quite huge, so it is important to design proper architecture for efficient information communication and management. We introduce the conventional client/server based network structure in UHGSN. Multi-layer DASs are deployed for aggregating the information, handling the requests from the human users and transferring the corresponding information back to the users. The network scale, along with the locations and numbers of SIPUs, decides the layer number for DASs required in the UHGSN. Figure 175.1 presents a typical model of a simple three-layer UHGSN where layer 1(L1)–layer 3(L3) DAS is responsible for the information communication in access, metro and backbone network, respectively. The users of UHGSN are preferred (but not required) to send their requests to the nearest L1 DAS through the Internet, as people are always most interested in the things around themselves. If the users try to get some information that far from themselves, the layer 2 (L2) or even the L3 DAS may be involved in the communication. The information in the UHGSN is being converged and aggregated, so the SIPUs are connected to the bottom layer L1 DASs in tree topology, so as to the connections between different layers of DASs. All of the DASs at the top layer connect in a mesh network for information exchange.

The internet protocol (IP) based network addresses (i.e., IPv4/IPv6) should be allocated to the SIPUs and DASs to realize the communication between UHGSN and the Internet. Such IP address only identifies who the host is without defining where the host is, however, the services in UHGSN is primarily based on the physical world location information of the SIPUs, so each SIPU needs another physical world localization address (e.g., GPS address) for user application. A network address mapping list (NAML) should be created which records and maintains the relationship between these two types of addresses. The user of UHGSN only need to provide the physical world address in the request to the DAS, and the DAS will use the NAML stored in its database to translate it into the IP address for the following message transfer.

The UHGSN is only the service plane of UHGSN and a network management system (NMS) which realizes the configuration, performance, fault and safety management is required as the management plane. To realize the long-haul communication with certain quality of service (QoS) requirements, both the generalized multi-protocol label switching control plane [5] and the optical fiber network transport plane are recommended to be deployed, whose functions are as same as those in current backbone networks such as rapid routing and provisioning, flexible protection and restoration and dynamic resource allocation. Figure 175.1 shows these four functional planes in the UHGSN architecture.

175.3 Characterized Searching in UHGSN

The basic application for the users of UHGSN is to search out valuable information from the huge amount of data captured by the sensors. Duo to the growing demands on various digital services, along with the large-scale deployments of MSNs, information in the UHGSN will be represented not only by text, but increasingly by audio, image, video and multi-media. Therefore it is necessary to develop novel characterized searching mechanisms and algorithms based on different kinds of "key word". The searching requirements from the users can be quite different, leading to different realization procedures. Here we define a standard format of user searching request message in UHGSN.

<UHGSN user searching request message>::=
<Hierarchical physical world localization address>
<Characterized searching "key word">
<User QoS requirements>

The *<Hierarchical physical world localization address>* object provides the real physical world localization addresses for the associated SIPUs involved in this request, which will be then translated to the network addresses using the NAML in DAS. This object also indicates that which layer of the DAS should handle this request and whether it is a uni-cast, multicast or broadcast request. The *<Characterized searching "key word">* object shows the type, length, and value of the searching "key word". Such information will then transfer to the SIPUs to finish the characterized searching algorithms. The *<User QoS requirements>* object presents the user QoS requirements, such as the operation delay, protection and restoration, the results format, etc. For example, if a user requires all the real-time information the MSN captured in a certain position, he may generate a searching request with the "minimal operation delay" requirements. Figure 175.2 shows the implementation of characterized searching for such a standard searching request in the UHGSN.

In practical applications, the DAS may receive repetitive searching requests from different users when something popular happens in the real world. To avoid the unnecessary repeated searching operations in SIPUs, the DAS should create a job-list which records the latest searching requests and the corresponding results it processed. When a new request arrives, the DAS searches it in its job-list. If it is a reduplicate request, the DAS will directly send the associated results to the users. This job-list based searching mechanism will enhance the searching efficiency and reduce the searching overhead in the repetitive requests situation. The length of the job-list may influence the searching efficiency, so it should be set according to the network scale, user number and the location of the DAS.

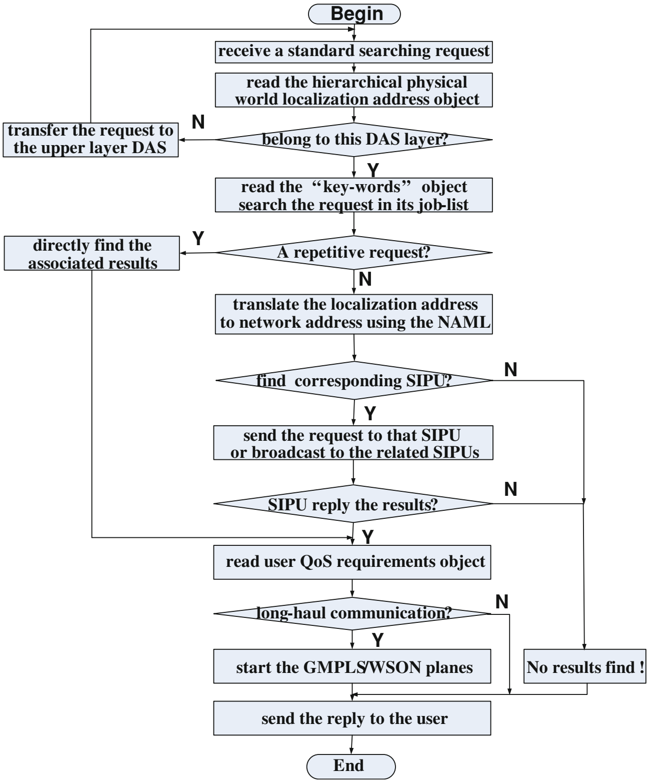

Begin
receive a standard searching request
read the hierarchical physical world localization address object
belong to this DAS layer?
N
transfer the request to the upper layer DAS
Y
read the "key-words" object search the request in its job-list
A repetitive request?
Y
directly find the associated results
N
translate the localization address to network address using the NAML
find corresponding SIPU?
N
Y
send the request to that SIPU or broadcast to the related SIPUs
SIPU reply the results?
N
Y
read user QoS requirements object
long-haul communication?
N
Y
start the GMPLS/WSON planes
No results find !
send the reply to the user
End

Fig. 175.2 The implementation of characterized searching procedure in UHGSN

175.4 Simulation Results and Performance Analysis

To evaluate the performance of the hierarchical DASs in the UHGSN architecture and the job-list based characterized searching mechanisms, we construct a UHGSN simulation environment which contains one NMS, 20 user clients, 5 DAS L3 nodes, 20 DAS L2 nodes, 60 DAS L1 nodes, and 120 SIPU nodes. The communication among the clients, the different layers of DASs, the SIPUs and the NMS are based on our user-defined user datagram protocol (UDP) messages. The simulation environment is shown in Fig. 175.3.

The first simulation is to analyze the performance of the hierarchical DASs in the UHGSN architecture. We compare the characterized searching time and the searching overhead in four kinds of network structures. In the first structure, no DAS is deployed, and both the clients and SIPUs are directly connected to the NMS. The NMS is responsible for the characterized searching operations. The five DAS L3 nodes are added in the second structure, and similarly, the 20 DAS L2 and 60 DAS L1 nodes are added in the third and fourth network structures respectively

Fig. 175.3 The UHGSN simulation environment

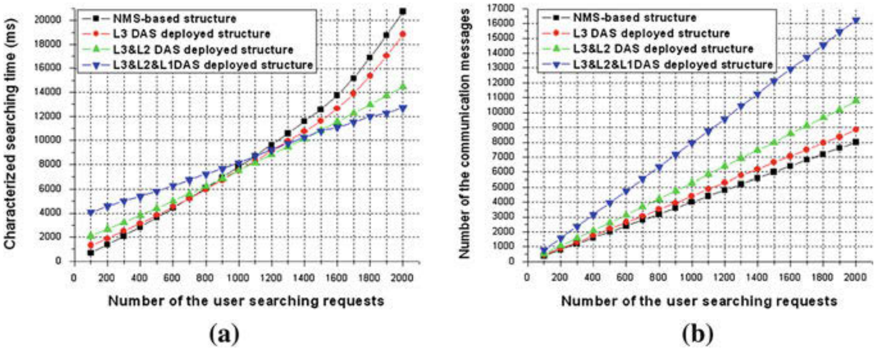


Fig. 175.4 Comparison on the characterized searching time and searching overhead among different UHGSN structures

(Fig. 175.1). In the simulation, the clients send searching requests randomly to the 120 SIPUs. The searching time is measured as the time from the instant that the first client sends out its first searching request to the instant that the last client receives the last searching result message. The searching overhead is measured by the number of the communication messages required throughout the network in the whole searching procedure. Figure 175.4 shows that the NMS-based centralized structure performs better when there are few searching requests, but the hierarchical DAS based distributed structures become increasingly efficient when there are more requests arrive. The one-, two-, and three-layer hierarchical DAS deployed structure performs faster than the NMS-based structure when the number of the searching requests reaches 700, 800 and 1,100, respectively. When the number of the requests become large enough, the three-layer DASs deployed

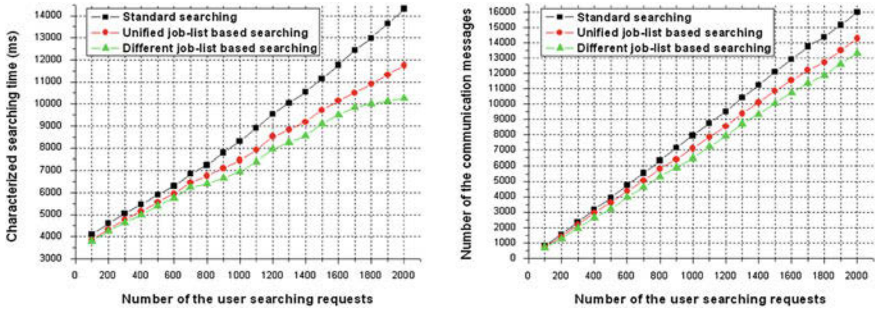


Fig. 175.5 Comparison on the searching time and searching overhead among different searching mechanisms in the repetitive requests situation

structure has the best performance. For example, when there are 1,900 requests, the NMS-based structure needs 18,750 ms to finish the searching operation, but the one-, two-, and three-layer hierarchical DAS deployed structure only needs 17,030, 13,840 and 12,260 ms respectively. On the other hand, the searching overhead in the hierarchical structure is larger than the centralized structure since it needs some more communication messages among different layers of DASs.

The second simulation is to analyze the performance of the presented job-list based searching mechanism in the repetitive requests situation among three kinds of mechanisms in the three-layer UHGSN. In the first mechanism, no job-list is created and one searching procedure should be generated for each searching request. In the second mechanism, DASs in different layers are deployed with the unified job-list which maintains the searching requests and the corresponding results it has processed in the latest 3,000 ms. In the third mechanism, DASs in L1, L2, and L3 are deployed with different job-lists which maintain the searching requests and the corresponding results in the latest 3,000, 5,000, and 8,000 ms, respectively. In the simulation, the clients send searching requests to the 120 SIPUs controlled by pseudo-random code generation software in NMS to create the repetitive requests cases. Figure 175.5 shows the results that the different job-list based mechanism performs best among these three mechanisms as it needs the shortest searching time and generates least searching overhead. For example, when there are 1,800 searching requests in the network (only averagely 8% are the repetitive requests), the different job-list based mechanism is 3,000 ms faster than the standard mechanism, and needs less than 2,500 communication messages.

175.5 Conclusion

In this paper, the UHGSN architecture is presented as a typical network model for the future unified communication environment. Two kinds of network elements, the SIPU and the hierarchical DAS are introduced in this architecture for the effective information communication. The combined “key word” based characterized searching mechanism is presented as the basic application in UHGSN, along with the corresponding protocol message definition for implementation. Simulation results show the advancements of the presented hierarchical architecture and the job-list based characterized searching mechanism. More applications and the related communication protocols are our future work.

Acknowledgments This research was supported by NSFC (No. 60932004), the National 973 Program (No. 2011CB302702), National 863 Program (No. 2009AA01A345), and Fundamental Research Funds for the Central Universities (BUPT2009RC0126), P. R. China.

References

1. Zorzi M, Gluhak A, Lange S, Bassi A (2010) From today’s intranet of things to a future internet of things: a wireless- and mobility-related view. IEEE Mag Wirel Commun 17:44–51
2. Pujolle G (2006) An autonomic-oriented architecture for the internet of things. IEEE JVA (06): 163–168
3. Saito H, Kagami O, Umehira M (2008) Wide area ubiquitous network: the network operator’s view of a sensor network. IEEE Commun Mag 46(12):112–120
4. Cho K, Hwang I, Kang S (2008) HiCon: a hierarchical context monitoring and composition framework for next-generation context- aware services. IEEE Netw Mag 22(4):34–42
5. Huang Z, Ji Y, Lu Y (2008) Sensor-based performance monitoring mechanism in GMPLS-controlled networks. IEEE Commun Lett 12(4):325–327
6. Akan O, Isik M, Baykal B (2009) Wireless passive sensor networks. IEEE Commun Mag 47(8):92–99

Chapter 176
Research of E-commerce Security Strategies Based on Cloud Computing Platform

Huang Hanyan

Abstract Cloud computing is a general term for anything that involves delivering hosted services over the Internet. Cloud computing was inspired by the cloud symbol that is often used to represent the Internet in flowcharts and diagrams. Electronic commerce, commonly known as e-commerce, consists of the buying and selling of products or services over electronic systems such as the Internet and other computer networks. The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage.

Keywords Data security · E-commerce · Cloud computing · Public key infrastructure

176.1 Introduction

Cloud computing is a general term for anything that involves delivering hosted services over the Internet. These services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). The name cloud computing was inspired by the cloud symbol that is often used to represent the Internet in flowcharts and diagrams [1]. Electronic commerce, commonly known as e-commerce, consists of the buying and selling of products or services over electronic systems such as the Internet and other computer networks. The amount of trade conducted electronically has grown extraordinarily with widespread Internet usage. The use of commerce is conducted

H. Hanyan (✉)
School of Information Management,
Jiangxi University of Finance and Economics,
Nanchang, China
e-mail: hyhuang2011@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_176,

in this way, spurring and drawing on innovations in electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems and automated data collection systems.

Cloud computing are parallel computing, distributed computing and grid computing development of a new business model. In cloud computing, users can use mobile phones, computers and other devices through the network to obtain the necessary hardware, software and other resources, sharing resources, convenient and fast access to the services they need. Cloud computing with its high reliability, versatility, scalability, low-cost advantage, quickly accepted by the majority of users [2]. Cloud computing is for the entire IT sector, especially e-commerce businesses to bring a new change.

Cloud computing can help companies quickly build an e-commerce platform, on-demand purchase and use, thus greatly reducing the cost of e-commerce website building and maintenance costs. Cloud computing requirements for lower business equipment, e-commerce companies can take advantage of existing enterprise computer and network equipment, cost very little money and you can enjoy the cloud computing services. In addition, cloud computing service providers provide professional teams with help in e-business software and hardware system maintenance. This business can focus more on primary business. Cloud computing can help companies quickly and easily anywhere in daily business activities. Users can use mobile phones and other devices through the network equipment at any time and any place with queries in goods, payment and other commercial activities. Employees can even take home to complete the transaction tasks. Cloud computing can help companies share information resources. Various e-commerce businesses can take advantage of cloud computing which provides a powerful product to achieve interoperability of information resources sharing.

Cloud computing platform performs dynamically according to the needs of e-business expansion, the number of e-commerce products and other information sharing, with a single e-commerce sites to reduce construction time and money. Cloud computing allows corporate data to maximize security. In cloud computing with virtualization, fault-tolerant multiple copies of data protection technologies exist such as cloud computing services with high reliability. Cloud computing service providers adopt e-commerce enterprise information such as the unified management of goods, load balancing, real-time monitoring tools to enable e-business enterprise data to maximize protection. Significant innovations in virtualization and distributed computing, as well as improved access to high-speed Internet and a weak economy, have accelerated interest in cloud computing.

176.2 E-Commerce and Security

E-commerce usually refers to a wide range of businesses around the world. The Internet is an open network environment, based on browser/server applications ways, where both buyers and sellers can meet on various business activities, to achieve online shopping consumer, online transactions between merchants and online electronic payment and a variety of business activities, trading activities, financial activities and activities related to integrated services [3]. E-commerce is the electronic and electronic technology means to business, as the core, original traditional sales, shopping channels, move onto the Internet, to break the national and regional visible and invisible barriers to globalization of production companies. It is a network of invisible, personalization integration. By way of electronic means to the main business activities, within the scope permitted by law, business activities are carried out by the process. E-commerce is the use of digital information technology for the activities of the enterprise continuous optimization process. A wide range of e-commerce in general can be divided into business to business, or business to consumer.

Platform-as-a-service in the cloud is defined as a set of software and product development tools hosted on the provider's infrastructure. Developers create applications on the provider's platform over the Internet. PaaS providers may use APIs, website portals or gateway software installed on the customer's computer. Force.com and Google Apps are examples of PaaS. Developers need to know that currently, there are no standards for interoperability or data portability in the cloud. Some providers will not allow software created by their customers to be moved off the provider's platform. As cloud computing model based on a large number of e-business information is stored in the cloud systems, transmission and processing, if there are problems, and its risks than traditional e-commerce model to be much higher. At present, security has become a business e-commerce implementation of cloud computing model based on the most important issue. Although the cloud computing model, data can be unified security management, and reliable real-time monitoring of e-business business information to get the maximum protection, but due to the complexity of the cloud itself, the user dynamics and other factors, so that e-commerce companies e-commerce platform built many new security problems.

As the cloud computing model based on e-commerce platform for all of the data is stored in the "cloud", enterprises are worried that during natural disasters when hardware failures and other facilities were damaged, and when they occur, cloud computing providers have appropriate measures to protect e-commerce normal operation of the platform. E-commerce businesses the biggest concern is that cloud computing security of data. In the cloud, most of the business information is stored in the "cloud", e-commerce companies will be unable to supervise the business sensitive information. As cloud computing uses virtualization technology, e-commerce companies using cloud computing services are not clear about where to store their own data, and do not even know in which country the data is located. Cloud computing environments and data are shared by multiple users, so, e-commerce companies often worry about whether their own data and other users of the data creating confusion. Electronic commerce that is conducted between businesses and consumers, on the other hand, is referred to as business-to-consumer or B2C. This is the type of electronic commerce conducted by companies such as Amazon.com. Online shopping is a form of electronic commerce where the buyer is directly online to the seller's computer usually via the internet. There is no intermediary service. The sale and purchase transaction is completed electronically and interactively in real-time such as Amazon.com for new books.

As cloud computing and e-commerce is something new, the cloud computing and e-commerce related law is not perfect [4]. Cloud computing service providers in the service agreement avoid most of the risk as much as possible, not promised to any data leaks and the data is liable for destruction or obligations. Hackers can take advantage of emerging technologies such as virtualization, virtual machine in the form of write for the spread of malicious software that allows users to be more difficult to detect and remove. At the same time, use is made of leased virtual machine to hide his true identity, making it difficult to be tracked. These data security e-commerce businesses bring great harm. Therefore, we must adopt effective measures to protect them. E-commerce businesses worry about cloud services in the event of termination of the service provider or another company, etc., your business and commercial data will be affected, how to take back their own data, and now the system is compatible with previous systems such as a series of questions.

176.3 Cloud Computing and Security Policy

Cloud computing refers to the delivery of IT infrastructure and usage patterns, to on-demand through the network, and a scalable way to obtain the necessary resources. Cloud computing refers to the mode of service delivery and use, to the network on-demand and an easy way to obtain the necessary expansion of services. This service can be IT and software, Internet-related, or any other services. It has a very large scale, virtualization, and reliable security and other unique effects. It aims at relatively low cost of the network for the calculation of multiple entities into one powerful computing capability with the perfect system, and use of SaaS, PaaS, IaaS, MSP business models and other advanced computing power of this powerful distribution to end users hands. The core idea of cloud computing is to a large number of computing resources with a network connection unified management and scheduling, to form a pool of computing resources on demand services to users. The basic principles of cloud computing is distributed computing by making a large number of distributed computers, rather than the local computer or remote server, enterprise data centers to run more like the Internet. This allows companies to switch to the required resources to applications, on demand access to the computer and storage systems.

With the rapid development of computer technology, information networks have become an important guarantee for social development. There is a lot of sensitive information, even a state secret. So inevitably it attract all kinds of people and attacks from around the world. Network security refers to the network hardware. Software and system data is protected, not because of accidental or malicious reasons, but due to damage suffered, change, disclosure, continuous and reliable system to run properly, the network service is not interrupted. Network security is its essence on the network information security. Broadly speaking, any information relating to the network confidentiality, integrity, availability, authenticity and control are related to technical and theoretical research on network security. System security and performance and functionality are a contradiction in the relationship. If a system does not provide any services to the outside world, the outside world cannot constitute a security threat. However, the company access to international interconnection network, providing services such as online stores and e-commerce, is equivalent to a closed network built within an open network environment. A variety of security including system-level security problem arises.

In the field of networking, the area of network security consists of the provisions and policies adopted by the network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and network-accessible resources. Network Security consists of a variety of computer networks, both public and private that are used in everyday jobs conducting transactions and communications among businesses, government agencies and individuals. Cloud computing provides the most reliable and secure data storage center, users do not have to worry about data loss, virus attack and other problems. Cloud security policy idea is: the more users, each user more security, because such a large user base, enough to cover every corner of the Internet, as long as a site to be linked to a new Trojan horse or virus appears will immediately be intercepted. In response to these security risks, we propose in the cloud computing platform built on the process of e-commerce platform for the following security policy. Use of safer storage techniques. When natural disasters, hardware failures and other facilities were damaged, and happens, the cloud service providers should adopt a safer and more effective storage technology once the main equipment was damaged or failure immediately switches to another set of mirrored devices. Meanwhile, e-commerce businesses need to regularly back up important data.

So far, Google, IBM, Microsoft, Amazon, including various cloud computing providers have their own set of standards not compatible with each other. For the healthy development of cloud computing it must be combined in various related organizations and enterprises to develop formal, open standards for the cloud computing public. Public key infrastructure (PKI) technology is the core information security technology and the key to electronic commerce and infrastructure technology. PKI technology can greatly reduce the use of e-business risk disclosure of sensitive information. The data in the "cloud" storage and transmission by non-authorized persons will not peek, illegal tampering, cannot be denied, thus ensuring the e-commerce of information confidentiality, integrity and effectiveness of the safe operation of e-commerce has provided a guarantee. In the software-as-a-service cloud model, the vendor supplies the hardware infrastructure and the software product, and interacts with the user through a front-end portal. SaaS is a very broad market. Services can be anything from Web-based e-mail to inventory control and database processing. Because the service provider hosts both the application and the data, the end user is free to use the service from anywhere.

The establishment of private clouds. Private cloud is a cloud services provider for enterprises in the construction of its internal proprietary cloud computing systems. E-commerce businesses can be non-critical business on the public cloud platform, which will be both business-critical core business is the part on the private cloud, so that enterprises can not only bring in more like authentication, data isolation, security technology to ensure data security, but retain the economies of scale of cloud computing systems. Cloud computing service provider choice. Cloud-based computing model in building e-commerce platform, before e-commerce businesses should be based on their own needs to choose the larger, better brand of cloud service providers, data storage and asked to confirm in case of cloud computing service termination of the service provider or another company such cases, their data will not be affected, how to take back their own data, and whether the data back into alternative applications and other issues.

176.4 Conclusions

Cloud computing technology is still in the early adopter phase for both providers and consumers. Cloud computing is a grid computing, distributed computing, parallel computing, utility computing, network storage technologies, virtualization, load balancing and other computer technology and traditional fusion of network technology development. Through the network, a number of lower cost computing entities, integrated into a powerful marketing ability of the perfect system. By improving the core concept “cloud” of coverage, and the “cloud” computing power between the logic to achieve the results in the system of marketing, it can reduce the financial burden of the user, eventually reduced to as long as the home user, a terminal, can get almost unlimited number of quality customers enjoy the “marketing cloud” to bring the powerful economic interests.

In summary, cloud computing applications will greatly improve the way electronic commerce services and functions, which greatly enhances the core competitiveness of enterprises, but also to the e-business which brings new challenges. In order to use Cloud Computing effectively, information and knowledge must flow easily and securely to and from the user, and these valuable assets must be saved securely and adequately backed up and protected from disasters. A comprehensive and rational security policy must take full account of users, e-commerce companies, cloud computing service providers, network operators, third-party certification bodies and other interests of all parties and their relationship in order to ensure the security of e-commerce operations.

Acknowledgments The authors thank the anonymous reviewers for their insightful comments and constructive suggestions that have improved the paper.

References

1. Jonathan FD, Sean RP (2006) E-Commerce: legal issues of the online retailer in Virginia. Richmond J Law Technol 13(2):215–223
2. Wang C-c (2010) Ai atmosphere cloud computing era of digital libraries of information security. Library 12(1):102–107
3. Graham M (2008) Warped geographies of development: the internet and theories of economic development. Geogr Compass 2(3):771–779
4. Sun L, Dai Z, Guo J (2010) Cloud computing key management framework. Telecommun Sci 7(9):65–73

Chapter 177
Interface Implementation for Web and EJB Layers in Enterprise Application

Jia Xiaoyun and Wang Xiaoxia

Abstract This paper introduces the interface implementation techniques for Web and Ejb layers by developing the management system of enterprise level based on J2EE. Firstly, it studies the hierarchy and functions of J2EE. Secondly, it discusses the steps to obtain EJB home interface through EJB home factory. Finally, this paper points out existing problems when it obtains the EJB home interface and its countermeasures. The management system of enterprise level is more rational by taking the steps. It makes more succinct of calling EJB procedure and it has reduced the load of consuming and network of systematic resources and improved the systematic speed too.

Keywords Interface · EJB home factory · Java naming directory interface

177.1 Introduction

Not only emphasizing the timing, the key to developing enterprise applications also requires program to deploy conveniently, to transplant flexibly and to upgrade easily, which require the application developers to deploy high quality applications based on rapid development. So how the applications choose the system structure or program model will be a major problem [1]. Using J2EE to develop the

J. Xiaoyun (✉) · W. Xiaoxia
College of Electrical and Information Engineering,
Shaanxi University of Science and Technology, Xi'an, 710021, China
e-mail: sunnyjingyi@126.com

W. Xiaoxia
e-mail: 332100987@qq.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_177,

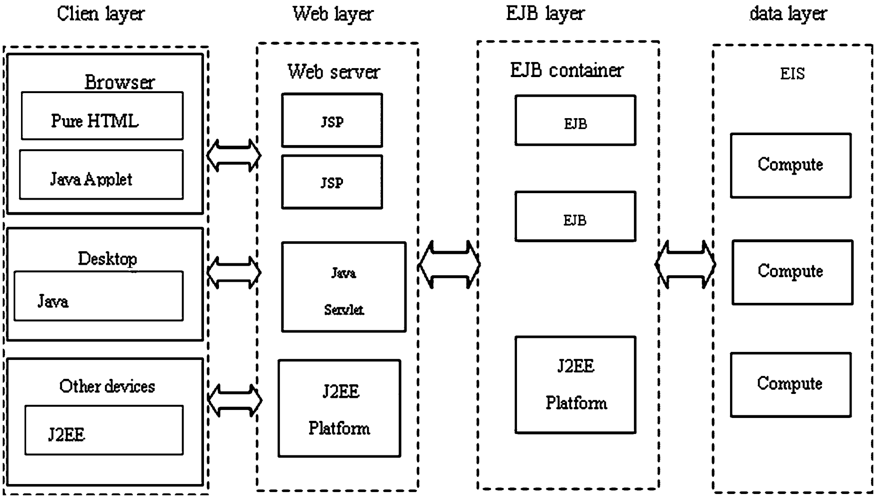


Fig. 177.1 J2EE Architecture

enterprise management system can be a very good choice to solve this problem in practical application. The J2EE multilayer structure is used as the system frame, including customer layer, WEB layer, EJB layer and data layer [2]. The merits of the frame are as follows: clear division, easiness to operate and maintain, strong independence, high security, simple programing, and cross-platform. So the interface among the layers becomes the primary problem, and the key to evaluating the system performance is the interface implementation between the web layer and the EJB layer.

177.2 J2EE System Hierarchy

J2EE has four layers as shown in the diagram.When we develop enterprise management system, we use four layers structure as Fig. 177.1 [3]:

First layer. Client layer

This is browser layer in charge of the interaction between system and client, for example, to show query result and collect the information input by client in HTML language.

Second layer. WEB application layer

This is composed by WEB components such as JSP, SERVIET, JavaBean and run by WEB container. It mainly takes charge of invocation of EJB layer and simple logic of some other client. It should be taken into consideration that WEB application layer should only comprise simple client logic, like effective simple judgment input, not application logic [2].

The third layer. EJB enterprise component layer

Enterprise component layer is run by EJB container to support EJB, JMS, JTA etc., service and technology. It is the core layer of the whole system, where the enterprise application logic is implemented. On the one hand, it transforms object for database record to analyze and design the data by object-oriented method, but on the other hand, it supplies invocation interface of application logic to WEB layer. Fortunately, we need not do many low-level programs owing to EJB container [5]. In order to ensure the high-efficiency and independence of the system, the layer is actually divided into several sublayers, and then we will discuss it more in the back. We choose Web Logic for the application server of this layer and then we shall describe Web Logic of EBA in detail.

The fourth layer. Data layer

It stores physical data, and provides relation data model to EJB or WEB layer. This layer is the lowest layer and realized by mature relational database system. Oracle 8i is used in practice [4].

177.3 The Interface of Web Layer and EJB Layer

177.3.1 Getting EJB Home Interface Through JNDI

The client that is relative to the EJB server calls procedure of the EJB is shown below. It is WEB layer [4].

Calling EJB has the following steps:

(1) Getting EJB through JNDI (Java Naming Directory Home with) namely step 1 and 2 in Fig. 177.2.

(2) Creating EJB object and getting its remote through EJB home interface, namely step 3, 4 and 5.

(3) Calling the EJB methods through the Remote Interface. Step 6, 7 in the Fig. 177.2. We will illustrate it by calling EJB of book session: the source program is as follows

```
    package ejb.book;
import Javax.naming.Context;
import Javax.naming.InitialContext;
import Java.util.Hashtable;
import Javax.ejb.*;
    import Java.rmi.RemoteException;
    import Javax.ejb.*;
    import Java.rmi.RemoteException;
    public class BookClient{
```

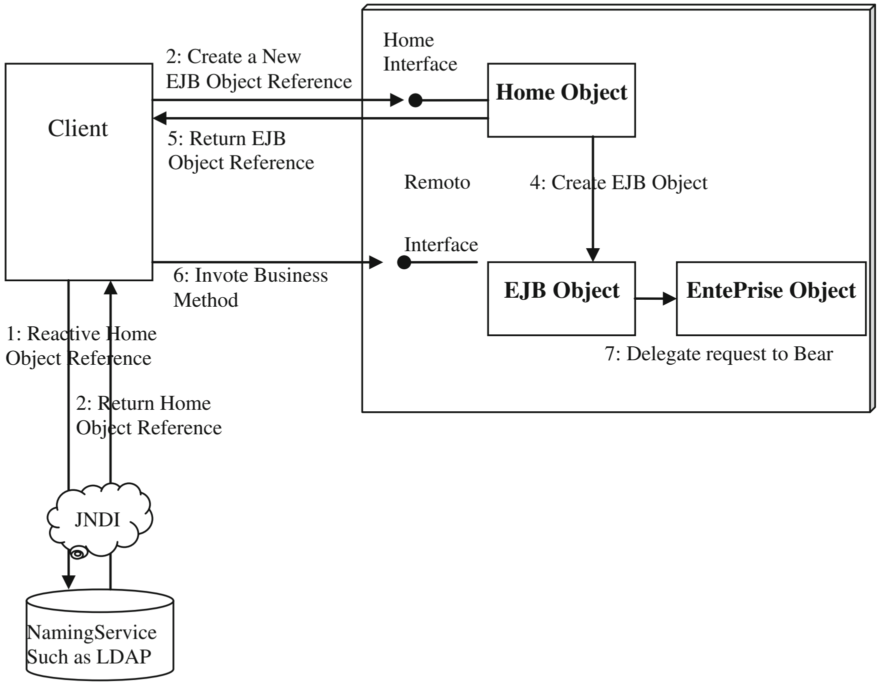


Fig. 177.2 The procedure of calling EJB

```
   public Book GetBook()
     { String url = “rmi : //localhost : 6888”;
        Context initCtx = null;
   BookSessionHome bookSessionhome = null;
      try{
   Hashtable env = new Hashtable();
   env.put(Context.INITIAL_CONTEXT_FACTORY,
   “com.apusic.jndi.InitialContextFactory”);
  env.put(Context.PROVIDER_URL, url);
   initCtx = new InitialContext(env);
                  }
catch(Exception e)
{
System.out.println(“Cannot get initial context: “ + e.getMessage());
System.exit(1);
                  }
```

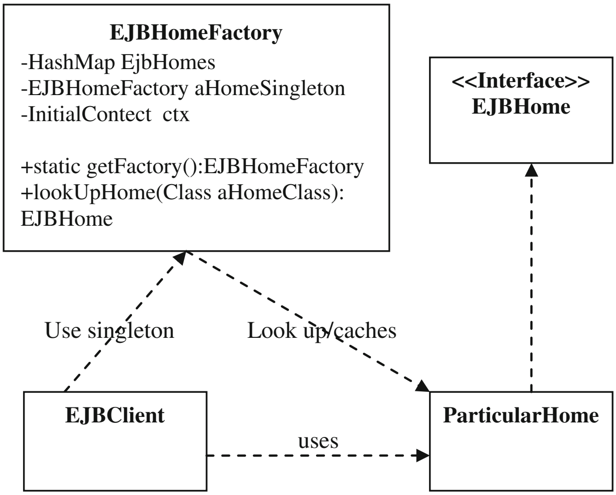


Fig. 177.3 The principles to get EJB home interface

```
try{//Getting EJB Home Interfac through JNDI
    BookSessionhome = (BookSessionHome)initCtx.lookup
     ("BookSessionHome");
BookSession bookSession = BookSessionhome.create();
 return bookSession.GetBook();
        }
catch(Exception e)
{
System.out.println(e.getMessage());
System.exit(1);
          }}}
```

177.3.2 Through EJB Home Factory to Get EJB Home Interface

From the above example we know that getting EJB home interface is the first step to calling EJB and also the necessary step to get EJB Home [4]. When the EJB is called each time, there may be some issues as follows:

First. Too much code duplication. The code to get EJB Home will spread around the whole WEB layer.

Second. Low efficiency. It will occupy many resources each time through JNDI to get EJB Home, and the speed of searching EJB object through the network is slow.

So a way should be adopted to extract all JNDI, and hide the initialization of Context, look up of Ejb object, modeling by narrow and violations of capture to make all the clients use the same code, reduce the complexity of the code, provide consistent control and use cache to improve performance [5]. Principles pursued are as Fig. 177.3. The source program is as follows:

```
public class EJBHomeFactory
{
private Map ejbHomes;
private static EJBHomeFactory aFactorySingleton;
Context ctx;
Private EJBHomeFactory() throws NamingException
{
ctx = new InitialContext();
this.ejbHomes = Collections.synchronizedMap(new HashMap());
}
public static EJBHomeFactory getFactory() throws HomeFactoryException
{try
  {
if (EJBHomeFactory.aFactorySingleton == null)
   {
EJBHomeFactory.aFactorySingleton = new EJBHomeFactory();}
  }
  catch (NamingException e)
{throw new HomeFactoryException(e);}
return EJBHomeFactory.aFactorySingleton;}
public EJBHome lookUpHome(Class homeClass) throws HomeFactoryException
{
EJBHome anEJBHome;
anEJBHome = (EJBHome) this.ejbHomes.get(homeClass);
try
{
if(anEJBHome == null)
{anEJBHome = (EJBHome) PortableRemoteObject.narrow(ctx.lookup (
homeClass.getName()),homeClass);
this.ejbHomes.put(homeClass, anEJBHome);
}}
catch (ClassCastException e)
{throw new HomeFactoryException(e);}
catch (NamingException e)
{throw new HomeFactoryException(e);}
return anEJBHome;}
public EJBHome lookUpHome(Class homeClass, String jndiName)
throws HomeFactoryException
{EJBHome anEJBHome;
anEJBHome = (EJBHome) this.ejbHomes.get(homeClass);
try
{if(anEJBHome == null)
{System.out.println("finding HOME for first time");
anEJBHome = (EJBHome) PortableRemoteObject.narrow
(ctx.lookup(jndiName), homeClass);this.ejbHomes.put(homeClass, anEJBHome);
}
}
catch (ClassCastException e)
{throw new HomeFactoryException(e);
}
catch (NamingException e)
{throw new HomeFactoryException(e);
}
return anEJBHome;}}
```

Thus, uniform and simple statement is used to connect EJB, and query times are reduced because of the EJBHome cache. For example, the statement of Home interface of Book Session is as follows:

```
BookSessionHome BookSessionHome = (BookSessionHome)
EJBHomeFactory.getFactory().lookUpHome(BookSessionHome.class);
```

177.4 Conclusions

This thesis introduced the interface implementation techniques for web and Ejb layers, and proposed improved interface implementation techniques, which made the management system of enterprise level more rational and calling EJB more succinct [5]. In this paper, for decreasing the consumption of resources and network load as well as increasing speed, Home interface need to be cached by Hash map class set.

Acknowledgments This work was supported by the special research projects of education department of shaanxi province (No.2010JK432), China.

References

1. Scotts Valley (2008) Enterprise javabeans developer's guide, Borland Company, [M]
2. Sestoft Peter (2007) Java precisely. [M] The MIT Press, London
3. ED Roman, Mastering EJB (2010) 2 New york, [M]Wiley, Longmen
4. BEA Systems (2010) Programming WebLogic Enterprise JavaBeans, [M]BEA Systems
5. Sun Microsystems Inc. (2009) Core Servlets and JSP, [M] Scotts Valley

Chapter 178
Web-Based Client-Side Multimedia QoS Information Model

Yang Song, Xiaoning Li and Xiaofeng Li

Abstract Qos is a comprehensive indicator to measure satisfaction with a service, or quality of service. This paper analyzes the web site client multimedia Qos, Qos technical model and IPQos the implementation mechanism, so as to establish a complete information model Qos.

Keywords Qos · Site customers · Information model · Multimedia

178.1 Lead Talk

Along with the multimedia technique [1] fly soon a development, multimedia applications on the Internet are piling up one after another, such as: the IP telephone, video frequency meeting and video frequency point sows. Long range education wait until a multimedia actually business. The lately applied emergence carries multimedia service quality to the customer [2] and also makes a new request, for example IP speech, video frequency point sow etc., is solid business to report a text of delivering and delaying and then having high request, if report the

Y. Song (✉)
Information Engineering College, Jilin Normal University, Changchun, China
e-mail: abc-cxf@163.com

X. Li
Changchun Teachers College, The Computer Science and Technology Institute,
Changchun 130032, Jilin Province, China

X. Li
Jilin University, Changchun 130032, Jilin Province, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_178,

text transmission postpone too long, the guest room carries a customer to can not accept.

It opposite but talk, traditional businesses, such as E-Mail and FTP, …, etc., to time delay sensitive not have dissimilarity service businesses such as demanding speech, video frequency and data etc., and request that network can classify a different type correspondence data for the sake of support, then for it provide homologous service. What traditional IP network provides is not transmission service for doing utmost, can identify and classify the correspondence data of various type in a network, more do not can provide different service for different correspondence. So making an effort service mode that says a traditional network has already canned not satisfy customer to carry an applied demand, customer in the web site carries the establishment of multi-media services quality QoS information models and then concentrates on to work out this problem.

178.2 Qos of WEB Network

Along with the development of Internet, appeared customer from station in a great deal of website to carry a multimedia solid business [3], because of solid the business postponed to the network's delivering, tremble while postponing etc., characteristic compare is sensitive, consequently these solid the emergence revelation of business two important blemishes of IP network technique: the traditional IP road is vomited not high and traditional IP of quantity by the swallowing of technique don't have service quality QoS to promise.

QoS can control various network application, satisfy various website customers to carry a multi-media application request, for example: can carry the bandwidth of FTP use according to the applied need restriction customer, can also visit for database with higher have the initiative class; can carry multi-media business to provide a bandwidth for the customer of the sensitive in time and low postpone assurance, and other business while using a network, cannot influence the business of these time sensitives, either.

When network occurrence is obstructive [4], one of network equipments connects under the situation that has QoS support and adopt priority queuing (PQ) brigade row of the strategy carry on a processing report text. Get into according to the category certificate report text belonged to by report text. So when every time sends out to report a text, always will have the initiative an of class Gao the text send out to first and promised that belonging to is higher the report text of row having the initiative a class brigade has lower of postpone [5].

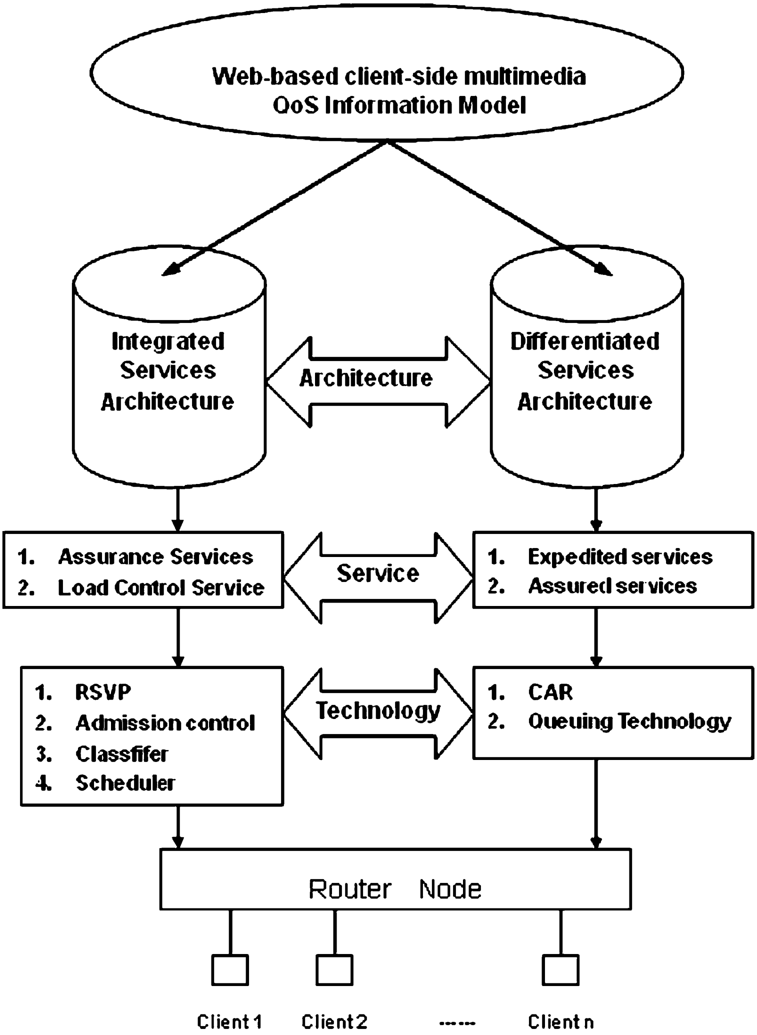


Fig. 178.1 Qos model

178.3 QoS Service Model

So far, different organization or unit of industry have already put forward some service models of QoSs [6], this XRM model for putting forward of HeidelbergQoS model, American the Columbia University COMET research set that includes IBM company and American guest Xi method Ni OMEGA system structure etc., of the second university. QoS then mainly has two kinds of system structure as follows, such as Fig. 178.1 show.

The underneath will report a little bit when the network takes place to obstruct text at have no QoS promise and have QoS to promise different treatment in the network.

When network occurrence obstruct, network equipments of a connect under the circumstance that have no QoS support, adoption tradition of go into first first the strategy of to carry on processing to report a text.While having no QoS to promise, it is all to want from should pick up a people output's report text, according to arriving order of sequence to get into FIFO of connecting [6] brigade row end of the tail, and connect one by one in order sends out to report a text while sending out to report a text, from the head beginning of FIFO brigade row, all report texts are in the process of sending out in, there is no differentiation, also not to report text transmission of the quality provide any assurance.

(1) Comprehensive service system structure, it can provide two kinds of service to promise service and load to control service as follows. Promise service to the bandwidth, postpone, the cent set throw to lose rate to provide a metered quality control and satisfy the request of applying the procedure. For example, can reserve a 5 Mbits to the VOIP application's bandwidth and 1 S of postpone a request. Load control service provides one category for customer under the situation that the network has never led to carry of service, it is one kind to settle sexual index sign and promise even if under the circumstance that network over carry, also ability to report text's providing Be looked like to a network have never led to carry similar service. For example, obstructive circumstances happen in the network under, promise that the report text of some application procedures can get low postpone and high service for passing.

The brigade row adjusts one degree machine mainly is according to definitely adjust one degree calculate way to carry on adjusting one degree service to the cent set brigade row in classification and familiarly adjust one degree calculate way to contain WFQ, WF2 Qses, SCFQ, VC, MD-SCFQ, and WRR, …, etc.

According to the rules of the Intserv model, before sending out and reporting a text, the applied procedure notifies relevant its own discharge parameter of network and particular service quality claim of demand, including bandwidth first, postpone etc. In the IntServ service model, be responsible for deliver Qos claim of the letter make is a resources to reserve agreement (RSVP), it notifies that the router applies the Qos need of procedure.

The network carries out a resources allotment check (admission control) after receiving the resources of applying the procedure and requesting, according to apply the resources application and network existing resources circumstance of procedure, judgment whether is applied procedure allotment resources [7]. Once the network confirmation was the report text that applies procedure to assign a resources, then wanted $\sim$ only to apply the report text of the procedure to control inside the scope that describes in the discharge parameter, network commitment satisfy apply the Qos need of procedure.

The applied procedure is generally receiving a network to really recognize an information, then confirming the network already for the report text that applies procedure reserve resources after, just beginning according to application of discharge parameter and particular service quality claim send out to report a text.

(2) Classify service system structure

Classify service system structure (Differentiated Services Architecture, DiffServ) is put forward in the RFC2474 by IETF, aim at soldier righteousness a kind of implementation IPQos and more easily expand of way. Classified service system structure to simplify letter to make, flow to the business of classification grain the degree is thicker. It passes to remit to gather (aggregate), can flow each close by business of Qos need to see into a big type and adjusts the brigade row that one degree calculate way handles number by decrease. On top of that, distinction service system structure passes to skip a line each time for (Per Hop Behavior, PHB) of gradually jump to forward way to provide definitely procedural Qos to promise, each PHB forwards way or Qos need toward houlding be a kind of.

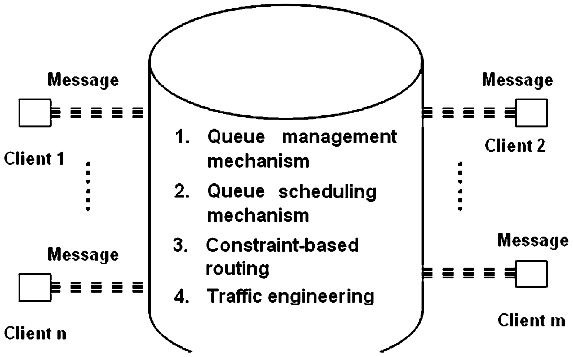
Lead the concept of going into the DiffServ area in the middle of classifying service system structure, a do not the DuffServ area can think to is a son net that can provide DiffServ business. The DiffServ area mainly constitutes to from some routers, and carried on a distinction to these routers, be located in the being called of DiffServ area boundary boundary router (edge router), but be located in core router (core router) in the being called of DiffServ area inner part. Boundary router completion to the classification, orthopedics, marking that the business flows and adjust a degree, core router completion according to remit to gather of thicker classification, adjust a degree. Currently, DiffServ main underneath of cash two kinds serve.

Qos that DiffServ generally uses to provide to carry to carry for the application of some importance. It passes following technique to carry out.

(1) CAR
It according to report text of Tos or Cos the value and IP report text of the 5 dollars set etc., information carry on reporting a text classification, the completion reports marking and discharge of text to take charge of.

(2) Brigade row technique
The brigade row techniques, such as WRED, PQ, CQ, WFQ and CBWFQ, …, etc., are to the report text obstructing carries on slowly saving and adjusting a degree, realization obstruct a management [8].

Fig. 178.2 Qoses carries out mechanism principle


Message
Client 1
Message
Client n
1. Queue management mechanism
2. Queue scheduling mechanism
3. Constraint-based routing
4. Traffic engineering
Message
Client 2
Message
Client m

178.4 Qos Carries Out a Mechanism

Qos carries out a mechanism to mainly have following four kinds: the brigade row manages mechanism, brigade row to adjust one degree mechanism, according to control of road from measure engineering with business. Qos carries out mechanism principle if Fig. 178.2 show:

(1) The brigade row manages a mechanism (queue management mechanism)

When the network takes place to obstruct [9], the router has to throw a set for some cents, solution of this problem is first lie in having to carry out effective brigade row a management mechanism (or buffer area management strategy)

The brigade row manages a mechanism (queue management mechanism)

When the network takes place to obstruct [9], the router has to throw a set for some cents, solution of this problem is first lie in having to carry out effective brigade row a management mechanism (or buffer area management strategy)

(2) The brigade row adjusts one degree mechanism (queue scheduling mechanism)

In spite of in IntServ is still DiffServ, all involve brigade row to adjust one degree problem. Chien talks it, brigade row's adjusting the function of degree is a router how choose from several (or a) brigade rows next treat to forward of cent set; this contains hypostatic differentiation with brigade row management mechanism. According to the different service rule, the brigade row adjusts one degree calculate way to is divided into a few kind is as follows: arrive first to serve (FCFS) first and circularly adjust a degree (RoundRobin), processing machine share (ProcessorSharing) and have the initiative class service, random service etc.

The brigade row that appear already currently adjusts one degree calculate way to mainly have the calculate way to adjust a degree according to the circulation and is two major types according to the GPS calculate way. An effective brigade row adjusts one degree function index sign that the calculate way should attain to mainly have equity, postpone characteristic and flow to the malice business of the utilization, and complexity, etc., of insulation ability, network bandwidth.

(3) According to the road of stipulation from (consrained-baxed routing)

According to control of road from (CBR) originate QoSRouting, just the Qos restriction parameter carried on a certain enlargement. The CBRr effective realization needs the mutual match of each router, for example mutually notify respectively some informations of network for knowing. The CBR crux lies in how to release and release of an frequency to obtain one to compromise in the precision of status information.

(4) The business measures engineering (traffic engineering)

The business measures the purpose of engineering to lie in how availably guided business to once circulate a network, for the purpose of cancelation because of business quantity the asymmetry distribute but the network resulting in is obstructive. Negotiate more marking commutation (MPLS) and according to is limited of road from is all useful tool of business quantity engineering, is also the topic that needs a further research currently.

178.5 Summarize Briefly

This text mainly put forward an establishment network the customer carry multimedia Qos model, and analyzed the mechanism of realization, can effectively of solution's serving in the traditional network can not satisfy customer to carry an applied demand [10], make the network report a delivering of text faster and convenient.

References

1. List H, Ma JW, Yang HZ (2009) From adapt to small wave of whole dimensions number picture watermark. Chin Hua university college journal (natural science version), 05 expect 749–754
2. Zer S (2005) The network educates medium problem and strategy. Calculator education, 06 expect, pp 62–65
3. Congratulates and thanks a Xiao orchid, Liu Ying (2009) Under the mesh environment of the research and application of the distribute type search processor. Computer knowledge and technique, 200906 expect 266–268, 271
4. Lines of Wei Sui, Huang Sheng Hua (2007) 1 kind quickly evades obstructive of road from calculate way. The calculator imitates really, 04 expect 136–138, 148
5. Lee are red and gorgeous (2005) Controls system according to the process of Internet of postpone repair. Science and technology university college journal (natural science version) in China, 04 expect 51–53

6. Beard chapters are even (2009) Virtual calculation method of number of times for saving to lack page interruption in the management. Computer knowledge and technique, 06 expect 272–273, 279
7. Tang light front (2005) Network resources localization:constuct the new mode that the network resources navigates a database. Intelligence report magazine, 01 expect 36–37
8. D expose, plum Ke (2010) The intelligence spreads a feeling machine to unite the application in the net realm in the thing. Information technique with standardize, 08 expect 22–25
9. Chen government Ye, Shi Yan Yuan (2008) bear Shu China, to the soldier, distance virtuous hero. Flow medium model to study according to CDN of wreath road node and P2 Ps. Information technique with standardize, 10 expect 37–39
10. Huang JC (2007) WebGIS that carries mode according to the customer carries out a technique. Soldier work automation, 04 expect 39

Chapter 179
A Web Service Composition Algorithm Based on Dependency Graph

Zhang Hua, Fu Yan and Gao Hui

Abstract Applications based on Web service technology have grown rapidly and it has become more and more necessary to select and composite appropriate services automatically to satisfy the user's complex requirements. To achieve this composition, the input–output interfaces relations and the concept of domain ontology between them were used to make the process flexible. Our proposed approach is based on abstracting atomic services specification, preprocessing the multiple interfaces of a Web service by combination and constructing a dependency graph including all service interfaces and web services themselves. By using this dependency graph, we perform a new bidirectional heuristic search algorithm from the desired input and output interfaces to find composite web services. Therefore the essence of the algorithm was that the problem of Web services composition was transformed into the research approach of directed graph with an improved bidirectional heuristic search algorithm used to realize the composition of services. Theoretical analysis and experimental results show that this algorithm is more efficient and effective than traditional searching algorithm.

Keywords Atomic service · Services dependency graph · Web services composition · OWL-S · Heuristic search

Z. Hua (✉) · F. Yan · G. Hui
Deptartment of Computer Science and Engineering,
University of Electronic Science and Technology, Chengdu 611731, China
e-mail: junnan321@163.com

F. Yan
e-mail: fuyan@uestc.edu.cn

G. Hui
e-mail: huigao@uestc.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_179,

179.1 Introduction

Currently, applications based on Web service technology have grown rapidly. The increasing number of available web services over the past few years has led to the development of new web services and complex applications by composing existing web services that other from single Web service and single functionality to satisfy the user's requirement.

Methodologies have been conducted to solve the problem [1, 2]. In Hu S's framework [3], a distributed approach based on input–output parameters of process that discovering and compositing services automatically was proposed. Another effective web services composition algorithm [4, 5] was put out based on global quality optimization and multi-object chaotic swarm optimization. Reference [6] put forward a heuristic algorithm based on object distance and selected the successor services dynamically at runtime according to runing status of each service. Reference [7] used the semantics similarity among interfaces of web services as connection index, and gave the highest priority to select the services with most closest semantic. However, the performance time of these schemes is still a matter of concern due to the large scale of web services.

In this chapter, we propose a model that supports the web services composition based on a web services dependency graph using ontological informations with an improved bidirectional heuristic search algorithm [8, 9].

179.2 Architecture

The essence of web services composition is discovering and selecting a series of related services and integrating them into a complexed web service to satisfy the requestor's composed demand by orchestration of them [10]. This section defines terms and models that can be used in this composition algorithm.

Definition 1 *Atomic Service* (*AS*)

An atomic service is an indivisible software component that is too granular and only executes fewer business or technical functionalities. Consider the functionality, an atomic service was abstracted as a two-tuples ws (I, O), where ws is the identifier of an atomic web service (like name or unique id of web service), I and O are the input provide to the web service ws and output produced respectively that specified by the concept of domain semantic ontology.The structure of a atomic web service is shown in Fig. 179.1.

Definition 2 *The Web Service Dependency graph* (*G*)

Formally, a web service dependency graph *G* (*V*, *E*), capturing input/output information of available web services and describing the relationship among them, is a directed graph and composed of two data sets *V* and *E*, where *V* is a vertex set of nodes of the graph which represents the set of input/output of all the atomic

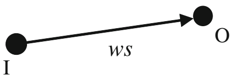


Fig. 179.1 Atomic service representation

services and the type of an input/output becomes the label of its corresponding node, *E* is a directed dependency edge set of all relevent atomic services,which is simply labeled by the identifier of a web service. The directed graph that connects all dependency edges is exactly the dependency model.

In Sect. 179.3 we briefly provide an algebra for specifying web services behavior and composition. The dependency graph is described in detail in Sect. 179.4. In Sect. 179.5 we present our composition approach. Finally, Sect. 179.6 concludes the chapter.

179.3 Web Service Dependency Model

To proceed with composition, the OWL-S specification language is briefly introduced. And the behavior of web services are mapped to a dependency graph, the problem mentioned above is converted to a general graph searching problem.

179.3.1 Extracting Web Service Specification

In this chapter we choose OWL-S from serveral web service specification languages, such as OWL-S and BPEL4WS, because it suits the problem at hand, and also it is partially meant to enable automated web services composition and interoperation [11].

In OWL-S, each web service is specified in three XML-based parts: service profile, which includes general information about the web service, such as the name, description, inputs, and outputs; service model, which using structures such as sequences, conditional statements, loops, and parallel constructs shows how the web service performs its functionality; and service grounding, which contains information on how the web service can be used in practice. We only concentrate on the service profile and service model constructs.

In this chapter, we assume that there is only one output for services.

179.3.2 Dependency Graph Construction

To construct a dependency graph, we need a local repository that contains required information for a set of existing web service. That is, set WS denotes dependency information between different inputs and outputs of the web services. (WS $= \{\mathrm{ws1},\ \mathrm{ws2},\ \mathrm{ws3}, \ldots, \ \mathrm{wsn}\}$) which is stored in the form of an $n \times n$ adjacency matrix N shown as below, and set SRIO collects its input and output, written as SRIO $= \bigcup_{\mathrm{ws} \in \mathrm{ws}} (ws.I \cup ws.O)$ The adjacency matrix is exactly the dependency graph.

$$N = \begin{bmatrix} n_{11} & n_{12} & \ldots & n_{1n} \\ n_{21} & .. & .. & n_{2n} \\ : & .. & .. & : \\ n_{n1} & n_{n2} & .. & n_{nn} \end{bmatrix} \quad n_{ij} = \begin{cases} ws_i & \text{(if there exists a dependency } i \to j \text{ labeled by } ws_i\text{)} \\ 0 & \text{(other)} \end{cases}$$

In general, when there exists more than one input, denote x1, x2, …, xk, we perform the combination operation to form them to one node {x1, x2, …, xk}.

The algorithm of constructing dependency graph is as follows:

Input: local repository, adjacency matrix N
Output: dependency graph G
Step 1. Get the nodes number and connections among nodes from N;
Step 2. If nij $\neq$ 0, then new two nodes I and O, and label the edge I → O as ws$i$;
Step 3. If there exists more than one edges pointed to same node g labeled by ws$i$, then perform combination operation and form them to one node {I i}, and label the edge $<$ {Ii} → $g >$ as ws$i$.
Step 4. If all connections between inputs and outputs are analyzed, then stop; Else, goto step3 and continue analyzing.

The time complexity of this algorithm is O (n2), thus huge consumption of system resources will be a problem in large scale number of input and output. However, the interfaces of web services hardly ever modify after designing. So we can construct the connection relationship beforehand, and only renew the service connection when the interfaces of a service changed or a new service appeared. Thus we can reduce the consumption of analyzing connection relationship dramatically.

179.4 Web Service Composition

Given service request r (Ir, Or), the essence of web service composition is to search for a subgraph G' $(V,'E')$ from the dependency graph G $(V,\ E)$, where $V' \in V,\ E' \in E$**,** That is, to find an effective path from start node s and goal node g, which makes the service composition satisfy the requirement goal.

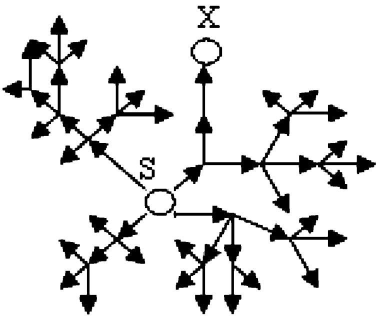


Fig. 179.2 A case in poor efficiency

179.4.1 Bidirectional Heuristic Search Algorithm (BdHS)

As the number of inputs and outputs is decided, more nodes in the graph are approached, more chances are there to get to goal node. We define the heuristic for a node n to be the number of nodes that are direct successors of n.

$h\ (n) =$ number of direct successors of node n
The search algorithm is given below:
Input: dependency graph $G,$ start node $s,$ goal node g
Output: service composition
Step 1. get all the successors of s and place in S;
Step 2. get all the successors of nodes in S' and place in S';
*Step*3. if $g \in (S \cup S')$ then success and stop;
Step 4. get all the successors of g and place in G;
Step 5. get all the successors of nodes in G and place in G';
Step 6. if $(S' \cap G')$ is not null, then return path from start node s to goal node g;
Step 7. otherwise, if $(|S| + |S'| + |G| + |G'|) \geq |\text{SRIO}|$ then stop report failure;
Step 8. Else
(a) Get the node x from S' such that h1 (x) is maximum;
(b) Remove x from S' and add it to S;
(c) Put all the successors of x to S';
(d) Get the node y, from G' such that h2 (y) is maximum;
(e) Remove y from G' and add it to G;
(f) Put all the successors of y to G';
(g) If $g \in (S' \cup G')$ then return the path from s to g; else GOTO step 7;
h1 (n) and h2 (n) are heuristic value for node n of set S' and G' respectively. In most cases, the algorithm can work fine with a shorter time than the execution time taken by breadth first search (BFS), depth first search (DFS).

Figure 179.2 shows a simple case of poor efficiency using our heuristic function,because $h\ (Y) = 1$ (Y is the parent of X), so the node x can not be expanded by heuristic search at the beginning no matter whether node X's parent is adjacent to a

Table 179.1 Execution time for dependency graph construction

Number of nodes	Number of atomic services	Execution time (ms)
15	19	140.3
50	72	2,366.2
100	167	10,674.7
200	352	23,774.6
400	567	40,032.7

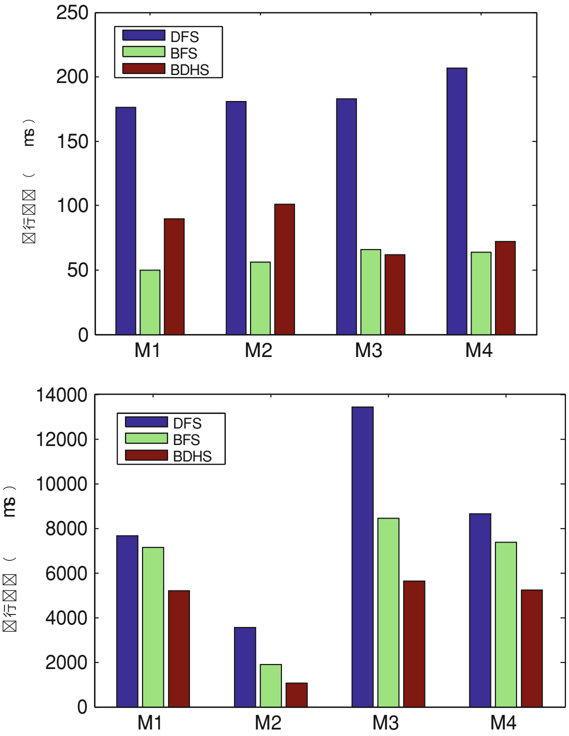


Fig. 179.3 **a** Execution time with 50 nodes; **b** execution time with 400 nodes

node in S'. So, the time taken to reach X using our heuristic is, in general, more than the time taken by other algorithms.

In order to overcome the defect, we propose an improvement that introduces a time_stamp to each nodes in set S' and G', which is initialized to all zeros. Then increment the value in cases that the node was placed in S' or the node was removed from S' and placed in S. Next, we choose a threshold min_limit for time_stamp, any nodes in S' and G' whose time_stamp is bigger than min_limit should be returned to expand despite whether its value of h (n) is maximum. Several experiments indicate that when the value of min_limit equals to $0.1 * |\mathrm{SRIO}|$, we can reach the best execution time. After that, the node X would be selected only after several steps due to its time_stamp.

179.5 Implementation and Experiments

Table 179.1 shows the time consumption to construct the dependency graph on different sets of web services and inputs and outputs. And the results show that more the inputs and outputs and services are, the longer time it consumes to construct the web service dependency graph. Therefore, for large scale services, the best method to reduce the dynamic connection setup time is to pre-establish links of dependency graph.

We choose the most classic search algorthms such as DFS and BFS in comparision with our proposed BdHS algorithm. Figure 179.3 clearly states the execution time for different search algorithms to finish the web service composition on different numbers of nodes in the dependency graph.

In Fig. 179.3, M1 stands for a batch of execution time of DFS, BFS and BdHS algorithm with the same start node and goal nodes, similarly it is applied to M2, M3 and M4. Blue column denotes the execution time of DFS algorithm, the green and brown ones stand for BFS and BdHS, respectively.

The results in Fig. 179.3 show that the execution time of service composition search time showed no significant differences among the algorithms when the count of nodes is not large in (a), while the difference is significant in (b) with the nodes number of 400. Above all, as the number of nodes and number of web services involved increases, the efficiency of our algorithm also increases. Thus for large scale graphs our algorithm can prove very efficient in terms of computational time and performance.

179.6 Conclusions and Future Work

In this chapter, we present an approach of repository web service model, called dependency graph, and introduce how it captures the properties of web services in terms of inputs and outputs, and input–output dependencies applying OWL-S as web service specification language. We mentioned the improvement of our dependency graph by combination of input–output dependencies where more than a single input are involved. Then the problem of web services composition is converted to the path searching in directed graph. We used this graph and BdHS algorithm to find a composite service for a given request. Experimental results demonstrate the advantage over some other search algorithm.

In our future work, we plan to focus on the quality of service and cost parameters in detail on Web services, as well as making use of preconditions and effects.

References

1. Oh SC, Lee DL, Kumara SRT (2007) Web service planner (WSPR) an effective and scalable web service composition algorithm[J]. Int J Web Serv Res 4(1):1–22
2. Oh SC, Lee DL, Kumara SRT (2008) Effective web service composition in diverse and large-scale service networks. IEEE Trans Serv comput 1(1):15–32
3. Hu S, Muthusamy V, Li G et al (2008) Distributed automatic service composition in large-scale systems. Proceedings of the second international conference on distributed event-based systems[C]. Rome ACM, pp 233–244
4. Qiqing F, Xiaoming P, Qinghua L, Yahui H (2009) A global QoS optimizing web services selection algorithm based on MOACO for dynamic web service composition. In: 2009 International forum on information technology and applications, pp 37–42
5. Deng S, Wu J, Li Y (2007) Automatic web service composition based on backward tree[J]. J Softw 18(8):1896–1910
6. Wen J, Chen J, Peng Y (2007) A method of heuristic web services composition based on goal distance estimate[J]. J Softw 18(1):85–93
7. Li M, Wang D, Du X (2005) Dynamic composition of web services based on domain ontology[J]. Chin J Comput 28(4):644–650
8. Hashemian SV, Mavaddat F (2006) A graph-based framework for composition of stateless web services. In: Proceedings of the European conference on web services, pp 75–86
9. Hashemian SV, Mavaddat F (2005) A graph-based approach to web services composition. In: Proceedings of the symposium on applications and the internet (SAINT), pp 183–189
10. Rao J, Su X (2005) A survey of automated web services composition methods. In: Cardoso J, Sheth AP (eds) SWSWPC 2004. LNCS vol 3387. Springer, Heidelberg, pp 43–54
11. Martin D (2003) Information on http://www.daml.org/services/owl-s.html

Part XVI
Wireless Network

Chapter 180
Architecture of Crossing Social Network System

Ruliang Xiao, Youcong Ni and Xin Du

Abstract How to organize crossing social network resources on a higher level of integration and address them to users' desktops is an important and difficult problem. Especially, there is lack of efficient approach to software architecture to build reusable system over crossing social network. From the view of point of temporal logic XYZ/E, this chapter proposes a kind of Architecture Description Language about Crossing Social Network system (CSN-ADL), which can be used to depict main key processes over the cross-social network system, formally defines some key concepts such as relation component, correlation component, override correlation connector, interaction connector and correlation network-oriented architecture, which provided a formally theoretical instruction for architecture reuse.

Keywords Social Network · ADL · Correlation

180.1 Introduction

In order to solve the problem about sharing the user's network resources on the Internet, social network established a new way which breaks through the traditional relationships on the further development of obstacles among people, and deeply affects business model. Most users can submit their data and share network resources simultaneously in several social networks [1] which are self-governed and separate but are across on the users desk. Along with the rapid development of

R. Xiao (✉) · Y. Ni · X. Du
Faculty of Software, Fujian Normal University, Fuzhou 350108, China
e-mail: xiaoruliang@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_180,

the semantic web technology, social network system rises various typical applications based on Web2.0 technology, such as famous social networks—Linkedin, Twitter, YouTube, MySpace, Flickr and so on [2]. After the year 2003, it successively appeared some new applications of Web2.0 in China, such as Renren, Kaixin, Sina Micro-Blog and so on. How to organize all social network resources on a higher level of integration and address them to users' desktops is an important and difficult problem [3]. Due to the different structures of various social network systems, problems such as interpersonal relationship, similarity of resources, change of associated state, accessibility of interactive behavior, interactions and activity, effective test and correctness inspection of results have become the bottleneck that influences crossing social network (CSN) system on providing popularization and application of sharing service. The traditional pattern of software design is not adapted to the construction of crossing social network system. At present, there are few researches on software architecture of CSN system.

Software Architecture Description Language is considered to be an important tool for reasoning the overall design scheme of software system structure in the early software development. It describes components of software system, and the interaction between the components and model of guiding component combination and relevant combination constraint mechanism on higher abstract level [4]. In the past decades there have been a large number of researches on software architecture description language. Currently main ADL take examples as Aesop, C2, SADL, Unicon Rapide, Wright, XYZ/ADL, ABC/AD and AC2-ADL. These ADLs emphasized on different sides of system structure and played an important role in system structure's research and application. For the social network system, because that there are lots of relationship, it is a new type of application forms, whose structural relationship showed crosscutting behavior and crosscutting characteristics as the main structure characteristic style. By depicting the main characteristics of the system structure formally, analyzing semantic and validation, it is an effective approach to solve social network bottleneck problem.

XYZ/ADL is an expansion of temporal logic language XYZ/E [5, 6]. It can be used in social network system structure for formal description and analysis through the top-down gradual refinement description method to provide formalized semantic guidance for interpersonal relationship, similarity of resources, change of associated state, accessibility of interactive behavior, interactions and activity, effective test and correctness inspection of implement results in the social network system and then to solve the crosscutting problem based on all kinds of relationships. This chapter firstly introduces the related research work. And, it defines the components of associated, relationship, relationship superposition, interactive etc., it then puts forward crossing social network system architecture description language CSN_ADL by which semantic analysis and formalism description in the three main processes of crossing social network system structure are done.

180.2 Related Work

So far, there has been a lot of work about the software architecture. These typical ADLs can be listed as Unicon, C2, Aesop, Wright, Darwin, Rapide, SADL, XYZ/ADL, ABC/ADL, AC2-ADL etc. Most of these languages come from research design and development in some special application. Therefore, they have different emphasis and characteristics. For example, Unicon supports heterogeneous type parts and connections and has a high-level compiler aiming at system architecture [7]; C2 supports for describing user interface system with an application message style [8]; Aesop supports application of software architecture style [7]; Rapide is specially designed for establishing a rapid prototype for system architecture based on events, concurrency, object-oriented languages and allows designing by simulating software architecture, and provides simulation results of tools analysis [8]; Wright supports definition and analysis of the interaction between the components, and connect-components is the most important part[8]; Darwin builds the model of system behavior by π calculus, uses its strong type system to do static checking [7], SADL provides the formal base about architecture [8]. In China, many scholars also put forward lots of characteristic ADLs. The most influence is a general architecture description language ABC/ADL, proposed by Professor Hong Mei of Beijing University Software Institute, which regards the software architecture description as component development framework and assembling system blueprint by referring programing language of type-instance relationship to distinguish type chart and instance configuration diagram. It is benefit for changing the architecture model to programing language, introducing the detail information such as type system, which is beneficial to translate from design to realization [7].

180.3 Concept Expansion on CSN_ADL

XYZ/E language includes core concepts about conditional element, unit, procedure and process and connection model about procedure-call and message-passing [5]. The most important difference between social network software architecture and currently typical software architecture is the four important components- relation component, correlation component, override correlation connector and interaction connector with users' participation. Correspondingly, the center of the software architecture is around these four components, with connector describing system operation process. Then, we can build the description framework of social network software architecture from these two aspects based on XYZ/E.

Definition 1 (*Relation*) Relation is a simple component with independent function, which represents a connection between people and people, sharing resources and sharing resources, or labels and labels. It encapsulates internal passage and process with standard, and be divided into two parts: interface and computation. Interface includes a set of explanation components and description of port with external environment interactive behaviors, and function specification of explanation component describes specific behaviors. It can define relational external interface based on XML, and describe the relational computation parts. Grammar of relation can be shown as below:

Relation::= RelationName[RelationDecPart]= =[Interface][WherePart]
RelationDecPart==[IntefaceDecPart][FunctionDecPart][ComputationDecPart]
IntefaceDecPart::=%PORT[port,…,port]
Port::=PortName[PortDec]==[DataType][PortBehavior]
FunctionDecPart::=FunctionName[Function specification]
ComputationDecPart::=ComputationName[Computation specification]
Interface::=InterfaceName[InterfaceDec]= =[Package][process]

where interface including package and processes can be considered as the refinement of interface, function, and declaration part of calculation for relation. In this definition, each port is a relation request to external conditions by the component, or a relational service provided for the environment expressed by channel. It includes two parts: relational data type declaration of channel and relational behavior description of channel. Among them, DataType should declare channel data types which can be accepted by relational ports. PortBehavior should depict behavior of the relational port, and also part behaviors of relational component (externally visible behaviors), and regulate how the external environment should interact with the relational components by the port and expressed by an XYZ/E unit. Functional specification explains what this component is doing.

Definition 2 (*Correlation*) Correlation is an autonomous software body which consists of one or more relations and has an independent function. It is established that new relation is iterated with the existing correlation. Correlation can be shown as below:

Correlation::=CorrelationName[CorrelationDecPart]= =[Correlation][Relation]
CorrelationDecPart::=[IntefaceDecPart][FunctionDecPart][ComputationDecPart]

In the current social network system, the characteristic between relation and correlation is expressed in Fig. 180.1.

In Fig. 180.1, correlation expansion presents the overrided relation form, and in the next we directly construct correlation connector contained relation overrided to the correlation. It defines "override" as the functional operation of connector between the relation and the correlation, and using XYZ/E to define override method. It includes three aspects: adding new relation when the original correlation has not such relation; updating into a new correlation according to the constraints when this relation has existed; deleting this relation in the original relationship. It is easy to get a theorem as follows.

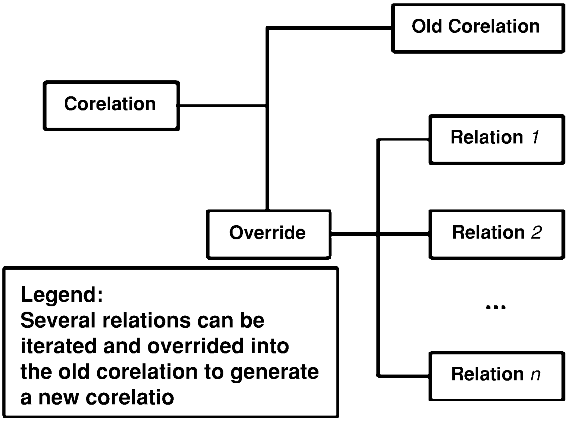

Old Corelation
Corelation
Relation 1
Override
Relation 2
...
Relation n
Legend:
Several relations can be iterated and overrided into the old corelation to generate a new corelatio

Fig. 180.1 Relation iterated to the old correlation to override new correlation

Theorem 1 *A correlation can be constituted by a variety of relations.*

From the macro of view, crossing social network system is a kind of relation network system and connects each correlation components with connector. It constitutes a frame structure of correlation network system.

According to description of structure about connector components (fittings) in XYZ/ADL [6], connector defines interactive modes and rules between components, to describe common characteristics of some interaction. It can be used as the medium of communication and collaboration between activities of some components, as well as the "glue" of the software architecture design, and combine all the components organically together. For the connection of crossing social network system, the system will conduct "connect-glue" within the category of communication protocols.

However, connector based on correlation components and relation components should not only take attention to the characteristics of basic connector, but also override methods. We support viewpoint of the literature [6] establishing such connector explicitly. It can improve efficiency of component reuse in the process of software development. Using aspect-oriented connector thoughts of AC2-ADL [7], we define connector of a special relation overrided to correlation as following.

Definition 3 (*Override*) Override is a correlation connector that represents functional operation of connector between the relation and the correlation.

```
Override::=OverridCorrelationConnetorName
                [OverrideCorrelationConnectorDecPart]
          = =[ConnectorInterface]
               [WherePart]OverrideCorrelationConnectorDecPart
          = =[OverrideDecPart]
OverrideDecPart::=OverrideDecPartName[Override specification]
                     OverrideDecPart
               = =%PORT[port,...,port]
```

Port::=PortName[PortDec]= =[RelationDataType][OverridBehavior]

ConnectorInterface:: = ConnectorIntefaceName[InterfaceDec]==[Package] [Override]

OverrideBehavior is the restriction to superposition behavior, expressed by a XYZ/E unit. Specification of override describes how to connect correlation components and relation components together to interact with each other, and then combines the abstract conductions of components that participated in collaboration. It is also the behavior description of connector and described by a XYZ/E unit. Its definition is as follows:

%OVERRIDE= =InternalProcess Protocol= =[%InternalProcess[OperationList]]

OperationList contains three override methods.

According to [6], we can divide relation connector into two parts: simple connector and composite connector. Composite connectors are similar to simple connector in interface and similar to compound components in structure, and obey the expression of the XYZ/E channel including such two parts as data type declarations of channel and behavior description of channel.

XYZ/E is a linear temporal logic system and can express state transitions mechanism [5]. In the terms of the social network system, an important abstraction to users is the abstraction of interactive mechanism: Interactive components.

Definition 4 (*Iterative Connector*) Interactive connector is a kind of components for the users' interaction. According to the direction of transmission of correlation, it can be divided into three different types of interactive forms, such as one-way interaction, two-way interaction, multiway interaction. Interactive connector is limit to the correlation between users and users, interactive behavioral characteristics between resources and resources do not exist. The definition based on XYZ/E is as follows:

Interaction::=%Interaction InteractionName
= =[[% InteractionAttributes[InteractionDeclPart]]
  [%InternalProcess[OperationList]]]
  InteractionDeclPart= =Direction[%PORT[port,…,port]]
Direction ::=OneWayl TwoWayl Multiway
Port::=PortName[PortDec]= =[RelationDataType][%Roles[RoleList]]
RoleList::=Role {;Role}
Role::=RoleName= =Direction; Type;□[ConditionalElementList]

So far, relationship components can construct a kind of network system architecture based on fittings and interactive components, and it can further define Correlation network-Oriented Architecture, as follows.

Definition 5 (*Correlation Network-Oriented Architecture, CNA*) Correlation network-Oriented Architecture is a distributed-correlation framework of relation system through a combination of relation connector between multiple correlation role object, relation and interactive components. Among them, the correlation role object refers to different types of characters that participate in correlation calls and updates and it includes three categories as requester of correlation, broker of correlation and provider of correlation requester.

Definition of correlation network-oriented architecture is shown as following.

COA::=COA InstanceName
= =[[CorrelationRoleObject[Correlation]][CorrelationConnector][Interaction]] [WherePart]

These five basic concepts above are expanded based on temporal logic XYZ/E, and lay a foundation to build a kind of software architecture of crossing social network system.

180.4 CSN_ADL: Crossing Social Network Architecture

It is necessary to organize the resources crossing social network (CSN) and provide a kind of standard distributed network model. Every user is peer, meantime every social network system is peer too. According to above formal definition, relation is a structured unit of correlation. User is the end provider and also the requester of the relation or correlation. Once users of the requester and provider can make a kind of connection of correlation, it produces a binding between them. Realization model of CSN reflects technical frame of COA.

In the process of the connecting or extending existing relation, using push/pull mechanism about message to connect relation provider and broker services. Provider sends relation to broker by invoking request process and in advance push relation to social network system SNi. Broker does some relation overriding by invoking override process, including relation storage and classification, and by calling for Provider to complete a communication by invoking provide process to send recommendation results including the message of overriding relation.

Now, the link of binding process of relation between requester and provider is established, as well as the relation override, correlation extension, correlation (relation) recommendation. The architecture of correlation establishment and the expanding process of correlation are divided into the following three main processes. We can build a new architecture description language CSN_ADL, and formally describe three processes of CSN Fig. 180.2.

First, the whole crossing social network system architecture of the framework is defined as follows.

CSNA::=Cross-Social-Network-Architecture
= =[[CorrelationRoleObject[Correlation];] Connector] [WherePart]
CorrelationRoleObject::=%Role[Requester,Provider,Broker,{SNi}]
Connector::=%CONN[CorrelationConnector, Interaction]
InternalProcess::=%PROCESS [Push,Pull, Recommending,Binding,P2P RPC]
WherePart::=$\Box$(SafetyΛLivenessΛCorrectness)

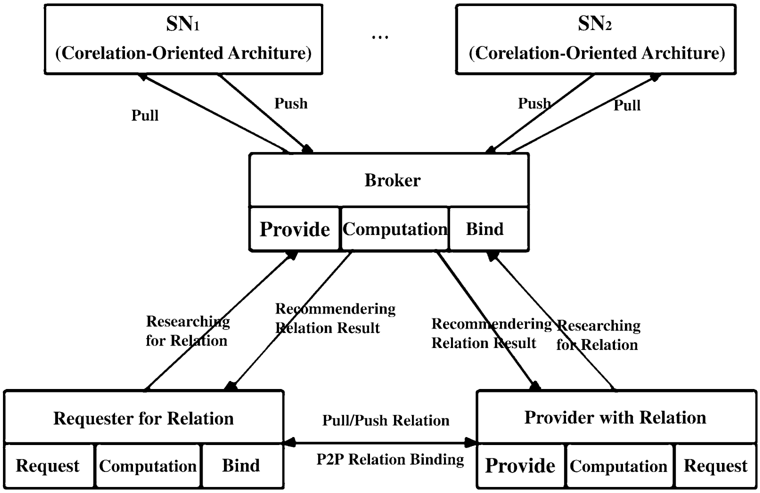


Fig. 180.2 Crossing social network system architecture based on CSN_ADL

In the meantime, Correlation Role Object, which act as a main body of the subject information communication within whole system structure encapsulates relations (correlation) and the corresponding override operation, and becomes a logical SN component equipped with the COA structure. For the main participants to the relation object, requester, provider, a broker and SNi, they can also be considered to be the four core logic components of CSNA. Under the conditions of safety, activity and correctness, correlation override connector and five interactive connectors (includes push, pull, recommending, binding, and P2P RPC) can organically be combined together and form a complete architecture.

180.5 Conclusions

Social network system is a new application based on Web2.0, and all social networks are independently distributed. One of the important differences between the crossing social network software architecture research and currently the most typical software architecture lies in its unique core components, such as relation components, correlation components, override correlation component and interaction components. This work is still going on our preliminary work of crossing social network architecture. In future, more research should be focused on developing the characteristics in heterogeneous distribution, and also paying attention to reuse characteristics to solve the current development problem of large complex crossing social network software.

Acknowledgment This work was supported in part by a grant from Fujian Province Science Foundation of China under the Grant 2009J01272, and the funding (type A) (JA09038) from the Education Department of Fujian Province.

References

1. Ofcom (2010) Social networking: a quantitative and qualitative research report into attitudes, behaviours, and use, http://www.ofcom.org.uk/advice/media_literacy/medlitpub/socialnetworking/report.pdf, 2008 (Last accessed 20 April 2010)
2. Xiao R, Xiong J (2009) An interest-based recommending framework of folksonomies, IEEE, ISA5
3. Kim J.-T, Lee J.-H, Lee H.-K, Paik E.-H (2009) Provision of the personalized social network service based on the locality/sociality relations, In: Proceedings of the 2009 fourth international conference on internet and web applications and services, pp 235–238
4. Hong M, Shen J-R (2006) Progress of research on software architecture[J]. J Softw 17(6):1257–1275 (in Chinese)
5. Tang ZS et al (2002) Temporal logic programming and software engineering. Science Press, Beijing[M] (in Chinese)
6. Zhu XY, Tang ZS (2003) A temporal logic-based software architecture description language XYZ/ADL[J]. J Softw 14(4):713–720 (in Chinese)
7. Shaw M, DeLine R, Klein DV, Ross TL, Young DM, Zelesnik G (1995) Abstractions for software architecture and tools to support them. IEEE Trans on Softw Eng 21(4):314–355
8. Medvidovic N, Mehta NR, Mikic-Rakic M (2002) A family of software architecture implementation frameworks. In: Proceedings of the 3rd IEEE/IFIP Conference on software architecture. Kluwer BV Press, Deventer, pp 221–235

Chapter 181
Decaying-Function-Based Cluster Algorithm of Sensed Data Stream for Wireless Sensor Network

Gao Feng, Yun Wu, Shangqiong Lu and Zhang Baiyu

Abstract Cluster analysis of sensed data stream is a hotpot in the field of wireless sensor network. Based on the characteristics of sensed data stream, some improvements are made in the cluster algorithm for data stream named CluStream, and a decaying-function- based cluster algorithm of sensed data stream for wireless sensor network named DFStream is proposed. DFStream partitions the data space into grids by the use of grid technology, then gets the dense grid using approximate method, lowers the weights of the outdated data through decaying function; and finds the clusters by depth-first search method. DFStream can discover the clusters of sensed data with arbitrary shape, and has high scalability of data flow and dimensionality of sensed data stream. The experimental results prove that, compared with CluStream, DFStream is more accurate, efficient and takes less memory.

Keywords Decaying function · Sensed data stream · Cluster algorithm · Wireless sensor network · Grid cell

G. Feng (✉) · S. Lu
Zhejiang A & F University, Lin'an 311300, China
e-mail: gaofeng@zafu.edu.cn

S. Lu
e-mail: zjcmxy@163.com

Y. Wu
Zhejiang University of Media and Communications, Hangzhou 310018, China
e-mail: lusq@zafu.edu.cn

Z. Baiyu
East China Normal University, Shanghai 200241, China
e-mail: ggzhangbaiyu@yahoo.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_181,

181.1 Introduction

Wireless sensor network (WSN) is a fully distributed system without central node. Numerous sensor nodes capable of communicating and computing are laid randomly in the monitored area, and through the layered network communication protocols and distributed algorithms, they rapidly get self-organized and construct intelligent sensor networks in the way of wireless communication. Sensor nodes have good collaboration ability, and through various transducers they integrated, they can automatically and in real time detect many physical phenomena of people's interest (such as temperature, humidity, light intensity and size/speed/direction of moving object) and deal with the information they receive before they send them back to the remote end users. All these features enable WSN to be used in a very wide range of fields, such as national defense, environmental monitoring, urban management, space exploration, intelligent agriculture, medical treatment, intelligent buildings, transportation, disaster warning and relief, warehousing/logistics management and manufacturing [1, 2]. All the typical application systems based on WSN, on the one hand, can realize a long-term, continuous, real–time, remote, automatic monitoring with the intelligent sensor networks constituted by a large number of sensor nodes, and on the other hand, would continuously generate a huge number of sensed data stream. This is because the monitoring scope and reliability of each sensor node are limited and thus requires sensor nodes to be laid dense enough to enhance the robustness of the entire network and accuracy of the monitored information, and sometimes even requires the monitoring scopes of several sensor nodes to be overlapped with each other, which causes information redundancy between adjacent nodes.

Sensed data stream is a new class of data objects (i.e. streaming data) featured by continuity, real-time, order, parallelism and rapid change [3]. It would be very difficult to use traditional data management techniques to manage and process sensed data flow. However, compared with traditional data management techniques, data stream clustering technology shows its great advantages in that. And therefore, the study on clustering algorithm of sensed data stream is much concerned by lots of many scholars at home and abroad. Currently, several clustering methods that can be used in sensed data stream are being proposed [3–15]. Of all these methods, CluStream algorithm is the most typical one [11], which provides a processing structure for analyzing data stream. There are two procedures for clustering data stream: the first one is online micro-cluster, which is to cluster the data stream for the first time; the second one offline macro-cluster, which is to analyze the clustering result of the first procedure as specifically requested by users. Though CluStream is very good, it still has some deficiencies: (1) Micro-cluster can have a satisfying result on clustering ball data stream, but as to other shapes (non-convex) of data stream, it cannot give good descriptions. (2) In the micro-cluster procedure, whether the data belongs to a certain subcluster depends on its distance from the center of the nearest subcluster and on the presupposed distance threshold. But as the subcluster is not strictly defined in space, its center would change with the arrival of every data, which lead the data close to the cluster center and does not necessarily belong to this cluster. This would result in two extreme cases: one the result may not be unique; the other is the intervals covered by each subcluster may overlap with each other. (3) High dimensional data cannot get clustered through micro-cluster procedure.

In order to solve the above-mentioned problems, decaying function based subspace cluster algorithm for high dimensional data streams (DFStream) is proposed in this chapter. The basic idea of this algorithm is as follows: first, considering the sensed data stream is dynamic, we use decaying function to reduce the weight of outdated data; second, adopt grid techniques to partition data space and estimate the statistical information of grid within a limited memory. Finally, use depth-first search method for clustering.

181.2 Related Concepts

Definition 1 Suppose $A = \{A_1, A_2, \cdots, A_k\}$ is the attribute set of Euclid space, $Sp = A_1 \times A_2 \times \cdots \times A_k$ is called $k$dimensional data space.

Definition 2 A sensed data stream *SDS* is defined as relation *R*, and its tuples $\bar{r}_i \in Sp(i = 1, 2, \cdots)$ comes in succession.

Definition 3 To partition all dimensions of sensed data stream into ξ intervals in average; as a result of that k dimensional data space is partitioned *Sp* into ξ^k independent hypercube units. Each hypercube unit is defined as a grid cell u.

Dynamic is a major feature of sensed data stream data. Through reducing the weight of outdated data with decaying function to gradually minimum the effect of its dynamic nature enables grid cell to better reflect the distribution situation of current data.

Definition 4 Named$d = a^{-(1/f)}(a > 1, f \geq 1, a^{-1} \leq d < 1)$ as decaying function.

Definition 5 Suppose when the tth data arrives, sensed data stream data would be $SDS = \{\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_t\}$ (in which data set belonging to the grid cell u would be $C_u = \{\bar{r}_{u1}, \bar{r}_{u2}, \ldots, \bar{r}_{uq}\}$) and the arriving order is $t_{u1}, t_{u2}, \ldots, t_{uq}$, then the $density(u)_t = count(u)_t / N_t$(in which $count(u)_t = \sum_{j=1}^{q} d^{t-t_{uj}}$, $N_t = \sum_{j=1}^{t} d^{t-j}$) would be the density of grid cell u in the moment the tth data arrives.

Definition 6 Given that the density threshold is τ and the error factor is ε, if$density(u) \geq \tau - \varepsilon$, it means that the grid cell u is dense; if $density(u) < (\tau - \varepsilon)$, it means u is not dense.

Definition 7 The largest set that connects dense grid cells is a class. Unit $u_1$and unit $u_2$are connected on condition that they have a common aspect or there exists a third unit joined $u_3$as their connecting media.

Property 1 Given that decaying function is

$$d = a^{-(1/f)}(a > 1, f \geq 1, a^{-1} \leq d < 1),$$

and when the tth sensed data arrives, the total weighting of data stream N_t is: $N_t = \begin{cases} 1 & t = 1 \\ N_{t-1} \times d + 1 & t \geq 2 \end{cases}$

With the increase of t, N_t would converge to $1/(1-d)$.

Proof When $t = 1$, then $N_1 = 1$ as there is no data in front needed to be decayed. When $t = 2$, then $N_2 = d + 1$. When $t > 2$, then

$$\begin{aligned} N_t = N_{t-1} \times d + 1 = (N_{t-2} \times d + 1) \times d + 1 = N_{t-2} \times d^2 + d + 1 \\ = d^{t-1} + \ldots + d^2 + d + 1 = (1 - d^t)/(1 - d) \end{aligned}$$

As $a^{-1} \leq d \leq 1$, so when t tends to infinity, N_t would converge to $1/(1-d)$.

According to Property 1, when t is big enough, N_t in definition 181.5 can be approximately represented as $1/(1-d)$.

181.3 DFStream Algorithm and its Structure

DFStream partitions data space with grid technology, maintains statistical information of grid cell online and dynamically and output clustering result offline on the request of users.

181.3.1 Maintain Statistical Information of Grid Cell On-Line

To maintain statistical information of grid cell dynamically within limited memory is an important part as well as difficult part of DFStream algorithm. The Grid Estimate of its subalgorithm is to estimate statistical information of grid cell approximately, realizing a compromise between space complexity and cluster accuracy. Since its space complexity has no correlation with the number of grid cells partitioned, Grid Estimate is good at improving the clustering accuracy and dimensional flexibility of sensed data stream. Also, to reduce the weight of outdated data on the basis of dynamic nature of sensed data stream can effectively reflect the changes of sensed data stream.

Based on the idea of Lossy Counting algorithm [16], Grid Estimate partitions sensed data stream several sensed data windows of constant-breadth and estimates every sensed data window, respectively. The sensed data stream can be represented as $SDS = ds_1 \cup ds_2 \cup \ldots \cup ds_h \cup \ldots$ (in which ds_h represents a sensed data stream window; the width of window can be $|ds_h| = w)$. This algorithm maintains data structure GEin the memory, and the structure of each tuple it saved is$(No, \overline{Count}, Latest, n)$(in which No is the identification of the corresponding grid cell u of the tuple, $\overline{Count}$ is the weighting estimation of number of data points included in u. $Latest$is the sequence number of the last arriving data point of u. n is the total weighting of data stream in the moment that GE is inserted in the tuple and the current window has not yet started). If n_{h-1}represents the total weighting of sensed data stream of the $(h-1)$th sensed data windows in the moment that the GE is inserted in the tuple at the hth time window, then $n_o = 0$.

The detailed algorithm is described as follows:

```
**Algorithm** Grid Estimate

**Input** $k$ dimension sensed data stream $SDS$; density threshold $\tau$; window size of sensed data stream $w$; error factor$\varepsilon$
**Output** $DenseGridSet$
Procedures
  **While** sensed data stream arrives $\{N = 0;\ M = 0;\ GE = \varphi;\ h = 1$
      **For** every datum $\bar{r}_i$ in $ds_h$/*$\bar{r}_i$ refers to the number $i$ sensed data*/
                  $\{N = N \times d + 1\,;\ M = M + 1;$
          **If** ($\bar{r}_i$in $GE$) $\{ge[\bar{r}_i].\overline{Count} = ge[\bar{r}_i].\overline{Count} \times d^{i-ge[\bar{r}_i].Latest} + 1;$
          /*$ge[\bar{r}_i]$refers to gain the corresponding tuple of$\bar{r}_i$from $GE$*/
                  $ge[\bar{r}_i].Latest = i;\}$
          **Else** {insert $\bar{r}_i$ into $GE$;
                  $ge[\bar{r}_i].\overline{Count} = 1;\ ge[\bar{r}_i].Latest = j;\ ge[\bar{r}_i].n = n_{h-1};\}$
          **If** $(M \bmod w = 0)\ \{n_h = N;$
              **For** every datum $\bar{r}$ in $GE\{$
                  **If** $(ge[\bar{r}].\overline{Count} + \varepsilon \times ge[\bar{r}].n \le \varepsilon \times n_h)\ \{$
                      delete $\bar{r}$ from $GE\}\}\}$
                  **If** requested output all $\bar{r}$ which
                  $ge[\bar{r}].\overline{Count} \times d^{M-ge[\bar{r}].Latest} \ge (\tau - \varepsilon) \times N;\}$

  $h = h + 1;\}$
```

Theorem 1 *When the* hth*window is full, tuple* $(No, \overline{Count}, T, n)$ *would be deleted if count* $\le \varepsilon N$, $h = 1, 2, 3, \ldots.$

Proof Inductive method is adopted. When$h = 1$, $count_{\omega} = \overline{Count}$, according to the deleting condition $\overline{Count} \le \varepsilon N_{\omega}$, we can get $count_{\omega} \le \varepsilon N_{\omega}$. Suppose when the $h$thwindow is full, tuple would be deleted if $count_{h\omega} \le \varepsilon N_{h\omega}$. Then we just need to prove that the tuple would be deleted when the lth $(l \ge h)$ window is full if $count_{l\omega} \le \varepsilon N_{l\omega}$. To prove that, we suppose that the tuple is deleted when the hth window is full, and is inserted at the mth$(h < m \le l)$ time window, and is deleted when the lth window is full. When the hth window is full, the total weighting of data stream would be $N_{h\omega} = (1 - d^{h\omega})\big/(1 - d)$. When the $(m - 1)$th window is full, the total data weighting would be $N_{(m-1)\omega} = (1 - d^{(m-1)\omega})\big/(1 - d)$. $count_{l\omega} = count_{h\omega} \times d^{(l-h)\omega} + \overline{Count}$. As we have supposed that $count_{h\omega} \le \varepsilon N_{h\omega}$, we can get that$count_{l\omega} \le \varepsilon N_{h\omega} \times d^{(l-h)w} + \overline{Count}$. And according to the deleting condition$\overline{Count} + \varepsilon N_{(m-1)\omega} \le \varepsilon N_{l\omega}$, if $N_{h\omega} \times d^{(l-h)\omega} \le N_{(m-1)\omega}$, then $d^{(l-h)\omega} - d^{l\omega} \le 1 - d^{(m-1)\omega}$, and thus the theorem has been proved. Actually, as $d^{(l-h)\omega} - d^{l\omega} = d^{(l-h)\omega}(1 - d^{h\omega})$, and $1 - d^{h\omega} < 1 - d^{(m-1)h}$ $(h < m \le l,\ 0 < d < 1)$, therefore, $d^{(l-h)\omega} - d^{l\omega} \le 1 - d^{(m-1)\omega}$. And finally the theorem got proved.

Theorem 2 *When the hth is full, if $(No,\ \overline{Count},\ T,\ n) \in GE$, then $\overline{Count} \le count_{h\omega} \le \overline{Count} + \varepsilon N_{h\omega}$ (in which $h = 1, 2, 3, \ldots$).*

Proof When $h = 1$, then $count_{\omega} = \overline{Count}$; when $h > 1$, then there would have two different situations: 1) if there exists no deletion among tuples, then $count_{h\omega} = \overline{Count}$. 2) if the tuple has already been deleted, then we might suppose the last deletion happened when the ith window was full and the tuple was inserted in the jth$(i < j \le h)$ window. Deducing from Theorem 1, we can get $count_{i\omega} \le \varepsilon N_{i\omega}$, and $count_{h\omega} = \overline{Count} + count_{i\omega} \times d^{(h-i)\omega} \le \overline{Count} + \varepsilon N_{h\omega}$, therefore, $\overline{Count} \le count_{h\omega} \le \overline{Count} + \varepsilon N_{h\omega}$.

Deduction 1 No matter when, as long as $(No,\ \overline{Count},\ T,\ n) \in GE$, then $\overline{Count} \le count \le \overline{Count} + \varepsilon N$.

Theorem 3 *Given density threshold τ and error factor ε, the output of Grid_Estimate satisfies the following conditions*:

(1) *cell u of all $density(u) \ge \tau$ are outputted*;
(2) *cell u of all $denstiy(u) < (\tau - \varepsilon)$ are not outputted*;
(3) *$denstiy(u) - \overline{denstiy(u)} \le \varepsilon$, $\overline{denstiy(u)}$ refers to the estimated value of density.*

Proof All the tuples outputted by Grid Estimate satisfy $\overline{Count} \ge (\tau - \varepsilon)N$, and $\overline{Count} \le count \le \overline{Count} + \varepsilon N$ which we can infer from Deduction 181.1, then we can get that $count \ge (\tau - \varepsilon)N$ and $count - \overline{Count} \le \varepsilon N$. In addition, defines $denstiy(u) = \frac{count(u)}{N}$, then we can deduce that $density(u) \ge (\tau - \varepsilon)N$ and $denstiy(u) - \overline{denstiy(u)} \le \varepsilon$. The theorem has got proved.

Property 2 [11] *Given the error factor* ε, *the number of tuples being kept in the data structure GE can be at most* $\frac{1}{\varepsilon}\log(\varepsilon N)$.

According to Theorem 181.3, we can infer that Grid Estimate can guarantee a relatively high accuracy for the result of output; while according to Property 181.2, we can know that the space complexity of Grid Estimate has no correlation with the number of grid cells partitioned, which is good for this algorithm to enhance its clustering accuracy and scalability of data stream.

181.3.2 OffLine Clustering

The result of offline cluster is relatively simple. Offline cluster adopts depth-first search algorithm method to connect Grid Estimate to get closely connected grid cells and output these cells in DNF form. All the initial dense grid cells are undefined (Undef), referring that they are not processed. The algorithm is described as follows:

```
Algorithm:   Offline Clustering
Input:       $DenseGridSet$; data stream dimension $k$;
Output:      $ClusterResult.$
Procedure:
      $Num = 0;$
While $(u \in DenseGridSet)$ and $(u$ is Undef) $\{Num = Num + 1;$
    $denfinecluster(u, Num);$/*$u$ belongs to the number$Num$ cluster
    For $(j = 1; j<k; j++)$ $\{u^l = \{[l_i, h_i), \ldots, [(l_j^l), (h_j^l)), \ldots, [l_k, h_k);$ $\{;$
    //check the left neighbor of $u$on dimension$A_j$.
    If $\left(u^l \in DenseGridSet\right)$ and $(u^l$ is Undef)
        $denfinecluster(u^l, Num);$
        $u^r = \{[l_i, h_i), \ldots, [(l_j^r), (h_j^r)), \ldots, [l_k, h_k);$
        //check the right neighbor of $u$ on dimension $A_j$.
    If $(u^r \in DenseGridSet)$ and $(u^r$ is Undef)
       $denfinecluster(u^r, Num);\}\}$
```

181.4 Analyses of Experimental Result and Performance of the Algorithm

This section is for testing the performance of DFStream algorithm. The experiment is to be conducted in the operating system of Windows 2000 Professional with 512 MB of main memory and 2.0 GHz of CPU. There are two kinds of data in the experiment: the first one is real data set, which comes from wireless sensor networks- based monitoring system for crop water regime (i.e. WSN-CWSM system) [1]. The data set, recorded as Greenhouse Environment data set, contains 581,012 data records, with each data consisting of 30 property fields; the second one is simulation data set. In this paper, "B", "C" and "D" stand for size of the set, cluster number of the set contains and data space dimensionality, respectively. For example, B100KC10D50 means that space dimensionality is 50, the size is 100 K and it is a sensed data set that contains 10 clusters. In the experiment, we consider $\varepsilon = 0.05\tau$ and $\xi = 15$, and as to the parameter of time span in the CluStream algorithm, we consider $h = 10$.

(1) **Accuracy of the algorithm**. In the experiment, the algorithm accuracy is defined as the proportion that the number of correctly clustered data takes in the whole sensed data set. Fig. 181.1 and Fig. 181.2 show the contrast of the accuracy of DFStream and CluStream in terms of real data set and simulate data set. DFStream, as it adopts grid technology, has advantage in clustering data of various shape, and therefore, it is more accurate than CluStream. In Fig. 181.1 and Fig. 181.2, $\tau = 0.005$.

(2) **Processing efficiency**. Figure 181.3 shows the contrast of processing efficiency between DFStream and CluStream. CluStream would frequently carry out such operations as micro-cluster, addition and deletion, which greatly lowers its processing speed; while DFStream, as it adopts grid technology, will simply calculate the density of grid and thus has high processing efficiency. In Fig. 181.3, we consider $\tau = 0.005$.

(3) **Memory usage**. Figure 181.4 shows the memory usage of simulation data set B300KC5D30 in DFStream. The necessary memory of this algorithm is kept in a relative small range, so when the class mode of sensed data stream is relatively stable, the memory it used also tends to be stable. The higher the density threshold is, the less the grid cells it needs and the less the memory it requires.

(4) **Scalability**. Figures 181.5 and 181.6 shows the scalability of data space dimensionality and the number of cluster it contains in DFStream algorithm. Since it adopts approximate estimation, this algorithm has good scalability. In Fig 181.5 and 181.6, we consider $\tau = 0.01$.

Fig. 181.1 Accuracy comparison (Stream speed = 100/s,Greenhouse Environment)

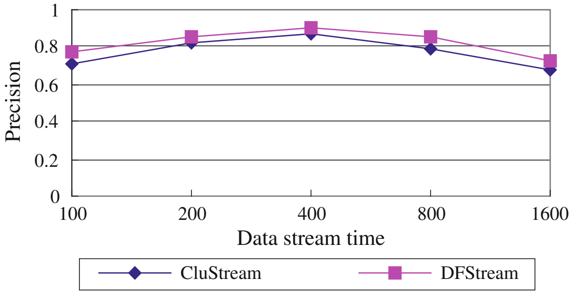


Fig. 181.2 Accuracy comparison (Stream speed = 400, B100kC10D50)

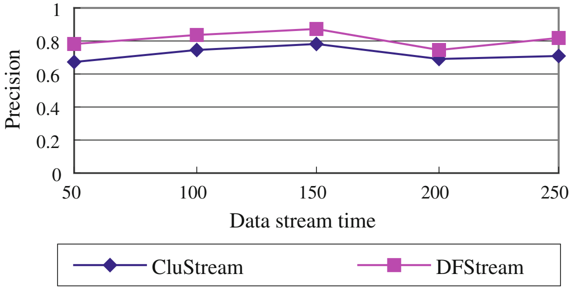


Fig. 181.3 Comparison of the Stream Processing Rate (Stream_speed = 200, Greenhouse Environment)

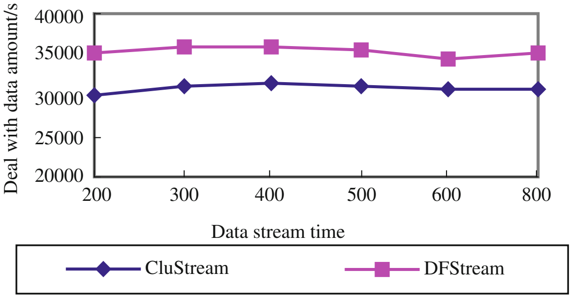
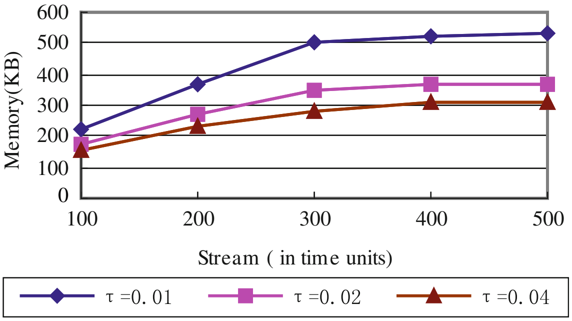


Fig. 181.4 Memory Usage (Stream_speed = 400)



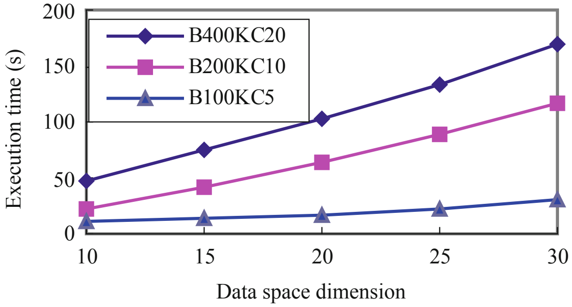


Fig. 181.5 Scalability with Data Dimensionality

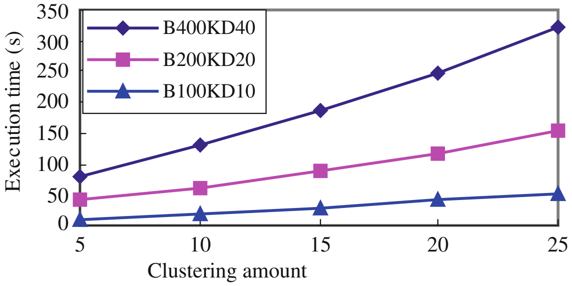


Fig 181.6 Scalability with Number of Clusters

181.5 Conclusion

This chapter studies the cluster problem of sensed data stream for wireless sensor network. In this chapter, DFStream is presented on the basis of the characters of sensed data stream. Through approximately estimating and time decaying grid statistical information, DFStream realized to cluster sensed data stream of various shapes, and has relatively good scalability of data quantity and dimensionality. The experiment proved that this algorithm is practicable and effective.

References

1. Gao F, Yu L, Zhang W, Xu Q, Yu L (2009) Research and design of crop water status monitoring system based on wireless sensor networks. Trans CSAE 25(2):107–112
2. Gao F, Lu S, Xu Q, Jiang Q (2010) Wireless sensor networks and its application in facility agriculture. J Zhejiang Coll 27(5):762–769
3. Zhu W, Yin J, Xie Y (2006) Arbitrary shape cluster algorithm for clustering data stream. J Softw 17(3):379–387
4. Sun Y, Lu Y (2006) A grid-based subspace clustering algorithm for high-dimensional data streams. Lect Notes Comput Sci 4256(2006):37–48
5. Gao Y, Huang Y (2008) A grid and density-based clustering algorithm for processing data stream. Comput Sci 35(2):134–137

6. Zheng Y, Ni Z, Wu S, Wang L (2009) Data stream cluster algorithm based on mobile grid and density. Comput Eng Appl 45(8):129–131
7. Chen H, Shi B, Qian J, Chen Y (2010) Wavelet synopsis based clustering of parallel data streams. J Softw 21(4):644–658
8. Shi J, Hu X (2010) Conceptual clustering based on data streams. Comput Eng 36(9):62–64
9. Ouyang Z, Luo J, Hu D, Wu Q (2010) An ensemble classifier framework for mining imbalanced data streams. Acta Electronica Sinica 38(1):184–189
10. Guha S, Meyerson A, Mishra N, Motwani R, O'Callaghan L (2003) Clustering data streams: theory and practice. IEEE Trans Knowl Data Eng 15(3):515–528
11. Aggarwal C, Han J, Wang J, Yu P (2003) A framework for clustering evolving data streams. In: Proceedings of the 29th international conference on very large data bases, Berlin, Germany, pp 81–92
12. Parsons L, Haque E, Liu H (2004) Subspace clustering for high dimensional data: a review. ACM SIGKDD Explor Newsl 6(1):90–105
13. Hou G, Yao R, Ren J, Hu C (2010) A clustering algorithm based on matrix over high dimensional data stream. Lect Notes Comput Sci 6318(2010):86–94
14. Ning H, Xu W, Zhou Y, Gong Y, Huang TS (2010) A general framework to detect unsafe system states from multisensor data stream. IEEE Trans Intell Transp Syst 11(1):4–15
15. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recogn Lett 31(8):651–666
16. Manku GS, Motwani R (2002) Approximate frequency counts over data streams. In: Proceedings of the 28th international conference on very large data bases, Hong Kong, China, pp 346–357

Chapter 182
A Genetic Algorithm for Channel Allocation Problem in Cognitive Radio Wireless Mesh Network

Jie Jia, Yanyan Li, Qufei Zhu, Zhaoyang Zhang and Jian Chen

Abstract In order to improve the communication reliability of the cognitive radio mesh network and increase spectrum utilization, a spectrum allocation algorithm based on genetic algorithm in consideration of the power control is proposed in this paper. A network utility model is built for joint power and channel allocation, and a corresponding encoding rule is designed. To ensure the validity of bodies and the fast convergence to the best solution, crossover, mutation operator and control mechanism are designed separately. Extensive simulation results are presented to verify this approach.

Keywords Cognitive radio · Wireless mesh network · Channel allocation · Power control · Genetic algorithm

182.1 Introduction

Wireless mesh network (WMN) is suitable for broadband wireless backbone transmission environment and can satisfy the “last mile” broadband access requirements with low cost. As the number of users increases, the demands for the quality of service are much higher. However, the spectrum resources are finite. It becomes one of the serious problems in wireless mesh network. Cognitive radio (CR) technology lets users sense and utilize available spectrum opportunistically. So it is the tendency to apply CR technology to WMN. In cognitive radio mesh

J. Jia (✉) · Y. Li · Q. Zhu · Z. Zhang · J. Chen
College of Information Science and Engineering, Northeastern University,
Shenyang 110000, China
e-mail: jiajie@ise.neu.edu.cn

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_182,

network, mesh nodes use CR technology to sense the unused spectrum of authorization system, and to access to the spectrum dynamically.

Spectrum allocation, which is the key technology of cognitive radio networks, affects the network utility directly. In consideration of the open features of spectrum, different communication channels must be assigned for the cognitive users in interference scopes to improve the reliability of the transmission. A color-sensitive graph coloring (CSGC) algorithm was proposed in [1] based on a graph-theoretic model, aiming at maximizing system effective bandwidth. With the network scalable increased, the computing quantity of CSGC was too large and the allocation time was too long. On account of these, a distributed local bargaining allocation algorithm was presented in [2], which compensated for the local small changes of the network and could complete channel allocation decisions quickly. But this local distribution would lead to deviating from the global optimal solution ultimately. A dynamic spectrum allocation algorithm was put forward in [3] on the basis of the spectrum shared pool strategy, in which the spectrum resources were integrated into a public subchannel spectrum pool to maximize spectrum efficiency. Although it took the price into account and spectrum efficiency as the design criteria, it lacked of flexibility. A dynamic spectrum allocation algorithm was proposed in [4], in which cognitive users could adjust their own strategies repeatedly according to spectrum demands. An asymmetric Nash consultation performance function was designed in [5] based on cooperative game theory, which allocated idle spectrum resources through the negotiation among users to maximize system throughput. In [6], the auction mechanism of game theory was applied to distribute spectrum sharing, and it obtained the Nash equilibrium.

As communication power has direct influence on the interference scope, power control should be considered in channel allocation to satisfy the communication demand of more users at the same time. A joint power and channel allocation algorithm was proposed in [7] based on dynamic interference graph. Combined with power control based on a graph-theoretic model, a downlink joint spectrum algorithm and a power assignment algorithm were presented in [8]. However, these two spectrum allocation algorithms did not take the spectrum utilization, synthetic interference and fairness into consideration at the same time. A distributed pricing approach for power and channel allocation was proposed in [9] based on a non-cooperative game model. CR users implemented price-based iterative water-filling (PIWF) algorithm to repeatedly negotiate their best transmission powers and spectrum, and finally it reached a Nash Equilibrium (NE). A potential game algorithm for joint channel and power allocation was proposed in [10]. Although it achieved good system throughput, the cost of message transmission was too high. Meanwhile, it had the risk of sinking into local optimal solution plane.

However, the existing joint power control and spectrum allocation algorithms have some problems, such as the allocation time is too long, or it is easy to fall into local optimum plane. In this paper, the users' interference, energy consumption, hidden receiving terminal and hidden sending terminal problems are considered. Aiming at maximizing system capacity, a genetic algorithm is proposed to solve this problem.

The remainder of the paper is organized as follows: in Sect. 182.2, we firstly introduce the research achievement on spectrum allocation, and then describe the network model for joint power and channel allocation. In Sect. 182.3, we describe the implementation details of the proposed genetic algorithm. In Sect. 182.4, we conduct a number of simulation experiments and provide simulation results. Finally, we draw conclusions in Sect. 182.5.

182.2 Network Model

WMN is usually used as broadband wireless access network. Assume that the whole network generally exists in the tree topology with mesh access router acting as the root. Every node in the mesh network is equipped with CR equipment, and can sense the available idle channels intelligently.

WMN is presented as graph $G = \{P, S, E\}$, in which P is the primary users set, S is the secondary users set and E is the set of edges between secondary users. Each primary user p_i in P has a corresponding coverage area, in which the user itself is the center and R_{pi} is the radius of the circle coverage. While each secondary user s_i in S has a corresponding circle interference area centering on itself, and I_{si} is the radius. A secondary user can use the same channel with a primary user when its interference range does not overlap the primary user's coverage area. Each secondary user s_i in S has a corresponding circle communication region centering on itself, and T_{si} is the radius. Only when user A is in the communication region of user B, A can receive the message sent by B. As for a link between secondary user pair (i_{SS}, i_{RS}), the receiver point i_{SS} must be located in the communication range of the sender point i_{SS} to ensure the link's validity. With regard of the link j between secondary user pair (j_{SS}, j_{RS}), if j_{RS} is in the interference range of i_{SS} or i_{RS} is in the interference range of j_{SS}, link i and j will interfere with each other, and they cannot use the same spectrum simultaneously.

In order to describe the system better, Fig. 182.1 gives a cognitive radio mesh network topology. There are 3 primary users p_i(i=1,2,3) and 6 secondary users s_i(i=1,2,3,4,5,6) deployed in the area randomly, and the available spectrum set of the system is C=$\{c_1,c_2,c_3,c_4,c_5\}$. The dotted line circles in Fig. 182.1 represent the coverage regions of primary users, and the solid line circles represent the communication ranges of secondary users. Assume that the interference radius is equal to the communication radius of every secondary user, so the interference range and communication range are the same. We can see that the interference range of secondary user s_2 overlaps with the coverage regions of primary p_1 and p_2. Therefore, s_2 can only use spectrums $\{c_3,c_4,c_5\}$. As the interference range of secondary user s_1 does not overlap with any primary users' coverage regions, spectrums $\{c_1,c_2,c_3,c_4,c_5\}$ are available to s_1. Moreover, secondary user s_1 is in the transmission range of s_2. So there is a link between s_1 and s_2, and the available spectrums are $\{c_3,c_4,c_5\}$.

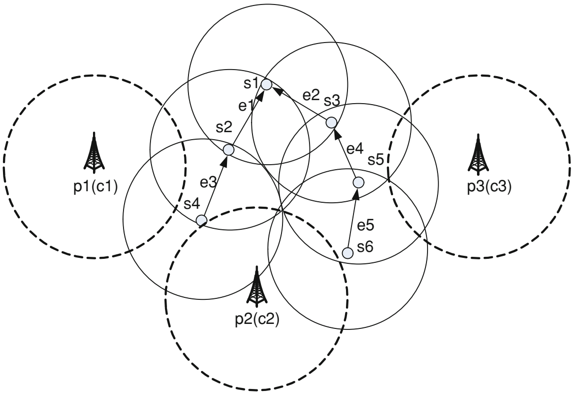


Fig. 182.1 Cognitive radio mesh network topology

Given an edge i in E, the channel gain $g_i = d_i^{-\gamma}$ can be defined according to the edge's length d_i, namely the distance from sender to receiver. Here, γ is the path loss index. If the received power p_i^R exceeds a threshold α, the transmission is successful. It is assumed that the sender's power is p_i^S, so the received power is,

$$p_i^R = p_i^S \times g_i. \tag{182.1}$$

Thus, the sender's power assigned to link i must satisfy $p_i^S \times g_i \geq \alpha$. Therefore, the transmission distance of the sender i_{SS} with power p_i^S, $T_i(p_i^S)$ can be obtained as,

$$T_i\left(p_i^S\right) = \left(\frac{p_i^S}{\alpha}\right)^{1/\gamma}. \tag{182.2}$$

When the transmission is successful, the minimum transmission power of link i is defined as p_i^{min}. In order to restrict the interference of secondary users on primary users, the maximum transmission power is defined as P_{max}. According to the above formulas, the minimum transmission power can be calculated by,

$$P_i^{\min} = \left(\frac{d_i}{T_i(p_i^S)}\right)^{\gamma} \times P_{\max} = \frac{\alpha}{d_i^{-\gamma}}. \tag{182.3}$$

Hence, the transmission power p_i^s assigned to edge i must satisfy$P_i^{\min} \leq p_i^S \leq P_{\max}$. Assume that the interference distance is only related to transmission power, distance and the path loss index. The interference from node i_{SS} with power p_i^s can be negligible when it exceeds a certain threshold β, then the interference distance is,

$$I_i\left(p_i^S\right) = \left(\frac{p_i^S}{\beta}\right)^{1/\gamma}. \tag{182.4}$$

When a node uses the maximum power P_{max}, the corresponding maximum transmission distance is T_{max} and the interference distance is I_{max}. The actual

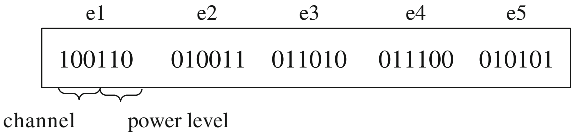


Fig. 182.2 Coding rule of an individual

transmission power is divided into a finite number of levels Q equably. Assume that there are C channels not interfering with each other, and link i on the nearby links will interfere with them. Hence, the network capacity is defined as:

$$V = \sum_{i \in E} \frac{1}{1 + N_i(c_i, q_i)} H_{ci} \log_2 \left(1 + \frac{p_i^R}{P_N} \right). \tag{182.5}$$

where c_i is the channel chosen by link i and q_i is the power level, $c_i \in C$, $q_i \in Q$. And $N_i(c_i, q_i)$ is the number of links which interfere with link i, H_{ci} is the bandwidth of channel c_i, p_i^R is the received power of node i_{RS}, and P_N is the noise power.

182.3 Algorithm Design and Implementation

The joint power and channel allocation problem of cognitive radio mesh network is a NP problem synthesizing many aspects to search for a certain objective solution. Genetic algorithm (GA) is a random search algorithm based on biological natural selection and genetic mechanism. Recently, GAs are recognized to be well qualified to tackle multi-objective optimization problems.

When GA is adopted to solve the joint power and channel allocation problem, the initial population is formed from a set of initial solutions generated randomly. Because GA cannot deal with parameters of the problem space directly, a joint binary coding scheme is presented in this paper. There are C channels having no interference with each other in the network, so $n_1 = \lceil \log_2 C \rceil$ binary bits are needed for the channel selection. Transmission power is divided into Q levels, so $n_2 = \lceil \log_2 Q \rceil$ binary bits are needed for the power level selection. Therefore, each link needs $n_1 + n_2$ binary bits. We take Fig. 182.1 for example, there are totally six secondary users, five links $\{e_1, e_2, e_3, e_4, e_5\}$, five available channels $\{c_1, c_2, c_3, c_4, c_5\}$ and eight power levels $\{q_1, q_2, .., q_8\}$. Each link uses six binary bits, the former three bits represent the channel, and the latter three bits represent the sending power level. Therefore, the code of an individual can be expressed as Fig. 182.2.

Crossover operator plays a kernel role in GA. The circulation single-point crossover operator is adopted in this paper. There are N individuals in a population, $2i-1$ and $2i$ $(1 \leq i \leq N/2)$ individuals are selected and a crossover point is set randomly. When executing crossover operation, the part codes of two bodies behind the crossover point are exchanged. Then two new individuals are generated. The operation of single-point crossover is shown in Fig. 182.3.

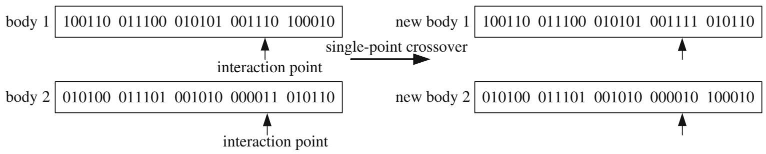


Fig. 182.3 Single-point crossover

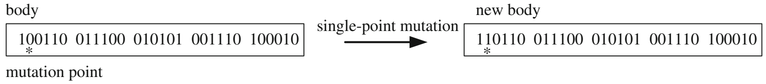


Fig. 182.4 Single-point mutation

Mutation operator provides opportunities for new individuals, and avoids sinking into the local optimal plane plight. The single-point mutation is used in this paper. The operation of basic mutation is shown in Fig. 182.4.

The fitness function is used to judge the merit degree of individuals. In order to improve spectrum utilization, ensure the fairness and make more users access the network. The network capacity in, 182.5 is defined as objective function. After crossover or mutation operation, the merit degree of the new produced individuals is judged. Henceforth, the better bodies are selected as the new farther bodies in next generation.

The algorithm terminates when the end condition is satisfied. Then we select the optimal individual from the last generation as the final allocation result.

When using GA to solve the joint power and channel allocation problem, all the nodes in the cognitive radio mesh network notify the information (such as sites, available channels, and so on) to the central controller (AP). According to the number of given links, the central controller firstly searches links. Then the central controller runs GA to solve the problem using the information of the nodes and links. And it distributes the allocation results to the responding sender nodes and receiver nodes.

182.4 Simulation

Matlab 7.5 is adopted as simulation platform. Three experiments are done in a square region of 2400 m×2400 m with different nodes deployments. There are 10 available channels. The maximum transmission power is set as 20*dBm* and the transmission power levels Q is set as 16. Under the maximum transmission power, the maximum transmission and interference distances are 250 and 500 m, respectively. The path loss index γ is set as 4. For simplicity, the bandwidth of each channel H_{c_i} is the same and normalized to 1 unit. In order to ensure that the

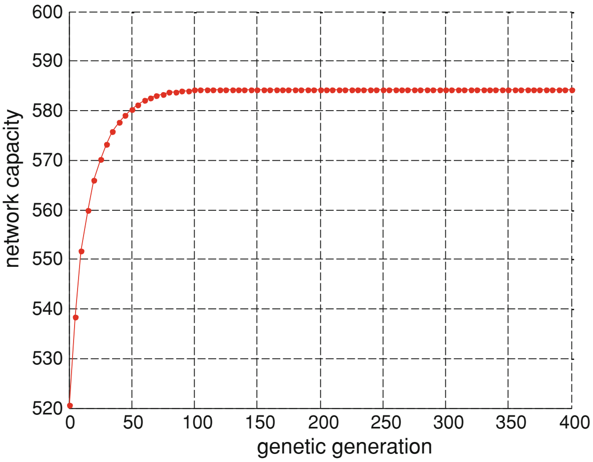


Fig. 182.5 Relationship between network capacity and the genetic generation

signal to noise radio SNR reaches 10 dB (received power threshold $\alpha = P_{\max} \times T_{\max}^{-\gamma}$, $\alpha/P_N = 10$) when communication is successful, the noise power is set as $P_N\ = -85.9\ dBm$.

In the first group of experiments, 200 secondary users with the same initial energy are distributed randomly in a square region of 2400 m × 2400 m. We take the situation with 100 links for illustration. The variation relationship between network capacity and the genetic generation is shown in Fig. 182.5.

From Fig. 182.5, we can see that the network capacity of WMN increases with the genetic process proceeding, and it almost remains constant after 100 generations.

In the second group of experiments, separately, 200 and 300 secondary users are distributed randomly to get the relationship between network capacity and the number of links. The simulation results are shown in Fig. 182.6.

In Fig. 182.6, MNC is the maximum network capacity of 200 instances, and ANC is the average one. It can be seen that the network capacity increases with more links, while the growing speed is becoming slow. When there are 300 secondary users, the network capacity is bigger because the choice space is enlarged.

In the third group of experiments, we deploy 200 and 300 secondary users, respectively in the network to know when the algorithm can reach stable situation with different number of links. The simulation results are shown in Fig. 182.7.

In Fig. 182.7, MSG and ASG are separately the maximum and average stable generation of 200 instances. It can be seen that the stable generation grows when the number of links increases. Further, the average stable generations of the two situations are almost the same. So we can come to the conclusion that the stable generation has less effect on the number of secondary users, while it is closely related to the number of links.

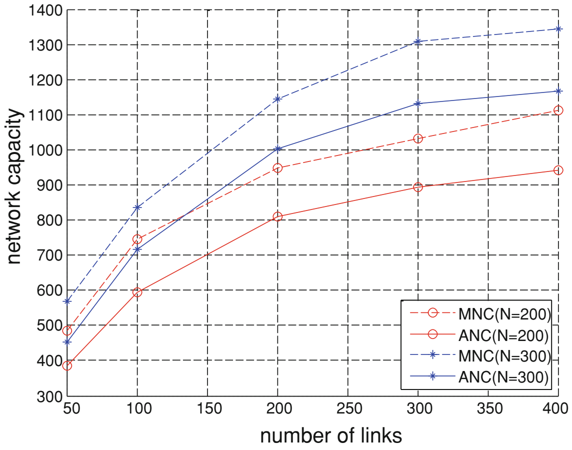


Fig. 182.6 Relationship between network capacity and the number of links

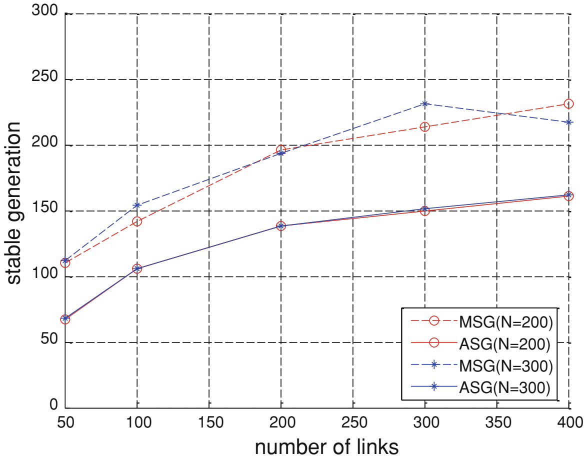


Fig. 182.7 Relationship between stable generation and the number of links

182.5 Conclusion

Channel allocation is an important problem in wireless mesh network. In this paper, a genetic algorithm is proposed to solve the joint power and channel allocation problem. Combined with the feature of cognitive radio mesh network, a network model is established. The encoding rule, crossover and mutation operator and corresponding control mechanism of genetic algorithm are designed. The experimental results show that the proposed algorithm can achieve optimal network capacity and provide better adaptability to cognitive radio network.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant Nos. 60903159, 61070162, 71071028, 60802023 and 70931001; the Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20070145017; the Fundamental Research Funds for the Central Universities under Grant No. N090504003 and No. N090504006.

References

1. Zheng HT, Peng CY (2005) Collaboration and fairness in opportunistic spectrum access. IEEE Int Conf Commun 5:3132–3136
2. Buddhikot MM, Kolodzy P, Miller S et al (2005) DIMSUMNet, New directions in wireless networking using coordinated dynamic spectrum access. In: 6th WoWMOM, pp. 78–85
3. Si P, Ji H, Richard FYu, Leung VCM (2010) Optimal cooperative internetwork spectrum sharing for cognitive radio systems with spectrum pooling. IEEE Trans Vehicular Technol 59(4):1760–1768
4. Mark F, Mario C, Jean-Pierre H (2009) Efficient mac in cognitive radio systems: a game-theoretic approach. IEEE Trans Wirel Comm 8(4):1984–1995
5. Tian F Yang Z (2007) A new algorithm for weighted proportional fairness based spectrum allocation of cognitive radios. In: 8th ACIS International conference on software engineering,artificial intelligence,networking,and parallel/distributed computing, pp 531–536
6. Bae J, Beigman E, Berry R (2008) Sequential bandwidth and power auction for distributed spectrum sharing. IEEE Sel Areas Comm 26(7):1193–1203
7. Hoang AT, Liang YC (2006) Maximizing spectrum utilization of cognitive radio networks using channel allocation and power control. IEEE Veh Technol Conf, pp 1–5
8. Hoang AT, LIANG YC (2008) Downlink channel assignment and power control for cognitive radio networks. IEEE Trans Wirel Comm 7(8):3106–3117
9. Wang Fan, Krunz Marwan, Cui Shuguang (2008) Price-based spectrum management in cognitive radio networks. IEEE J Sel Top Signal Process 2(1):74–87
10. Canales M, Gallego JR (2010) Potential game for joint channel and power allocation in cognitive radio networks. Electr Lett 46(24):1632–1634

Chapter 183
A Heuristic Algorithm for Link Scheduling with Different Slot Demands in Wireless Mesh Networks

Jian Chen, Jie Jia, Yingyou Wen and Dazhe Zhao

Abstract Link scheduling plays an important role for the performance of TDMA-based wireless mesh networks. In order to maximize the entire network throughput, theoretically the interference-aware link scheduling model is analyzed. As for the link list with fixed sequences and different slot demands, a heuristic algorithm based on expanded graph model is proposed. Simulation results demonstrate that the proposed scheme can converge to the optimal schedule length more rapidly and efficiently, thus having a better transfer efficiency and a lower implementation complexity than most existing algorithms.

Keywords Wireless mesh network · Link scheduling · Slot demand · Heuristic algorithm

183.1 Introduction

Wireless mesh network is a multi-hop wireless network consisting of a large number of wireless nodes, some of which are called gateway nodes and connected with a wired network [1]. It has attracted much research attention with its potential applications, including last-mile broadband Internet access, neighborhood gaming, VoD (Video-on-Demand), distributed file backup, video surveillance and so on.

J. Chen (✉) · J. Jia · Y. Wen · D. Zhao
School of Information Science and Engineering, Northeastern University,
Shenyang 110006, China
e-mail: chenjian_2002cn@163.com

Y. Wen · D. Zhao
Key Laboratory of Medical Image Computing of Ministry of Education,
Northeastern University, Shenyang 110006, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_183,

Due to the limited channel capacity, the influence of interference, the large number of users and the emergence of real-time multimedia applications, supporting guaranteed quality-of-service (QoS) has become one of the key issues in wireless mesh networks.

The new multi-hop MAC protocols based on time division multiple access (TDMA), such as the 802.16 mesh protocol [2] and the 802.11s mesh deterministic access (MDA) protocol [3], can provide guaranteed link bandwidth with scheduled access to the wireless channel with connection-oriented services. Each link should be assigned a set of time slots $\subset [0, T]$ on which it will transmit, where T is the scheduling period. As the number of bytes transmitted per time slot is only related with the modulation, and the transmission quality is mainly influenced by the network interference, a interference-aware schedule with the minimum scheduling period should be found, thus to improve the overall network throughput.

If a scheduled transmission on a link $a \to b$ does not result in a collision with the conflicting links, the schedule is interference-aware. The interference-aware link scheduling has received a great attention from both networking and theory fields in the past few years [4, 5]. As the wireless conflicts can be modeled with interference (conflict) graphs [4], the work using graph coloring is presented based on the conflict graph and finding cliques in the complement of the compatibility graph [5]. For the relay characteristics in WMN, each router has to transmit its own packets as well as to forward those of its children. Thus, each node with the same packet arriving rate is impossible. For this reason, the above-mentioned scheduling mechanism with the same slot demand cannot be directly applied to WMN environment. Some literatures have studied the joint routing and link scheduling problem in WMN, and proposed some cross-layer design mechanisms for link scheduling [6]. However, as the link selection criteria will influence the scheduling performance greatly, these algorithms, assuming routers far from BS nodes first in slot allocation, have a limited spatial reuse [7]. To this problem, some link selection criteria, such as minimum interference degree priority [7], nearest hop first [8] and largest hop first [9] have been proposed for link scheduling. All of these algorithms are aiming at maximizing the concurrent throughput of end-to-end flows. However, in order to reduce the synchronization complexities, each link can transmit only once during a scheduling period in these algorithms, which will require a longer scheduling period to transmit the same traffic flow, and cause the decline of network throughput eventually.

In this chapter, a heuristic algorithm is devised to find the near-optimal scheduling periods. By importing the concept of node decomposition, the modified heuristic algorithm can also adapt to WMN environment with relay modes. The remainder of the chapter is organized as follows: in Sect. 183.2, we introduce the network model and the interference model. In Sect. 183.3, we describe the problem in this chapter. In Sect. 183.4, we describe the implementation details of the proposed algorithm. In Sect. 183.5, we conduct a number of simulation experiments and provide simulation results. Finally, we draw conclusions in Sect. 183.6.

183.2 System Model

183.2.1 Network Model

Consider a wireless mesh network consisting of n static mesh routers (MRs). If two nodes are in the range of each other, they will establish links in MAC layer. So wireless mesh network can be represented as a directed connectivity graph $G(V, E)$, where vertices $V = \{v_1, v_2, \ldots, v_n\}$ represents the MRs and edges $E = \{e_1, e_2,\ldots, e_m\}$ represents the links. In particular, an edge $e = (u, v)$ means that the traffic on the link e is transmitted from u to v, $\|u - v\| \leq R_c$. Assume that some routes between nodes are established by a routing protocol, and a routing tree is formed with all the paths using a subset of the links.

Since wireless mesh network is a stop-and-go queuing system, each link is equivalent to a server in stop-and-go queuing. Assuming that each mesh nodes has the initial rate demand $G = \{g_1 ,\ldots, g_n\}$, we can obtain final rate demand $\widehat{r}_i$ for each node i by the routing tree and the routing algorithm.

$$\widehat{r}_i = g_i + \sum_{j \in \mathrm{child}(i)} g_j \tag{183.1}$$

Similarly, for any link $e_j \in E$, the rate demand $\widehat{r}_{e_j}$ is,

$$\widehat{r}_{e_j} = \sum_{p_l \in p} g_l I\big(e_j \in P_l\big) \tag{183.2}$$

where P is the set of all paths found by the routing algorithm, $I(.)$ is the indicator function. If $I(.) = 1$, its argument is true. If $I(.) = 0$, its argument is false. And g_l is the requested end-to-end rate of routing path p_l, which is equal to the initial bandwidth of the source node.

For link e_j, we use r_{e_j} to denote the corresponding time slot demand of $\widehat{r}_{e_j}$. Assuming T is the scheduling period and b_j is the number of bits in each time slot, then we have

$$r_{e_j} = \left\lceil \frac{r_{e_j} T}{b_j} \right\rceil \tag{183.3}$$

Calculating the time slots demand of each link is an important prerequisite in link scheduling algorithm. The slot demand and the specific slot allocation strategy determine the actual number of slots to possess the channel. Note that our link scheduling algorithm is independent of the existing routing algorithms, of which the decision-makers can choose the routing strategy based on its real requirement, and calculate the time slot demand for each node and link finally.

183.2.2 Interference Model

Due to the broadcast transmission nature of wireless communication, a transmission between node *i* and *j* may block all the transmissions within R_I away from *i*. A pair of nodes using the same channel and within the interference range may interfere with each other, even if they cannot directly communicate. To schedule two links at the same time slot, we must ensure that the schedule will avoid the interference. Furthermore, the protocol model is employed to describe interference [6]. In a TDMA system, interference is dependent on time. Therefore, we address interference on the basis of link-slot pair. The set of link-time pairs is denoted as *A*. Two link-slot pairs (e, t) and (e', t') ($e = (u, v)$ and $e = (u', v')$) are said to interfere with each other, if $t = t'$ and $\|u - v'\| \le RI$, $\|u - u'\| \le RI$, or $\|v - v'\| \le R_{\mathrm{I}}$. Note that we adopt a symmetric interference model which requires both the receiver and the transmitter to be free of interference.

Given the interference model discussed above, the pairs of communication links that interfere with each other can be represented by a conflict graph [6]. To define a conflict graph, firstly, we create a set of vertices V_c, corresponding to the communication links in the network. In particular,

$$V_c = \{l_{\mathrm{ij}} | e(v_i, v_j) \text{ is a communication link}\}. \tag{183.4}$$

Now, the conflict graph $G_c(V_c, E_c)$ is defined over the set V_c as vertices, and a conflict edge $(l_{\mathrm{ij}}, l_{\mathrm{ab}})$ is used to signify that the communication links $e(v_i, v_j)$ and $e(v_a, v_b)$ interfere with each other. The conflict graph can be used to represent any interference model.

183.3 Problem Formulation

Generally, link scheduling is to assign each slot in a frame to a link in *G*. Since the schedule repeats from frame to frame, it is sufficient to represent this assignment as a mapping matrix $I : E \times M \to \{0, 1\}$, where *E* is the link list and $M = [0,1, \ldots, T\text{-}1]^T$ is the index of slot in scheduling period.

$$I(e(i), t) = \begin{cases} 1 & \text{if } e(i) \text{isactive in time slot } t, \\ 0 & \text{otherwise.} \end{cases} \tag{183.5}$$

Once a schedule is decided, the mapping matrix is repeated in every frame until a new schedule is determined. So if $I(e(i), t) = 1$, link *e*(*i*) is active at all times in the set $\{t + z_i T, z_i \in Z\}$.

A mapping matrix *I* defines a valid schedule, if the allocated number of slots is no more than the required,

$$\sum_{i=0}^{T-1} I(e(i), t) \leq r_{e_j} \tag{183.6}$$

and the schedule have to make sure that no conflicting links do not transmit at the same time,

$$I(e(i),\, t) + I(e(j),\, t) \leq 1, \quad \forall e_i,\, e_j \in E,\, \text{and}\, e_c : \left(e_i,\, e_j\right) \in E_c \tag{183.7}$$

Given the communication graph $G(V, E)$, the conflict graph $G_c(V_c, E_c)$ and link-slot demand vector $R = [r_{e1}, r_{e2}, \ldots, r_{em}]^T$, the link scheduling problem seeks to find an optimal map I, such that the corresponding network throughput is maximum. Formally, it can be described as follows,

$$\begin{aligned} & Min\, T \\ s.t. \quad & \sum_{i=0}^{T-1} I(e(i), t) = r_{e_i}, \forall e_i \in E \\ & \sum_{i=0}^{T-1} I(e(i), t) \leq T, \forall e_i \in E \\ I(e(i), t) + I(e(j), t) \leq 1, \quad & \forall e_i, e_j \in E, \text{and}\, e_c : \left(e_i, e_j\right) \in E_c \end{aligned} \tag{183.8}$$

183.4 Link Scheduling Algorithm

From the mathematical definition of the link scheduling problem, we observe that it is a complex optimization problem to search the minimum scheduling period T. Further, the scheduling problem is splinted into two parts. First, an expanded graph model based on node decomposition is proposed. Hence, in order to schedule the link with different slot demands in wireless mesh network, the expanded graph model incorporated with the heuristic algorithm for fixed transmission orders and equal slot demand is presented.

Given a directed connectivity graph $G(V, E)$. Assume each node $v_i \in V$ is set with final slot demand w_i. Its expanded graph $G_A(V_A, E_A)$ is constructed by node decomposition as follows, for any node $u \in V$ with the slot demand $w_u \geq 0$, V_A has the w_u nodes after decomposition of node u, namely $u^1, \ldots, u^{w_u}$, and each decomposed node u^i has the same location as the node u. For any link $(u, v) \in E$, if u is the mth child of v, E_A has w_u mutually interfered links after decomposition of node u and v, namely $\left(u^1, v\sum_{i=1}^{m-1} w_i + w_v + 1\right), \ldots, \left(u^{w_u}, v\sum_{i=1}^{m-1} w_i + w_v + w_u\right)$. Note that each node can use its own forwarding node after the decomposition of its farther node. In addition, when applying routing algorithm on the extended graph

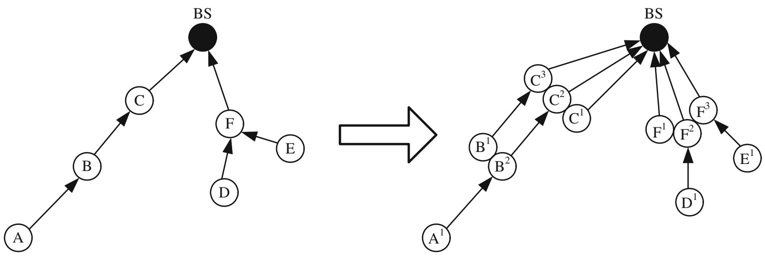


Fig. 183.1 The expanded graph based on node decomposition

$G_A(V_A, E_A)$, each link in G_A can only belong to one path, and have the same packet arriving rate with the source node of its route.

We illustrate the concept of expanded graph and node decomposition in Fig. 183.1, where the initial slot demand of each node is 1. As node *C* has to relay the traffic of its children node *A* and node *B*, its actual slot demand is 3. In expanded graph G_A, node *C* is decomposed into three nodes, C^1, C^2 and C^3. Node *E* is the 2nd child of *F*, then the corresponding link (*E*, *F*) in *G* is changed into (E^1, F^3) in G_A. The routing path $P = \{l_{A,B}, l_{B,C}, l_{C,BS}\}$ in *G* is used to transmit packet from node *A*, *B* and *C* to*BS*. While in G_A, it has been changed into three separate routing paths, $P1 = \{l_{A^1,B^2}, l_{B^2,C^2}, l_{C^2,BS}\}$, $P2 = \{l_{B^1,C^3}, l_{C^3,BS}\}$ and $P3 = \{l_{C^1,BS}\}$, all of which have the same slot demand.

As to the heuristic algorithm to find the optimal schedule length, the common approach is to assign the possible number of slots to each link by greedy algorithm. Based on the expanded graph model, the heuristic algorithm for scheduling link list with different slot demands is described in Fig. 183.2.

The details of the algorithm are shown from line (4) to (17). The conflict graph $G'_c(V'_c, E'_c)$ and link list *O* are given as input, the link-time vector *A* and the scheduling period *T* are given as output. Each time, a link *e* is selected for the *left* link set, if *e* does not conflicts any link in the *current* link set, and the time slot *s*(*e*) is set as the smallest link time in *current* set. Or else the smallest time slot will be allocated from time interval [0, *s*(*e*′) + 1] not yet to assigned to any of its neighbors in G_c, where *s*(*e*′) is maximum end time of link $e' \in$ *currrent*. The time interval is set as [0, *s(e′)* + 1] means that if the slot in [0, *s*(*e*′)] cannot be found, we have to set link-time *s*(*e*) as *s*(*e*′) + 1. Each time if the new obtained *s*(*e*) is larger than the current scheduling period *T*, *T* is updated as *s*(*e*), which ensures that *T* will always meets the whole needs of every link.

183.5 Performance Evaluation

At First, 36 mesh routers are deployed in a grid topology in the area of 500 × 500 grid units, the BS node is placed at the center of the area. Each node has a fixed transmission range of 100.

```
Algorithm: Link Scheduling with different slot demand
  Using the routing algorithm on G(V,E) and node $i$'s initial slot $w'_i$ to calculate node $i$'s final slot demand $w_i$;
  Build the expanded graph $G_A(V_A, E_A)$ based on graph G and slot demand $w_i$;
  Construct a new conflict graph $G'_c(V'_c, E'_c)$ based on $G_A(V_A, E_A)$ using the interference model;
  $T$← 0;
  left ←0;
  current ←ϕ ;
  for each $e \in left$ in $E_A$
    if e do not conflict with all the links in current then
      find $e' \in current$ in $E_A$ which has the minimum $s(e')$;
      $s(e) \leftarrow s(e')$;
    else find $e' \in current$ in $E_A$ which conflicts with e and has the maximum $s(e')$;
      Searching the smallest time slot $s_{temp}$ not yet assigned to any of its neighbors in $G'_c(V'_c, E'_c)$ in time interval [0,$s(e')$+1];
      $s(e) \leftarrow s_{temp}$;
    T  s(e)>T)?S(e):T
    $left \leftarrow left - \{e\}$;
    $current \leftarrow current + \{e\}$;
  endfor
Assign link (u,v) all the slots that are assigned for $(u^1, v^{\sum_{i=1}^{m-1} w_i + w_v + 1})$, … , $(u^{w_u}, v^{\sum_{i=1}^{m-1} w_i + w_v + w_u})$ in $G'_c(V'_c, E'_c)$.
```

Fig. 183.2 The link scheduling algorithm with different slot demand

Table 183.1 shows the performance comparison of *T* with different traffic load using non-preemptive scheduling and our algorithm. As shown in Table 183.1, we can see that by importing node decomposition, our algorithm achieves much better results. But its solution space is too discrete, this will lead to a great synchronization overhead.

Finally, no more than 120 mesh routers are distributed in a grid topology in the area of 100 × 100 grid units, the BS node is placed at the center of the area. Each node has a fixed transmission range of 20 and an initial slot demand of 1. We compare the performance of our algorithm with the largest first link scheduling [9] and nearest first scheduling algorithm [8].

Figure 183.3 shows the performance comparison of *T* with different number of nodes. From Fig. 183.3, we can see that the nearest algorithm gets a much better *T* than largest first algorithm because it uses multiple transmission technology for each link. Since our algorithm gives a more detailed view of the link scheduling, it gets a much better time slot reusability.

Table 183.1 Performance comparison using different algorithms

Slot demand	Rand(3)		Rand(5)		Rand(9)	
Scheduling algorithm	Non-preemptive	Our algorithm	Non-preemptive	Our algorithm	Non-preemptive	Our algorithm
Scheduling Period	27	23	65	53	105	92

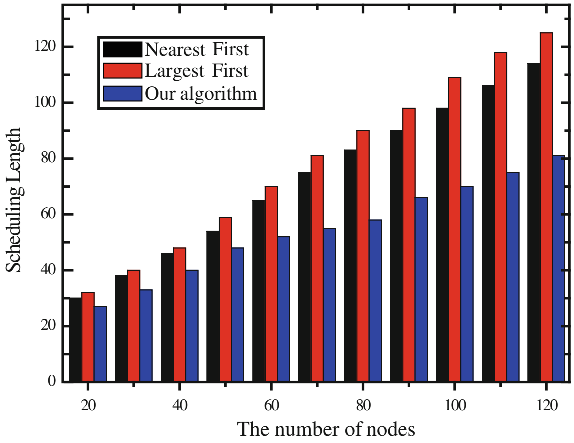


Fig. 183.3 Performance comparisons with TH and Neareast

183.6 Conclusion

In this chapter, we considered the link scheduling problem to maximize the throughput of the network. The analysis of the relationship between link scheduling with network throughput is given. As a contribution, a heuristic algorithm to solve the near- optimal scheduling period with equal time slot demand is proposed. And then, a expanded graph model based on node and link decomposition is presented. Incorporated with the expanded graph model, the heuristic algorithm can also apply to WMN environment with relay models. Simulation results are presented to verify these approaches. In the future, we intend to design distributed algorithms and implementations for the proposed problem.

Acknowledgments This work is supported by the National Natural Science Foundation of China under Grant No. 60903159, No. 61070162, 71071028, 60802023 and 70931001; the Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 20070145017; the Fundamental Research Funds for the Central Universities under Grant Nos. N090504003 and N090504006.

References

1. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
2. Akyildiz IF, Wang X, Wang W (2005) Wireless mesh networks: a survey. Comput Netw 47(4):445–487
3. IEEE standard for local and metropolitan area networks part 16, Air interface for fixed broadband wireless access systems, 802.16 2004
4. Jain K, Padhye J, Padmanabhan VN, Qiu L (2005) Impact of interference on multi-hop wireless network performance. Wirel Netw 11:4471–4487
5. Nelson R, Kleinrock L (1985) Spatial TDMA: a collision-free multi-hop channel access protocol. IEEE Trans Commun 33:934–944
6. Cao Y, Liu Z, Yang Y (2006) A centralized scheduling algorithm based on multi-path routing in WiMAX mesh network, In: IEEE international conference on wireless communications, Networking and mobile computing, pp 1–4
7. Chen L, Tseng Y, Wang D, Wu J (2007) Exploiting spectral reuse in resource allocation, scheduling, and routing for IEEE 802 16 mesh networks, VTC, pp 1608–1612
8. Liu S, Feng S, Ye W (2009) Slot allocation algorithm in centralized scheduling scheme for IEEE 802.16 based wireless mesh networks. Comput Commun 32:942–953
9. Yiqun W, Yingjun Z, Zhisheng N (2008) Non-preemptive constrained link scheduling in wireless mesh networks, In: Proceeding of IEEE GLOBECOM, New Orleans, IEEE press pp 5286–5291

Chapter 184
A Middleware Supporting Multiple Application Tasks for Wireless Sensor Networks

Zhi Hu, Yingyou Wen and Hong Zhao

Abstract Wireless Sensor Networks (WSNs) are becoming more and more popular and used in a variety of applications. Since developing the applications have many technical challenges, WSNs middleware is an effective solution to shield the hardware and operation system, and provide APIs for user. However, WSNs application systems are more complex and need to support more tasks in recent years, and the early and sample WSNs middleware cannot meet the requirements of multiple tasks executed simultaneously. So it is necessary to have a middleware supporting complex and multiple application tasks. This chapter presents the architecture of a middleware supporting multiple application tasks in WSNs. It is based on the architecture to design and implement the entire system and all relative components. By using this WSNs middleware, the different application tasks may be developed and executed concurrently to meet the diverse requirements.

Keywords Wireless sensor networks · Middleware · Multiple application tasks

Project supported by the National Nature Science Foundation of China (No. 60803131).

Z. Hu (✉) · Y. Wen · H. Zhao
School of Information Science and Engineering,
Northeastern University, Shenyang, China
e-mail: huzhi@neusoft.com

Y. Wen
e-mail: wenyy@neusoft.com

H. Zhao
e-mail: zhaoh@neusoft.com

Z. Hu · Y. Wen
Neusoft Group Research, Shenyang, China

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_184,

184.1 Introduction

In recent years, wireless sensor networks (WSNs) are attracting increasing number of researchers from academic and industrial communities because WSNs have been used in many industries and life environments, including habitat monitoring, military applications, disaster prevention, infrastructural security and automation control. WSNs, which consists of sensing, data processing and communication module, collects the data from environment and transports them to sink by ad hoc networks protocol [1, 2]. Compared to traditional networks, WSNs have their own unique characteristics, such as strained energy, dynamic network topology, high rates of node failure, limited processing ability, heterogeneous node hardware, low bandwidth and so on. This poses considerable impediments on WSNs application and makes application development nontrivial. Middleware is ability of hiding heterogeneity of node hardware, operation system and network platform, which eases application developing and executing. WSNs middleware is a kind of middleware that provides service platform for sense-based WSNs application on the node's hardware and operation system [3, 4].

The traditional solutions to WSNs middleware are developed for simple and single WSNs application. The main functions are focusing on hiding underlying hardware and operation system. The entire WSNs system is an application for some environment. With the requirement increasingly developing from the diversified domains, where multiple application tasks need to be executed for the different function at the same time in WSNs, this kind of simple approach has been not suitable to complex user cases and large-scale deployment. It is becoming more common to be capable of supporting multiple and different application tasks run simultaneously in WSNs. So it is essential to design a novel middleware model for multiple application tasks (M-MAT). The M-MAT is proposed in this paper. M-MAT resolves the above problems by abstraction of system function and unified management of applications. First, low level components abstract the interfaces of the various sensor and hardware, while application tasks cover only the requirement of users. Second, but most important, a scalable architecture is proposed and implemented. This architecture manages uniformly all application tasks, and any application tasks that accords with its rules can be added into it.

The rest of this chaper is structured as follows. Section 184.2 presents the existing WSNs middleware. The architecture of M-MAT is proposed in Sect. 184.3. Section 184.4 describes the implementation of M-MAT in detail. The results of experiments and evaluation are explained in Sect. 184.5. The conclusion is drawn in Sect. 184.6.

184.2 Relate Work

The term middleware is a board definition, but generally speaking, WSNs middleware is a kind of system software that is located between operation system and application, which abstracts the low level interface from operation system and hardware, and provides APIs for user application developing. And WSNs middleware is capable of providing services for supporting the WSNs development, maintenance, deployment and execution of WSNs application.

There are currently a number of WSNs middleware solutions. Maté [5] is virtual machine middleware, which abstracts the underlying node and network to provide runtime support for application. Maté has a byte code interpreter that is implemented on TinyOS, and provides a set of programing primitives. In TinyDB [6], the entire sensor network is seen as a queried database. TinyDB is a query processing system that provides the users with a SQL-like interface to query network-wide data. The processing operation of data is executed by TinyDB, including extracts data from network, filters it, aggregates it and transports it to gateway and users. MiLAN [7] is a middleware that provides the required QoS for the application of user. MiLAN allows applications to specify a policy to operate network and nodes, so that it can continuously adapt to the various requirement. Agilla [8] is a middleware framework for mobile agents. There are mobiles agents, tuple space and neighbor list on each node. Mobile agents may access the local and remote tuple space, and migrate carrying the code and state.

These middleware above are application specific or designed to fit some projects. They cannot be adapted to all applied scenarios.

184.3 Design of M-MAT

184.3.1 Architecture of M-MAT

WSNs system usually consists of hardware, operation system and applications. Senor board, radio frequency (RF) modules and low-power micro controller unit (MCU) are used for nodes hardware of WSNs. Operation system and applications are software of WSNs system. Middleware is a sort of system software operated between operation system and application.

M-MAT is a kind of WSNs middleware between WSNs operation system and application, which is made up of many functional components. These functional components offer services for the developing and executing applications. A functional component may offer a lot of service. A set of functional components also provide a sort of service. And a functional component may offer several kinds of services. By application programing interfaces (APIs) that functional components provide in WSNs middleware, users may develop and operate application tasks. Figure 184.1 shows the architecture of M-MAT.

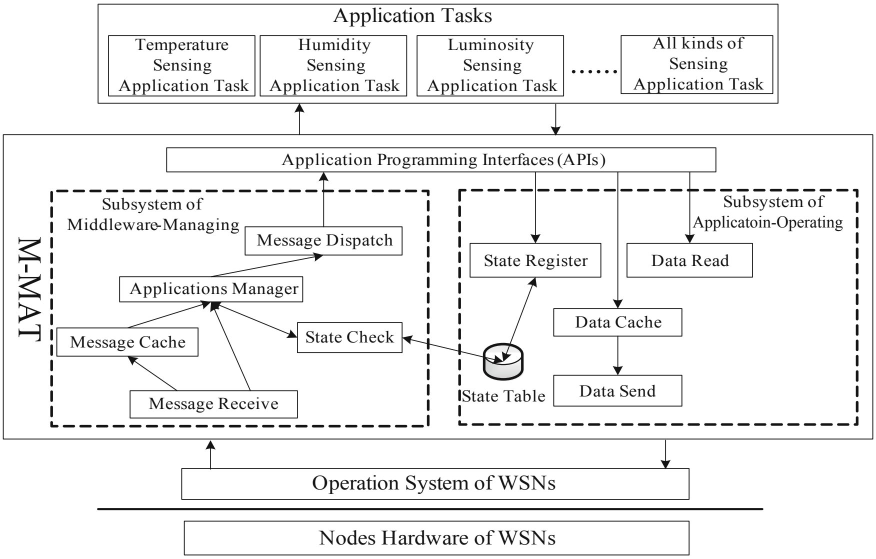

Application Tasks
Temperature Sensing Application Task
Humidity Sensing Application Task
Luminosity Sensing Application Task
......
All kinds of Sensing Application Task
M-MAT
Application Programming Interfaces (APIs)
Subsystem of Middleware-Managing
Message Dispatch
Applications Manager
Message Cache
State Check
Message Receive
Subsystem of Applicatoin-Operating
State Register
Data Read
Data Cache
State Table
Data Send
Operation System of WSNs
Nodes Hardware of WSNs

Fig. 184.1 Architecture of M-MAT

In WSNs system, the main function of nodes is sensing and collecting of environment data for the gateway and users. Meanwhile, application tasks may be dynamically reconfigured according to the requirements of users. So there are a lot of relative functional components created in WSNs middleware, which cooperate to accomplish the relevant function and provide service for application tasks executed. The entire M-MAT is divided into two parts in logic level: Subsystem of Middleware-Managing and Subsystem of Application-Operating. The Subsystem of Middleware-Managing interacts with gateway and users to obtain the goals for the utilization of WSNs, and adjust the application tasks operated based on the message from gateway and users. The Subsystem of Application-Operating abstract the interfaces of sensing data and sending data, and it can store the current state of application tasks into the information table of application tasks. The APIs that M-MAT provides are abstracted from these two subsystems.

184.3.2 Functional Components

The Subsystem of Middleware-Managing consists of Message Receive, Message Cache, Applications Manager, State Check and Message Dispatch component. On the other hand, Data Send, Data Cache, Data Read, State Register and State Table component belong to the Subsystem of Application-Operating. To better understand all components of M-MAT, their function are explained here.

Message Receive. This component receives the messages from gateway and users. These messages are divided into node-level messages and application-level messages according to the components that handle messages. For node-level messages, Applications Manager component usually deals with them to adjust some application tasks from node's view, while each application task treat with the application-level messages dispatched from M-MAT. On the other hand, these messages are also divided into high priority messages and general messages according to response time that the messages are processed. The high priority messages need to be handled immediately, which are often forwarded into the relevant application tasks. The general messages may be stored in cache, and handled in succession.

Message Cache. The messages received will be stored in this component. Since there is a MCU in node hardware, only one message can be handled each time. By caching these received message, the Applications Manager component dispatches them into the relevant application tasks according to FCFS rule.

Applications Manager. This component is in charge of dispatching and dealing with the messages. The node-level messages are mainly disposed of in it, while application-level messages are forwarded into the relevant application tasks. When receiving high priority messages, it handles and forwards them by interrupt mode. By checking the application tasks information in State Table, it can control application tasks in node, according to the requirement of message received.

Message Dispatch. By pre-defined APIs in M-MAT, the message received can be forwarded into the relevant application tasks. When developing application tasks, the interfaces of receiving messages must be added into application tasks. The parameters of interfaces have been defined in manual of M-MAT.

State Check. Applications Manager gets the information of application tasks by this component. When every application task running, they store the state of start, stop and sampling frequency into State Table. So the relevant state of application tasks may be queried by checking State Table.

State Table. The relative information of application tasks, such as start, stop and sampling frequency, is saved in this component.

State Register. This component provides a set of APIs for application tasks. When application tasks are started, stopped and re-configured using sampling frequency, they may modify and maintain the information in State Table.

Data Read. This component is primarily focused on providing an abstraction of underlying reading operation of sensor. Since there are a lot of products and types of sensor, reading operation may be divided into some categories—readTemperature, readHuminity, readLuminosity and so on. An operation is capable of abstracting a kind of the sensors of different manufacturers, such as readTemperature operation may abstract the interface of temperature sensor in SHT11 sensor chip and MTS300 [9] sensor board. These operations are provided for users in APIs mode.

Data Cache. When WSNs system is running, application tasks can generate a large number of data sensed. These data have no way to be sent out at the same time, since there is a unique RF module in node. The data sensed need to be firstly cached, and then sent to sink node by network.

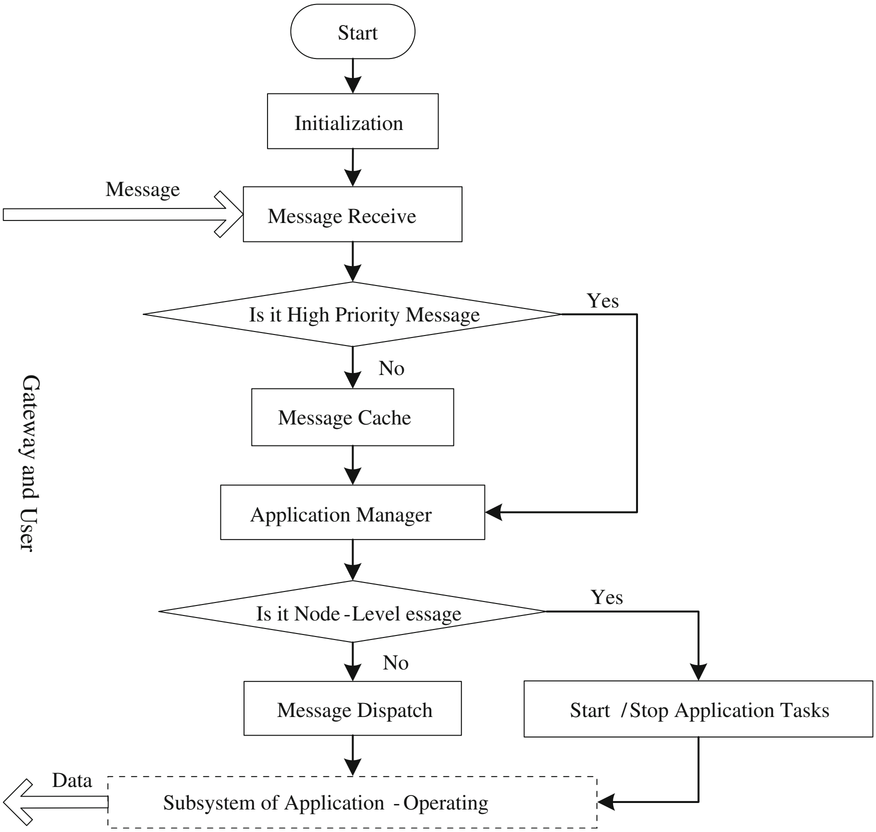


Fig. 184.2 Workflow in Subsystem of Middleware-Managing

Data Send: There are a lot of network routing protocols in WSNs, and they vary in operating mode and interface. So this component provides mainly an abstraction of underlying routing protocols of WSNs.

184.4 Implementation of M-MAT

184.4.1 Subsystem of Middleware-Managing

Subsystem of Middleware-Managing provides the main function to receive and handle the messages from gateway and users, and rectifies the various functions by the requirement of messages. As is shown, Fig. 184.2 illustrates the workflow in Subsystem of Middleware-Managing.

When WSNs system boots, Message Receive and Message Cache component are initialized, and they wait for the messages from gateway and users. When the messages arrive, based on response time that the messages need to be processed, these messages may be divided into high priority messages and general messages. General messages are stored in Message Cache component, while high priority messages need to be immediately handled by Applications Manager. Applications Manager deals with general messages by FCFS rule. For high priority messages, the relative operation is executed by interrupt mode, such as data convergence operation.

In the light of the object that treats with the messages in fact, these messages may also be divided into node-level messages and application-level messages. Applications Manager is the object that deals with node-level messages, which is about starting or stopping application tasks. Applications Manager makes decision based on management policy. Application tasks are the objects disposing of application-level messages. Application-level messages are defined by users.

184.4.2 Subsystem of Application-Operating

Subsystem of Application-Operating focus mainly on the function that supports the operation of multiple application tasks at the same time. Since there are multiple application tasks executing simultaneously and concurrently, and the RF module is unique in node, a large deal of data sensed cannot be sent out at one time. The sensed data that multiple application tasks produce is firstly put into the queue in Data Cache component, and then sent out to gateway and users in success.

Based on M-MAT, there are many application tasks developed, such as temperature, humidity, luminosity application task and so on. These application tasks may be composed together in order to accomplish an intricate work. They may also be converged into an application task that implements a complex case. The workflow is shown in Fig. 184.3.

When WSNs system boots, default application tasks are started, such as Temperature Sensing Application Task. Default application tasks collect the sensed data, and then put them into the queue in Data Cache. Data in queue is sent out to gateway and users by FCFS rule. Meanwhile, all application tasks may also be started, stopped and re-configured by Subsystem of Middleware-Managing.

184.5 Experiment and Evaluation

Based on TinyOS and MICAz platform [9], a proto system that multiple application tasks are executed on M-MAT is done. The sensor network that includes 10 nodes and 1 gateway is set up in the laboratory. Figure 184.4 shows the

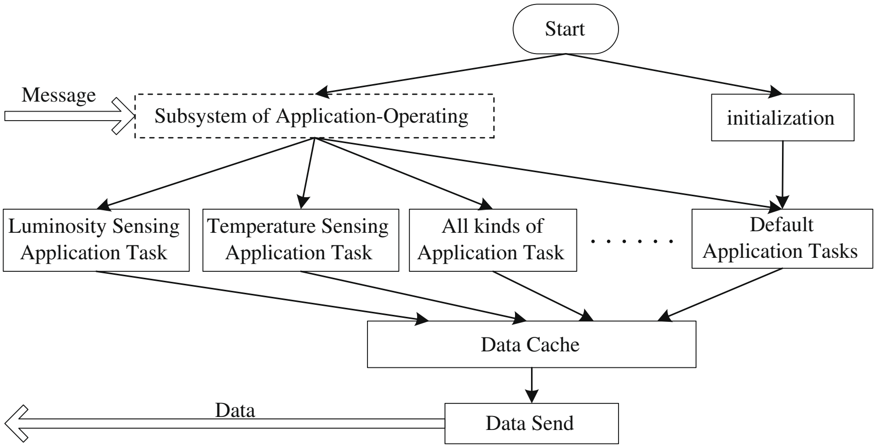
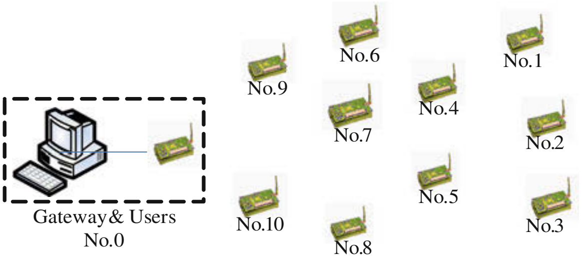

Start
Message
Subsystem of Application-Operating
initialization
Luminosity Sensing Application Task
Temperature Sensing Application Task
All kinds of Application Task
. . . . . .
Default Application Tasks
Data Cache
Data Send
Data

Fig. 184.3 Workflow in Subsystem of Application-Operating

Fig. 184.4 Sensor Network Scenarios



topology of network. There are some different application tasks on every node, including Temperature, Luminosity Sensing Application Task et al., which form an intricate WSNs application system.

184.5.1 Function Tests

The various application tasks are deployed on 10 nodes, for implementing the complex application system that each node collects the sensed data, according to the diversified requirement. Based on M-MAT, application tasks that sense the physical and environmental factors may not only be developed, but also application tasks that collect context information of network topology may be implemented, such as Neighbor Found, Dynamic Routing Path and so on.

The context data is shown in Table 184.1, in which Time Interval is at 9–10 AM.

The data is the average value in the table during the time interval. The type of sensor board is MTS300 from Crossbow Corporation [9], which includes

Table 184.1 Data Sensed

NodeID	Temperature	Light	Neighbor Found	Route Path
1	0x01C6	Off	Off	1 → 4→7 → 9→0
2	0x01C6	Off	Off	2 → 4→7 → 9→0
3	Off	0x037D	Off	Off
4	0x01C5	Off	6, 1, 5, 2, 7	Off
5	Off	0x0369	4, 3, 2, 8, 7	5 → 7→9 → 0
6	0x01C4	Off	Off	Off
7	Off	Off	4, 5, 9, 10	Off
8	Off	0x037A	Off	8 → 10 → 0
9	Off	Off	Off	Off
10	Off	0x0373	Off	Off

thermistor, light sensor and so on. The thermistor is a highly accurate and stable sensor element. The resistance of the thermistor varies with temperature. The value of its resistance may be converted to the relevant Celsius scale by the formulation in MTS300 manual. The thermistor value of Node 1 is 0x01C6, which may be calculated and equal to 20.27°C. In the same way, the temperature of Node 2, 4 and 6 are 20.27, 20.19 and 20.11°C, respectively.

Application task of Neighbor Found is deployed on Node 4, 5 and 7, which transports Node's neighbor information to the gateway and users. Application task of Routing Path is started on Node 1, 2, 5 and 8. According to path information from these nodes, it is concluded that Node 4 is the parent of Node 1 and 2. Node 7 is the parent of Node 4 and 5. As the parent of Node 7, Node 9 relays the packet to gateway Node 0. So the entire network is divided into the two trees. By experiment, it is found that all application tasks may stably run in the long time.

184.5.2 Memory Statistic

The RAM and ROM size of M-MAT and application tasks is crucial to the whole WSNs system. After the single M-MAT is compiled, which merely includes TinyOS and network protocol stack, RAM size is 2,314 bytes and ROM size is 33,780 bytes. The size of RAM and ROM will be increased with application tasks to be added. The size of RAM and ROM that is utilized is shown in Fig. 184.5, and the combinational state of application tasks started is described in Table 184.2.

In the diversified environment, since application tasks are added in the light of the requirement, the size of RAM and ROM fluctuates in the WSNs system. In the Fig. 184.5a, when only Route Path added on the base of M-MAT, the RAM size requires 2,328 bytes. Route Path need merely to initialize packet and add IDs into it. When Neighbor Found added, 2,604 bytes is taken. The system size that includes Temperature or Light application task is 2,380 bytes. The above scenarios that one application task is in WSNs system is described. The RAM size of

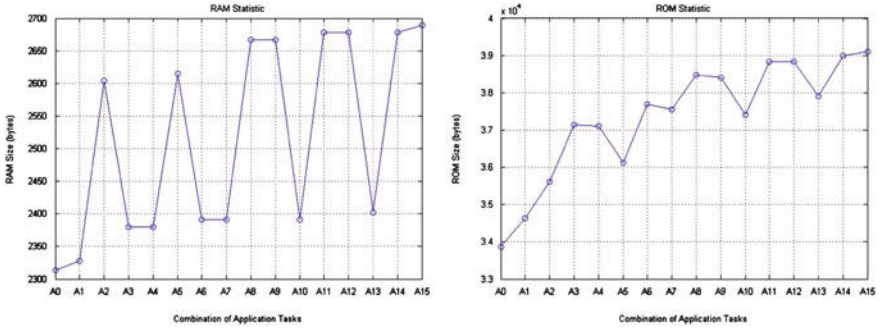


Fig. 184.5 Size of RAM and ROM **a** RAM statistic **b** ROM statistic

Table 184.2 Combination of Application Tasks

Label	Description	Label	Description
A0	Only M-MAT, no application task	A8	M-MAT, Temperature and Neighbor Found
A1	M-MAT and Route Path	A9	M-MAT, Light and Neighbor Found
A2	M-MAT and Neighbor Found	A10	M-MAT, Temperature and Light
A3	M-MAT and Temperature	A11	M-MAT, Route Path, Temperature and Neighbor Found
A4	M-MAT and Light	A12	M-MAT, Route Path, Light and Neighbor Found
A5	M-MAT, Route Path and Neighbor Found	A13	M-MAT, Route Path, Temperature and Light
A6	M-MAT, Route Path and Temperature	A14	M-MAT, Temperature, Light and Neighbor Found
A7	M-MAT, Route Path and Light	A15	M-MAT, Route Path, Temperature, Light and Neighbor Found

combinations from A5 to A10 is 2,615, 2,391, 2,391, 2,667, 2,667 and 2,391 bytes, respectively, when two application tasks are opened. The combinations from A11 to A14 are the scenarios that three application tasks are started. The RAM size is 2,689 bytes after four application tasks are added at the same time.

In Fig. 184.5b, the ROM size is 34,622 bytes when Route Path is only added, while the WSNs system takes 35,610 bytes when Neighbor Found is merely added. Since the development of Temperature and Light application task need to be supported by the library of sensor reading in M-MAT, their ROM size goes up to 37,188 and 37,100 that cover the library memory. When two application tasks are operated, it is viewed from A5 to A10 that the relevant ROM size is increased. The combinations, which three application tasks are in the system, rise similar to two application tasks opened. Lastly, the ROM size gets to 39,108 bytes when four application tasks are added synchronously.

184.6 Conclusion

With the advance in intricate WSNs application, multiple application tasks need to be simultaneously operated in WSNs. The simple middleware, which is developed for some applied domains, cannot meet the current requirement. Therefore, the M-MAT WSNs middleware is described in this paper. It can support multiple application tasks executing at the same time, according to the various requirement. M-MAT can dynamically adjust the relevant application tasks to adapt to the various environments. Any application task, which is not confined to sensing application task, may also be added into M-MAT. The prototype system of M-MAT is demonstrated and evaluated. And the experiments testify the stability, feasibility and practicability of M-MAT. The design and implementation of M-MAT provides the foundation for the further research in the theory and technology of WSNs system.

Acknowledgments This work was supported by Neusoft Group Research (NSR). The author would like to thank NSR and research colleagues for support.

References

1. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) Wireless sensor networks: a survey. Comput Netw 38:393–422
2. Karl H, Willig A (2003) A short survey of wireless sensor networks. Technical report TKN-03–018. Telecommununication networks group, Technical University Berlin
3. Hadim S, Mohamed N (2006) Middleware challenges and approaches for wireless sensor networks. IEEE Distrib Syst Online 7(30):853–865
4. Wang M, Cao J, Li J, Dasi SK (2008) Middleware for wireless sensor networks: a survey. J Comput Sci Technol 23:305–326
5. Levis P, Culler D (2002) Maté: a tiny virtual machine for sensor networks. In: Proceedings of 10th international conference on architectural support for programming languages and operating system (ASPLOS-X), ACM Press, pp 85–95
6. Madden SR, Franklin MJ, Hellerstein JM, Tiny DB (2005) An acquisitional query processing system for sensor networks. ACM Trans Database Syst 30(1):122–173
7. Heinzelman W, Murphy A, Carvalho H, Perillo M (2004) Middleware to support sensor network application. IEEE Netw 18(1):6–14
8. Fok C, Roman G, Lu C (2005) Mobile agent middleware for sensor networks: an application case study. In: Proceedings of 4th international conference on information processing in sensor networks (IPSN 05), IEEE Press, pp 382–387
9. Crossbow Corporation, Product. http://www.xbow.com

Chapter 185
Complexity Analysis of Configuration for a Wireless Sensor Network

Jie Ding and Zhen-Xin Zhang

Abstract This chapter presents complexity analysis of network configuration for a single-hop wireless sensor network which employs the slotted ALOHA protocol, mainly in terms of the time and energy consumed for the network establishment. The results are used to analyse the periodical refreshment of cluster-head in LEACH protocol. In addition, the probability of misjudgement for network configuration is also given in this chapter.

Keywords Complexity analysis · Wireless sensor network · Slotted ALOHA

185.1 Introduction

The study on energy conservation in wireless sensor networks has exploded recently [1–3], where the research interests focus on MAC and routing technologies. But reducing the complexity of network configuration can also save much

Partial of this research was carried out in 2005 when the first author worked with Prof. Q. Wang and L. Yu in the School of Information Science and Engineering, Southeast University, China. The work is finally supported by Natural Science Foundation of Jiangsu Province of China under Grant BK2010313 and BK21010311.

J. Ding (✉)
School of Information Engineering, Yangzhou University,
Yangzhou 225009, China
e-mail: jieding@yzu.edu.cn

Z.-X. Zhang
School of Mathematics, Yangzhou University, Yangzhou 225009, China
e-mail: zhenxinok@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_185,

energy and enhance network lifetime in many situations. For example, some hierarchical protocols such as LEACH [4, 5] have to refresh cluster-heads after every certain period of time. This means that the network must be configured periodically. So the lower the complexity is, the more the energy will be saved.

Recently, first passage time analysis of Slotted ALOHA protocol [6, 7] in event-driven process in the context of wireless sensor networks has been given by Yu et al. [8]. Although some considerations about energy dissipation were presented, there is a lack of deep analysis in their work. In this chapter, we will present a comprehensive analysis for establishing a single-hop network, mainly in terms of time and energy consumption. Moreover, we assume that the network scale is unknown to each sensor node in the network and presents a probability analysis of stopping the network configuration by misjudgement. In addition, these results will be applied to analyse the periodical refreshment of cluster-head in LEACH protocol.

The remainder of this chapter is organised as follows: Section 185.2 presents the network model and some assumptions. Analysis of time, energy, as well as judgement to finish the network configuration, has been demonstrated in Sect. 185.3, and been applied to LEACH protocol in Sect. 185.4. We finally conclude the paper in Sect. 185.5.

185.2 Network Model and Assumptions

This section will present a network model. We assume:

- N Sensor nodes scattered in a region can communicate with each other, which means the network is single-hop. Time is divided into slots of the same certain period, and time synchronisation technology is employed for each node.
- Nodes operate under the slotted ALOHA mechanism. During each time slot, each node either sends packets with the probability ρ, or listens or receives packets with the probability ρ. The activity of each node in each time slot is independent.
- A successful sending in a slot means that only one node sends its packet and all other nodes receive it. If two or more nodes send packets simultaneously in a slot, the other nodes can receive these packets but this sending is considered as unsuccessful because the packets conflict. If there is no node sending, then all nodes are listening.
- It makes a node to consume energy $e_1,\ e_2,\ e_3$ for sending once, receiving once and listening once, respectively.
- The term of the sending set is the set of nodes that can listen, receive and send. A node quits from the sending set as long as this node can listen and receive but cannot send.

The configuration strategy of the network is proposed as the follows:

- *Step 1:* All N nodes are activated. The packet sent by each node contains its ID. We denote by A_1 the first successfully sending node at this step.
- *Step 2:* After A_1 successfully sending, each packet sent by the other $N-1$ nodes contains A_1's ID and the ID of the node who sends. We denote by A_2 the first non-A_1 successfully sending node at Step 2. So A_1 can read its ID from the received packet sent by A_2 and quits from the sending set immediately.
- *Step* $i+1(i=1,\ 2,\ldots,\ N-2)$: After A_i successfully sending, all $N-i+1$ nodes other than $\{A_1,\ A_2,\ldots,\ A_{i-1}\}$ continue sending packets under the slotted ALOHA mechanism. Each packet sent by these $N-i+1$ nodes contains A_i's ID and the ID of the node who sends. We denote by A_{i+1} the first non-A_i successfully sending node. As long as A_i reads its ID from the received packet sent by A_{i+1}, it quits from the sending set.
- *Step* N:

  - *Step* $N(a)$: After A_{N-1} successfully sending, only A_{N-1}, A_N can send packets. A_{N-1}'s packet contains A_{N-1}'s ID and its own. A_{N-1} reads its ID and then quits from sending once it receives A_N's packet.
  - *Step* $N(b)$: After A_N successful sending, only A_N continues sending its packet with the probability ρ in each slot. But A_N will never know whether its sending is successful, and thus it will continue sending all along until some rules make it to give up. Here is a stopping rule: if it hears nothing within L's listened slots, i.e., the slots that it does not send, then this node concludes that it is the last one and sends a message to finish the network configuration.

185.3 Complexity and Misjudgement Probability Analysis

185.3.1 Preliminary Results

Lemma 1 *(I) The probability of successfully sending among* N *nodes during one slot is* $p_1 = C_N^1\rho(1-\rho)^{N-1}$. *The time (i.e. the number of slots) needed for the occurrence of this event satisfies geometry distribution with the parameter* p_1 *and the expectation* $\frac{1}{p_1}$.

(II) The probability of the event that there is successful sending among $N-i+1$ *nodes and the sending node is not the specific one, and is*

$$p_{i+1} = (1-\rho)C_{N-i}^1\rho(1-\rho)^{N-i-1} = C_{N-i}^1\rho(1-\rho)^{N-i},$$

where $i=1,\ 2,\ldots,\ N-1$. *The time for the occurrence of this event satisfies geometry distribution with the parameter* p_{i+1} *and the expectation is* $\frac{1}{p_{i+1}}$.

We denote by γ_i $(i = 1,\ 2, \ldots,\ N)$ the time for completing Step i, during the configuration of the network. By Lemma1, γ_i satisfies the geometry distribution with parameter p_i and the expectation is $\frac{1}{p_i}$. Let ε_i be the average energy consumed by the whole network in one slot before A_i sends successfully, and E_i be the whole energy dissipation at Step i. Then for $i = 1,\ 2, \ldots,\ N - 1$, we have

$$E_i = [E(\gamma_i) - 1]\varepsilon_i + [e_1 + (N - 1)e_2] = \left(\frac{1}{p_i} - 1\right)\varepsilon_i + [e_1 + (N - 1)e_2].$$

In order to determine E_i, we must specify ε_i. The following will present an analysis of energy consumption for the network configuration.

Analysis for Step 1: As the first term of Lemma 1 shows, the probability of the event that a node (namely, A_1) successfully sends is p_1. So the nonoccurrence of the event has the probability $1 - p_1$. Under the condition of the nonoccurrence of this event, only one of the following two things will happen during each slot:

(I) There is no node sending with the conditional probability $\frac{(1-\rho)^N}{1-p_1}$, which makes the network to consume energy Ne_3;

(II) There are j $(2 \le j \le N)$ nodes sending simultaneously, which makes the network to consume energy $je_1 + (N - j)e_2$. The conditional probability of this event is $\frac{C_N^j \rho^j (1-\rho)^{N-j}}{1-p_1}$.

Therefore, under the condition of unsuccessful sensing, the expectation of the energy dissipated by the whole network in one slot is

$$\varepsilon_1 = \sum_{j=2}^{N} \left(\frac{C_N^j \rho^j (1-\rho)^{N-j}}{1-p_1}\right)[je_1 + (N-j)e_2] + Ne_3\left(\frac{(1-\rho)^N}{1-p_1}\right).$$

Analysis for Step $i + 1$: At Step $i + 1$ $(i = 1,\ 2, \ldots,\ N - 2)$, there are *Nminus*; $i + 1$ nodes in the sending set. Here we have included the case of Step 2. Under the condition of the nonoccurrence of that a non-A_i node (namely, A_{i+1}) successfully sends in the network, one of the following three cases will happen in each slot:

(i) All $N - i + 1$ nodes do not send packets. The conditional probability of this case is $\frac{(1-\rho)^{N-i+1}}{1-p_{i+1}}$ and the energy dissipation is Ne_3.

(ii) There are j $(2 \le j \le N - i + 1)$ nodes sending packets simultaneously and the probability is $\frac{C_{N-i+1}^j \rho^j (1-\rho)^{N-i+1-j}}{1-p_{i+1}}$. The energy consumption in this case is $je_1 + (N - j)e_2$.

(iii) Only A_i sends a packet. The probability is $\frac{\rho(1-\rho)^{N-i}}{1-p_{i+1}}$ and the consumed energy is $e_1 + (N-1)e_2$.

So the average of the energy dissipation during one slot is

$$\varepsilon_{i+1} = \frac{1}{1-p_{i+1}}(1-\rho)^{N-i+1}Ne_3 + \frac{\sum_{j=2}^{N-i+1} C_{N-i+1}^{j}\rho^j(1-\rho)^{N-i+1-j}[je_1+(N-j)e_2]}{1-p_{i+1}}$$
$$+\frac{1}{1-p_{i+1}}\left[\rho(1-\rho)^{N-i}(e_1+(N-1)e_2)\right].$$

Analysis for Step N: At Step $N(a)$, by a similar argument, the average time needed for completing Step $N(a)$ is $\frac{1}{p_N}$. At Step $N(b)$, because the probability of A_N not sending in a slot is $1-\rho$, so in order to accumulate L listened slots, it averagely has to wait for $\frac{L}{1-\rho}$ slots. After that it needs another slot to broadcast the configuration being over. Therefore, the expected time for completing Step N is $E(\gamma_N) = \frac{1}{p_N} + \frac{L}{1-\rho} + 1$. Denote by $E_{\{N(a)\}}, E_{\{N(b)\}}$ the expectations of the energy consumed by the whole network at Steps $N(a), N(b)$, respectively. We have denoted by ε_N the energy consumed before A_N successful sending. It is clear to see that

$$\varepsilon_N = \frac{\rho^2}{1-p_N}[2e_1+(N-2)e_2] + \frac{(1-\rho)^2}{1-p_N}Ne_3 + \frac{\rho(1-\rho)}{1-p_N}[e_1+(N-1)e_2],$$

$$E_{N(a)} = \left(\frac{1}{p_N}-1\right)\varepsilon_N + [e_1+(N-1)e_2],$$

$$E_{N(b)} = \left[\frac{\rho L}{1-\rho}[e_1+(N-1)e_2] + LNe_3\right] + [e_1+(N-1)e_2],$$

and $E_N = E_{\{N(a)\}} + E_{\{N(b)\}}$.

185.3.2 Time and Energy Expectations

By $E(\gamma) = \sum_{i=1}^{N} E(\gamma_i), E = \sum_{i=1}^{N} E_i$, and simple computation, we have

Theorem 1 *Under the condition of nonoccurrence of misjudgement, time and energy expectations are given as follows.*

$$E(\gamma) = \frac{1}{N\rho(1-\rho)^{N-1}} + \frac{1}{\rho}\sum_{i=1}^{N-1}\frac{1}{i(1-\rho)^i} + \frac{L}{1-\rho} + 1,$$

$$E = J_1e_1 + J_2e_2 + J_3e_3,$$

where

$$J_1 = K_2 + \frac{1+\rho}{\rho(1-\rho)^{N-1}} - \frac{1}{\rho(1-\rho)} + \frac{1+L\rho+2\rho}{1-\rho} + 3,$$

$$J_2 = \frac{N-\rho}{\rho}(K_2 - K_3) - \frac{1}{(1-\rho)^{N-1}} + \frac{N-1}{\rho} + \frac{L(N-1)+\rho(N-2)}{1-\rho} + 4(N-1),$$

$$J_3 = \frac{1-\rho}{\rho}(1+NK_3) + \frac{N}{\rho} + (L-1)N,$$

and

$$K_2 = \sum_{i=1}^{N-1} \frac{1}{i(1-\rho)^i}, \quad K_3 = \sum_{i=1}^{N-1} \frac{1}{i}.$$

Notice that $K_3 \le 1 + \ln N$, then $K_2 \le \frac{1+\ln N}{(1-\rho)^{N-1}}$. Thus, if we let $\rho = \frac{1}{N}$, then by $\left(1 + \frac{1}{1+x}\right)^x < e$ we have

$$J_1 \le (N + \ln N + 2)e + \frac{N+L+2}{N-1} + 4,$$

$$J_2 \le eN^2(1+\ln N) + N(N+L+2),$$

$$J_3 \le N^2(2+\ln N) + LN.$$

185.3.3 Misjudgement Probability

At Step N, the rule can make A_N to stop the network configuration, but it may bring the problem of misjudging. For example, during the configuration a node will announce the establishment being over as long as there are more than two nodes in the sending set. The probability of misjudging of course depends on L. So a proper L is essential to the rule.

The sending set is composed of $N - i + 1$ nodes at Step $i + 1$ $(1 \le i \le N-2)$. Misjudgement implies L slots, during each of which there is no node sending. Clearly, the probability of the occurrence of this event at Step $i + 1$ is $(1-\rho)^{(N-i+1)L}$. Note that $\gamma_{i+1} - 1$ is the number of slots, for which A_{i+1} waits before it sends successfully at the step $i + 1$. So the probability of no misjudgement at this step is $1 - (1-\rho)^{(N-i+1)L}$. Note that $1 - (1-\rho)^{NL}$ is the probability of no misjudgement at Step 1. If we denote by P_C the probability of no misjudgement before Step $N(b)$, then

$$\begin{aligned} P_C &= \left(1 - (1-\rho)^{NL}\right) \prod_{i=1}^{N-1} \left[1 - (1-\rho)^{(N-i+1)L}\right] \\ &\approx (1 - (1 - NL\rho)) \prod_{i=1}^{N-1} [1 - (1 - (N-i+1)L\rho)] \\ &= NN!(\rho L)^N, \end{aligned}$$

where we have applied the formula $(1-x)^n \approx 1 - nx$. Obviously, increasing L can improve the correction probability, but will result in the increase of time and energy dissipation as Theorem 1 shows. Therefore, there is a trade off between the correct judgement probability and the time and energy consumption.

185.4 Application in LEACH

In this section we consider a concrete example: LEACH protocol [4, 5]. In this protocol, as pointed out in [5], "cluster heads, which are elected from all nodes, broadcast their status to the other sensors in the network and each node determines to which cluster it wants to belong by choosing the cluster-head that requires the minimum communication energy". Further, it is required in LEACH that the cluster-heads must be selected periodically for the purpose of saving energy of heads. Therefore, its corresponding needs network configuration periodically. Here we only consider the time and energy consumed for broadcasting by the cluster-heads since cluster-heads drains much more energy than other ordinary nodes. We suppose that the network is composed of n nodes in which about $h = \frac{N}{n} \times 100\%$ nodes are expected to be elected as cluster-heads at each round. But $\hat{N}$, the number of cluster-heads selected at each round, is a random variable which depends on the cluster-head election algorithm. Here we assume $\hat{N}$ to satisfy binomial distribution with the parameters n and h. Thus $E(\hat{N}) = nh = N$. We also denote by γ, E the time and energy consumed in broadcasting by cluster-heads at each round , respectively.

According to LEACH protocol, when a cluster-head is broadcasting all other head nodes can hear it. So all cluster-head nodes can be considered as a single-hop subnetwork. Moreover, the assumptions presented in Sect. 185.2 are naturally satisfied. So the previous configuration model can be applied to this subnetwork. By Theorem 1 and the discussion in the previous section,

$$E(\gamma|\hat{N} = i) \leq 1 + \frac{L}{1-\rho} + \frac{1/i + 1 + \ln i}{\rho(1-\rho)^{i-1}}, \quad i = 1,\ 2, \ldots,\ n.$$

Suppose $E(\gamma|\hat{N}=0)=0$, then

$$\begin{aligned} E[\gamma]=E[E(\gamma|\hat{N})] &\le \sum_{i=1}^{n} C_n^i\left[1+\frac{L}{1-\rho}+\frac{1/i+1+\ln i}{\rho(1-\rho)^{i-1}}\right]h^i(1-h)^{n-i} \\ &\le +\frac{L}{1-\rho}+\frac{1-\rho}{\rho}\sum_{i=1}^{n} C_n^i(1+1+\ln n)\left(\frac{h}{1-\rho}\right)^i(1-h)^{n-i} \\ &\le +\frac{L}{1-\rho}+(2+\ln n)\frac{1-\rho}{\rho}\left(1+\frac{\rho h}{1-\rho}\right)^n. \end{aligned}$$

So we have the following

Theorem 2 *Under the previous assumption, the time cost for refreshing cluster-heads in LEACH is*

$$E[\gamma]\le 1+\frac{L}{1-\rho}+(2+\ln n)\frac{1-\rho}{\rho}\left(1+\frac{\rho h}{1-\rho}\right)^n.$$

If we let $\rho=\frac{1}{N}$, *then*

$$\begin{aligned} E[\gamma] &\le 1+\frac{L}{1-\rho}+(2+\ln n)\frac{1-\rho}{\rho}\left(1+\frac{\rho h}{1-\rho}\right)^n \\ &\approx 1+\frac{L}{1-\rho}+(2+\ln n)\frac{1-\rho}{\rho}\left(1+\frac{n\rho h}{1-\rho}\right) \\ &= 1+\frac{NL}{N-1}+2N(2+\ln n). \end{aligned}$$

Note that $\sum_{i=1}^{n} C_n^i i x^{i-1} y^i(1-y)^{n-i}=ny(1+xy-y)^{n-1}$. By a similar argument, we have

Theorem 3 *Under the previous assumption, the average energy consumption for refreshing cluster-heads in LEACH is* $E=J_1e_1+J_2e_2+J_3e_3$, *where*

$$J_1\le\left(2+\ln n+\frac{1}{\rho}\right)\frac{\left(1+\frac{\rho h}{1-\rho}\right)^n}{1-\rho}-\frac{1}{\rho(1-\rho)}+\frac{1+L\rho+2\rho}{1-\rho}+3,$$

$$J_2\le\frac{N(1+\ln n)}{\rho}\left(1+\frac{\rho h}{1-\rho}\right)^n+\left(\frac{1}{\rho}+\frac{L+p}{1-\rho}+4\right)N-\frac{1}{\rho}-\frac{L+2\rho}{1-\rho}-4,$$

$$J_3\le\frac{1-\rho}{\rho}+\left(\frac{1}{\rho}+L+\ln n\right)N.$$

185.5 Conclusions

In this chapter we have presented the complexity analysis of network configuration for single-hop wireless sensor networks, mainly in terms of the time and energy needed for the network establishment. The results have been applied to analyse the the cluster-head management in LEACH protocol. In addition, the probability of misjudgement for network configuration has been established, which is suggested to be traded off with the time and energy consumption.

References

1. Akyildiz IF, Su WL, Sankarasubramaniam Y, Cayirci E (2002) A survey on sensor networks. Commun Mag IEEE 40(8):102–114
2. Al-karaki JN, Kamal AE (2004) Routing techniques in wireless sensor networks: a survey. IEEE Wirel Commun 11:6–28
3. Yick J, Mukherjee B, Ghosal D (2008) Wireless sensor network survey. Comput Netw 52(12):2292–2330
4. Heinzelman WB, Chandrakasan A, Balakrishnan H (2002) An application-specific protocol architecture for wireless microsensor networks. IEEE Trans Wirel Commun 1:660–670
5. Heinzelman WR, Chandrakasan A, Balakrishnan H (2000) Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Hawaii international conference on system sciences, vol 8 HICSS '00, Washington, IEEE Computer Society
6. Abramson N (1970) The aloha system: another alternative for computer communications. In: Proceedings of the November 17–19, 1970, fall joint computer conference, AFIPS '70 (Fall), pp 281–285, New York, ACM
7. Abramson N (1977) The throughput of packet broadcasting channels. IEEE Trans Commun 25(1):128
8. Yu L, Wang Q (2009) First passage time analysis in wireless sensor networks for spr and mpr systems. Wirel Pers Commun 48:531–550

Chapter 186
A New Wireless Communication Technology Based on Neural Network

Xiaomin Chen and Haitao Li

Abstract An artificial neural network usually called neural network, is a mathematical or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons which processes information using a connectionist approach to computation. Critical technical bottlenecks in a wireless link are the capacity of the radio channel, its unreliability due to adverse time-varying, multipath propagation and severe interference from other transmissions, in the neighboring cells.

Keywords Artificial neural networks · Wireless communication · Access control layer · Short-range communication

186.1 Introduction

As the national economic and social development occurs, information technology is useful for people who want pass information to open up new ways of working, management, business and financial methods, the exchange of ideas, the cultural teaching methods, health care methods, and consumer and lifestyle [1]. Wireless communication means the development of fixed-mobile way. With the development of communication technology, wireless communication environment becomes increasingly complex communications signals over a wide frequency band using a variety of modulation. Line communication has increased a variety of

X. Chen (✉) · H. Li
Basic Department of Jiaozuo University, Jiaozuo 454003, Henan, China
e-mail: xmchen2011@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_186,

distinct complementary technologies. This was reflected in different access technologies with different coverage for different regions, technical characteristics and access rates such as 3G and WLAN, UWB, etc. [2, 3], which can achieve complementary effects. For the wide area 3G coverage seamless roaming and strong demand for mobility, WLAN resolves the longer-distance high-speed data access, while UWB can achieve ultra high-speed wireless access at close range. Therefore, we should be integrated in policy to promote the development of wireless access and promote networking in the integration process, through the diversification of means of access network construction, to achieve the coverage needs of different user groups, market segmentation and business diversities to address the uneven development of mobile communication situations.

Multiple integrations in the trend such as, 3G, WLAN and other wireless technologies in the competition learn from each other, both by the emergence of a new type of wireless technology using radio frequency technology, such as MIMO and OFDM technologies. Meanwhile, the ITU and 3GPP/3GPP2 led the 3G cellular mobile communications to the E3G, then the evolution path towards B3G/4G, and the IEEE led the wireless broadband access from a wireless personal area network to wireless LAN, wireless MAN and the evolution of wireless wide area network on the road.These have started to increase each other's content, such as: mobile communications continue to strengthen the transmission performance of broadband and wireless broadband access, roaming performance and continuously to enhance the safety performance.

The original inspiration for the term Artificial Neural Network came from examination of central nervous systems and their neurons, axons, dendrites, and synapses, which constitute the processing elements of biological neural networks investigated by neuroscience. In an artificial neural network, simple artificial nodes, variously called "neurons", "processing elements" (PEs) or "units", are connected together to form a network of nodes mimicking the biological neural networks—hence the term "artificial neural network" [4]. In modern software implementations of artificial neural networks, the approach inspired by biology has been largely abandoned for a more practical approach based on the statistics and signal processing. In some of these systems, neural networks or parts of neural networks are used as components in larger systems that combine both adaptive and non-adaptive elements.

We first discuss the applications for adhoc wireless networks, including data networks, home networks, device networks, sensor networks, and distributed 389 controls. Next, we consider cross-layer design in adhoc wireless networks: what it is, why it is needed, and how it can be done. Link layer design issues are discussed next, followed by consideration of the medium access control (MAC) layer design issues [5], including the trade-offs inherent to frequency/time/code channelization and the assignment of users to these channels via random access or scheduling. This section also describes the role, power control can play in multiple access. Networking issues such as neighbor discovery, network connectivity, scalability, routing and network capacity are outlined next. Last, we describe techniques for the network to adapt to the application requirements and the application to adapt to network capabilities. One of the biggest challenges in providing multimedia wireless services is to maximize the efficient use of the limited available bandwidth. Cellular systems exploit the power fall off with the distance of signal propagation to reuse the same frequency channel at spatially separated locations.

186.2 Artificial Neural Networks

An artificial neural network (ANNs), usually called neural network (NN) [6], is a mathematical or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are nonlinear statistical data modeling tools. ANNs is a neural network model of animal behavior characteristics, the distributed parallel algorithm for information processing model.

ANNs is a neural network of the human brain or natural (Natural Neural Network) and some basic properties of the abstract simulation. ANNs to the physical brain-based research, which aim to simulate some of the brain mechanism, and mechanisms to achieve a functional aspect. This network relies on the complexity of the system, by adjusting the connection between the numbers of nodes within the relationship so as to achieve the purpose of processing information. ANNs is built by hand in order to have a picture that shows the topology of the dynamic system, which on continuous or intermittent input for the state of information processing to be carried out accordingly. People think that simulation of ANNs is the second way . This is a nonlinear dynamic system, that characteristic is the information distributed in storage and parallel co-processing. Although the structure of individual neurons is extremely simple, limited functionality, a large number of neurons in a network system can realize the behavior is very colorful. The results show that in many areas of the technology that have broad potential applications, such as forecasting, pattern recognition, automatic control and other fields of intelligent simulation and information processing; a number of studies: neural network technology with massively parallel processing, distributed storage, adaptability, fault tolerance and other significant advantages are included. Present methods in neural network research has formed a number of genres, the most fruitful research work include: multi-layer network BP algorithm, Hopfield network model, adaptive resonance theory and self-organizing feature mapping theory. ANNs is based on modern neuroscience initiative. Although it reflects the basic characteristics of human brain function, but is far from realistic description of the natural neural network, rather it is a simplified abstraction and simulation.

ANNs features and benefits, mainly are present in three aspects:

- First, a self-learning function; for example, in pattern recognition, only the first sample many different images need to be identified and the corresponding results of ANNs input. The network will be through self-learning function, slowly learns to identify similar images for the forecasts of particular importance. Expected future human ANNs computer will provide economic forecasting, market forecast, forecast efficiency, its application is very ambitious future.
- Second, with the association storage, ANNs feedback network can achieve this association. In applications where the goal is to create a system that generalizes well in unseen examples, the problem of overtraining has emerged. This arises in convoluted or over-specified systems when the capacity of the network significantly exceeds the needed free parameters. There are two schools of thought for avoiding this problem: the first is to use cross-validation and similar techniques to check for the presence of overtraining and optimally select hyper parameters such as minimizing the generalization errors. The second is to use some form of regularization.
- Third, is to find the optimal solution with high capacity. In order to find the optimal solution of a complex often requires a large amount of computation, using a problem to the design of a feedback type ANNs, high-speed computing power to play the computer, may soon find the optimal solution.

ANNs have several outstanding advantages that caused a great concern in the recent years: (1) Can fully approximate any complex nonlinear relationship; (2) all quantitative or qualitative information such as potential distribution are stored in the neurons within the network, it has strong robustness and fault tolerance; (3) the parallel distributed processing, makes it possible to quickly carry out large operations; (4) can be learning and adaptive system does not know for uncertain; and (5) can handle both quantitative and qualitative.

186.3 Wireless Communications Technology

Wireless Communication is the use of electromagnetic waves in free space signal propagation characteristics in an exchange of information communication. Wireless communications include microwave and satellite communications [7]. Microwave is a radio wave, which is sent only a few dozen kilometers away from the general. But the microwave frequency band is very wide, and has large-capacity communications. Intervals of tens of kilometers required for microwave communication to build a microwave relay station. Satellite communications is the use of communications satellites relay, standing on the ground between two or more earth stations or mobile microwave communication links between the bodies.

Radio spectrum is a natural resource, but one with rather unusual properties. As noted above, it is non-homogeneous, with different parts of the spectrum being best used for different purposes. It is finite in the sense that only part of the electromagnetic spectrum is suitable for wireless communications, although both the available frequencies and the carrying capacity of any transmission system depend on technology. The radio spectrum is non-delectable; using spectrum today does not reduce the amount available for future use, but it is non-storable.

IEEE 802.11 standard wireless LAN standard developed, mainly for the network physical layer (PH) and medium access control layer (MAC) were provided; the key includes the provisions of the MAC layer. In the MAC layer below, 802.11 provides for sending and receiving of three technologies: Spread Spectrum technology; IR (Infared) technology; narrowband (Narrow Band) technology. The direct sequence spread spectrum is divided into Direct Sequence Spread Spectrum (DSSS) technology and Frequency Hopping (FH) spread spectrum technology. DSSS technology, will often combine with Code Division Multiple Access (CDMA) technology.

The first mobile system for cars was built in St. Louis, US in 1946. Only six calls could be made at the same time and the switching with the PSTN (Public Switching Telephone Network) was done manually by operators. To enhance the capacity, the AT&T Bell Lab promoted the idea of cellular system in 1947. But in the 1950s, the satellite communication was the main research direction for the radio communications. When it was possible to make the communication devices practically potable, the cellular system gained interest again in the 1970s. During this period, a series of studies on radio propagation, Doppler spectrum, fading statistics and other quantities were formulated. Motorola, AT&T, Ericsson and NTT were pioneers in this area.

Currently using a wider range of short-range wireless communications technology is Bluetooth (Bluetooth), Wireless LAN 802.11 (Wi-Fi) and infrared data transfer (IrDA). At the same time there are some short-range potential with wireless technology standard, they are: Zigbee, UWB (Ultra WideBand), short-range communication (NFC), WiMedia, GPS, DECT, Wireless 1394 and private wireless systems. They all have their characteristics based on transmission speed, distance, power consumption of the special requirements; or to focus on the functional expandability, meet some special requirements of a single application,or to create competitive differentiation and other technology.

186.4 Conclusions

Wireless networks and cellular modems are examples of devices that use wireless communication. Such devices may be restricted in some situations or environments, such as when traveling in an airplane. Neural network can be used for wireless channel modeling and simulation, as well as the channel parameter extraction. In future for wireless multimedia networks or communication networks for Intelligent Transport Systems, the multiple access issue becomes substantially more important than it is for circuit-switched voice communication. The ALOHA, CSMA and ISMA protocols all allow multiple users to share radio communication resources. How these protocols perform differs substantially for wired and unguided channels; performance highly depends on the physical propagation characteristics of the channel.

Acknowledgments The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions that have improved the paper.

References

1. Ibnkahla M, Bershad NJ et al (1998) Neural network modeling and identification of nonlinear channels with memory: algorithms, applications, and analytic models[J]. IEEE Trans Signal Process 46(5):1208–1220
2. Bershad NJ, Ibnkahla M, Blauwens G et al (1999) Fluctuation analysis of a two-layer backpropagation algorithm used for modeling nonlinear memoryless channels[J]. IEEE Trans Signal Process 47(5):1297–1303
3. Turchetti C, Conti M et al (1998) On the approximation of stochastic processes by approximate identity neural networks[J]. IEEE Trans Neural Netw 9(6):1069–1085
4. Balabin RM, Lomakina EI (2009) Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. J Chem Phys 131(7):74–104
5. Caillaud B, Jullien B (2001) Competing cybermediaries. Eur Econ Rev 45:797–808
6. Cramton P, Kwerel E, Williams J (1998) Efficient relocation of spectrum incumbents. J Law Econ 41:647–675
7. Gans JS, King SP (2000) Mobile network competition, customer ignorance and fixed-to-mobile call prices. Inf Econ Policy 12:301–327

Chapter 187
Multi-agent Technology Applied to Mobile Communication

Li Haitao and Chen Xiaomin

Abstract Wireless communication technology is diffusing around the planet faster than any other communication technology to date. Wireless communication provides a powerful platform for political autonomy on the basis of independent channels of autonomous communication, from person to person. The communication networks that are enacted by mobile telephony can be formed and reformed instantly, and message is received from a known source, enhancing their credibility.

Keywords Mobile agent · Mobile communication · Wireless communication technology

187.1 Introduction

Wireless communication technology is diffusing around the planet faster than any other communication technology to date [1]. Because communication is at the heart of human activity in all domains, the advent of this technology, allowing multimodal communication from anywhere to anywhere where there is the appropriate infrastructure, is supposed to have profound social effects. Wireless networks are the fastest growing communications technology in history.

With the development of computer science, Agent in artificial intelligence and computer science is becoming an increasingly important position. Agent Systems theory by academic and industry researchers is gaining more and more attention

L. Haitao (✉) · C. Xiaomin
Basic Department of Jiaozuo University, Jiaozuo 454003, Henan, China
e-mail: htli2011@126.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_187,

and applied research. Agent can mimic human behavior, with self-government, social, adaptability, intelligence and other human characteristics. With the establishment of information infrastructure and improvement of people's increasing demanding applications Agent Application involving various industries is applied to human social life. High intelligence, networking, high reliability and fast adaptability is application systems goal. And that goal must be just and consistent with the characteristics of Agent, so as to promote the theoretical research and applied research of Agent.

Modern wireless technologies started in the 1980s in China with the introduction of the pager in 1984 and cellular phones in 1987. While initial adoption was slow and restricted to a very small circle of high-end business users, the speed of growth has been extraordinary since 1990. Pager subscription took off throughout the 1990s to peak with 49 million users in 2000. It then started to decline, becoming a technology used largely by migrant workers. The change happened at the same time as cell phone penetration started to surge strongly together with the rapid spread of short message systems.

The next wave of technological innovation must integrate linked organizations and multiple application platforms [2]. Developers must construct unified information management systems that use the World Wide Web and advanced software technologies. Software agents, one of the most exciting new developments in computer software technology, can be used to quickly and easily build integrated enterprise systems. The idea of having a software agent that can perform complex tasks on our behalf is intuitively appealing. An agent is simply another kind of software abstraction, an abstraction in the same way that methods, functions, and objects are software abstractions. An object is a high-level abstraction that describes methods and attributes of a software component. An agent, however, is an extremely high-level software abstraction which provides a convenient and powerful way to describe a complex software entity. Rather than being defined in terms of methods and attributes, an agent is defined in terms of its behavior. This is important because programing an agent-based system is primarily a matter of specifying agent behavior instead of identifying classes, methods and attributes. Agent technology plays a role in two fundamental ways: resources are described, located and tasked using semantic descriptions based on ontologism and semantic services; tracking, fusion and decision-making logic is implemented using agent objects and semantic descriptions as well.

187.2 Agent Technology

Agent Technology is looking for small and medium suppliers in the food and grocery sector in Tasmania who supply to Tasmanian-based stores such as IGA stores. Along with a core group of initial suppliers, Agent Technology seeks a further reference group for the purposes of reviewing and trialing its innovative software that improves access, communication and supply chain transactions

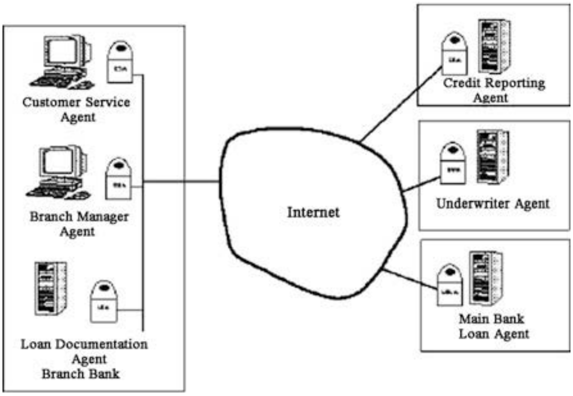


Fig. 187.1 Multiple agents for enterprise applications

between suppliers and retailers. Agent Technology's peer-to-peer technology offers supply chain participants for more powerful connectivity and supply chain communication compared with EDI, electronic data interchange [3]. EDI is poor at handling one-to-many and many-to-many communication and collaboration. In other words EDI is not good for the business requirements of the twentyfirst century.

When agents interact, for instance, to cooperate, negotiate or even to compete, they should be able to communicate. Some important research topics here are: defining the semantics of communication primitives, dealing with different vocabularies, specifying interaction protocols and verifying correctness properties of such protocols [4]. Mobile agent is a special kind of Agent, Agent in addition to its most basic characteristics of intelligent—autonomy, responsiveness, initiative and interaction, but also mobile, the network that it can autonomously move from one host to another host, on behalf of the user to complete the assigned task. Mobile Agent in heterogeneous hardware and software in the mobile network environment. Mobile Agent computing model can effectively reduce the network load in distributed computing. Improve communication efficiency, support for offline computing, support for asynchronous self-interaction, dynamically adapt to the network environment, with security and fault tolerance. Mobile Agent and other traditional centralized computing model of distributed technologies (such as client/server model, distributed object technology and mobile code technology) combined with the advantages of distributed artificial intelligence technology to provide a general, open. Figure 187.1 is the multiple agents for enterprise applications.

The ability of a single agent is limited and incomplete, and needs to keep learning, especially in the increasingly common limited resources, where the dynamic development of the environment was. In a multi-agent environment, in order to complete the task, an independent agent may need through a number of intelligent behaviors such as collaboration, negotiation, competition co-operation of the entire multi-agent system ,and coordination. It basically has the following technical characteristics: agents, autonomy, initiative, and intelligence. Agent is mainly reflected to do some work on behalf of the user or proxy user software to communicate with other software and links. Agent itself is an autonomous computing entity, which can independently find and use various information resources and services, solve problems independently, and provide services for users. Initiative is the Agent according to the needs of users, adapt to changes in the environment initiative to provide services for users. Agent intelligence can sense the surrounding environment, with reasoning and intelligent computing, can analyze the needs of users, and continue to accumulate experience in order to enhance their problem solving skills.

187.3 Mobile Communication Technology

Mobile technology has already created hype throughout the world. Today’s mobiles networks supports features such as SMS, GPRS, MMS, emailing facility on mobile, Bluetooth, WAP, and many more depending upon how reputed and bigger the mobile network company is, most of the networks world wide provide these features as they have become the standard features in mobile communication between their customers and of course one cannot neglect how sophisticated mobiles phones are available now. These mobile phones carry many features which some times are not supported by mobile networks [5]. Mobile phones of today’s age are now equal to portable PCs.

Mobile Technology has groomed a lot in past few years. Major reasons for rapid advancements in mobile network technology is requirements for being mobile or connected on the move. Latest mobile handsets offer features which one had never thought off. Ultimately it forces mobile network companies to bring these features in practice use to take commercial advantages. The first generation mobile communication systems (1G) in the early twentieth century, made 80, which was completed in the 20th century, early 90s, such as NMT and AMPS, NMT put into operation in 1981. The first generation mobile communication system is based on analog transmission, which is characterized by the volume of business small, poor quality, poor pay the whole. There is no encryption and the speed is low.

The second generation mobile communication systems (2G) originated in the early 90s. European Telecommunications Standards Institute in 1996 aims to expand and improve the GSM Phase 1 and Phase 2 in the original business and performance. It includes CMAEL (customized applications for mobile network enhanced logic), S0 (support optimal routing), the immediate billing, GSM 900/1800 dual-band work, etc., and also includes fully compatible with the full-rate voice codecs enhanced technology, making the voice quality a qualitative improvement; half-rate GSM codec can provide nearly double the capacity of the system. In the GSM Phase2 + stage, a more intensive frequency reuse, multiple re-use, repeated use of multi-structure technology, the introduction of smart antenna technology, dual-band technology, to effectively overcome the surge in business volume with the GSM system capacity caused by lack of defects; adaptive speech coding (AMR) technology, which greatly improves the system call quality; GPRs/EDGE technology, the introduction of the GSM communication with the computer/Internet combination of organic phase, the data transfer rate of up to 115/384 kbit/s, so that the GSM functions are growing, initially with the ability to support multimedia services.

The third generation mobile communication system (3G) is being fully developed system. Its basic features is the intelligent signal processing, intelligent signal processing unit will be the basic functional module to support voice and multimedia data communications, it can provide the first two generations products and cannot provide information on a variety of broadband services, such as high-speed data, slow images, and television images. Third generation mobile communication system communication standard total WCDMA, CDMA2000, and TD-SCDMA three branches. 4G is the fourth generation mobile communication and technology, referred to, is a 3G and WLAN in one and is able to transmit high quality video images and image transmission quality is comparable to high-definition television technology products [6].

The fourth generation mobile systems network architecture can be divided into three layers: the physical network layer, middle layer of the environment, and the application network layer. Provide access to the physical network layer and routing functions, which by a combination of radio and core network to complete the format. Intermediate environment layer QoS mapping functions, address change, and complete management. Physical network layer and the middle layer and its environment, the interface between the application environments is open, it makes the development and delivery of new applications and services easier to provide seamless high data rate wireless service, and run on multiple band. Cellular companies use AMPS, D-AMPS, CMMA2000, UMTS, GSM, EVDO etc. AMPS however pretty much vanished from the scene, AMPS network system was based on analog communication technology, latest features were not supported by AMPS therefore all cellular networks world wide have adopted above-mentioned digital communication methodologies to meet the need of consumers. GSM remains the highly used mobile communication methodology worldwide. Cellular networks and mobile phones vary from geographical locations and providers to providers, but still standard communication methods are more or less same every where. Basic communication takes place using electromagnetic microwaves with cellular base stations. Cellular networks have huge antennas normally located in the middle of certain area to provide optimum signal broadcasting. These antennas are known as Base Transceiver Station (BTS). Mobile handsets have low powered transceivers which transmit voice data to the closed BTS which can usually be with in 5 to 8 miles radius.

187.4 Conclusion

With a rapid expanding network of sensors in the battlefield, one critical challenge is how to integrate dynamically networked sensors with multi-level information fusion processes to support real-time sensing, exploitation, and decision-making. Wireless communication provides a powerful platform for political autonomy on the basis of independent channels of autonomous communication, from person to person. The communication networks that are enacted by mobile telephony can be formed and reformed instantly, and message is received from a known source, enhancing their credibility. The networking logic of the communication process makes it a high volume communication channel, but with a considerable degree of personalization and interactivity. In this sense, the wide availability of individually controlled wireless communication effectively bypasses the mass media system as a source of information, and creates a new form of public space.

In general, there is considerable social differentiation in the adoption of wireless technologies. The adoption pattern of mobile phones clearly varies along the dimension of age. While the gender gap and socio-economic differentiation are diminishing, especially in more saturated markets, they remain significant in parts of the world.

Acknowledgments The authors would like to thank the anonymous reviewers for their insightful comments and constructive suggestions that have improved the paper.

References

1. Klusch M (2001) Information agent technology for the Internet: a survey. Data Knowl Eng 36:337–372
2. Liu B, Wang H, Feng A (2001) Applying information agent in open bookmark service. Adv Eng Softw 32:519–525
3. Alden J (2002) Competition policy in telecommunications: the case of the United States of America. In: ITU workshop on competition policy in telecommunications, Geneva, 20–22 Nov (2002)
4. Beaubrun R, Pierre S (2001) Technological developments and socio-economic issues of wireless mobile communications. Telem Inform 18:143–158
5. Kim D-Y (2002) The politics of market liberalization: a comparative study of the South Korean and Philippine telecommunications service industries. Contemp Southeast Asia 24(2):337–370
6. Rafael V (2003) The cell phone and the crowd: Messianic politics in the contemporary Philippines. Pop Culture 15(3):399–425

Chapter 188
A Simple Power Control Approach in Wireless Ad Hoc Network

Shaoping Jiang and Yi Wang

Abstract Ad hoc networks are a new paradigm of wireless communication for mobile hosts, in which there are no fixed infrastructure such as base stations or mobile switching centers. Mobile nodes that are within each other's radio range communicate directly via wireless links, while those that are far apart rely on other nodes to relay messages as routers Node mobility in an ad hoc network causes frequent changes of the network topology. Power control basically deals with the performance within the system. The intelligent selection of the transmit power level in a network is very important for good performance for minimizing the traffic carrying capacity, reducing the interference and latency. In this chapter, a simple approach is present to deal with the selection of the transmission by the cross-layer design between MAC and PHY. The experimental result shows the new approach has good performance on throughout, delay and energy consumption.

Keywords Ad Hoc network · Power control · MAC · PHY

188.1 Introduction

Power control is the intelligent selection of lowest common power level in an ad hoc network in which the network remains connected. The power optimal route for a sender receiver pair is calculated and the power level used for this transmission is

S. Jiang (✉) · Y. Wang
The School of Information Science and Technology,
Hunan Agricultural University Changsha, Changsha, China
e-mail: 23281776@QQ.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_188,

set as the lowest power level for that particular transmission. In case of multiple nodes, power optimal route for each transmission is calculated. The importance of power control arises from the fact that it has a major impact on the battery life and the traffic carrying capacity of the network.

Power control has received increasing attention over the recent past [1–5, 12] where attention has been focused on developing joint congestion control. Typically, the approach consists of formulating the network resource allocation problem as a convex optimization problem (by approximating the wireless physical layer [6]), and cross-layer solutions either are based on primal–dual algorithms for convex optimization[5, 7] and/or by means of a per-time-slot scheduling combined with a queue-length-based back-pressure algorithm[2, 8–10].

In multi-hop network with the IEEE 802.11/802.11b MAC protocol, there is only one common channel. Each mobile terminal accesses the channel through a CSMA/CA competition mechanism, i.e., a four frame RTS–CTS–ATA–ACK handshake to realize a data transmission. In the scheme, each mobile terminal gets access to the medium on a contention basis. Before a data transmission begins, the sender and receiver must have a RTS–CTS signaling handshake to "reserve" the channel. When a sender has a packet to transmit, it senses the channel by detecting the air interface (in the physical layer) and looking up its Network Allocation Vector (NAV). If the channel is busy, the terminal waits until the channel becomes free, in which case it sends a Request to send (RTS) to the destination terminal. On successfully receiving the RTS, the destination replies the source with Clear to send (CTS). The source can begin data transmission after the CTS is received. After the data is received at the destination, the destination sends an acknowledgment (ACK) to the source, confirming the success of a data reception. This is an ideal case of a four-way handshake. If the source fails to receive CTS or ACK (collision at source or destination), it backs off for a random period of time. To deal with the power control and to not affect the MAC Protocol, a simple approach is present to handle the emission power. Note that the size of DATA packet is more large than the control packet such as CTR, RTS and ACK. How to set the power for transmitting DATA packet is the mail goal in this chapter.

188.2 A Simple Power Control Approach

In wireless AD Hoc network based on IEEE 802.11 protocol, the sender requests to construct a wireless link connection with receiver via sending a RTS control packet and then monopolizes the wireless channel. The receive sends the CTS to acknowledge the wireless link. After receiving the data packet successfully, the receiver should reply a ACK control packet. The size of the three packets is less than data packet. Based on the point, the transmission power of the data packet is taken into account in this chapter. In multi-hop wireless AD Hoc network, the basic MAC protocol does not need to be modified and has a simple approach to handle the power control.

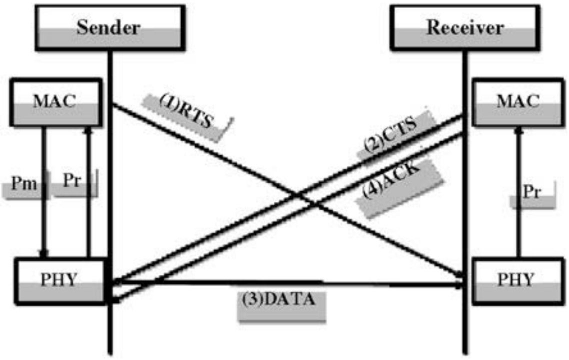


Fig. 188.1 Cross-layer design

188.2.1 Basic Power control

How to decide the optimal emission power in the power control approaches is present in the chapter [11]. According Friis Formula in Free–space propagation loss model, the receiver's power can get as follow.

$$P_r = \frac{P_t G_t G_r \lambda^2}{(4\pi)^2 d^2 L} \tag{188.1}$$

In (188.1), P_t is the emission power of the sender, G_t, G_r is the gain of the transmit/receive antenna, respectively, d is the distance between sender and receiver and L is the system loss factor.

The node judges whether the signal is transmitted in the wireless channel by listening the power strength in the channel. The listened power strength is higher than the received threshold P_t, the receiver can receive the data packet correctly. The least emission power for the sender can get as follows.

$$P_m = \frac{R_t \mathrm{Ld}^2 (4\pi)^2}{G_t G_r \lambda^2} \tag{188.2}$$

In (188.2), R_t is the threshold of the receiver, and P_m is the least emission power for the sender.

Combined (188.1) with (188.2), P_m is calculated as follow.

$$P_m = \frac{P_t R_t}{P_r} \tag{188.3}$$

Usually, a reliable transmission can be obtained by multiplying P_m with a adjust parameter.

$$P = c * P_m (c \geq 1) \tag{188.4}$$

To get the above parameters, cross-layer information must be exchanged between the MAC layer and PHY layer (as seen in Fig. 188.1).

To handle the Hidden terminal problem, the broadcast packets are transmitted by the max power. Due to this point, packets can be transmitted on multi-hop. After the sender and the receiver exchange the essential information, the DATA and ACK packets can use the optimal power transmit.

188.2.2 New Approach

In (188.4), the adjust parameter is usually a constant. In a dynamic wireless Ad Hoc network, the node is often moved. So the adjust parameter cannot ensure the quality of the transmission. When the node is moved, the ACK fail count indicates the variation of the wireless channel. When the condition of wireless channel is changed, the adjust parameter must be adjusted. The adjust parameter is calculated dynamically according the ACK fail count.

$$c = C + \frac{ACKfailCount + 1}{Threshold} \tag{188.5}$$

In (188.5), C is a constant, *Threshold* is the parameter in 802.11 MAC protocol to show the total retransmission count, and *ACKfailCount* indicates the number of retransmission.

188.3 Experimental result and analysis

To evaluate the performance after modifying the basic power control approach, the ns platform (http://www.isi.edu/nsnam/ns/tutorial/) is used. A total of 100 nodes are distributed on an area sized 1000 m × 1000 m. The maximum transmission range of the nodes is 250 m which needs 0.2818 W to transmit. The speed of the nodes is randomly defined. The traffic type is TCP traffic flow and the number is 10. And the size of the packet is increases from 128 Kb to 1440 KB.

As seen from Fig. 188.2, the average throughout of the network with power control is slightly larger than without power control. When the ACK fail count is bigger, the transmission power is larger. So the chance for successful transmission is more. And the modified approach does not cause the Hidden terminal problem and do not need a global information.

From the Fig. 188.3, it can be seen that the average power consumption with power control is less than without power control. Notice that the size of DATA Packet is larger than the other broadcast packets, the power control under the transmission of DATA packet is more efficient. Without power control, more extra power is consumed.

Fig. 188.2 Average Throughout

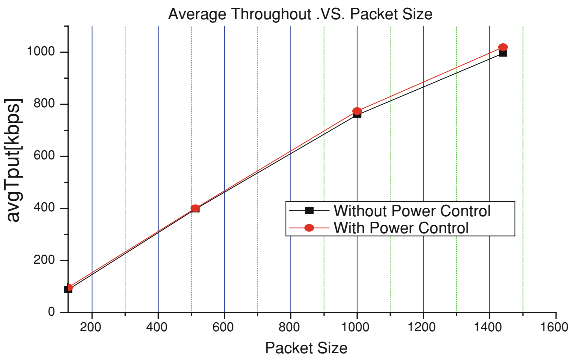


Fig. 188.3 Average consumed energy

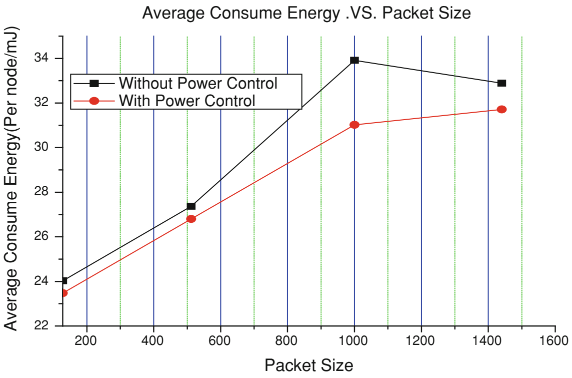
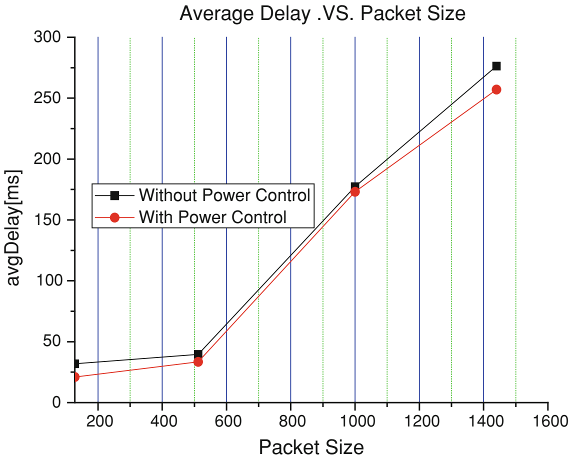


Fig. 188.4 Average delay



It can be seen from Fig. 188.4, the transmission of a TCP flow effects the other flows under the power control. And the transmission of the DATA packet needs less time. After the ACK fail happens, the transmitted power is increased. And then the delay is decreased.

188.4 Conclusion

In this chapter, a simple approach is present to handle the power control in wireless AD Hoc network. The new approach does not need global information and only deal with the power to transmit the DATA and ACK packets. After power control, the new approach shows good performance on throughout, power consumption and delay than unmodified original.

References

1. Lin X, Shroff NB (2005) The impact of imperfect scheduling on crosslayer rate control in multihop wireless networks. In: Proceedings of the IEEE INFOCOM, Miami, Florida, Mar
2. Eryilmaz A, Srikant R (2005) Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control. In: Proceedings of the IEEE INFOCOM, Miami, Florida, Mar
3. Chen L, Low S, Doyle J (2006) Joint congestion control, routing and mac for stability and fairness in wireless networks. IEEE J Selected Areas Commun 24(8):1514–1524
4. Johansson M, Xiao L (2004) Scheduling, routing and power allocation for fairness in wireless networks. In: IEEE VTC-Spring, Milan, Italy, May
5. Chiang M (2005) Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control. IEEE J Selected Areas Commun 23(1):104–116
6. Julian D, Chiang M, O'Neill D, Boyd S (2002) Qos and fairness constrained convex optimization of resource allocation for wireless cellular and adhoc networks. In: Proceedings of IEEE INFOCOM, June, pp 477–486
7. Xi Y and Yeh EM (2006) Node-based optimal distributed power control, routing, and congestion control in wireless networks. In: Proceedings of the conference on information sciences and systems, Princeton, NJ, Mar
8. Tassiulas L, Ephremides A (1992) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. IEEE Trans Autom Control 37(12):1936–1948
9. Modiano E, Shah D, Zussman G (2006) Maximizing throughput in wireless networks via gossip. In: Proceedings of ACM SIGMetric/performance, Saint Malo, France, June
10. Georgiadis L, Neely MJ, Tassiulas L (2006) Resource allocation and cross-layer control in wireless networks. Found Trends Netw 1(1):1–144
11. Nar PC, Cayirci E (2005) PCSMAC A power controlled sensor-MAC protocol for wireless sensor networks. In: Cayirci E, (ed) In: Proceedings of the IEEE EWSN 2005. Piscataway: IEEE Comput Society. 81–92
12. Sarkar M, Gupta A (2010) Power management in wireless ad-hoc networks with directional antennas. Int J Comput Netw Security 2(6):20–29 June

Chapter 189
Application of DAS in Cold Chain Logistics Warehousing System Based on WMS

Sanyou Ji and Xingxing Niu

Abstract To meet the JIT, variety and high frequency requirements of modern cold chain logistics, and application of digital assorting system (DAS) in cold chain logistics warehousing system is studied in this paper. The facilities requirements and its layout are suggested and the key data form is designed to support the DAS system. Wireless weighing system is added to get the weight data opportunely for the unstable weight of commodities in cold chain logistics system. The process of DAS is designed based on WMS and the rush order picking process is provided to achieve an accurate, safe, green and efficient cold chain logistics warehousing system.

Keywords DAS · Cold chain logistics · WMS · Wireless weighing system · RFID

189.1 Introduction

Cold chain logistics references the refrigerated products which is under low temperature throughout manufacture, storage, transportation and marketing. Its a type of system engineering which can ensure the product quality and reduce product loss [1]. Compared to the normal logistics system, cold chain logistics has higher demand, more construction investment and more complicated restriction.

S. Ji (✉) · X. Niu
College of Logistics Engineering, Wuhan University
of Technology, Wuhan, Hubei, China
e-mail: jisanyou@126.com

X. Niu
e-mail: clover_1@163.com

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2_189,

Timeliness requirement of refrigerated products needs high organization and coordination; thus the development of cold chain logistics is closely related to effectively control and operation. Therefore, using barcode technology and RFID technology is of great significance to cold chain logistics. Considering the JIT, variety and high frequency requirements of modern cold chain logistics, RFID assisted picking system is a beautiful solution to achieve a safe, green and efficient picking process in cold chain logistics warehousing system.

189.2 Key Technologies

RFID assisted picking system is a part of WMS. It is a type of paperless picking system. By a series of display electronics unit—RFID tag which installed on departments of rack is used to point the category and quantities. Then picking workers can work with its assistance. It can shorten the operating time significantly, and enhance the picking accuracy. Besides, it can keep the inventory data synch with the picking operation. Thanks to the highly developed technology, RFID tag has been proved stable in low temperature environment.

There are two main types of RFID assisted picking system: digital assorting system (DAS) and digital picking system (DPS).

DPS is based on order processing units. RFID tag is installed on departments in which there are commodities of a certain category. That is to say, a RFID tag represents a category of commodities. When the system is working, RFID tags will be lit on if the corresponding category exists in the order. Picking workers should take out the commodities from the department according to the light and the number displayed on the LED screen [2].

DAS is an efficient and intelligent picking support system [3]. Compared to DPS, DAS can handle bulk orders with higher efficiency and accuracy. The occupancy area and walking distance are smaller. DAS is more applicable to circumstances such as relatively fixed clients, variety or similar categories of commodities, changeful storing place of commodities and so on. Therefore, DAS is more suitable for cold chain logistics warehousing system. Here are some short introductions of the key technologies this paper will use.

WMS. warehouse management system, is a warehouse business operating system for modern enterprises to manage and process commodities [4]. It is a real-time software system which is used to support all the functions of warehouse including goods receiving, putting on shelves, inventory management, replenishment, picking, packing, shipping and so on.

WCS. warehouse control system, is the middle layer software between WMS and equipment electrical control. The main function of WCS is to transfer the tasks generated by WMS into the equipment operation orders at the appropriate time and send messages to WMS according to the actions of the equipment.

Wireless weighing system. Wireless weighing system consists of pressure sensor, data collection terminal, wireless data communication module, digital

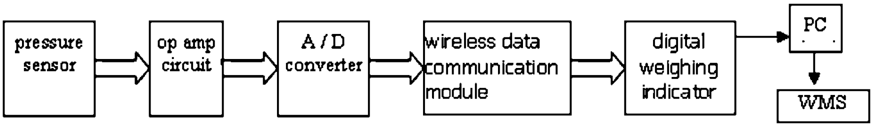

pressure sensor
op amp circuit
A / D converter
wireless data communication module
digital weighing indicator
PC
WMS

Fig. 189.1 Flow chart of wireless weighing system

weighing indicator, printer, PC and large screen display [5]. Weighing plate is placed on the table trolley, when the object is put on the weighing plate, the pressure sensor will turn the pressure signal into analog signal. The op amp circuit will magnify the signal and then transfer it to the A/D converter which can convert the analog signal to digital sample value. Then the digital sample value will be send to the digital weighing indicator through the wireless data communication module. After manipulated in digital weighing indicator, the digital sample value will turn to be dynamic weighing information and be send to WMS through serial ports. The process is shown in Fig. 189.1.

189.3 System Design

189.3.1 Facilities Layout

For the convenience of discharging commodities from warehouse, dynamic rack is used in picking area. The rack is arranged surrounding three sides and temporary storage area is in the middle of the picking warehouse. The dynamic rack is divided into many small departments.In each of the department, there is an RFID tag, or vividly; it can be called as light module. The layout plan is shown in Fig. 189.2.

Dynamic rack. The dynamic rack has two sides of channels, one side of the channels is for depositing commodities, and the other side is for removing. Commodities can be put on the roller. The rack has a tiny camber angle. So the commodities can slide down under gravity [6]. In the picking system, implication of dynamic rack can realize the picking process in the inside and the removing process in the out side. It can make sure the picking process can work with the removing process without disturbing each other.

Light module. It is a kind of RFID tag. In this paper, there are two types of light module. One is red light module; the other is green light module (as shown in Fig. 189.2). The red light module which is installed in each of the departments can be divided into two parts: light and LED display. The light is used to express that commodities should be delivered to this department, and the LED display shows the demanding quantity. The green light module, on the other hand, is installed in the department at the end of the corridor and has only the light to show whether picking process of this category is done. Both of the two types of light module can be lit on in accordance with the orders and will go out with one pat.

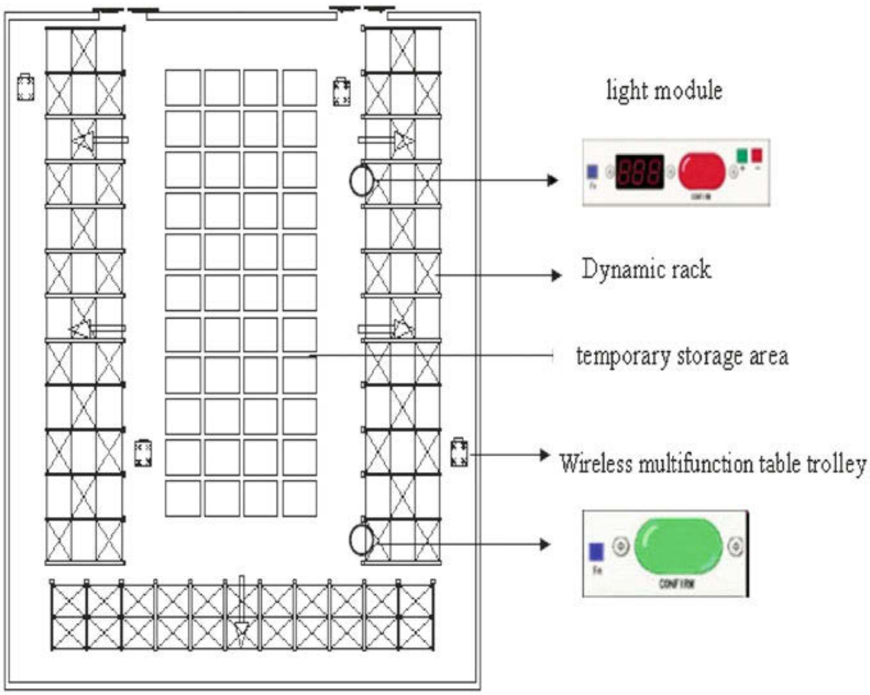


Fig. 189.2 The layout plan of the of the picking warehouse

Wireless multifunction table trolley. Multifunction table trolley is including a platform and a fork. Pressure sensor is placed on the platform. Picking worker can keep the commodity on the fork while picking, then walking along the rack, following the light module to take out the commodities and weigh it on the pressure sensor.

Under low temperature, the weight of refrigerated products is fluctuant with the change of temperature and environment, it is necessary to record the weight data opportunely. Considering this demand, the picking area, the conveying belt does not exist as the picking warehouse in this case. Instead, the multifunction table trolley with weighing system is used for uploading the weight data of commodities in time.

189.3.2 Data Form Design

DAS gather multiple orders; group them by the category of commodities, and then for each category, organizing the assorting process like sowing. The system requires summarizing and handling large number of orders with high speed. Thus

Table 189.1 Part of the table Ex-warehouse

Name	Field name	Type	Null	Constraint
Assorting ID	Ass_ID	VARCHAR2(32)	N	Primary key
Order ID	Order_ID	VARCHAR2(32)	N	–
Compartment ID	Comp_ID	VARCHAR2(32)	Y	Foreign key
Category ID	Goods_ID	VARCHAR2(12)	N	–
Category name	Goods_name	VARCHAR2(60)	Y	–
Measuring mode	Measure_M	VARCHAR2(20)	N	Check in (“quantity”, “weight”)
Weight needed	Order_Weight	NUMBER	Y	–
Assorting weight	Ass_Weight	NUMBER	Y	–
Assorting done	Ass_done	VARCHAR2(20)	N	Check in (“yes”,”no”)

has high requirement to the computer hardware, software and processing algorithms. The main relevant data table is including table Ex-warehouse and table storing compartment information. Part of the important relevant data field of table Ex-warehouse is shown in Table 189.1.

In the table, the assorting ID is generated by the system automatically, it is unique. The initial value of compartment ID is null, while the assorting process beginning, WMS will assign compartment ID for each order according to table storing compartment information. The same order ID corresponding to the same compartment ID. The initial value of assorting weight is null too, it will be fulfilled with the number which is achieved by the wireless weighting system during the assorting system. Assorting done refers to a sign to express the assorting task of that row is done, its initial value is no.

189.3.3 Process Design

While the assorting process begins, WMS will assign compartment ID for each order according to the order ID. Through WCS, the conjunction of order ID and DAS is built. Under normal assorting situation, picking worker scan turnover box by using RFID handheld terminal to get category ID. Through WCS, the category ID will be uploaded to WMS. Then WMS will execute data select command. The relevant data field in table Ex-warehouse will get and displayed in DataGridView in the interface. The key query string is as follows:

```
public void tb_ThGoodsFind(string textID, int intFalg, Object DataObject)
{string strSecar;
switch(intFalg)//enquiries tag
{
case 1://under normal assorting situation
strSecar = “select * from Ex-warehouse where GoodsID = ” +textID + “and Ass_done = no order by Order_ID ”;
break;
case 2://under rush order picking process
strSecar = “select * from Ex-warehouse where orderID = ” + textID;
}
```

Under normal assorting situation, turnover box of the certain category of commodities is put on the fork of table trolley. The light module on the departments will be lit on in accordance with the assorting order through WCS. Picking worker will walk along the rack, take out the commodities from turnover box according to the quantitative information presented on the LED panel of light module and then weigh it on the pressure sensor placed on platform of table trolley. The weight information will be uploaded to WMS through wireless weighting system and trigger the text input events of WMS interface. Then WMS assorting weight will be recorded in data base. The key code is as follows:

```
private void txtWeigh_TextChanged(object sender, EventArgs e)
{string str_Update;
str_Update += “select * from Ex-warehouse_Info go update Ex-warehouse_Info set Ass_Weight = ” + “ txtWeigh.text” + “Ass_ID = ” + textAss_ID.text;
getSqlConnection getConnection = new getSqlConnection();//SQL sever connection
conn = getConnection.GetCon();
cmd = new SqlCommand(str_Update, conn);
conn.Dispose()
}
```

After weighing, the commodities of certain weight and quantity will be put on the department. Then assorting worker offs the light of the light module with a pat. Meanwhile, the value of assorting done bycorresponding item in table Ex-warehouse will be changed into yes by WMS. While each of the items in DataGridView is done, the green light module will turn on for conducting assorting work going to another species of commondities. The system flow chart is shown in Fig. 189.3.

189.3.4 Rush Order Process Module

To increase system flexibility, the rush order process is designed in response to different situations. A department of the dynamic rack is reserved for the rush order in the very beginning of assorting process. While rush order is arising, the controller would click the pause button on the interface of DAS in WMS, and then chose the rush order process option. The dialog box will be popped on in which the controller could input the order ID. Then the enquiries tag intflag will be revised into 2. Clicking the confirm button and back to the main interface of DAS. The interface will be refreshed and the Ex-warehouse information of the rush order will be displayed on the DataGridView instead. The RFID handheld terminal would be used to scan each category of commodities, while for a category existing in the DataGridView, light module on the department will be lit on. Other steps are just

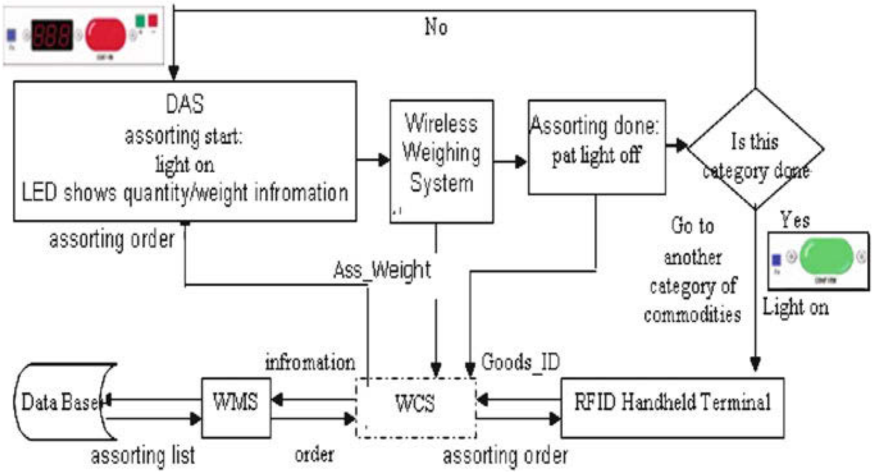


Fig. 189.3 The flow chart of DAS

the same as under normal assorting situation. When the rush order is done, the green light will turn on. Then the controller can back to the suspended normal assorting process or go to deal with another rush order.

189.4 Summary

With the rapid development of national economy and the increasingly strict food requirements of people, the importance the cold chain logistics has taken more and more attention [7]. The picking process, as detailed as it is, is always the most error-prone process. To address this problem, application of DAS in cold chain logistics warehousing system is studied in this paper. Based on WMS, the process of DAS is designed and the rush order picking is provided. Wireless weighting system is used to make sure an accurate real-time data transfer. Through the DAS, an accurate, safe, green and efficient picking process of cold chain logistics warehousing system can be achieved.

References

1. Zhou X (2008) Web information systems. VLDB Database School. China
2. Che X, Jin L (2010) Comparing of DPS and DAS. Logist Technol Appl 8:112–115
3. Information on http://www.transway.com.cn/daili/l-pick/content-1.htm
4. Li B (2005) Study and exploitation on the automated system of WMS. Zhengzou University, China

5. Zu J,Tian Z, He L (2007) Design of wireless weighting system based on RF technology. Process Automation Instrum 2:10–15
6. Wu X, Li L (2005) Warehousing and distribution management. Fu Dan University Press, Sanghai
7. Bao C (2007) On operational management of cold chain. Tongji Unerversity, China

Author Index

Y. Yang and M. Ma (eds.), *Green Communications and Networks*,
Lecture Notes in Electrical Engineering 113, DOI: 10.1007/978-94-007-2169-2,