Lecture Notes on Data Engineering and Communications Technologies 6

Leonard Barolli
Mingwu Zhang
Xu An Wang *Editors*

Advances in Internetworking, Data & Web Technologies

The 5th International Conference on Emerging Internetworking, Data & Web Technologies (EIDWT-2017)

Springer

Lecture Notes on Data Engineering and Communications Technologies

Volume 6

Series editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain
e-mail: fatos@cs.upc.edu

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It publishes latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series has a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

More information about this series at http://www.springer.com/series/15362

Leonard Barolli · Mingwu Zhang
Xu An Wang
Editors

Advances in Internetworking, Data & Web Technologies

The 5th International Conference on Emerging Internetworking, Data & Web Technologies (EIDWT-2017)

Springer

Editors
Leonard Barolli
Fukuoka Institute of Technology
Fukuoka
Japan

Mingwu Zhang
School of Computer Sciences
Hubei University of Technology
Wuhan
China

Xu An Wang
Department of Electronic Technology, Key Engineering University of CAPF
Xi'an, Xizang
China

ISSN 2367-4512 ISSN 2367-4520 (electronic)
Lecture Notes on Data Engineering and Communications Technologies
ISBN 978-3-319-59462-0 ISBN 978-3-319-59463-7 (eBook)
DOI 10.1007/978-3-319-59463-7

Library of Congress Control Number: 2017940631

© Springer International Publishing AG 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Welcome Message of EIDWT-2017 International Conference Organizers

Welcome to the 5th International Conference on Emerging Internetworking, Data and Web Technologies (EIDWT-2017), which will be held from June 10 to June 11, 2017, in Wuhan, China.

The EIDWT is dedicated to the dissemination of original contributions that are related to the theories, practices, and concepts of emerging internetworking and data technologies yet most importantly of their applicability in business and academia toward a collective intelligence approach.

In EIDWT-2017 will be discussed topics related to information networking, data centers, data grids, clouds, crowds, mashups, social networks, security issues, and other Web 2.0 implementations toward a collaborative and collective intelligence approach leading to advancements of virtual organizations and their user communities. This is because current and future Web and Web 2.0 implementations will store and continuously produce a vast amount of data, which if combined and analyzed through a collective intelligence manner will make a difference in the organizational settings and their user communities. Thus, the scope of EIDWT-2017 includes methods and practices which bring various emerging internetworking and data technologies together to capture, integrate, analyze, mine, annotate, and visualize data in a meaningful and collaborative manner. Finally, EIDWT-2017 aims to provide a forum for original discussion and prompt future directions in the area.

For EIDWT-2017 International Conference, we received 254 papers, and out of them 76 were accepted for presentation (about 30% acceptance ratio) during the conference days.

An international conference requires the support and help of many people. A lot of people have helped and worked hard for a successful EIDWT-2017 technical program and conference proceedings. First, we would like to thank all authors for submitting their papers. We are indebted to Program Area Chairs, Program Committee members, and reviewers who carried out the most difficult work of carefully evaluating the submitted papers. We would like to give our special thanks to Honorary Chairs of EIDWT-2017, for their guidance and support. We would like to express our appreciation to our keynote speakers: Prof. Kouichi Sakurai, Kyushu University, Japan, Prof. Jian Weng, Jinan University, China, Prof. Lein Harn, University of Missouri-Kansas City (UMKC), USA and Prof. Qian Wang, Wuhan University, China, for accepting our invitation and delivering very interesting keynote talks at the conference.

We would like as well to thank the Local Arrangements Chairs for making excellent local arrangements for the conference. We hope you will enjoy the conference and have a great time in Wuhan, China.

EIDWT-2017 Steering Committee Chair

Leonard Barolli	Fukuoka Institute of Technology (FIT), Japan

EIDWT-2017 General Co-chairs

Mingwu Zhang	Hubei University of Technology, China
Santi Caballé	Open University of Catalonia, Spain

EIDWT-2017 Program Committee Co-chairs

Debiao He	Wuhan University, China
Elis Kulla	Okayama University of Science, Japan
Wei Ren	China University of Geosciences (Wuhan), China

EIDWT-2017 International Conference Organizers

Honorary Co-chairs

Yingjiang Zhang	Hubei University of Technology, China
Shijie Dong	Hubei University of Technology, China

General Co-chairs

Mingwu Zhang	Hubei University of Technology, China
Santi Caballé	Open University of Catalonia, Spain

Program Co-chairs

Debiao He	Wuhan University, China
Elis Kulla	Okayama University of Science, Japan
Wei Ren	China University of Geosciences (Wuhan), China

Workshops Co-chairs

Xu An Wang	Engineering University of CAPF, China
Giuseppe Fenza	University of Salerno, Italy
Chunzhi Wang	Hubei University of Technology, China

International Advisory Committee

Makoto Takizawa	Hosei University, Japan
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Vincenzo Loia	University of Salerno, Italy

Publicity Co-chairs

Hua Shen	Hubei University of Technology, China
Makoto Ikeda	Fukuoka Institute of Technology, Japan
Admir Barolli	Alexander Moisiu University of Durres, Albania

International Liaison Co-chairs

Yunfei Cao	CETC 30, China
Tomoya Enokido	Rissho University, Japan
Evjola Spaho	Polytechnic University of Tirana, Albania
Anabela Mesquita	ISCAP/IPP and Algoritmi Centre, University of Minho, Portugal
Chia-Wen Tsai	Ming Chuan University, Taiwan
Ghazi Alkhatib	The Hashemite University, Jordan

Local Organizing Co-chairs

Ou Yuan	Hubei University of Technology, China
Yuanyuan Zhang	Hubei University of Technology, China

Web Administrators

Shinji Sakamoto	Fukuoka Institute of Technology (FIT), Japan
Donald Elmazi	Fukuoka Institute of Technology (FIT), Japan
Yi Liu	Fukuoka Institute of Technology (FIT), Japan
Yonghui Chen	Hubei University of Technology, China

Finance Chair

Makoto Ikeda	Fukuoka Institute of Technology (FIT), Japan

Steering Committee Chair

Leonard Barolli	Fukuoka Institute of Technology (FIT), Japan

Track Area Co-chairs

1. Internetworking Issues and Challenges

Chairs

Zhe Xia	Wuhan University of Technology, China
Takahiro Hara	Osaka University, Japan

2. Mobile and Wireless Networks

Chairs

Elis Kulla	Okayama University of Science, Japan
Evjola Spaho	Polytechnic University of Tirana, Albania

3. Network Protocols, Modelling, Optimization and Performance Evaluation

Chairs

Fagen Li	University of Electronic and Science Technology, China
Fabrizio Messina	University of Catania, Italy

4. P2P and Grid Computing

Chairs

Rongmao Chen	National University of Defense Technology, China
Dianhua Tang	CETC 30, China
Kun Ma	University of Jinan, China

5. Distributed and Parallel Systems

Chairs

Giovanni Moranai	C3DNA, USA
Yong Ding	Guilin University of Electronic Technology, China

6. Ontologies and Metadata Representation

Chairs

Chingfang Hsu	Central China Normal University, China
Trina Myers	James Cook University, Australia

7. Knowledge Discovery and Mining

Chairs

Zhenhua Liu	Xidian University, China
Jugappong Natwichai	Chiang Mai University, Thailand

8. Databases and Data Warehouses

Chairs

Jia Yu	Qingdao University, China
Agustinus Borgy Waluyo	Monash University, Australia

9. Data Centers and IT Virtualization Technologies

Chairs

Baocang Wang	Xidian University, China
Omar Hussain	UNSW Canberra, Australia

10. Web Science and Business Intelligence

Chairs

Li Yang	Xidian University, China
Natalia Kryvinska	Comenius University in Bratislava, Slovakia

11. Data Analytics for Learning and Virtual Organisations

Chairs

Licheng Wangi	Beijing University of Posts and Telecommunications, China
Marcello Trovati	Edge Hill University, UK
Lei Zhang	East China Normal University, China

12. Data Management and Information Retrieval

Chairs

Qiong Huang	South China Agricultural University, China
Ziqiang Yu	University of Jinan, China

13. Machine Learning on Large Data Sets & Massive Processing

Chairs

Zhiwei Ye	Hubei University of Technology, China
Giuseppe Fenza	University of Salerno, Italy

14. Data Modeling, Visualization and Representation Tools

Chairs

Guan Gui	Nanjing University of Posts and Telecommunication, China
Bogdan Manate	West University of Timisoara, Rumania

15. Nature Inspired Computing for Emerging Collective Intelligence

Chairs

Yixin Su	Wuhan University of Technology, China
Aboul Ella Hassanien	Cairo University, Egypt

16. Data Sensing, Integration and Querying Systems and Interfaces

Chairs

Youwen Zhu	Nanjing University of Aeronautics and Astronautics, China
Miguel Wister	Universidad Juárez Autonoma de Tabasco, Mexico

17. Data Security, Trust and Reputation

Chairs

Yong Yu	Shaanxi Normal University, China
Hiroaki Kikuchi	Meiji University, Japan

18. eScience Data Sets, Repositories, Digital Infrastructures

Chairs

Jian Shen	Nanjing University of Information Science and Technology, China
Jose Luis Jodra Luque	University of the Basque Country, Spain

19. Energy-Aware and Green Computing in Data Centers

Chairs

Huaqun Wang	Nanjing University of Posts and Telecommunication, China
Tomoya Enokido	Risso University, Japan
Matteo Cristani	University of Verona, Italy

20. Emerging Trends and Innovations in Inter-networking Data Technologies

Chairs

Jie Chen	East China Normal University, China
Danda Rawat	Howard University, Washington, USA

21. Bitcoin, Blockchain Techniques and Security

Chairs

Sheng Gao	Central University of Finance and Economics, China
Shuai Xue	Engineering University of CAPF, China

Program Committee Members

Admir Barolli	Aleksander Moisiu University of Durres, Albania
Akio Koyama	Yamagata University, Japan
Baodong Qin	Southwest University of Science and Technology, China
Bhed Bista	Iwate Prefectural University, Japan
Chandra Bajracharya	Capitol Technology University, USA
Chengyu Hu	Shandong University, China
Chi Cheng	China University of Geosciences (Wuhan), China

Chong Zhang	Armed Police College of CAPF, China
Chotipat Pornavalai	King Mongkut's Institute of Technology Ladkrabang, Thailand
Corrado Santoro	University of Catania, Italy
Danda Rawat	Howard University, USA
David Taniar	Monash University, Australia
Ding Wang	Peking University, China
Douglas Macedo	Federal University of Santa Catarina (UFSC), Brazil
Elis Kulla	Okayama University of Science, Japan
Eric Pardede	La Trobe University, Australia
Eric Renault	Telecom SudParis, France
Evjola Spaho	Polytechnic University of Tirana, Albania
Faliu Yi	UT Southwestern Medical Center, USA
Fang-Yie Leu	Tunghai University, Taiwan
Farookh Khadeer Hussain	University of Technology Sydney, Australia
Feng Xiong	Zhongnan University of Economics and Law, China
Filipe Portela	University of Minho, Portugal
Fushan Wei	PLA Information Engineering University, China
Giancarlo Fortino	University of Calabria, Italy
Giovanni Morana	University of Catania, Italy
Giuseppe Fenza	University of Salerno, Italy
Gongjun Yan	University of Southern Indiana, USA
Haowen Tan	Chosun University, Korea
Hector Gabriel Allende Cid	Escuela de Ingeniería Informática (PUCV), Chile
Houbin Song	West Virginia University, USA
Hui Xia	Qingdao University, China
Ilias Savvas	TEI of Thessaly, Greece
Jordi Conesa	Open University of Catalonia, Spain
José A. Hernández-Nolasco	Juarez Autonomous University of Tabasco, Mexico
Juana Canul-Reich	Juarez Autonomous University of Tabasco, Mexico
Jugappong Natwichai	Chiang Mai University, Thailand
Keiichi Yasumoto	Nara Institute of Science and Technology, Japan
Keita Matsuo	Fukuoka Institute of Technology, Japan
Kun Ma	University of Jinan, China
Li-Ling Hung	Aletheia University, Taiwan
Liangli Ma	Naval University of Engineering, China
Luca Pilosu	Istituto Superiore Mario Boella, Italy
Lucas Nussbaum	University of Lorraine, France
Luis A. Castro	Sonora Institute of Technology (ITSON), Mexico
Makoto Ikeda	Fukuoka Institute of Technology, Japan
Matthias Steinbauer	Onlinegroup, Austria
Meijiao Duan	Central University of Finance and Economics, China

Min Lei	Beijing University of Posts and Telecommunications, China
Minoru Uehara	Toyo University, Japan
Mubashir Husain Rehmani	COMSATS Institute of Information Technology, Pakistan
Olivier Terzo	Istituto Superiore Mario Boella, Italy
Omar Hussain	UNSW Canberra, Australia
Pablo Pancardo	Juarez Autonomous University of Tabasco, Mexico
Pietro Ruiu	Istituto Superiore Mario Boella, Japan
Ping Xiong	Zhongnan University of Economics and Law, China
Pruet Boonma	Chiang Mai University, Thailand
Qi Liu	Nanjing University of Information Science and Technology, China
Qingsong Yao	Xidian University, China
Ran Liu	China University of Geosciences (Wuhan), China
Rong Hao	Qingdao University, China
Ruilin Tan	Engineering University of CAPF, China
Sai Kiran Mukkavilli	Tennessee State University, USA
Siyong Huang	China University of Geosciences (Wuhan), China
Syed Hassan Ahmed	Kyungpook National University, Korea
Tetsuya Shigeyasu	Prefectural University of Hiroshima, Japan
Tetsuya Oda	Okayama University of Science, Japan
Wei Zhang	Engineering University of CAPF, China
Wenhai Sun	Virginia Tech, USA
Wenny Rahayu	La Trobe University, Australia
Xiang Lu	Institute of Information Engineering, Chinese Academy of Sciences, China
Xiaofeng Chen	Xidian University, China
Ximeng Liu	Singapore Management University, Singapore
Xiong Li	Hunan University of Science and Technology, China
Xuefeng Liu	Xidian University, China
Yan Xu	Anhui University, China
Yanbing Zheng	Guilin University of Electronic Technology, China
Yang Lei	Engineering University of CAPF, China
Yanping Li	Shaanxi Normal University, China
Yi Ren	National Chiao Tung University, Taiwan
Yongjun Ren	Nanjing University of Information Science and Technology, China
Yongli Tang	Henan University of Science and Technology, China
Youliang Tian	Guizhou University, China
Yu Yang	Beijing University of Posts and Telecommunications, China
Yuechuan Wei	Engineering University of CAPF, China
Zhenhua Chen	Xi’an University of Science and Technology, China

EIDWT-2017 Reviewers

Ali Khan Zahoor
Barolli Admir
Barolli Leonard
Bessis Nik
Bista Bhed
Caballé Santi
Castiglione Aniello
Chellappan Sriram
Chen Hsing-Chung
Chen Xiaofeng
Cui Baojiang
Di Martino Beniamino
Dobre Ciprian
Durresi Arjan
Enokido Tomoya
Ficco Massimo
Fiore Ugo
Fun Li Kin
Gotoh Yusuke
Hachaj Tomasz
He Debiao
Hussain Farookh
Hussain Omar
Javaid Nadeem
Ikeda Makoto
Ishida Tomoyuki
Kikuchi Hiroaki
Kolici Vladi
Koyama Akio
Kulla Elis
Lee Kyungroul
Loia Vincenzo
Matsuo Keita
Koyama Akio
Kryvinska Natalia
Nishino Hiroaki
Oda Tetsuya
Ogiela Lidia
Ogiela Marek
Palmieri Francesco
Paruchuri Vamsi Krishna
Pop Florin
Rahayu Wenny
Rawat Danda
Ren Wei
Shibata Yoshitaka
Sato Fumiaki
Spaho Evjola
Sugita Kaoru
Takizawa Makoto
Taniar David
Terzo Olivier
Uchida Noriki
Uehara Minoru
Venticinque Salvatore
Waluyo Agustinus Borgy
Wang Xu An
Woungang Isaac
Xhafa Fatos
Yim Kangbin
Zhang Mingwu

EIDWT-2017 Keynote Talks

A Security Management with Cyber Insurance—Event Study Approach with Social Network Sentimental Analysis for Cyber Risk Evaluation

Kouichi Sakurai

Kyushu University, Fukuoka, Japan

Abstract. Since the recent security breach requires the intensification of security management, the documents, describing the best practice of security management, are published by experts. However, the implementations of all best practices are tough because of the cost and the difficulty of cost-effective security investment. In this talk, I will discuss the security management theory with cyber risk insurance, especially the effectiveness of cyber risk insurance by Monte Carlo simulation approach. Once organizations have the security incident and breaches, they have to pay tremendous costs. Although visible cost, such as the incident response cost, customer follow-up care, and legal cost are predictable and calculable, it is tough to evaluate and estimate the invisible damage, such as losing customer loyalty, reputation impact, and the damage of branding. In this talk, I also will present a new method, called Event Study Methodology with Twitter Sentimental Analysis, to evaluate the invisible cost. This method helps to assess the impact of the security breach and the impact on corporate valuation.

Look Back! Earlier Versions will Reveal Weaknesses in Android Apps

Jian Weng

Jinan University, Guangzhou, China

Abstract. Nowadays, Android platform gains explosively growing popularity. A considerable number of mobile consumers are attracted to varieties of Android Apps, which leads developers to invest resources to maintain the upward trajectory. In the early stage, the developers usually pay more attention to the functionality of Android Apps than the security matters. Unfortunately, it makes Android Apps a hot target for attackers. For the sake of resolving the attacks, developers attach great importance to improve the security of Apps and upgrade them to new versions, whereas leave their earlier versions diffuse through the network. In this paper, we indicate how to attack newly versions of popular Apps, including Facebook, Sina Weibo and Qihoo360 Cloud Driven, by using the weaknesses existing in their earlier versions. We design and implement an App weaknesses analysis tool named “DroidSkynet” to analyze the security weakness on widespread applications. Among 900 mainstream Apps collected from real world, our DroidSkynet indicates that 36.3% Apps are suffering from such weaknesses.

Secret Sharing and Its Applications

Lein Harn

University of Missouri-Kansas City (UMKC), Kansas City, USA

Abstract. Secret sharing is one of the most popular cryptographic tools in network applications used to protect data. For example, secret sharing has been used in cloud to strengthen data security. Shamir's threshold secret sharing scheme which was proposed originally in 1979 is the most popular scheme in the literature. In this talk, I will briefly introduce Shamir's scheme and point out some interesting properties. Then, I will introduce some related research problems to the secret sharing, including secure and fair secret reconstruction, verifiable secret sharing, multi-secret sharing, cheater detection and identification in the secret reconstruction. Applications using the secret sharing will also be discussed, such as group key establishment in group communications and group authentication. Finally, I will briefly introduce my recent research paper on the design and implementation of a general secret sharing.

New Short-Range Communication Technologies over Smartphones: Designs and Implementations

Qian Wang

Wuhan University, Wuhan, China

Abstract. With the ever-increasing popularity of smartphones in our daily lives, people more and more heavily rely on them to share and spread a wide variety of information. Because of the limitations of traditional short-range communication technologies (e.g., complex network configuration and troublesome authentication process), some new short-range communication technologies over smartphones have been proposed recently. In this talk, I will discuss barcode-based and acoustics-based short-range communication systems over off-the-shelf smartphones. First, we introduce Dolphin, a novel form of real-time acoustics-based dual-channel short-range communication, which uses a speaker and the microphones on smartphones to achieve concurrent audible and hidden communication. By leveraging masking effects of the human auditory system, Dolphin ensures real-time unobtrusive speaker-microphone communication without affecting the primary audio-hearing experience for human users, while, at the same time, it overcomes the main limitations of existing unobtrusive screen-camera links. Then, we introduce RainBar, a new and improved color barcode-based NFC system for achieving lightweight real-time streaming for smartphones, which features a carefully-designed high-capacity barcode layout design to allow flexible frame synchronization and accurate code extraction.

Contents

An Image Steganalysis Algorithm Based on Rotation Forest Transformation and Multiple Classifiers Ensemble

Zhen Cao[1(✉)], Minqing Zhang[1], Xiaolong Chen[2], Wenjun Sun[1], and Chun Shan[3]

[1] Key Laboratory of CAPF for Cryptology and Information Security,
Engineering University of CAPF, Xian 710086, China
18729207342@163.com

[2] Jinhua Polytechnic, Jinhua 321017, China

[3] Guangdong Polytechnic Normal University,
Guangzhou 510665, Guangdong, China

Abstract. In order to enhance the detection rate of ensemble classifiers in steganalysis, concern the problems that the accuracy of basic classifier is low and the kind of basic classifier is single in typical ensemble classifiers, an algorithm based on rotating forest transformation and multiple classifiers ensemble is proposed. First, some feature subsets generated randomly merger with training sample to generate sample subsets, then the sample subset is transformed by rotating forest algorithm and train some basic classifiers, which is made of fisher linear discriminate, extreme learning machine and support vector machine with weighted voting. At last, the majority voting method is used to integrate the decisions of base classifiers. Experimental results show that by different steganography approaches and in different embedding rate conditions, the error rate of proposed method decreased by 3.2% and 1.1% in compared with the typical ensemble classifiers and ensemble classifiers of extreme learning machines, therefore demonstrating the proposed method could improve the detection accuracy of ensemble classifier.

1 Introduction

The goal of steganalysis is to detect the presence of secretly hidden date in an object. Today, the most accurate steganalysis methods for digital media are built as supervised machine learning on feature vectors extracted from the media [1]. The supervised machine learning mainly consists of three phases: feature extracting, feature selection and classification [2, 3]. In order to detect the highly secure steganography methods based on STC (Syndrome Trellis Coding, STC) [4] and minimizing embedding impact in steganography, steganalysts have to use feature spaces of high dimensionality. For example, the state-of-the-art feature set for image steganalysis, “Rich Models” [5] even has a dimensional of 34,671. Although the highly feature is effective to detect the highly secure steganography methods, but the large feature dimensions cause the curse of dimensionality in steganalysis. To address this problem, the ensemble classifiers based on random forests were proposed in [6]. This method uses many simple basic classifiers.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_1

Each basic classifier is a Fisher Linear Discriminant (FLD). These classifier are trained on different parts of the training dataset and feature subsets. The final decision is taken by combining the decisions of all basic classifiers by majority voting. The ensemble classifiers based on AdaBoost were proposed in [7], but the effect is worse than the typical ensemble classifier proposed in [6]. The ensemble classifiers based on selective ensemble are proposed in [8, 9]. The method improves the accuracy but training time is much higher than typical ensemble classifiers. A cognitive ensemble of extreme learning machines for steganalysis were proposed in [10]. The method used ELM (Extreme Learning Machine, ELM) as basic classifier and gradient descent algorithm to give every ELM a weight. The performance of this method is better, but the training time is higher than typical ensemble classifiers. Cogranne models and extends the ensemble classifiers for steganalysis using hypothesis testing theory in [11]. This method could allow steganalysts to train the ensemble classifiers with desired statistical properties, such as the false-alarm probability, but this method can't enhance the detection rate of ensemble classifiers.

In order to enhance the detection rate of ensemble classifiers, concerning on the accuracy and diversity of basic classifiers, an algorithm based on rotating forest transformation [12] and multiple classifiers ensemble is proposed in this paper. First, the method use rotation forest algorithm to transform the feature subsets. The Principal Component Analysis (PCA) algorithm in rotation forest have the effort of pretreatment which could eliminate the redundant features and improve the accuracy of basic classifier. Meanwhile, the transform could generate different subsets and enhance the diversity of basic classifiers. In the course of training basic classifier, the method trains three different kinds of classifiers as basic classifier instead of only one kind of basic classifier in [6–11]. This method could not only improve the accuracy of basic classifier, but also enhance the diversity of basic classifiers. Experimental results show that the accuracy of proposed method is much higher than that of [6, 10]. And, the training time of this method is smaller than [10].

This paper is organized as follows. In Sect. 2, we will briefly describe the ensemble classifiers proposed in [6] and rotation forest algorithm. In Sect. 3 we present the proposed method. Experimental results are shown in Sect. 4. The performance of the proposed method is tested on three steganography methods: S-UNIWARD [13], HILL [14] and WOW [15]. Furthermore, we compare our method with linear SVM (L-SVM), ensemble classifiers [6] and cognitive ensemble of ELM [10]. Finally, the conclusions are summarized in Sect. 5.

2 Basic Theory

2.1 Ensemble Classifiers

In [6], multiple diverse basic classifiers are combined together to overcome high computational complexity of modern steganalysis methods. In this method, image database is divided into two subsets: training set N^{trn} and testing set N^{tst}. The method use FLD as basic classifier. For each basic classifier, the training set samples are selected using a bootstrap algorithm and each basic classifier is trained on a subset of the whole features set (its size is d_{sub}).

Cardinality of the whole feature dimensionality is d, the basic classifier is B_l and the number of basic classifiers is L. The following is the training process of ensemble classifiers.

```
Algorithm 1：Ensemble Classifiers
Input：Training set : $N^{trn}$
For $l$=1 ：$L$
{
 （1）Form a random feature subspace $d_{sub} << d$
 （2）Form a bootstrap sample $N_l = Bootstrap(N^{trn})$
 （3）Train a base classifier $B_l$ on bootstrap sample $N_l$ with random subspace
      $d_{sub}$
}
Output：The basic classifiers $\{B_1, B_2, \cdots B_L\}$
```

The algorithm proposed in [6] uses majority voting to make the final decision. Cardinality of x is the feature of unnamed picture, $B_l(x)$ is the result of basic classifier. If $B_l(x) = 1$, the picture is a stego picture, else if $B_l(x) = -1$, the picture is a cover picture. The finally decision of ensemble classifiers is

$$B(x) = \begin{cases} 1 & \sum_{l=1}^{L} B_l(x) > 0 \\ -1 & \sum_{1}^{L} B_l(x) < 0 \\ \text{Random,} & \sum_{1}^{L} B_l(x) = 0 \end{cases} \tag{1}$$

2.2 Rotation Forest

In ensemble learning, the diversity of basic classifiers and the powerful of basic classifier are two important properties of ensemble classifiers [16]. Rotation forest transformation is one kind of feature transformation methods. This method could through transforming the training set to improve the powerful and diversity of basic classifiers [17].

First, the rotation forest use Bootstrap algorithm to generate the training subsets, then the method will segment the feature set and use PCA (Principal Component Analysis, PCA) to transform the feature set. This transformation can be as a kind of data pretreatment to improve the powerful and diversity of basic classifiers.

Let x = $[x_1, x_2, \cdots\cdots, x_n]^{\mathrm{T}}$ be a data point describe by n features and let X be the data set containing the training objects in a form of an $N \times n$ matrix. D_1, D_2, …, D_{L} are classifiers and F is the feature set. The following is the process of rotation forests.

```
Algorithm2：Rotation Forests
For $i$=1：$L$
  （1）Split F randomly into K subsets（the subsets are disjoint），the feature’s number is M=F/K.
  For $j$=1:K
    （2）Denote by $F_{ij}$ the $j$th subset of features for training set of classifier $D_i$. And $X_{ij}$ is the data set $X$ on the features in $F_{ij}$
    （3）Select a bootstrap sample from $X_{ij}$ of size 75% of the number of objects in $X_{ij}$. Denote the new set by $X'_{ij}$.
    （4）Apply PCA on $X'_{ij}$ to obtain coefficients in a matrix $C_{ij} = [a_{ij}^1, a_{ij}^2, \cdots, a_{ij}^{M_j}]$
  END
  （5）Arrange the $C_{ij}(j = 1 \cdots K)$ in a rotation matrix $R_i$
      $R_i = \begin{bmatrix} a_{ij}^1, a_{ij}^2, \cdots, a_{ij}^{M_1} & 0 & \cdots & 0 \\ 0 & a_{ij}^1, a_{ij}^2, \cdots, a_{ij}^{M_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{ij}^1, a_{ij}^2, \cdots, a_{ij}^{M_K} \end{bmatrix}$
  （6）Construct $R^a{}_i$ by rearranging the columns of $R_i$ so as to match the order of features in F.
  （7）Build classifier $D_i$ using the $X \times R_i^a$ as the training set.
END
```

At last, we get L basic classifiers and every classifier has a R_i^a. For given an unknown sample x, we use basic classifier to judge the $x' = x \times R_i^a$ and use appropriate method to aggregate the classifiers.

3 The Proposed Algorithm

The key to improve the generalization of ensemble classifiers are accuracy and diversity of basic classifiers. The rotation forest could generate different subsets and use PCA algorithm to improve diversity and accuracy. However, the dimensionality of feature usually is very high. It will consume much time if we use rotation forests directly. And we also need much space to storage R_i^a (If the dimensionality of feature is n, the dimensionality of R_i^a is $n \times n$). To solve this problem, we first generate some feature subsets randomly and use rotation forest to them. This method could optimize the features which are used to train basic classifier directly. And this method could get more diversity than method in paper [6].

In steganalysis, ensemble classifiers all use one kind of classifiers as basic classifier, which causes the insufficient of accuracy and diversity of basic classifier. Instead of it, we propose to train three different classifiers to composite a basic classifier. This strategy has two advantages: 1. Every basic classifier is a simple ensemble classifiers, so it can improve the accuracy of basic classifier. 2. This strategy could enhance the diversity because the different classifiers have different mechanisms and there are different internal weights in basic classifiers.

3.1 Basic Classifier

We train three different classifiers to instead of only one kind of basic classifier in typical ensemble classifiers. The three classifiers are correlated and use weighted voting method to get the decision. The three different classifiers are FLD, ELM which ‘sigmoid’ is active function, SVM with RBF kernel. The following is the process to train basic classifiers.

```
Algorithm 3：The training process of basic classifiers
Input：Feature with $d$ dimensionality, Training samples $N$.
 For $i$=1:$L$
  （1）Generate a feature subset $F_i$ and dimensionality is $d_{sub} << d$
  （2）Extract 75% training samples randomly from $N$ to generate training set $N_1$ and
  the last is $N_2$ which is used for calculating weights.
  （3）Run rotation forest to training set $N_1$ in features subset $F_i$ and then train the
  FLD classifier $B_{\mathrm{iF}}$.
  （4）Run rotation forest to training set $N_1$ in features subset $F_i$ and then train the
  ELM classifier $B_{\mathrm{iE}}$.
  （5）Use FLD to distinguish the training set $N_1$ and record the wrong samples in
  $N_{\mathrm{R1}}$. Use ELM to distinguish the training set $N_1$ and record the wrong samples in
  $N_{\mathrm{R2}}$.
  （6）Let $N_{\mathrm{R}}= N_{\mathrm{R1}} \cap N_{\mathrm{R2}}$. Then run rotation forest to training set $N_{\mathrm{R}}$ in features
  subset $F_i$ and train SVM $B_{\mathrm{iS}}$.（the reason why we use the wrong samples of FLD
  and ELM to train SVM is:1. The training time of SVM is longer than FLD and
  ELM，the small samples could reduce the training time of SVM. 2.The SVM
  could become a special classifier to deal with the difficult samples.）
  （7）Use training samples in $N_2$ to calculate weights for three classifiers
Output：The basic classifier and weights $D_i = [\{\mathrm{FLD,ELM,SVM}\},\{w_1, w_2, w_3\}]$
END
```

3.2 Weight Assignment for Classifiers

We use the method proposed in [18] to calculate weights for classifiers. There are two methods in [18], one is for classifiers that are independent and another is for depend. In this paper, we use method for depend case.

In our proposed method, let X_1, X_2, X_3 stand for B_{iF}, B_{iE} and B_{iS}. e_1, e_2, e_3 are error rate on N_2 and w_1, w_2, w_3 are weights for B_{iF}, B_{iE} and B_{iS} respectively. Each X_i takes on two possible values, 1 with probability e_i, which represents misclassification, and 0 otherwise. The objective is to find a weight assignment w_i, satisfying $w_i \geq 0 \sum_{i=1}^{3} w_i = 1$ and $e = P\{\sum_{i=1}^{3} w_i X_i > 1/2\}$ is minimized. Because the joint distribution of (X_1, X_2, X_3) is unknown, this problem does not lead to easy numerical solutions.

For solving this problem, the paper [18] formulate this problem from an alternative perspective that can be solved easily and has some statistical justifications. We cast our goal as to minimize expectation of the square loss $\mathrm{E}(\sum_{i=1}^{3} w_i X_i)^2$. According to the relationship between expectation and variance, we can use the formula (2) as the object function.

$$\begin{aligned} &\min\{\sum_{i=1}^{3} (EX_i^2) w_i^2 + 2 \sum_{1<i<j\leq 3} (EX_i X_j) w_i w_j\} \\ &\sum_{i=1}^{3} w_i = 1 \end{aligned} \tag{2}$$

In this function, we can use error rate to stand for the probability of $X_i = 1$, such as $P\{X_i = 1\} = e_i$. If we know the numerical of EX_iX_j, the object function will become a quadratic programming problem.

We use statistical method to get the pair-wise correlations between classifiers. For example, let B_{iF} and B_{iE} to judge the samples in N_2, we denote a as the number of samples both classifiers make correct results, b as the number of samples both classifiers make wrong results, c and d as the number samples where they have different results. Then, we can get: $P\{X_1 = 0,\ X_2 = 0\} = a/N_2$, $P\{X_1 = 1,\ X_2 = 1\} = b/N_2$, $P\{X_1 = 0, X_2 = 1\} = c/N_2$, $P\{X_1 = 1, X_2 = 0\} = d/N_2$. And the EX_1X_2 can be calculated as b/N_2. We also can get other EX_iX_j with the same method. Then, the object function will become a quadratic programming by standardizing this problem and we can obtain the weights by solving it. In this paper, we use Matlab optimization toolbox to solve this problem.

We can get w_1, w_2 and w_3 for B_{iF}, B_{iE} and B_{iS} respectively. Then we use formula (3) to get the decision of basic classifier.

$$D_i = \begin{cases} -1 & w_1 B_{i\mathrm{F}} + w_2 B_{i\mathrm{E}} + w_3 B_{i\mathrm{S}} < 0 \\ 1 & w_1 B_{i\mathrm{F}} + w_2 B_{i\mathrm{E}} + w_3 B_{i\mathrm{S}} > 0 \\ Rand & w_1 B_{i\mathrm{F}} + w_2 B_{i\mathrm{E}} + w_3 B_{i\mathrm{S}} = 0 \end{cases} \tag{3}$$

After we get the decisions of every basic classifier, our method use majority voting to aggregate the classifiers and get the finally result. The whole process of proposed method is in Fig. 1.

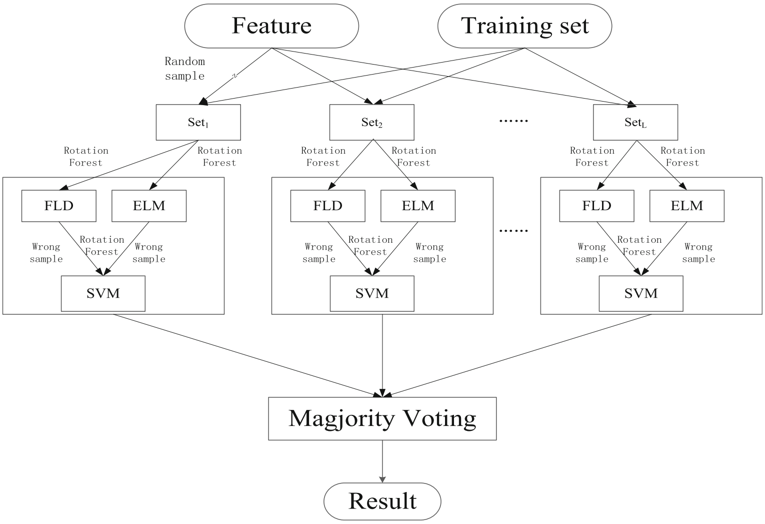


Fig. 1. The flow chart of proposed method

3.3 Determination of d_{sub} and L

From Fig. 2 we can know the error rate of ensemble classifiers will become stability with the number of basic classifiers increase gradually when the numerical of d_{sub} is certain. The Table 1 gives the stability of error rate and the number of basic classifier in different d_{sub}. From the Table 1, we know that the different d_{sub} has different stable

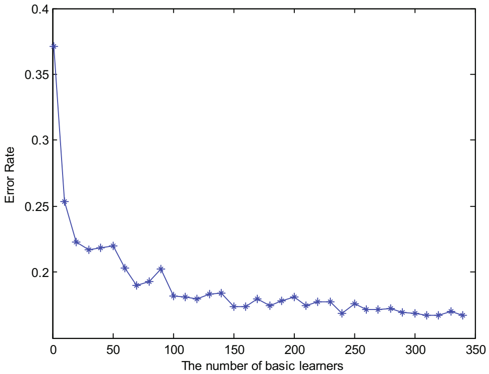


Fig. 2. Error rate with different number of basic classifiers

Table 1. The error rate and L of different d_{sub}

d_{sub}	L	P_{E}
100	109	0.2096
200	253	0.1763
250	262	0.1849
300	312	0.1751
325	304	0.1814
350	298	0.1920

error rate and L. When $d_{sub} = 300$ and $L = 312$, we get the lowest error rate. And according to the Table 1 and the correlation between L and error rate. We chose the numerical of d_{sub} is 300 and the number of basic classifiers is 320 at last.

4 Experiment Results

4.1 Experimental Environment

In this section, we show the performance of proposed method to detect three steganography methods S-UNIWARD, HILL and WOW. We compare it with that of linear SVM, typical ensemble classifiers in paper [6] and ensemble ELM classifiers in paper [10].

The experiments are executed on the BossBaseV1.01 [19]. This database contains 10000 grayscale raw images. The whole dataset is randomly divided into two equal size subsets as training and testing set. In this paper, we use the 34671 dimensional SRM feature set. The dimensionality of feature subsets d_{sub} is 300 and it will be divided into 5 different sets when we use rotation forest algorithm. The number of basic classifiers is 320. The experimental environment is Windows Sever 2012 (Inter Xeon E5620 16 GB memory) and the simulation software is MATLAB_R2012a.

4.2 Comparison of Error Rate

The error rate could reflect the power of classifiers quantitatively. In this paper, we compared the error rate of different classifiers. The experiment results are shown in Table 2.

As Table 2 show, the error rate of ensemble classifiers is lower than SVM and our method has the lowest error rate in all classifiers when we detect S-UNIWARD HILL and WOW with different relative payloads. To compare the typical ensemble classifiers [6] and ensemble ELM classifiers [10], the error of our method is lowered by 3.21% and 1.12% on average, respectively.

Table 2. The error rate of different methods

Steganography	Payload	L-SVM	Paper [6]	Paper [10]	Proposed algorithm
S-UNIWARD	0.1	0.4315	0.4143	0.3912	0.3778
	0.2	0.3306	0.3162	0.2951	0.2833
	0.3	0.2679	0.2519	0.2323	0.2221
	0.4	0.2186	0.2034	0.1836	0.1733
HILL	0.1	0.4607	0.4530	0.4318	0.4197
	0.2	0.4106	0.3912	0.3712	0.3611
	0.3	0.3665	0.3490	0.3293	0.3195
	0.4	0.3190	0.3030	0.2841	0.2757
WOW	0.1	0.4645	0.4540	0.4294	0.4148
	0.2	0.4179	0.3990	0.3766	0.3646
	0.3	0.3372	0.3265	0.3054	0.2937
	0.4	0.3024	0.2853	0.2653	0.2553

4.3 Comparison of ROC Curve

ROC curve (Receiver Operating Characteristic ROC) is an important indicators which can reflect the power of classifiers intuitively. The more sleek and closer to the top left corner the curve is, the more powerful the classifier is. AUC (Area Under ROC Curve, AUC) stands for the area of the ROC curve which can evaluate ROC curve quantitatively. The larger the number of AUC is, the better the curve is.

The Figs. 3, 4 and 5 show the ROC curve and AUC of three different classifiers when we detect the three steganography methods. From the three figures, the ROC curve of our method is more sleek and closer to the top left corner than other two methods. To compare the typical ensemble classifiers and ensemble ELM classifiers, the AUC of our method is higher by 3% and 1.4% on average, respectively. From the ROC curve, we can get the conclusion that our method is better than other classifiers in steganalysis.

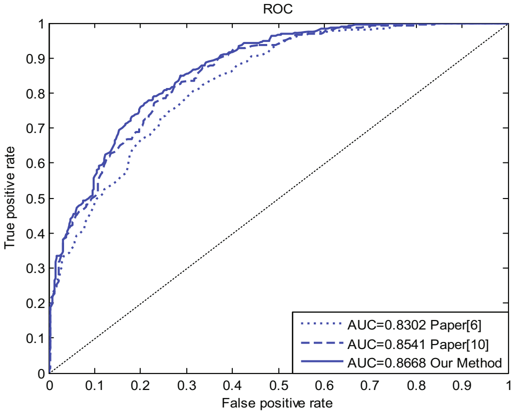


Fig. 3. The ROC curve of S-UNIWARD (0.4 bpp) detection

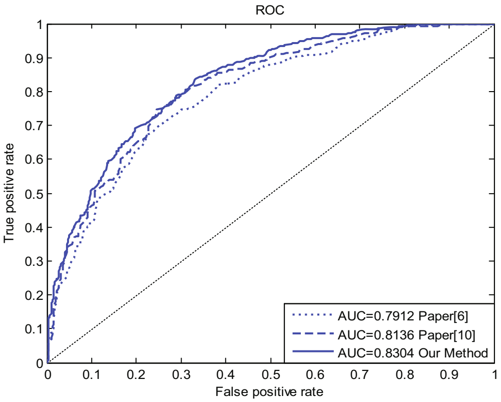


Fig. 4. The ROC curve of HILL (0.4 bpp) detection

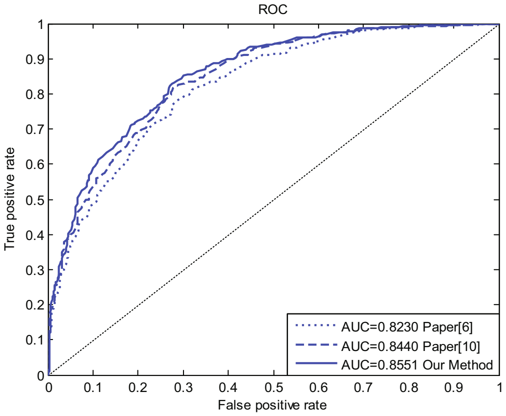


Fig. 5. The ROC curve of WOW (0.4 bpp) detection

4.4 Comparison of Training Time

The training time is another an important indicator. This paper compares the training time of our method to other classifiers when we detect the three steganography methods with 0.4 bpp.

The Table 3 shows that the training time of ensemble classifiers is lower than L-SVM because the former can overcome the curse of dimensionality. The training time of ensemble ELM classifiers is higher than typical ensemble classifiers because it need use gradient descent algorithm to give every ELM a weight. Because we use rotation forest algorithm and train three different classifiers to combine a basic classifier, the training time is also higher than typical ensemble classifiers, but it is lower than ensemble ELM classifiers. And the accuracy of our method is better than them.

Table 3. The training time of different methods

Steganography	L-SVM	Paper [6]	Paper [10]	Proposed algorithm
S-UNIWARD	18.0 min	2.2 min	9.0 min	5.8 min
HILL	19.1 min	3.0 min	11.2 min	8.0 min
WOW	18.6 min	2.8 min	10.3 min	6.8 min

5 Conclusion

In this paper, we propose an algorithm based on rotation forest transformation and multiple classifiers ensemble. This method uses rotation forest to transform features and trains three different classifiers as basic classifiers. As experimental results show, this algorithm is more powerful than typical ensemble and ensemble ELM classifiers. In the future works, we will study how different training and test image sources affect the accuracy and decrease the training time.

Acknowledgments. This work was supported in part by the National Nature Science Foundation of China (Grant Nos. U1636114, 61402531, 61572521, 61379152) and the Nature Science Basic Research Plan in Shaanxi Province of China (Grant Nos. 2014JM8300, 2014JQ8358, 2015JQ6231, 2016JQ6037) and the Public Welfare Research Project of Guangdong Province (Grant Nos. 2014A010103031).

References

1. Sedighi, V., Fridrich, J.: Effect of saturated pixels on security of steganographic schemes for digital images. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 2747–2751. IEEE, September 2016
2. Tang, W., Li, H., Luo, W., Huang, J.: Adaptive steganalysis based on embedding probabilities of pixels. IEEE Trans. Inf. Forensics Secur. **11**(4), 734–745 (2016)
3. Li, F., Wu, K., Lei, J., Wen, M., Bi, Z., Gu, C.: Steganalysis over large-scale social networks with high-order joint features and clustering ensembles. IEEE Trans. Inf. Forensics Secur. **11** (2), 344–357 (2016)
4. Filler, T., Judas, J., Fridrich, J.: Minimizing additive distortion in steganography using syndrome-trellis codes. IEEE Trans. Inf. Forensics Secur. **6**(3), 920–935 (2011)
5. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. **7**(3), 868–882 (2012)
6. Kodovsky, J., Fridrich, J., Holub, V.: Ensemble classifiers for steganalysis of digital media. IEEE Trans. Inf. Forensics Secur. **7**(2), 432–444 (2012)
7. Kodovský, J.: Steganalysis of digital images using rich image representations and ensemble classifiers. Doctoral dissertation, State University of New York (2012)
8. Zhang, M.Q., Di, F.Q., Liu, J.: Universal steganalysis based on selective ensemble classifier. J. Sichuan Univ. **47**(1), 36–44 (2015)
9. Li, F.Y., Zhang, X.P.: Steganalysis for color images based on channel co-occurrence and selective ensemble. J. Image Graph. **20**(5), 609–617 (2015)

10. Sachnev, V., Ramasamy, S., Sundaram, S., Kim, H.J., Hwang, H.J.: A cognitive ensemble of extreme learning machines for steganalysis based on risk-sensitive hinge loss function. Cogn. Comput. **7**(1), 103–110 (2015)
11. Cogranne, R., Fridrich, J.: Modeling and extending the ensemble classifier for steganalysis of digital images using hypothesis testing theory. IEEE Trans. Inf. Forensics Secur. **10**(12), 2627–2642 (2015)
12. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: a new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. **28**(10), 1619–1630 (2006)
13. Denemark, T., Fridrich, J., Holub, V.: Further study on the security of S-UNIWARD. In: IS&T/SPIE Electronic Imaging, p. 902805. International Society for Optics and Photonics, February 2014
14. Li, B., Wang, M., Huang, J.: A new cost function for spatial image steganography. In: Proceedings of IEEE International Conference on Image Processing, pp. 27–30 (2014)
15. Holub, V., Fridrich, J.: Designing steganographic distortion using directional filters. In: 2012 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 234–239. IEEE, December 2012
16. Wang, X.X.: Research on classifier selective ensemble method and their diversity measurement. Master dissertation, Lanzhou University of Technology (2011)
17. Mao, S.S., Xiong, L., Jiao, L.C., Zhang, S., Chen, B.: Isomerous multiple classifier ensemble via transformation of the rotating forest. J. Xidian Univ. **41**(5), 48–53 (2014)
18. Li, X., Zhao, H.: Weighted random subspace method for high dimensional data classification. Stat. Interface **2**(2), 153 (2009)
19. Pevný, T., Filler, T., Bas, P. (eds.): Break Our Steganographic System. http://boss.gipsa-lab.grenobleinp.fr

Separable and Three-Dimensional Optical Reversible Data Hiding with Integral Imaging Cryptosystem

Liu Yiqun(✉)

Key Laboratory of CAPF for Cryptology and Information Security,
Department of Electronic Technology, Engineering University of Chinese Armed
Police Force, Xi'an, Shaanxi 710086, China
wjliuyiqun@126.com

Abstract. Reversible data hiding in encrypted domain (RDH-ED) is an important and effective technical approach for security data management of cloud computing, big data and privacy protection. This paper proposes a three-dimensional (3D) optical reversible data hiding (RDH) with integral imaging cryptosystem. The secret data is encrypted and embedded into the cover image. The receivers can decrypt the cover image and secret data with a reversible or lossless manner, respectively. The simulation experiment and results show that the data embedding rate can be increased to one. Besides, the quality of image decryption is quite high. The technique boasts the advantages of high data embedding rate, security level and real-time capability.

Keywords: 3D-ORDH-InImC · Reversible data hiding in encrypted domain (RDH-ED) · Three-dimensional optical information hiding

1 Introduction

The security of information systems is increasingly crucial in our lives, as everything is going to be connected to the Internet [1–4]. Barton presented the concept of reversible data hiding (RDH) for the first time in his patent in 1997 [5]. He adopted the lossless compression technology to create more redundant space in image and realized reversible hiding of carrier image and secret information. After that, RDH gradually becomes a new hot spot of information hiding research field. According to the current research situations, RDH can be divided into six categories [6], i.e., RDH of spatial domain, RDH of compressed domain, semi-fragile RDH, RDH of cipher-text domain, RDH of audio and video, and RDH of contrast enhancement type. The existing reversible information methods of spatial domain mainly include RDH of difference expansion, RDH of histogram shifting, RDH of lossless image compression and RDH of contrast enhancement.

RDH-ED is an important and effective technical approach for security data management of cloud computing, big data and privacy protection. The existing methods include RDH method based on private key cipher [7–9] and RDH method based on public key cipher [10, 11]. Zhang Xinpeng et al. [7] jointed encryption and information

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_2

hiding technology, and put forward a RDH algorithm in cipher images, which owned the advantage of convenient operation and could meet the reversible requirements. However, the encryption algorithm was too simple and image decryption was required before the secret image was extracted. As a result, steganography load and steganography quality were greatly limited by cipher image. The literatures [10–12] utilized encryption carrier data of public key cipher and homomorphic encryption embedding information. The algorithm led to obvious expansion of encrypted data volume, complex computation and low embedding capacity. The literatures [7, 8, 13, 14] carried out pre-processing to compress partial data before image encryption and then hided information. It could guarantee the reversibility. However, it couldn't be regarded as a method for the encryption domain. The true information hiding of encryption domain shall be carried out completely in the encryption domain, and no characteristic should be public when the carrier data is in plaintext state. But many algorithms for the cipher-text domain fail to deal with this problem.

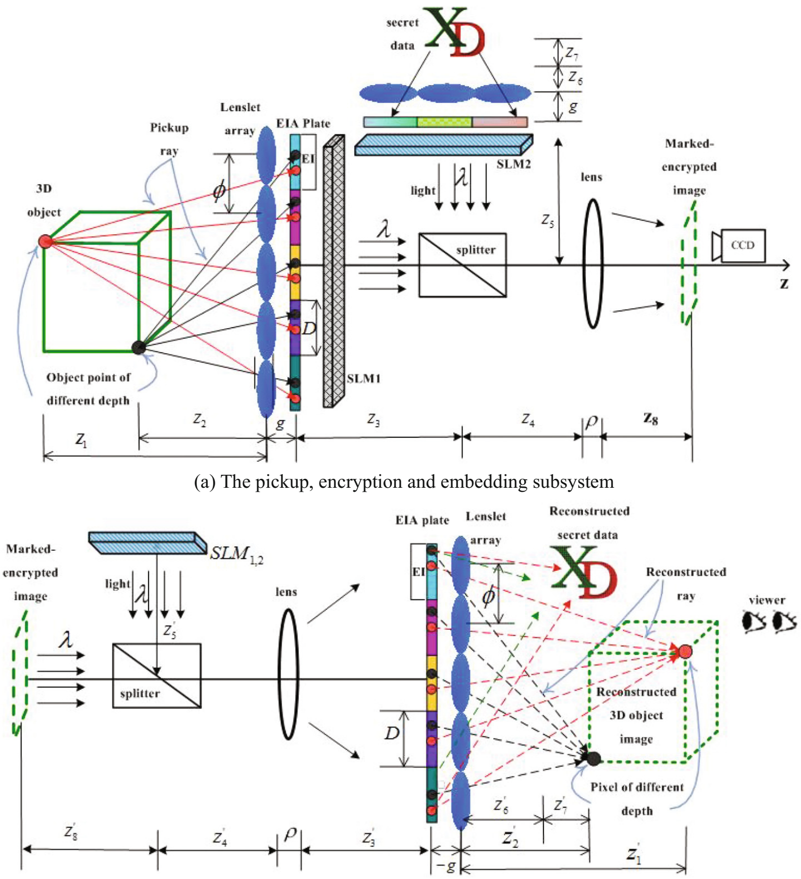
The main motivation for using optical technology of optics and photonics for information security is that optical waveforms possess many complex degrees of freedom such as amplitude, phase, polarization, nonlinear transformations, quantum properties of photons, and multiplexing that can be combined in many ways to make information encryption more secure and more difficult to attack [15, 16]. Among published research reports, patents and literatures [1–4, 6, 17–21], there are few researches on jointing optical technologies, integral imaging and RDH, not to mention three-dimensional multimedia RDH. Therefore, it's worthy taking full advantage of optical technologies for three-dimensional RDH technology based on integral imaging cryptosystem.

This paper proposes a three-dimensional optical reversible data hiding with integral imaging cryptosystem (3D-ORDH-InImC). We have researched on the technical principles, implementation algorithm and implement workflows of 3D-ORDH-InImC. They are introduced in detail. Finally, the simulation experiment was done, and the results proved that the data embedding rate could be approximately increased to 1, which was higher than that of the existing methods by about 60%. Besides, the image decryption quality was quite high. We also further analyzed influence of change in key space element on image decryption quality. The system boasts the advantages of high data embedding rate, computation efficiency, security level and real-time capability, and thus can meet the performance requirements of RDH. The current state of the arts, it is the first scheme on jointing the integral imaging and RDH in cipher-text domain.

2 The Principle of the Proposed Scheme

The 3D-ORDH-InImC is designed and shown in Fig. 1. It is divided into two subsystems. Figure 1 (a) is the pickup, encryption and embedding subsystem and Fig. 1 (b) is the extraction, decrypted and display subsystem.

Therein, A and B represent the two planes which separate spatially in the direction of propagation. s, t, z_{AB} and λ denote the sampling number of two adjacent orthogonal pixels, the spacing between the planes, the wavelength of incident light, respectively. We define correlation sampling lengths of the input plane along the x and y axes as Δx



(a) The pickup, encryption and embedding subsystem

(b)The extraction, decryption and display subsystem

Fig. 1. The schematically of 3D-ORDH-InImC

and Δy, and the Fourier plane along the ξ and η axes in Fresnel transform domain (FTD) as $\Delta\xi$ and $\Delta\eta$, respectively. C is a complex constant whose value may be calculated by the formula (2).

$$\begin{aligned} DFD[A,B,s,t;z_{AB},\lambda] &= \frac{\exp[j2\pi z_{AB}/\lambda]}{j\lambda z_{AB}} \times \exp[j\frac{\pi}{\lambda z_{AB}}(s^2\Delta\xi^2 + t^2\Delta\eta^2)] \\ &\times \sum_{q=0}^{N-1}\sum_{l=0}^{N-1} U_A(q,l)\exp[j\frac{\pi}{\lambda z_{AB}}(q^2\Delta x_0^2 + l^2\Delta y_0^2)] \times \exp[-j2\pi(\frac{qs}{N} + \frac{lt}{N})] \end{aligned} \tag{1}$$

Where

$$C = \frac{\exp[j2\pi z_{AB}/\lambda]}{j\lambda z_{AB}} \tag{2}$$

As we all know, since $DFD[A, B, s, t; z_{AB}, \lambda]$ is complex value in Formula (1), it pickups both the amplitude and phase information of the result signal in the optical implementation described in Fig. 1.

In order to improve security of the cryptosystem, the EIA images are encrypted by Discrete 2D-logistic [22] algorithm. The most important characteristics of chaos encryption algorithm are efficiency and high speed in encryption. It is especially applicable to real-time communication [22]. Therefore, the plain-text images are encrypted with Discrete 2D-logistic algorithm and these cipher-text images are marked as $W_1(x, y), W_2(x, y)$.

$$\begin{aligned} G_{RDH}(\omega, \gamma) = \{&\alpha_1 DFD[W_1(x, y), L(x, y), s, t; z_{W_1}, \lambda] \\ &+ \alpha_2 DFD[W_2(x, y), L(x, y), s, t; z_{W_2}, \lambda] + \alpha_3 DFD[R_1(x, y), L(x, y), s, t; z_{R_1}, \lambda] \\ &+ \alpha_4 DFD[R_2(x, y), L(x, y), s, t; z_{R_2}, \lambda]\} \times T(s, t; f) \end{aligned} \tag{3}$$

Assume that $W_1(x, y)$,$W_2(x, y)$, $R_1(x, y)$, $R_2(x, y)$,$L(x, y)$ represent planes in different position, respectively. s,t are the number of pixel samples. Z_j represents the distance between the different planes, where $j = 1, 2, \cdots 9, W_1, W_2, R_1, R_2$. These symbols $z_{W_1} = z_{R_1} = z_3 + z_4$, $z_{W_2} = z_{R_2} = z_5 + z_4$ and g represent the distance between the lenslet array and elemental image plane. D represents the size of elemental images. ϕ represents lenslet spacing. The focal length of the imaging lens ρ is f, $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are encryption weighting factors, where $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$. They are used to adjust the energy ratio among the DFD transforms of the cover image, 3D digital secrete image and RPMP. The optical transfer function is $T(s, t; f)$. $G_{RDH}(\omega, \gamma)$ is marked encrypted image that is encrypted cover images containing secret data or additional bits.

The three-dimensional decryption, extraction and display subsystem is shown in Fig. 1(b). The legal user can receive these marked encrypted images transmitted by the secure communication network, implement subtraction through contribution in embedding of RDH of two random phase mask plates, and then obtain cipher-text image according to Fourier transform and inverse Fresnel diffraction transformation theory, as shown in Eq. (5). The mathematical model of decryption can be represented with Eqs. (4)–(6). Implement decryption of EIA images of carrier image and secrete image with the original value set by Discrete 2D-logistic. Finally, three-dimensional images can be displayed by the 3D-ORDH-InImC.

$$\begin{aligned} EW' = G_{RDH}(\omega, \gamma) - DFD\{&\alpha_3 DFD[R_1(x, y), L(x, y), s, t; z_{R1}, \lambda] \times T(s, t; f)\} \\ -DFD\{&\alpha_4 DFD[R_2(x, y), L(x, y), s, t; z_{R2}, \lambda] \times T(s, t; f)\} \end{aligned} \tag{4}$$

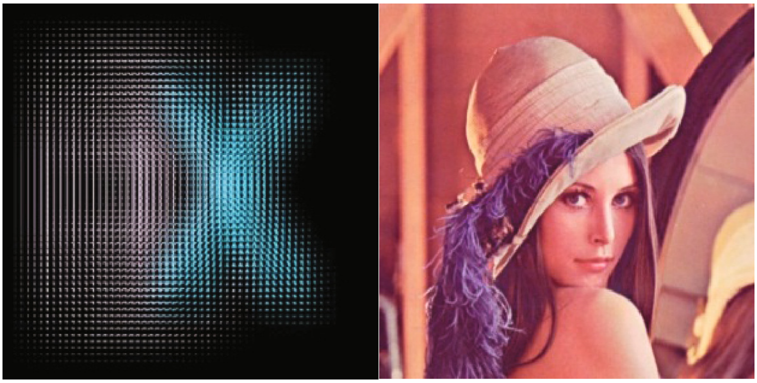
$$W' = IDFD[EW']|_{z = z'_{W_{1,2}}} \tag{5}$$

Among them, the diffraction distance $z'_{W_{1,2}}$ can be calculated by the following formula:

$$\frac{1}{z_{W_{1,2}}} + \frac{1}{z'_{W_{1,2}}} = \frac{1}{f} \tag{6}$$

3 Experimental Results and Discussions

As shown in Fig. 2, Lena image and XD EIAs are chosen as the cover images and secrete images. Their sizes are all in 512 × 512 pixels. Joint Photographic Experts Group (JPEG) is the image format.

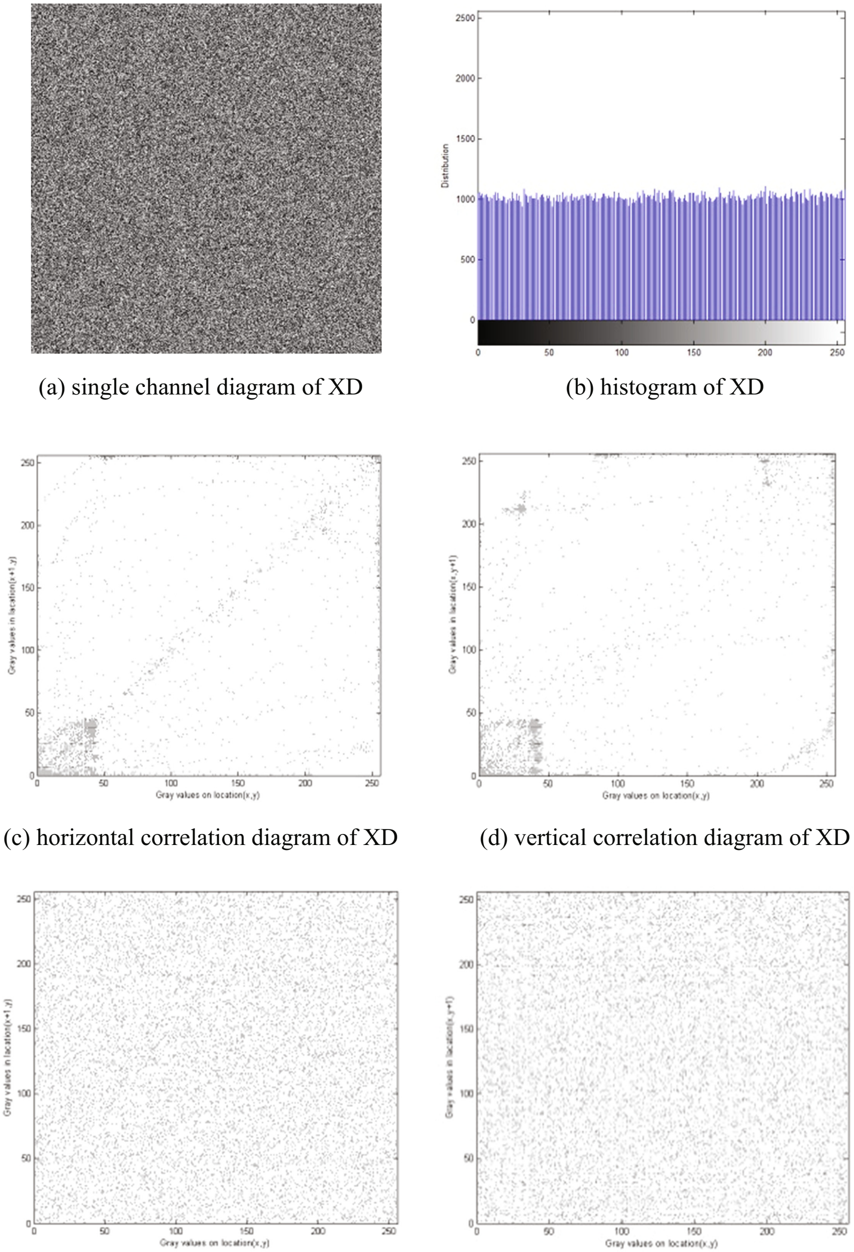


(a) XD EIA (b) Lena

Fig. 2. Sample images of the experiments

XD EIAs are encrypted by discrete 2D-logistic algorithm. For analyzing the correlation of neighboring pixels, we chose scatter diagrams in the horizontal and vertical directions to characterize the correlation. Figure 3 shows correlation between neighboring pixels of plaintext images and cipher-text images. According to the above figures, neighboring pixels of plaintext images have large correlation, and distribution diagrams of cipher-text images are relatively uniform. They indicate the correlations of neighboring pixels in two directions are relatively small. Therefore, the encryption algorithm can greatly reduce pixel correlation of cipher-text images, improve the capability to resist the statistical analysis attacks, and meet the cipher-text image evaluation index requirements of histogram statistics and neighboring pixel correlation.

The security and robustness are improved by the proposed method. The linear characteristics of the 4*f* system are enhanced with these techniques such as discrete 2D-logistic, DFD and double random phase coding, etc. That is to say, Plain images can be encrypted by scrambling the pixel position or pixel value replacement. Thus the integral security of the optical cryptosystem is improved. Discrete 2D-logistic



(a) single channel diagram of XD

(b) histogram of XD

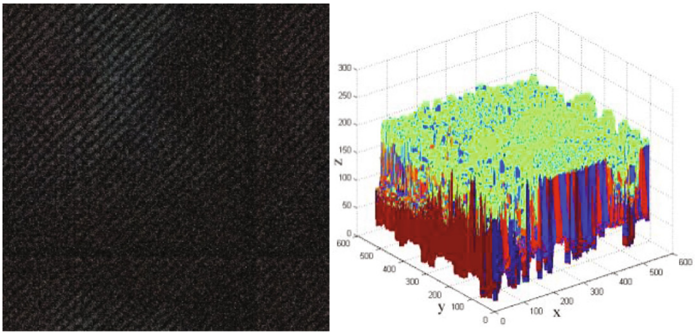
(c) horizontal correlation diagram of XD

(d) vertical correlation diagram of XD

(e) horizontal correlation diagrams of encrypted XD

(f)vertical correlation diagrams of encrypted XD

Fig. 3. The correlation diagrams of XD and encrypted XD



(a) marked encrypted image (b) pixel distribution diagram

Fig. 4. The pixel distribution diagram of the marked encrypted image

algorithm is a kind of digital image encryption technology based on the chaos theory, and it can enhance the system security. First, the higher the sensitivity of the original value of chaotic mapping, the smaller the correlation between neighboring pixels will be. After these images are scrambled, the better ergodic property, the larger randomness of scrambling image will be. Therefore, during the pixel scrambling, the sensitivity of the original value of chaotic mapping and ergodic property determine the scrambling intensity. Second, during the pixel scrambling and the pixel replacement, the more number of iterations, the higher the encryption intensity will be. At the same time, the exhaustion become more difficult than additional technologies and computing complexity is higher. It's necessary to trade off the security and computing complexity when users decide the number of iterations. Finally, the key sensitivity is determined by the parameter sensitivity when mapping parameters are used as the scrambling key. While key sensitivity will resolve security of the whole system. 3D-ORDH-InImC can improve the encryption performance and practicability. Figure 4(a) and (b) are marked encrypted image and pixel distribution diagram.

In the new method, the optical device parameters and functionality parameters of integral imaging system can be used as keys. These modulation parameters for secrete data embedding coefficient can also be used as keys. Multiple keys synthesizing the above mentioned keys will enhance security. It could be more difficult to crack. By redistributing signal energy and diffusing hidden signal energy embedded into transformation coefficients in spatial and temporal domains, the DFD embedding and extraction algorithm applied in the new method effectively resolves the contradiction between imperceptibility and robustness of information hiding. Thus, the new method can meet robustness and security requirements.

4 Conclusion

A three-dimensional (3D) optical reversible data hiding (RDH) with integral imaging cryptosystem is proposed. The secret data is encrypted and embedded into the cover image that is encrypted by 3D-ORDH-InImC. The receivers can decrypt the cover image and secret data with a reversible or lossless manner, respectively. The simulation experiment and results prove that the data embedding rate can be increased to one. The technique boasts the advantages of high data embedding rate, security level and real-time capability. The proposed method can be used in such fields as three-dimensional new media information hiding and multimedia information security. This paper is a powerful new example for optical information security theory. In the future work, we'll research on characteristics of three-dimensional image cryptosystem from the perspectives of cryptanalysis and information theory, and then improve strategies for the system security. To the best of our knowledge, three-dimensional multimedia information security is developing forward from scientific theoretical research to engineering technology in spite of many challenges.

References

1. Liu, Y., Wang, X., Zhang, J., Zhang, M., Luo, P., Wang, X.A.: An improved security 3D watermarking method using computational integral imaging cryptosystem. Int. J. Technol. Hum. Interact. (IJTHI) **12**, 1–12 (2016)
2. Pereira, R., Pereira, E.G.: Future internet: trends and challenges. Int. J. Space-Based Situated Comput. (IJSSC) **5**, 159–167 (2015)
3. Akase, R., Okada, Y.: WebGL-based 3D furniture layout system using interactive evolutionary computation and its user evaluations. Int. J. Space-Based Situated Comput. (IJSSC) **4**, 143–164 (2014)
4. Moore, P., Thomas, A., Tadros, G., et al.: Detection of the onset of agitation in patients with dementia: real-time monitoring and the application of big-data solutions. Int. J. Space-Based Situated Comput. (IJSSC) **3**, 136–154 (2013)
5. Barton, J.M.: Method and apparatus for embedding authentication information within digital data. In: US Patent. US: 1997
6. Shi, Y.-Q., Li, X., Zhang, X., et al.: Reversible data hiding: advances in the past two decades. IEEE Access 3210–3237 (2016)
7. Zhang, X.: Reversible data hiding in encrypted image. IEEE Signal Process. Lett. **18**, 255–258 (2011)
8. Lian, S., Liu, Z., Ren, Z., Wang, H.: Commutative encryption and watermarking in video compression. IEEE Trans. Circuits Syst. Video Technol. **17**, 774–778 (2007)
9. Cancellaro, M., Battisti, F., Carli, M., et al.: A commutative digital image watermarking and encryption method in the tree structured Haar transform domain. Signal Process. Image Commun. **26**, 1–12 (2011)
10. Memon, N., Wong, P.W.: A buyer-seller watermarking protocol. IEEE Trans. Image Process. **10**, 643–649 (2001)
11. Kuribayashi, M., Tanaka, H.: Fingerprinting protocol for images based on additive homomorphic property. IEEE Trans. Image Process. **14**, 2129–2139 (2005)
12. Jiayong, C., et al.: A secure image steganographic method in encrypted domain. J. Electron. Inf. Technol. **34**, 1721–1726 (2012)
13. Zhang, X., Long, J., Wang, Z., Cheng, H.: Lossless and reversible data hiding in encrypted images with public key cryptography. IEEE Trans. Circuits Syst. Video Technol. **26**, 1622–1631 (2015). 1
14. Ma, K., Zhang, W., Zhao, X., et al.: Reversible data hiding in encrypted images by reserving room before encryption. IEEE Trans. Inf. Forensics Secur. **8**, 553–562 (2013)

15. Markman, A., Carnicer, A., Javidi, B.: Security authentication with a three-dimensional optical phase code using random forest classifier. J. Opt. Soc. Am. A **33**, 1160–1165 (2016)
16. Javidi, B., Carnicer, A., Yamaguchi, M., et al.: Roadmap on optical security. J. Opt. **18**, 1–39 (2016)
17. Wang, Y., Du, J., Cheng, X., Lin, Z.L.K.: Degradation and encryption for outsourced PNG images in cloud storage. Int. J. Grid Util. Comput. **7**, 22–28 (2016)
18. Honarvar, A.R., Sami, A.: Extracting usage patterns from power usage data of homes' appliances in smart home using big data platform. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**, 39–50 (2016)
19. Alamareen, A., Al-Jarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**, 1–14 (2016)
20. Mola, L., Rossignoli, C., Carugati, A., Giangreco, A.: Business Intelligence System Design and its Consequences for Knowledge Sharing, Collaboration, and Decision-Making: An Exploratory Study. Int. J. Technol. Hum. Interact. (IJTHI) **11**, 1–25 (2015)
21. Balusamy, B., Krishna, P.V.: Collective advancements on access control scheme for multi-authority cloud storage system. Int. J. Grid Util. Comput. **6**, 133–142 (2015). Special Issue on Intelligent Grid and Cloud Computing
22. Sun, F., Liu, S.: Cryptographic pseudo-random sequence from the spatial chaotic map. Chaos, Solitons Fractals **41**, 2216–2219 (2009)

Credit Risk Assessment of Peer-to-Peer Lending Borrower Utilizing BP Neural Network

Zhengnan Yuan[1], Zihao Wang[2], and He Xu[2,3](✉)

[1] Glasgow College, UESTC,
University of Electronic Science and Technology of China, Chengdu, China
916260629@qq.com
[2] School of Computer,
Nanjing University of Posts and Telecommunications, Nanjing, China
1390424429@qq.com, xuhe@njupt.edu.cn
[3] Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, China

Abstract. This paper proposes an innovated approach of risk assessment of borrowers based on the BP neutral network model. Specifically, firstly, referring to the empirical data published by the website 'peer-to-peer lender' and the indicators of personal credit risk assessment from commercial bank is an efficient method to pick several valid values through data processing, classification and quantification, then the final modeling indicators are selected by information gain technology. Secondly, the new credit risk assessment model is formed after training the modeling indicators. Meanwhile, several strings of collected testing data would be substituted to find out the default rates which are supposed to be compared with the practical ones on the website and the calculated ones from existing credit risk assessment evaluating models. Last but not the least, the effect of this new method is evaluated.

1 Introduction

P2P (Peer-to-peer) Website which is under an admirable escalation with the promotion from the comprehensive regulatory and the new international marketing environment has been operating for over a decade [1–3]. However, as a result of the limitation in scale, comparing with traditional commercial banks, peer-to-peer lending companies demonstrate a weak controllability when facing the varying risks, ranging from credit risk, legal risk, regulatory risk, information asymmetry, investment risk, discipline risk, settlement risk and information security risk. Among those the most serious problem is the credit risk – most of the platforms have not polish up their model of risk assessment whereas present typical credit assessment model of lending has innovated into a totally different state. Simply applying those models that are merely suitable for classical traditional finance model could not be practical.

The most innovative part of this study is utilizing BP neutral network, information gain technology and introduction of values to evaluate the credit risk assessment of P2P

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_3

lending and getting a desired result, conclusion and testing results. On top of that, some expected advancements which root in the conclusion of this study on personal credit assessment of peer-to-peer lending borrower based on BP neural network are proposed at this paper.

The internet financial is becoming different from not only indirect financing of traditional commercial banks, but also the capital market whose indirect fund is raised by direct fund new financial model with the improvement of modern information technology especially the internet. Starting from 2007, P2P lending came to China and gradually becomes the delegate of internet financial models [4–7]. P2P network lending is designed to match the requirement of individual borrowers and the loans of small or medium-sized entrepreneurs. As the intermediary platform, P2P network platform allow the individuals and small or medium-sized entrepreneurs that have idle funds or are willing to loan to publish the information of loan with rate and due date selected by themselves. As a result, more deals could be done by this kind of method.

2 BP Neutral Network

BP neutral network (Background Propagation), also called error back propagation network, was firstly introduced by Werbos in 1974. In 1985, Rumelhart and other scholars did effort to develop the theory and propose clear and strict algorithms [8]. BP algorithms is applied to forward network. It applies the training forms which the tutors' help is involved, and it can also provide both the input and the output vector product simultaneously. By using the back propagation learning algorithms and adjusting the link weight of network, the network output is expected to be approximate to the expected output to the greatest extent under the condition of least mean square error. The progress of backward learning consists of forward and backward propagation. Specifically, in the process of forward propagation, the input information transfers to output layer after the testing of hidden neutron, if the output layer could not receive the output as expected, then the information will transfer to the backward propagation process where the error of the actual output compared to the expected one will be sent back in the former connected channel [9]. Eventually, through modifying authority of the connecting of each layer neurons, the errors can be reduced, and then it can transfer to the forward propagation process where a recycle is formed until the error is less than a given value.

3 Application Flow Chart

Firstly, being a new product of the modern information society, P2P lending has not established efficient systems that is related to the information risk evaluation mechanism so far, and the evaluation for the information risk of borrower is not impeccable. However, BP neutral network has the characteristics of self-adjusting, high self-study and high flexibility which can adjust itself merely according to the variations of the environment, then find regulations for large amount of data and provide relatively correct inference results based on those regulations. Thus, BP neutral network has strong practical feasibility on the P2P lending's defects which includes uncertainty of information, lacking of efficient systems that is related to the information risk evaluation mechanism and the evaluation for the information risk of borrower. In addition, BP neutral network can display the professors' knowledge, experience and thoughts, thus it can get rid of the subjective evaluation as much as possible. Then it is obvious that the credit evaluation can be more precise. In the end, BP neutral network model is a nonlinear modeling process which is not necessary to learn the nonlinear relationship between data. Technically, this indicates that it can effectively overcomes the difficulties of choosing the suitable model functions in the traditional modeling process, and it can establish the modeling speedily. As a result, it can be applied in various fields. The algorithm flow chart for using BP is shown in Fig. 1.

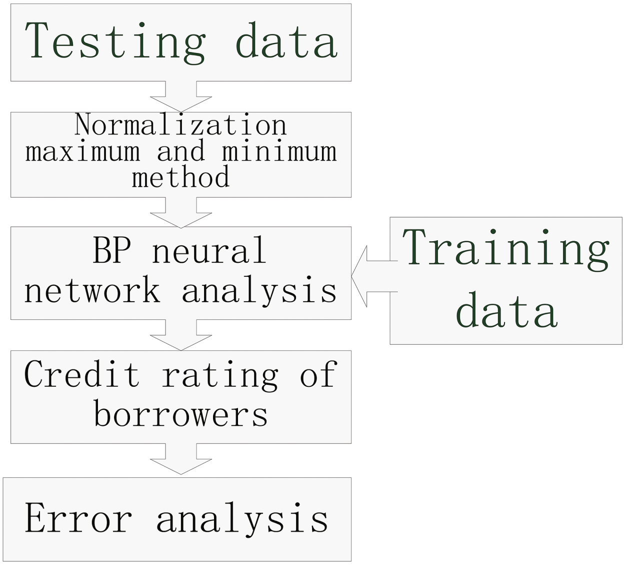

Testing data
Normalization maximum and minimum method
BP neural network analysis
Training data
Credit rating of borrowers
Error analysis

Fig. 1. Algorithm flow chart

4 Model Construction

4.1 Target Selection

P2P network lending platform generally requires the borrower to provide personal information including the identity, occupation, property and other personal basic conditions. After this state is complicated, the information borrowers provide is judged through the field of authentication, video authentication, and so on, to ensure the authenticity of the information. According to the certification, followed by the assessment of the credit rating of the borrowers, the information and credit rating results will be published on the website as a reference for lenders. Therefore, according to the characteristics of P2P lending on the website, the data in Renren Loaning Website and the selection principle of the traditional commercial bank personal credit rating index, the basic information of the borrower is concluded into five aspects: demographic characteristics, occupation status, income and property, credit history operation and certification. Considering of the personal credit rating of commercial banks which is combined with the characteristics of P2P network platform, various indicators of the importance of the credit rating and five aspects of credit evaluation index are quantitative. The reasons for the selection of the index and the value are as follows:

Demographic characteristics: demographic characteristics include age, marital status and educational level. For age, there is a significant difference in default rates among borrowers of different ages. Generally, the default rate of individuals from 35 to 50 years-old who own a stable job and in satisfied economic condition is low. On top of that, for 26- to 35-year-old borrowers, although their income may experience an increase, the pressure comes from the families are serious problem of rising the default rate to a general level. Borrowers of whose ages are below 25, their low income, the lack of mercury spending habits and mostly no saving capabilities all contribute to the high risk of defaulting. Those older than 50 years old, whose level of income begin to decline, are more likely to have sudden consumption. As a result, the default risk is relatively large. Apart from that, marital status is also a key to this, married borrowers are more reliable and divorces or unmarried borrowers may be in low credit status. Lastly, the education is also essential, overall, the higher the education level is, the lower the probability of occurrence of default is.

Occupational status: Occupational status, ranging from unit type, type of jobs and years of services. As for unit type, generally speaking, people who work for the government could have a more stable source of income and have less likelihood of default. Therefore, they are of the highest value. The larger the firm sizes, the more stable income level they have. As a result, the default rate is smaller. For post type, higher post has higher level of earnings with small risk of default. Moreover, for work experience, the longer years of working leads to a more reliable income levels. Those individuals also seem to have lower risk of default.

Income property: First of all, the high the income is, the smaller the default is. Secondly, as for properties, especially that of China, housing conditions normally represent individual economic capacity, so there are existing a smaller risk of default than that of non-real state. Last but not the least, cars can also represent the economic capacity of people, that is to say, the car owners are less likely to default.

Credit history: Credit history is mainly reflected by the number of successful repayment (i.e. Successful repayment times: The more number of successful repayment, the better the credit inertia is and the less likely to violate rules) and the number of overdue repayment (i.e. overdue repayment times: the more number of overdue repayment, the worse the history of credit history is and the easier to violate rules).

Operation certification: The more kinds of authentication, the more reliable the information is, then the more complete and the smaller the possibility of default is. According to the standard of the personal credit rating, combined with the characteristics of P2P network platform, the higher index value and the higher credit rating, the smaller the possibility of default is.

4.2 Data Processing

This paper extracts transaction data from a number of online trading platforms (i.e. a total of 14 k borrower information) as a sample of P2P personal borrowers' credit risk assessment. Then those data would be transformed into quantitative data according to the personal credit indicators in qualitative indicator [10].

In general, the input sample values of the neural network are required to be normalized. In this paper, the maximum and minimum method is utilized to normalize the quantitative data of personal credit indicators, that is to say, using the formulas below to normalize. Maximum and minimum method is a kind of linear transformations that will not cause too much loss according to formula (1). And our code for BP network to process the data is shown as the following.

$$\mathbf{u}_i = \frac{\boldsymbol{u} - \boldsymbol{min}(\boldsymbol{u})}{\boldsymbol{max}(\boldsymbol{u}) - \boldsymbol{min}(\boldsymbol{u})} \tag{1}$$

```
pseudo-code:
Program prepare data for BP network (Input)
{Gain main information}Repeat
i=42:1:50:[data,text]=xlsread(strcat(''dataintegration/traini
ng data/platform',num2str(i),'.csv'));
count=size(text,1);
i=1:10303,strcmp(listO(i),'installment')==1:n_listO(i)=1;
else : strcmp(listO(i),'one-off payment')==1: n_listO(i)=2;
else : n_listO(i)=3;end;
i=1:10303, strcmp(listQ(i),'male')==1:n_listO(i)=1;
else : n_listO(i)=2;end;
i=1:10303,strcmp(listS(i),'underguaduate')==1||strcmp(listS(i
),'graduate')==1: n_listS(i)=1;
else : n_listO(i)=2;end;
i=1:10303, strcmp(listT(i),'married')==1: n_listS(i)=1;
else : strcmp(listT(i),'unmarried')==2;
else : n_listO(i)=3;end;
i=1:10303, strcmp(listW(i),'\N')==1: n_listW(i)=1;
else : n_listW(i)=2;end;
i=1:10303, strcmp(listX(i),'yse')==1: n_listX(i)=1;
else : n_listX(i)=2;end;
i=1:10303, strcmp(listY(i),'yse')==1: n_listY(i)=1;
else : n_listY(i)=2;end;
Program training data for BP network (Input)
{Process for F data including yuan or RMB}
y1=strfind(S, 'yuan');r1=strfind(S, 'rmb')；
r2=strfind(S, 'RMB');Repeat
if there exists any these key word: transform data to lisfF
{Process for J data including loan rate}
listJ=data(:,10);
money_rate: aJ=data(:,10); if listJ(2)<=1 listJ=listJ*100;
listO=text(:,14);%repay_type
listQ=text(:,16);%borrower_sex
listR=data(:,18);%borrower_age
if there exists any these key word: get digital information
in the string one by one and transform data to lisfF
{Input data}i=length(aF); b=text{i,5};
A=isstrprop(b,'digit');B=b(A); C=str2num(B);listF(i)=C;
listJ=[listJ;data(:,10)];if listJ(2)<=1;listJ=listJ*100;
{Input list information}
i=1:length(listO):S={'a'};S=listO(i);
{Input the kinds of loan} n_listO(i)=n(n=1~4);
{Input gender} n_listQ(i)=n(n=1~3);
{Input education} n_listS(i)=n(n=1~5);
{Input marriage state}n_listW(i)=n(n=1~2);n_listT(i)=n(n=1~4);
{Input house information state} n_listX(i)=n(n=1~3);
{Input car information} n_listY(i)=n(n=1~3);end
If i<=27553:state(i)=1;
Else  state(i)=-1; end;
```

5 Model Processing

5.1 Model Description

In this paper, the personal credit risk assessment process of P2P internet lending platform is simulated by the three-layer neural network [11]. Input layer mode number is 11.

The output layer is the credit rating of the individual borrower of the P2P platform which refers to the classification of platforms. The number of nodes in the output layer is 1 and the selected values are 10, 8, 6, 4 and 2, corresponding to the five credit levels, respectively. Specifically, the highest credit rating is 10 (i.e. the least likely to default), a minimum of 2 credit rating value of 1 (i.e. the lowest level which is the most likely breach of contract and cannot repay in time). The approximate range of the number of nodes in the hidden layer is firstly determined by the golden section method, and then the optimal number of nodes in hidden layers is determined through experiment.

5.2 Model Simulation

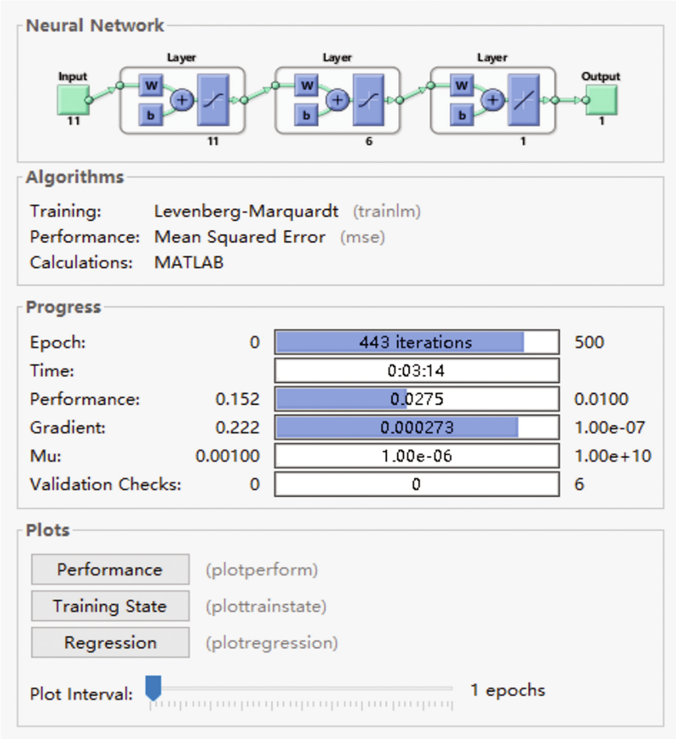
Before the simulation, 14000 sets of data from 10 different platforms are simulated and integrated as training data including 2000 of defaulted and 12000 of clean loan records. Training function is used to build BP neural network with epoch set to 500, mean squared error to 0.001 and the number of hidden layer to 5. All those are shown in Fig. 2.

This program is aimed to evaluate the risk of P2P station. There are 40 sets of training data including 16 sets of normal data which are provided by working stations and 24 sets of abnormal data which are provided by bankrupt stations. And the given 10 sets of predicting data are used to evaluate the risk.

Each station contains a number of records of loan, and the risk of P2P station is relevant to the evaluation of every record, so we decide to evaluate the risk of P2P station by evaluating the risk of each record of loan. It is obvious that the station is considered as a bankrupt station when the number of the risky records in this station exceed the threshold.

Referring to the information provided by experience and cogitate the weight of each property, 11 properties are chosen as input:

column F ITEM_AMOUNT: the amount of the loan money, it is a key property.
column J MONEY_RATE: the rate of this loan, it is a key property.
column O REPAYTYPE: the repay type of this loan, it is a key property.
column Q BORROWER_SEX: the gender of the borrower, it is borrower's personal information.



Algorithms

Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MATLAB

Progress

```
Epoch:                 0     443 iterations     500
Time:                        0:03:14
Performance:       0.152     0.0275             0.0100
Gradient:          0.222     0.000273           1.00e-07
Mu:              0.00100     1.00e-06           1.00e+10
Validation Checks:     0     0                  6
```

Fig. 2. Settings before training

column R BORROWER_AGE: the age of the borrower, it is borrower's personal information.
column S BORROWER_EDUCATION: the degree of the borrower, it is borrower's personal information.
column T BORROWER_MARRIAGE: the marriage of the borrower, it is borrower's personal information.
column W BORROWER_INCOME: the income of the borrower, it is borrower's personal information.
column X BORROWER_HOUSE: the house of the borrower, it is borrower's personal information.
column Y BORROWER_CAR: the car of the borrower, it is borrower's personal information.
column Z REWARD: the reward of this loan.

6 Data Analysis

Since the information of the P2P network platform is entered by the borrowers and is not mandatory, it is possible that some information is missed from the borrowers or the borrowers intentionally conceal the information which causes some mistakes in the information. The results are shown in Fig. 3. Therefore, when an individual borrower evaluates a credit risk, there are some missing or invalid information. One of the characteristics of the BP neural network model is offering a more accurate grading result by the training result in the case of partial data missing. This paper excluded some indicators, ranging from state condition, passenger vehicles, however, the output of the model and the output of the target are the same. In the absence of unit type, job type and income range, the difference between the model output and the target output is large. On top of that, if the number of successful borrowing, out-of-date repay and lack of confirming information would cause a huge gap between the model output and the target output. However, there is no reversal result which means that the borrower with low credit risk will not be regarded as a borrower that has high credit risk. In a nutshell, this data can still be regarded as the basis for credit risk evaluation of borrowers in P2P network credit. Thus, in the absence of fuzzy information, BP neural network model still has the ability of P2P network credit borrower credit prediction and the assessment accuracy rate is still high. Our data code is shown in the following.

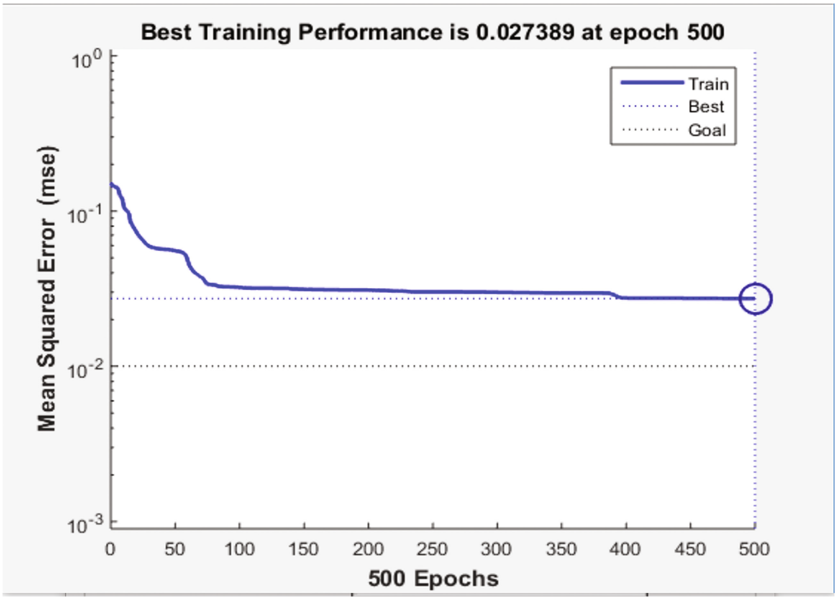


Fig. 3. The accuracy of the valuation of data from 10 different platforms is 80%.

```
pseudo-code:
Program Credit Risk Assessment (Output)
{Load train data and test data with normalization};
P=[listF';listJ';n_listO;n_listQ;listR';n_listS;n_listT;n_lis
tW;n_listX;n_listY;listZ'];
[p1,minp,maxp,t1,mint,maxt]=premnmx(P,state);
a=[listF';listJ';n_listO;n_listQ;listR';n_listS;n_listT;n_lis
tW;n_listX;n_listY;listZ'];
normalization:aa=tramnmx(a,minp,maxp);
output: b=sim(net,a);
predicting data: c=postmnmx(b,mint,maxt);
{Build and set network}
net=newff(minmax(P),[11,6,1],{'tansig','tansig','purelin'},'t
rainlm');
net.trainParam.epochs =500;
net.trainParam.goal=0.01;
net.trainParam.lr = 0.05;
  repeat
i=1:10: cnt(i)=0;
i=1:10303, i<634,  c(i)>0:cnt(1)=cnt(1)+1;
i=1:10303, 634<i<1249, c(i)>0:cnt(2)=cnt(2)+1;
i=1:10303, 1249<i<1266, c(i)>0:cnt(3)=cnt(3)+1;
i=1:10303, 1266<i<3915, c(i)>0:cnt(4)=cnt(4)+1;
i=1:10303, 3915<i<4124, c(i)>0:cnt(5)=cnt(5)+1;
i=1:10303, 4124<i<4752, c(i)>0:cnt(6)=cnt(6)+1;
i=1:10303, 4752<i<5285, c(i)>0:cnt(7)=cnt(7)+1;
i=1:10303, 5285<i<7809, c(i)>0:cnt(8)=cnt(8)+1;
i=1:10303, 7809<i<9036, c(i)>0:cnt(9)=cnt(9)+1;
i=1:10303, 9036<i<10303,c(i)>0:cnt(10)=cnt(10)+1;
number=[300 300 10 1300 400 340 270 1400 800 800];
i=1:10, cnt(i)>number(i)：pt(i)=1;
i=1:10, cnt(i)<number(i)：pt(i)=0; end;
```

7 Conclusion

The credit risk assessment model of credit borrower of P2P network based on the BP neural network in this study works well since the credit risk of the personal borrower can still be measured accurately even though some information is missing or ambiguous. Specifically, this study that has a certain applicability is expected to be popularized and used. The reason why the BP neural network credit risk assessment model in this study has a reliable ability of evaluation is that the BP neural network itself is good at the discovery of the knowledge and extraction of characteristic values which is suitable for the credit assessment. Overall, the results of the whole experiment demonstrate that BP neural network is a desirable selection for the credit risk assessment of individual borrower in the P2P network. In a nutshell, according to the conclusion and results of the experiment above, this paper would like to propose the countermeasures and suggestions to not only improve the credit risk evaluation of P2P borrowers but also facilitate a healthy processing of P2P platform:

Firstly, strengthen the information authentication of the P2P network lending platform to ensure the accuracy of personal information. As is known to all, the credit rating of the borrower is based on the information provided by the borrowers, so the authenticity and accuracy of the information are the key to the rating system. To improve this, the website is supposed to carry out real-time authentication or certification of the information to avoid misleading information which could do harm to the profit of the lenders.

Secondly, increase the disclosure of P2P network lending platform information. Since more personal information could help the lenders have a more comprehensive understanding for the borrowers and so as the internet to adjust the credit rating of borrowers.

Thirdly, P2P network lending platform is expected to disclosure the overdue repayment list on time. On the one hand, the number of overdue repayments is critical for the credit rating. The rating could be adjusted properly through the on-time disclosure. If this aim is achieved, the lender is able to know the real condition of the borrowers. On the other hand, under this invisible pressure from the disclosure, the borrower could repay the money on time and pay attention to the credit.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P.R. China (No. 61373017, No. 61572260, No. 61572261, No. 61672296, No. 61602261), the Natural Science Foundation of Jiangsu Province (No. BK20140886, No. BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No. 12KJA520002), China Postdoctoral Science Foundation (No. 2014M551636, No. 2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1302090B, No. 1401005B), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks (WSNLBZY201509), NUPTSF (Grant No. NY214060, No. NY214061) and the STITP projects of NUPT (No. XZD2016032 and No. XYB2016532).

References

1. Chen, Y.F., Wu, C.J.: Influence of website design on consumer emotion and purchase intention in travel websites. Int. J. Technol. Hum. Interact. (IJTHI) **12**(4), 15–29 (2016)
2. Sula, A., Spaho, E., Matsuo, K., et al.: A new system for supporting children with autism spectrum disorder based on IoT and P2P technology. Int. J. Space-Based Situated Comput. **4** (1), 55–64 (2014)
3. Di Stefano, A., Morana, G., Zito, D.: QoS-aware services composition in P2PGrid environments. Int. J. Grid Util. Comput. **2**(2), 139–147 (2011)
4. Sawamura, S., Barolli, A., Aikebaier, A., et al.: Design and evaluation of algorithms for obtaining objective trustworthiness on acquaintances in P2P overlay networks. Int. J. Grid Util. Comput. **2**(3), 196–203 (2011)
5. Takeda, A., Oide, T., Takahashi, A.: Simple dynamic load balancing mechanism for structured P2P network and its evaluation. Int. J. Grid Util. Comput. **3**(2–3), 126–135 (2012)
6. Eftychiou, A., Vrusias, B., Antonopoulos, N.: A dynamically semantic platform for efficient information retrieval in P2P networks. Int. J. Grid Util. Comput. **3**(4), 271–283 (2012)

7. Higashino, M., Hayakawa, T., Takahashi, K., et al.: Management of streaming multimedia content using mobile agent technology on pure P2P-based distributed e-learning system. Int. J. Grid Util. Comput. **5**(3), 198–204 (2014)
8. Holyoak, K.J.: Parallel distributed processing: explorations in the microstructure of cognition. Science **236**, 992–997 (1987)
9. Rochester, N., Holland, J., Haibt, L., et al.: Tests on a cell assembly theory of the action of the brain, using a large digital computer. IRE Trans. Inf. Theory **2**(3), 80–93 (1956)
10. Hoskins, J.C., Himmelblau, D.M.: Process control via artificial neural networks and reinforcement learning. Comput. Chem. Eng. **16**(4), 241–251 (1992)
11. Ciresan, D., Giusti, A., Gambardella, L.M., et al.: Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in Neural Information Processing Systems, pp. 2843–2851 (2012)

Implementation of a GA-based Simulation System for Placement of IoT Devices: Evaluation for a WSAN Scenario

Miralda Cuka[1], Kosuke Ozera[1], Ryoichiro Obukata[1], Donald Elmazi[1],
Tetsuya Oda[2](✉), and Leonard Barolli[2]

[1] Graduate School of Engineering, Fukuoka Institute of Technology (FIT),
3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811–0295, Japan
mcuka91@gmail.com, kosuke.o.fit@gmail.com, obukenkyuu@gmail.com,
donald.elmazi@gmail.com

[2] Department of Information and Communication Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811–0295, Japan
oda.tetsuya.fit@gmail.com, barolli@fit.ac.jp

Abstract. A Wireless Sensor and Actor Network (WSAN) is a group of wireless devices with the ability to sense physical events (sensors) or/and to perform relatively complicated actions (actors), based on the sensed data shared by sensors. In order to provide effective sensing and acting, a coordination mechanism is necessary among sensors and actors. This coordination can be distributed-local coordination among the actors or centralized coordination from a remote management unit, usually known as sink in Wireless Sensor Networks (WSNs). In this work, we propose a simulating system based on Rust for actor node placement problem in WSAN, while considering different aspects of WSANs including coordination, connectivity and coverage. We describe the implementation and show the interface of simulation system. We evaluated the performance of the proposed system by a simulation scenario considering WSANs. The simulation results show that the constructed WSAN could cover both events.

1 Introduction

Wireless Sensor Networks (WSNs) can be defined as a collection of wireless self-configuring programmable multihop tiny devices, which can bind to each other in an arbitrary manner, without the aid of any centralized administration, thereby dynamically sending the sensed data to the intended recipient about the monitored phenomenon [1].

Wireless Sensor and Actor Networks (WSANs), have emerged as a variation of WSNs. WSANs are capable of monitoring physical phenomenons, processing sensed data, making decisions based on the sensed data and completing

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_4

appropriate tasks when needed. WSAN devices deployed in the environment are sensors able to sense environmental data, actors able to react by affecting the environment or have both functions integrated [2]. For example, in the case of a fire, sensors relay the exact origin and intensity of the fire to actors so that they can extinguish it before spreading in the whole building or in a more complex scenario, to save people who may be trapped by fire.

Unlike WSNs, where the sensor nodes tend to communicate all the sensed data to the sink1 by sensor-sensor communication, in WSANs, two new communication types may take place. They are called sensor-actor and actor-actor communications. Sensed data is sent to the actors in the network through sensor-actor communication. After the actors analyse the data, they communicate with each other in order to assign and complete tasks. To provide effective operation of WSAN, is very important that sensors and actors coordinate in what are called sensor-actor and actor-actor coordination. Coordination is not only important during task conduction, but also during networks selfimprovement operations, i.e. connectivity restoration [3,4], reliable service [5], Quality of Service (QoS) [6,7] and so on.

Actor-Actor (AA) coordination helps actors to choose which actor will lead performing the task (actor selection), how many actors should perform and how they will perform. Actor selection is not a trivial task, because it needs to be solved in real time, considering different factors. It becomes more complicated when the actors are moving, due to dynamic topology of the network.

In this paper, we propose and implement a simulation system for Internet of Things (IoT) device placement. The system is based on Genetic Algorithm (GA). We describe the implementation of proposed system and show its interface. We evaluated the performance of the proposed system by a simulation scenario considering WSANs. As evaluation metrics, we considered Size of Giant Component (SGC) and Number of Covered Events (NCE).

The remainder of the paper is organized as follows. In Sect. 2, we describe the basics of IoT and WSANs including architecture and research challenges. In Sect. 3, we present the overview of GA. In Sect. 4, we show the description and design of the simulation system. Simulation results are shown in Sect. 5. Finally, conclusions and future work are given in Sect. 6.

2 IoT and WSAN

2.1 Internet of Things (IoT)

The term IoT has recently become popular to emphasize the vision of a global infrastructure of networked physical objects [8–10]. IoT is a new type of Internet application which enables the things/objects in our environment to be active participants with other members of the network, by sharing their information on a global scale using the same Internet Protocol (IP) that connects to the Internet. The descriptive models for Internet of Things are introduced based on two attributes ("being an Internet", "relating to thing's information") and four different features (only for thing's information, coded by UID or EPC, stored in RFID electronic tag, uploaded by non-contact reading with RFID reader).

The IoT creates human-machine or machine-to-machine communications. In this way the things/objects are capable of recognizing events and changes in their surroundings and are acting and reacting autonomously largely without human intervention in an appropriate way. The major objectives for IoT applications and services are the creation of smart environments/spaces and self-aware things for smart transport, products, cities, buildings, energy, health, social interaction and living applications (see Fig. 1).

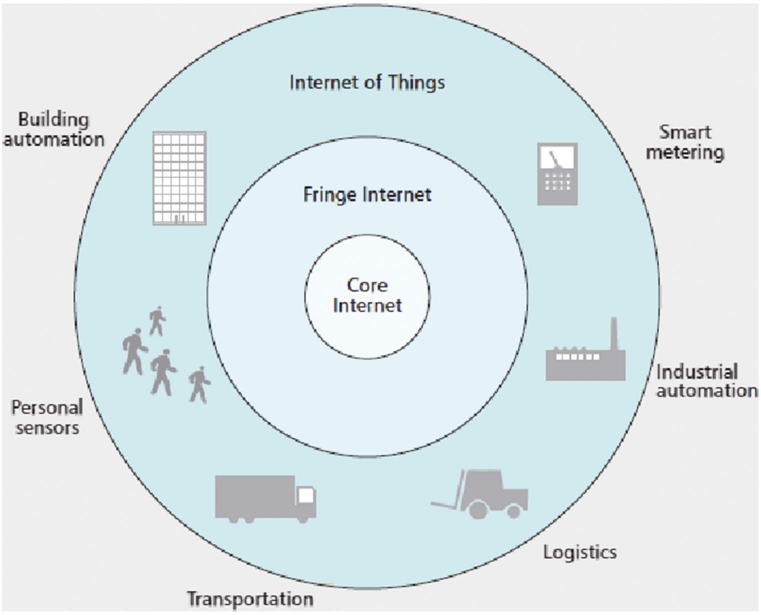


Fig. 1. Simulation system structure.

2.2 WSAN Architectures

The main functionality of WSANs is to make actors perform appropriate actions in the environment, based on the data sensed from sensors and actors. When important data has to be transmitted (an event occurred), sensors may transmit their data back to the sink, which will control the actors tasks from distance, or transmit their data to actors, which can perform actions independently from the sink node. Here, the former scheme is called Semi-Automated Architecture and the latter one Fully-Automated Architecture. Obviously, both architectures can be used in different applications. In the Fully-Automated Architecture are needed new sophisticated algorithms in order to provide appropriate coordination between nodes of WSAN. On the other hand, it has advantages, such as low latency, low energy consumption, long network lifetime [2], higher local position accuracy, higher reliability and so on.

2.3 Node Placement Problems and Their Applicability to WSANs

Node placement problems have been long investigated in the optimization field due to numerous applications in location science (facility location, logistics, services, etc.) and classification (clustering). In such problems, we are given a number of potential facilities to serve to costumers connected to facilities aiming to find locations such that the cost of serving to all customers is minimized [11]. In traditional versions of the problem, facilities could be hospitals, polling centers, fire stations serving to a number of clients and aiming to minimize some distance function in a metric space between clients and such facilities. One classical version of the problem is that of p-median problem, defined as follows.

Definition 1. Given a set $\mathscr{F}$ of m potential facilities, a set $\mathscr{U}$ of n users, a distance function $d : \mathscr{U} \rightarrow \mathscr{F}$, and a constant $p \leq m$, determine which p facilities to open so as to minimize the sum of the distances from each user to its closest open facility.

The problem, which is known for its intractability, has many application not only in location science but also in communication networks, where facilities could be servers, routers, etc., offering connectivity services to clients. In WSANs node provide network connectivity services to events. The good performance and operability of WSANs largely depends on placement of nodes in the geographical deployment area to achieve network connectivity, stability and user coverage. The objective is to find an optimal and robust topology of the nodes network to support connectivity services to events.

Facility location problems are thus showing their usefulness to communication networks, and more especially from WSANs field. In a general setting, location models in the literature have been defined as follows. We are given:

(a) a universe $\mathscr{U}$, from which a set $\mathscr{E}$ of event input positions is selected;
(b) an integer, $\mathscr{N} \geq 1$, denoting the number of facilities to be deployed;
(c) one or more metrics of the type $d : \mathscr{U} \times \mathscr{U} \rightarrow \mathscr{R}_+$, which measure the quality of the location; and,
(d) an optimization model.

The optimization model takes in input the universe where facilities are to be deployed, a set of client positions and returns a set of positions for facilities that optimize the considered metrics. It should be noted that different models can be established depending on whether the universe is considered: (a) continuous (universe is a region, where clients and facilities may be placed anywhere within the continuum leading to an uncountably infinite number of possible locations); (b) discrete (universe is a discrete set of predefined positions); and, (c) network (universe is given by an undirected weighted graph; in the graph, client positions are given by the vertices and facilities may be located anywhere on the graph).

For most formulations, node placement problems are shown to be computationally hard to solve to optimality and therefore heuristic and meta-heuristic approaches are useful approaches to solve the problem for practical purposes.

3 Overview of GA

As an approach to global optimization, GA have been found to be applicable to optimization problems that are intractable for exact solutions by conventional methods [12,13]. It is a set-based search algorithm, where at each iteration it simultaneously generates a number of solutions. In each iteration, a subset of the current set of solutions is selected based on their performance and these solutions are combined into new solutions. The operators used to create the new solutions are survival, where a solution is carried to the next iteration without change, crossover, where the properties of two solutions are combined into one, and mutation, where a solution is modified slightly. The same process is then repeated with the new set of solutions. The crossover and mutation operators depend on the representation of the solution, but not on the evaluation of its performance. They are thus the same even though the performance is estimated using simulation. The selection of solutions, however, does depend on the performance. The general principle is that high performing solutions (which in genetic algorithms are referred to as fit individuals) should have a better change of both surviving and being allowed to create new solutions through crossover. The simplest approach is to order the solutions $J(\theta_{[1]}) \leq J(\theta_{[2]}) \leq \ldots \leq J(\theta_{[n]})$, and only operate on the best solutions. If a strict selection of the top k solutions were required, this would complicate the issue significantly in the simulation optimization context, and considerable simulation effort would have to be spent to obtain an accurate ordering of the solutions.

4 Design and Implementation of IoT Device Placement Simulation System

In this section, we present design and implementation of a simulation system based on GA for IoT device placement in WSANs. The simulation system structure is shown in Fig. 2. The proposed simulating system is based on Rust [14,15]. Rust is a system programming language focused on three goals: safety, speed, and concurrency [16]. Rust supports a mixture of programming styles: imperative procedural, concurrent actor, object-oriented and functional.

Our system can generate instances of the problem using different distributions of events, sensor nodes and actor nodes. For the network configuration, we use: distribution of events, number of events, number of sensor nodes, number of actor nodes, area size, radius of communication range and radius of sensing range. For the GA parameter configuration, we use: number of independent runs, GA evolution steps, population size, crossover probability, mutation probability, initial placement methods, selection methods.

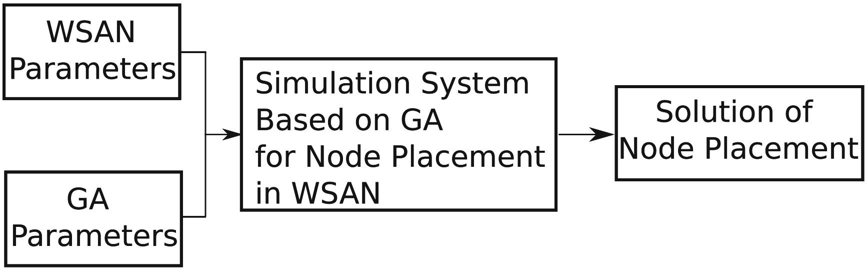

WSAN Parameters
GA Parameters
Simulation System Based on GA for Node Placement in WSAN
Solution of Node Placement

Fig. 2. Simulation system structure.

We explain in details the GA operations in following.

Selection Operator

As selection operator, we use roulette-wheel selection [12,13,17]. In roulette-wheel selection, each individual in the population is assigned a roulette wheel slot sized in proportion to its fitness. That is, in the biased roulette wheel, good solutions have a larger slot size than the less fit solutions. The roulette wheel can obtain a reproduction candidate.

Crossover Operator

The crossover operators are the most important ingredient of GAs. Indeed, by selecting individuals from the parental generation and interchanging their *genes*, new individuals (descendants) are obtained. The aim is to obtain descendants of better quality that will feed the next generation and enable the search to explore new regions of solution space not explored yet.

There exist many types of crossover operators explored in the evolutionary computing literature. It is very important to stress that crossover operators depend on the chromosome representation. This observation is especially important for the WSAN nodes problem, since in our case, instead of having strings we have a area of nodes located in a certain positions. The crossover operator should thus take into account the specifics of WSAN nodes encoding. We have considered the following crossover operator, called *intersection operators* (denoted `CrossRegion`, hereafter), which take in input two individuals and produce in output two new individuals.

Mutation Operator

Mutation operator is one of the GA ingredients. Unlike crossover operators, which achieve to transmit genetic information from parents to offsprings, mutation operators usually make some small local perturbation of the individuals, having thus less impact on newly generated individuals.

Crossover is "a must" operator in GA and is usually applied with high probability, while mutation operators when implemented are applied with small probability. The rationale is that a large mutation rate would make the GA search to resemble a random search. Due to this, mutation operator is usually considered as a secondary operator.

In the case of WSAN node placement, the matrix representation is chosen for the individuals of the population, in order to keep the information on WSAN nodes positions, events positions, links among nodes and links among nodes and events. The definition of the mutation operators is therefore specific to matrix-based encoding of the individuals of the population. We consider *SingleMutate* mutation operator which is a move-based operator. It selects a WSAN node in the problem area and moves it to another cell of the problem area.

5 Simulation Results

In Fig. 3 is shown visualization interface of implemented simulation system. We show a simulation scenario for WSANs where the number of actor nodes is 5, the number of sensor nodes is 15, and the number of events is 2. For simulation, we also consider the communication range of sensor and actor nodes, and sensing range of sensor and actor nodes. In Fig. 4 is shown simulation results of implemented simulation system. As evaluation metrics we use Size of Giant Component (SGC) and Number of Covered Events (NCE). The simulation results show that SGC is 16 and NCE is 2. So, the constructed WSAN could cover both events.

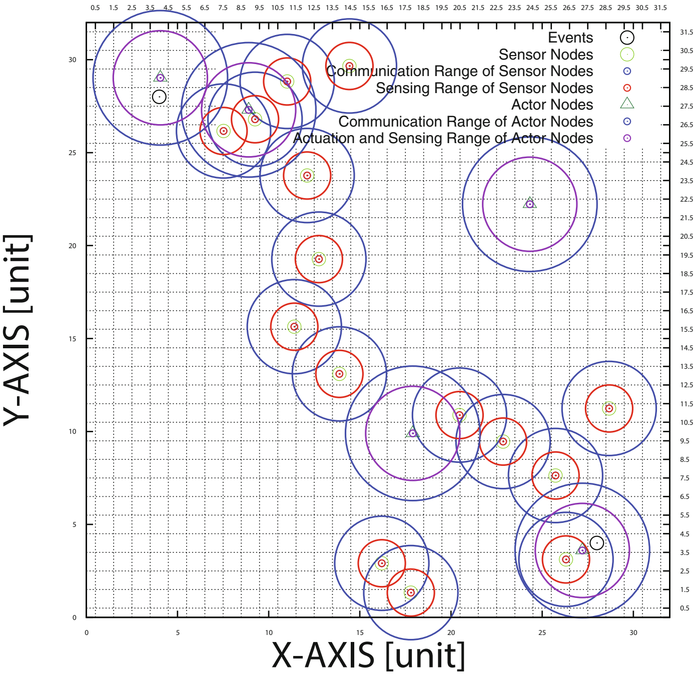
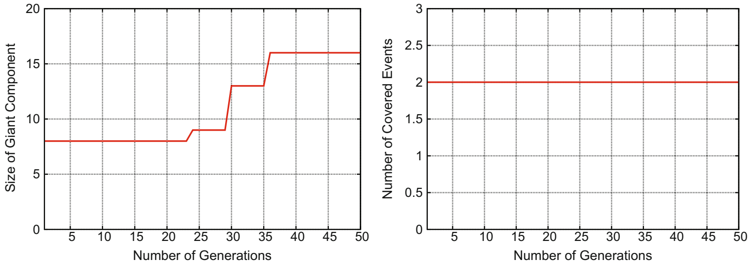


Fig. 3. Visualization interface.



(a) SGC (b) NCE

Fig. 4. Simulation results of sensor node optimization.

6 Conclusions

In this work, we designed and implemented a GA-based simulation system for IoT device placement in a WSAN scenario. We presented the implementation of the proposed simulation system and have shown also its interface and a simulation scenario. The simulation results show that 2 events were covered by the constructed WSAN. In the future, we would like to make extensive simulations for different simulation scenarios.

References

1. Akyildiz, I., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Comput. Netw. **38**(4), 393–422 (2002). Elsevier
2. Akyildiz, I.F., Kasimoglu, I.H.: Wireless sensor and actor networks: research challenges. Ad Hoc Netw. J. **2**(4), 351–367 (2004). Elsevier
3. Haider, N., Imran, M., Saad, N., Zakariya, M.: Performance analysis of reactive connectivity restoration algorithms for wireless sensor and actor networks. In: IEEE Malaysia International Conference on Communications (MICC-2013), pp. 490–495, November 2013
4. Abbasi, A., Younis, M., Akkaya, K.: Movement-assisted connectivity restoration in wireless sensor and actor networks. IEEE Trans. Parallel Distrib. Syst. **20**(9), 1366–1379 (2009)
5. Li, X., Liang, X., Lu, R., He, S., Chen, J., Shen, X.: Toward reliable actor services in wireless sensor and actor networks. In: IEEE 8th International Conference on Mobile Adhoc and Sensor Systems (MASS), pp. 351–360, October 2011
6. Akkaya, K., Younis, M.: Cola: a coverage and latency aware actor placement for wireless sensor and actor networks. In: IEEE 64th Conference on Vehicular Technology (VTC-2006), pp. 1–5, September 2006
7. Kakarla, J., Majhi, B.: A new optimal delay and energy efficient coordination algorithm for WSAN. In: IEEE International Conference on Advanced Networks and Telecommuncations Systems (ANTS), pp. 1–6, December 2013
8. Zanella, A., Bui, N., Castellani, A., Vangelista, L.: Internet of Things for smart cities. IEEE Internet Things J. **1**(1), 22–32 (2014)
9. Atzori, L., Iera, A., Morabito, G.: The Internet of Things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)

10. Bellavista, P., Cardone, G., Corradi, A., Foschini, L.: Convergence of MANET and WSN in IoT urban scenarios. IEEE Sens. J. **13**(10), 3558–3567 (2013)
11. Oda, T., Barolli, A., Xhafa, F., Barolli, L., Ikeda, M., Takizawa, M.: WMN-GA: a simulation system for WMNs and its evaluation considering selection operators. J. Ambient Intell. Humanized Comput. (JAIHC) **4**(3), 323–330 (2013). Springer
12. Holland, J.H.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor (1975)
13. Goldberg, D.E.: Genetic Algorithm in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
14. The Rust Programming Language. https://www.rust-lang.org/
15. GitHub - rust-lang/rust: a safe, concurrent, practical language. https://github.com/rust-lang/
16. ‘rust’ tag wiki - Stack Overflow. http://stackoverflow.com/tags/rust/info/
17. Sastry, K., Goldberg, D., Kendall, G.: Genetic algorithms. In: Burke, E.K., Kendall, G. (eds.) Search Methodologies - Introductory Tutorials in Optimization and Decision Support, Techniques, pp. 97–125. Springer, Heidelberg (2005)

A Cryptographically Secure Scheme for Preserving Privacy in Association Rule Mining

Hufsa Mohsin(✉)

Comsats Institute of Information Technology, Islamabad, Pakistan
hufsamohsin@gmail.com

Abstract. In this research, the primary focus is on privacy preservation in data mining. In particular, the problem of privacy preservation is addressed when the data is to be provided for applications or association rule mining is to be carried out on the datasets shared among two parties, i.e. the two party case. These scenarios are complex to address since privacy issues also lead to the non availability of correct data; also one must meet privacy requirements accompanied by valid data mining results. A system is proposed that is capable of hiding the sensitive information in the given set of data with the help of cryptographic algorithms. The encrypted data is then analyzed using Apriori algorithm for finding frequent itemsets that can lead to vital business decisions. Results reveal that our system provides strong privacy, guarantees accurate data mining while protecting sensitive information during association rule mining.

1 Introduction

Data mining, also known as Knowledge-Discovery in Databases (KDD), is quite useful for finding hidden patterns and to predict future behavior using tools such as classification, association rule mining, clustering etc. but with strong analysis tools and advancement in technology it also poses a threat to the business and legal privacy of individuals as well as organizations that are much concerned with the privacy of their data [19, 20]; credit card transactions, for instance. This leads to the problem of privacy preserving data mining (PPDM) [16], i.e. producing valid mining models and patterns without disclosing private information [2, 3, 12]. It also leads to the issue of veracity; (one of the 3 V dimensions in big data) a challenge of big data [17, 18].

PPDM comes into play when data is required in the development of software that processes the data, or in a situation where two companies carrying data mining operations on a joint set of data; a part being contributed by either company or when the data is outsourced to a data mining company for observing patterns. These situations introduce challenges of locating data as well as understanding and synchronizing data relationships with other databases and files. So, there is an ultimate urge to find some scheme capable of providing data privacy in testing and implementation environments [13–15, 21].

Key contribution of this paper is a privacy preserving data mining system that is capable of analyzing the given set of data for association patterns using Apriori algorithm and preserves privacy by means of encryption techniques. The privacy

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_5

question is addressed with the help of two encryption algorithms, DES and AES. The results produced by each are compared in terms of privacy and efficiency.

The aim of this paper is to identify the threats to privacy and the situations where privacy concerns arise. It will primarily be dealing with the case where test data is to be provided for applications and also discuss the two party case, that is the case where two organizations want to have a collective analysis of the data provided privacy of each of them is not compromised. This is done by first sensitizing the given dataset with DES or AES algorithm and then applying the apriori algorithm on this sensitized data to figure out the frequent itemsets so that the business decisions can be formulated.

The paper is organized as follows. Section 2 provides a review of related work. The overview of association rule mining and encryption techniques with the rationale behind their selection is discussed in Sect. 3. Subsequently, our proposed approach for privacy preservation of association rule mining by means of encryption is discussed in Sect. 4. The experimental results and discussion are presented in Sect. 5. Finally, Sect. 6 presents conclusion and a discussion of future work.

2 Related Work

One possible solution for two party case through encryption is discussed in [2], Z. Yang, S. Zhong, and Rebecca have proposed a privacy-preserving method of frequency mining and applied it to naive Bayes learning in a fully distributed setting. The proposed system guarantees the efficiency with no tradeoff between privacy and accuracy. Lindell and Pinkas in [8] also discussed the two party scenario for preserving privacy of privileged data and its use for research purposes. In [7], S.R. Oliveira focuses primarily on privacy issues for association rule mining and clustering when data are shared before mining.

In [4], B. Pinkas discussed the two party case with examples of secure computations. Another way in which the two party case can be dealt with is to hide, not the information but the knowledge patterns, this is discussed in [3], S.M. Oliveira and O.R. Zaiane have introduced new algorithms for balancing privacy and knowledge discovery in association rule mining.

To support efficient privacy preserving data mining, techniques and algorithms are presented in this section. Apriori Algorithm is used for association rule mining that provides high reliability and refers to a large dataset as compared to decision trees that tend to find few simple and small set of rules; most rules have somewhat low reliability [8, 9, 11].

3 Preliminaries

Before carrying out the association rule mining on the data sets, knowledge hiding is to be done by means of encryption algorithms (AES, DES). Knowledge hiding is usually fast, generally computationally inexpensive and memory efficient, and tend to lead to good overall solutions. An important aspect in knowledge hiding is that a solution always exists. This means that any itemset can be hidden before sharing of datasets while minimizing the impact of sanitization process on the insensitive knowledge [10].

Association rules represent a promising technique to find hidden patterns in a data set. These patterns may lead to some vital business decisions, e.g. these patterns can affect the policies or marketing campaigns by to a great deal [9]. Some transaction behaviors will lead to some association patterns, e.g. the buying behavior of customers, weather effects, website frequentation etc. Data mining searches for these association patterns in the database yielding association rules.

Advanced Encryption Standard (AES) capable of protecting information up to the top secret level and Data Encryption Standard (DES) algorithms are used to maintain data privacy. Also the results obtained by both algorithms are compared in terms of efficiency. When knowledge hiding is done by means of encryption [4], anyone with the key can have access to the real data and also complete privacy not affected by the data partitioning as in the case of randomization [1, 5, 9]. The approach in this research provides an option that any column or any field can be encrypted, depending upon the secrecy requirements keeping the relationships and the referential integrity intact. However, changing numeric to string data upon encryption is yet complex issue to deal with. Also, the key and the encryption algorithm will ensure production of original values there is no need of additional data base and relationship identification as in case of data perturbation [6].

4 Proposed Algorithm

The steps of the algorithm proposed for privacy preserving data mining in this paper is shown in Fig. 7. Below is the description of each step.

- *Select the dataset D_o on which the data mining algorithm is to be carried out.*
- *Select the sensitive attribute/column, s, which is to be hidden.*

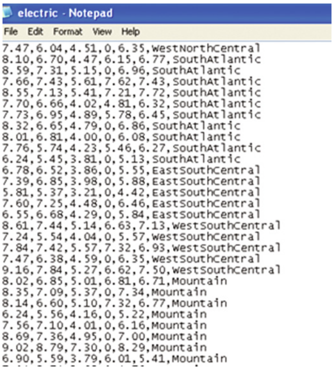
Figure 1 displays the sample dataset that is taken as the input by the algorithms. The tabular data has been converted into file format where each column is separated by ‘,’ the data represents the electricity consumption in different regions. The last column displays the regions which we want to encrypt.

- *Choose an algorithm for encryption i.e. DES or AES. Encrypt the selected string/column either through AES(D_s) or DES(D_s).*

DES and AES algorithms produce the encrypted output of the selected attribute ‘mountain’ as shown in the Figs. 2 and 3. The whole column can also be encrypted depending upon the user choice, whether to encrypt a column, selected attribute or a single record.

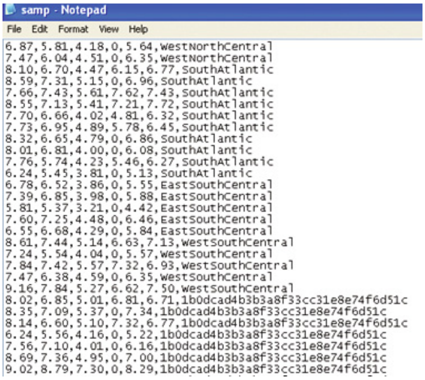
- *Apply necessary transformations τ on $D_{s^{-1},o}$ to convert it to format on which Apriori algorithm can be applied i.e. D_t. Store the original and the replaced values in a file/database.*

Hex → *Int*



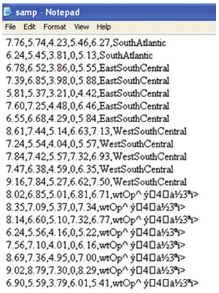
```
7.47,6.04,4.51,0,6.35,WestNorthCentral
8.10,6.70,4.47,6.15,6.77,SouthAtlantic
8.59,7.31,5.15,0,6.96,SouthAtlantic
7.66,7.43,5.61,7.62,7.43,SouthAtlantic
8.55,7.13,5.41,7.21,7.72,SouthAtlantic
7.70,6.66,4.02,4.81,6.32,SouthAtlantic
7.73,6.95,4.89,5.78,6.45,SouthAtlantic
8.32,6.65,4.79,0,6.86,SouthAtlantic
8.01,6.81,4.00,0,6.08,SouthAtlantic
7.76,5.74,4.23,5.46,6.27,SouthAtlantic
6.24,5.45,3.81,0,5.13,SouthAtlantic
6.78,6.52,3.86,0,5.55,EastSouthCentral
7.39,6.85,3.98,0,5.88,EastSouthCentral
5.81,5.37,3.21,0,4.42,EastSouthCentral
7.60,7.25,4.48,0,6.46,EastSouthCentral
6.55,6.68,4.29,0,5.84,EastSouthCentral
8.61,7.44,5.14,6.63,7.13,WestSouthCentral
7.24,5.54,4.04,0,5.57,WestSouthCentral
7.84,7.42,5.57,7.32,6.93,WestSouthCentral
7.47,6.38,4.59,0,6.35,WestSouthCentral
9.16,7.84,5.27,6.62,7.50,WestSouthCentral
8.02,6.85,5.01,6.81,6.71,Mountain
8.35,7.09,5.37,0,7.34,Mountain
8.14,6.60,5.10,7.32,6.77,Mountain
6.24,5.56,4.16,0,5.22,Mountain
7.56,7.10,4.01,0,6.16,Mountain
8.69,7.36,4.95,0,7.00,Mountain
9.02,8.79,7.30,0,8.29,Mountain
6.90,5.59,3.79,6.01,5.41,Mountain
```

Fig. 1. Original file. Containing sample data set where each column is separated by comma.



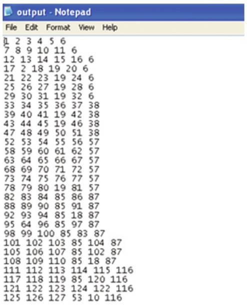
```
6.87,5.81,4.18,0,5.64,WestNorthCentral
7.47,6.04,4.51,0,6.35,WestNorthCentral
8.10,6.70,4.47,6.15,6.77,SouthAtlantic
8.59,7.31,5.15,0,6.96,SouthAtlantic
7.66,7.43,5.61,7.62,7.43,SouthAtlantic
8.55,7.13,5.41,7.21,7.72,SouthAtlantic
7.70,6.66,4.02,4.81,6.32,SouthAtlantic
7.73,6.95,4.89,5.78,6.45,SouthAtlantic
8.32,6.65,4.79,0,6.86,SouthAtlantic
8.01,6.81,4.00,0,6.08,SouthAtlantic
7.76,5.74,4.23,5.46,6.27,SouthAtlantic
6.24,5.45,3.81,0,5.13,SouthAtlantic
6.78,6.52,3.86,0,5.55,EastSouthCentral
7.39,6.85,3.98,0,5.88,EastSouthCentral
5.81,5.37,3.21,0,4.42,EastSouthCentral
7.60,7.25,4.48,0,6.46,EastSouthCentral
6.55,6.68,4.29,0,5.84,EastSouthCentral
8.61,7.44,5.14,6.63,7.13,WestSouthCentral
7.24,5.54,4.04,0,5.57,WestSouthCentral
7.84,7.42,5.57,7.32,6.93,WestSouthCentral
7.47,6.38,4.59,0,6.35,WestSouthCentral
9.16,7.84,5.27,6.62,7.50,WestSouthCentral
8.02,6.85,5.01,6.81,6.71,1b0dcad4b3b3a8f33cc31e8e74f6d51c
8.35,7.09,5.37,0,7.34,1b0dcad4b3b3a8f33cc31e8e74f6d51c
8.14,6.60,5.10,7.32,6.77,1b0dcad4b3b3a8f33cc31e8e74f6d51c
6.24,5.56,4.16,0,5.22,1b0dcad4b3b3a8f33cc31e8e74f6d51c
7.56,7.10,4.01,0,6.16,1b0dcad4b3b3a8f33cc31e8e74f6d51c
8.69,7.36,4.95,0,7.00,1b0dcad4b3b3a8f33cc31e8e74f6d51c
9.02,8.79,7.30,0,8.29,1b0dcad4b3b3a8f33cc31e8e74f6d51c
```

Fig. 2. File encrypted with DES. Showing encrypted attributes in the last column and encryption is done by means of DES



```
7.76,5.74,4.23,5.46,6.27,SouthAtlantic
6.24,5.45,3.81,0,5.13,SouthAtlantic
6.78,6.52,3.86,0,5.55,EastSouthCentral
7.39,6.85,3.98,0,5.88,EastSouthCentral
5.81,5.37,3.21,0,4.42,EastSouthCentral
7.60,7.25,4.48,0,6.46,EastSouthCentral
6.55,6.68,4.29,0,5.84,EastSouthCentral
8.61,7.44,5.14,6.63,7.13,WestSouthCentral
7.24,5.54,4.04,0,5.57,WestSouthCentral
7.84,7.42,5.57,7.32,6.93,WestSouthCentral
7.47,6.38,4.59,0,6.35,WestSouthCentral
9.16,7.84,5.27,6.62,7.50,WestSouthCentral
8.02,6.85,5.01,6.81,6.71,wtOp^ ý□4□ä½3ª¹>
8.35,7.09,5.37,0,7.34,wtOp^ ý□4□ä½3ª¹>
8.14,6.60,5.10,7.32,6.77,wtOp^ ý□4□ä½3ª¹>
6.24,5.56,4.16,0,5.22,wtOp^ ý□4□ä½3ª¹>
7.56,7.10,4.01,0,6.16,wtOp^ ý□4□ä½3ª¹>
8.69,7.36,4.95,0,7.00,wtOp^ ý□4□ä½3ª¹>
9.02,8.79,7.30,0,8.29,wtOp^ ý□4□ä½3ª¹>
6.90,5.59,3.79,6.01,5.41,wtOp^ ý□4□ä½3ª¹>
```

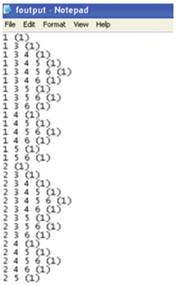
Fig. 3. File encrypted with AES. Showing the encrypted attributes in the last column and encryption is done by means of AES



```
1 2 3 4 5 6
7 8 9 10 11 6
12 13 14 15 16 6
17 2 18 19 20 6
21 22 23 19 24 6
25 26 27 19 28 6
29 30 31 19 32 6
33 34 35 36 37 38
39 40 41 19 42 38
43 44 45 19 46 38
47 48 49 50 51 38
52 53 54 55 56 57
58 59 60 61 62 57
63 64 65 66 67 57
68 69 70 71 72 57
73 74 75 76 77 57
78 79 80 19 81 57
82 83 84 85 86 87
88 89 90 85 91 87
92 93 94 85 18 87
95 64 96 85 97 87
98 99 100 85 83 87
101 102 103 85 104 87
105 106 107 85 102 87
108 109 110 85 18 87
111 112 113 114 115 116
117 118 119 85 120 116
121 122 123 124 122 116
125 126 127 53 10 116
```

Fig. 4. Encrypted file converted to integer format. Where each attribute is converted into integer value so that this file can be used as input to apriori algorithm.

Tokens of each attribute are made with ‘space’ as delimiter and then these tokens are replaced with the actual values as shown in Fig. 4. These output files are then converted to integer format as shown in Fig. 4. It can be noticed that single integer value is assigned to the same attributes e.g. *WestNorthCentral* has been assigned a value 6. This integer file is then given as the input to the Apriori algorithm to produce the frequent itemsets (Fig. 5).



```
1 (1)
1 3 (1)
1 3 4 (1)
1 3 4 5 (1)
1 3 4 5 6 (1)
1 3 4 6 (1)
1 3 5 (1)
1 3 5 6 (1)
1 3 6 (1)
1 4 (1)
1 4 5 (1)
1 4 5 6 (1)
1 4 6 (1)
1 5 (1)
1 5 6 (1)
2 (1)
2 3 (1)
2 3 4 (1)
2 3 4 5 (1)
2 3 4 5 6 (1)
2 3 4 6 (1)
2 3 5 (1)
2 3 5 6 (1)
2 3 6 (1)
2 4 (1)
2 4 5 (1)
2 4 5 6 (1)
2 4 6 (1)
2 5 (1)
```

Fig. 5. Frequent itemsets found by apriori. Showing the combinations of attributes found frequent by apriori algorithm.

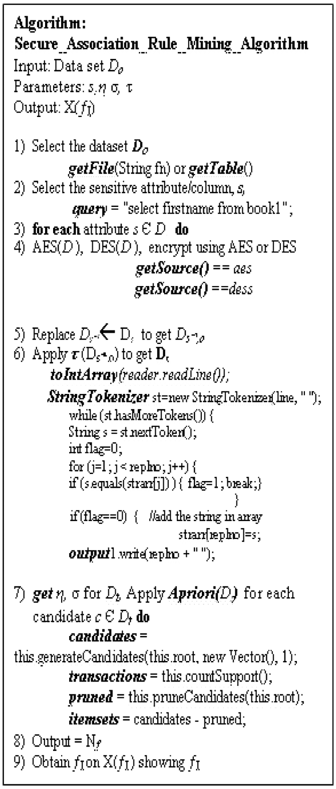
- *Get η, σ for D_t, Apply Apriori algorithm for association rule mining on this data* ***Apriori(D_t).***
- *The number of frequent itemsets* $\boldsymbol{N_f}$ *will be given by Apriori and* Obtain f_{I}, frequent itemsets, are stored in file X(f_{I}) showing f_{I} (Table 1).

Finding frequent itemsets is a stepwise procedure, candidate itemsets are generated as the first step then, the infrequent itemsets are pruned and finally the set of frequent itemsets are found as explained in Sect. 4 and pseudocode is shown in Fig. 6.

Figure 7 shows the actions performed by the user and the related output provided by the system. To start with user has to select the format for input and output i.e. Database format or File format. After that names of the algorithms (AES, DES) are displayed. Selecting a particular algorithm displays the user interaction window where

Table 1. Symbol table. Showing symbols used in proposed algorithm along with its description.

Symbol	Description
D_o	The original dataset
s	Sensitive Attribute/column to be encrypted
D_s	Dataset containing sensitive Attribute/Column
s^{-1}	Encrypted Attribute/Column
$D_{s^{-1}}$	Dataset containing encrypted Attribute/Column
$D_{s^{-1},o}$	Dataset containing original and encrypted Attribute/Column
D_t	Dataset containing the transformed data for applying apriori
F_e	Selection function of encryption algorithm



```
Algorithm:
Secure_Association_Rule_Mining_Algorithm
Input: Data set $D_o$
Parameters: $s, \eta\ \sigma,\ \tau$
Output: $X(f_I)$

1) Select the dataset $D_o$
        getFile(String fn) or getTable()
2) Select the sensitive attribute/column, $s$,
        query = "select firstname from book1";
3) for each attribute $s \in D$ do
4) AES($D$), DES($D$), encrypt using AES or DES
                getSource() == aes
                getSource() ==dess

5) Replace $D_{s^{-1}} \leftarrow D_s$ to get $D_{s^{-1},o}$
6) Apply $\tau(D_{s^{-1},o})$ to get $D_t$
      toIntArray(reader.readLine());
      StringTokenizer st=new StringTokenizer(line, " ");
        while (st.hasMoreTokens()) {
        String s = st.nextToken();
        int flag=0;
        for (j=1 ; j < repln0; j++) {
        if (s.equals(strarr[j]) ) { flag=1 ; break;}
                                          }
        if(flag==0) {   //add the string in array
                              strarr[repln0]=s;
        output1.write(repln0 + " ");

7) get $\eta$, $\sigma$ for $D_t$, Apply Apriori($D_t$) for each
   candidate $c \in D_t$ do
        candidates =
this.generateCandidates(this.root, new Vector(), 1);
        transactions = this.countSupport();
        pruned = this.pruneCandidates(this.root);
        itemsets = candidates - pruned;
8) Output = $N_f$
9) Obtain $f_I$ on $X(f_I)$ showing $f_I$
```

Fig. 6. Proposed algorithm. Showing the steps of the proposed secure association rule mining algorithm

certain parameters are required as input file, string/column to be encrypted. Related output is displayed and stored in a file which is then taken as the input by Apriori after converting into the acceptable format by *covttoInt.java* algorithm to be executed

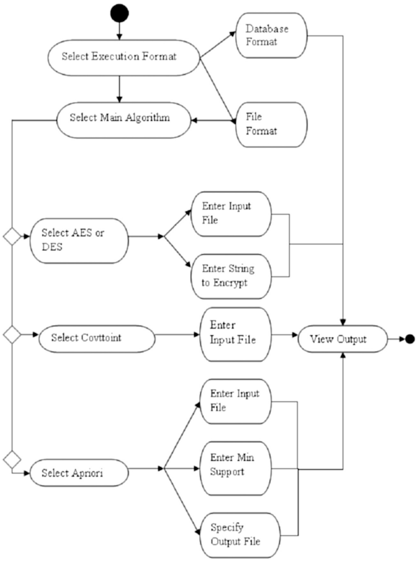


Fig. 7. Activity diagram. Showing actions performed by user and output provided by the proposed system

successfully, user have to enter the input three parameters; input file name, $X2(s^{-1})$, Minimum support value based on a support, σ set by the user, frequent itemsets are determined through consecutive scans of the database. Last Parameter is the output file name *X3* to store and view the results.

5 Results

The proposed system is tested by giving datasets as the input to the system. Results with different parameters are shown with the help of graphs that show the relation between time and minimum support parameter. It shows that when the support is high the time required to produce the output is also high and vice versa. Also the number of frequent items mined is large when support is minimum.

Figure 8 shows that the time taken for finding frequent itemsets using apriori is greater when the minimum support is given the value 1 and is less for the value 2. This shows that by increasing the support parameter the time can be reduced. The number of itemsets mined is greater when the support has the minimum value. As the value of minimum support increases the number of frequent itemsets decreases and reaches to zero at a certain value of support. So this value needs to be adjusted carefully and it also depends on the dataset used.

Table 2 shows the output of Apriori algorithm. The first pass simply scans the database to find the large-1 itemsets. Second pass generates the candidate itemsets that can lead to frequent itemsets from the large itemsets, it also depends on the value of support. The pruned candidate column shows the number of infrequent or unfruitful candidates.

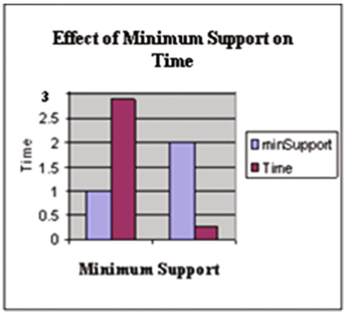


Fig. 8. Effect of minimum support on time. Shows effect of minimum support parameter on time i.e. Time also increases when support value increases.

Table 2. Output of apriori. Showing the total no. of frequent itemsets found for the given no. of pass

No of pass	Candidates	Pruned candidates	Support value	No of frequent itemsets found	Time (Sec)
1	378	0	1	615	2.8
2	96691	96349			
3	139	244			
1	378	371	2	15	0.2
2	21	13			
3	0	0			

615 frequent itemsets were found when the minimum support was 1. increasing this parameter will result in more strict finding of itemsets. In this case the third pass has generated no candidates and the total no. of frequent itemsets is only 15 with support value 2.

Figure 9 depicts the comparison between the time consumed for encryption of data by the two algorithms i.e. DES and AES. It is clear that DES is faster than AES.

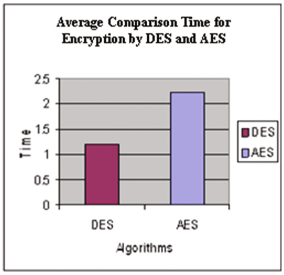


Fig. 9. Average comparison time for encryption by DES and AES. Shows that time for encryption by AES is greater as compared to time for encryption by DES

Figures 10 and 11 show the total time taken for finding frequent itemsets from original dataset is less Where as time taken to encrypt data and then finding the frequent itemsets is higher in view of the fact that privacy is achieved at the cost of encryption by AES and DES. This encryption cost has a considerable effect on the time for lower values of minimum support value as shown in Figs. 10 and 12.

The time for the encryption of data adds to the cost of the system that has to be paid for preserving privacy but where security of data is concerned this can be negligible.

AES has less resource consumption as compared to DES and TDES. While the Figures discussed above show that DES is faster as compared to AES. Therefore, when speed is the major concern during PPDM the encryption has to be done using DES while AES will be the choice when the security of data is not to be compromised at any cost.

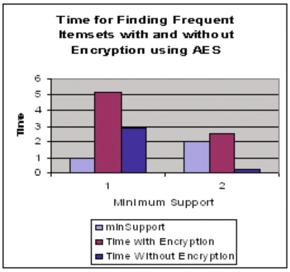


Fig. 10. Comparison of time for finding frequent itemsets from encrypted and non encrypted data using AES. Represents the comparison of time required to find frequent itemsets while performing encryption with the time it takes to mine frequent itemsets without performing encryption, using AES algorithm

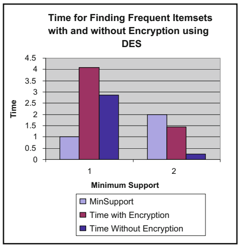


Fig. 11. Comparison of time for finding frequent itemsets from encrypted and non encrypted data using DES. Demonstrating the comparison of time required to find frequent itemsets while performing encryption with the time it takes to mine frequent itemsets without performing encryption, using DES algorithm

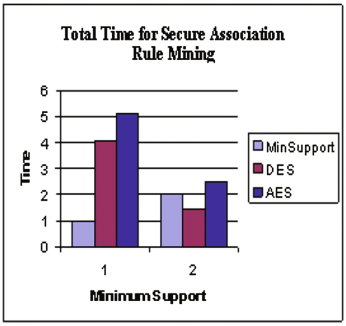


Fig. 12. Comparison of total time for secure association rule mining using AES and DES encryption algorithms. Presents comparison of total time for association rule mining taken by AES and DES Algorithms and the effect of minimum support.

6 Conclusions and Future Work

In this research, a privacy-preserving system for association rule mining is proposed that provides strong privacy and accurate results by means of cryptography. The key contribution is a privacy preserving method that allows computing frequent itemsets in a transaction database while hiding the sensitive part of the data Using DES and AES. The results show that the Apriori Algorithm is more efficient when the support parameter is increased provided it can be able to mine frequent itemsets properly. A large number of frequent itemsets is given when the support is minimum. This number is reduced considerably when the support parameter is increased even by one. For the above values it may not even find the frequent itemsets but this depends on the nature and attributes of data. The efficiency of the algorithm can be improved if this parameter is tuned properly. Results also reveal that the DES algorithm can be used for faster computation, whereas, AES algorithm is to be used where highly sensitive data hiding is the major issue.

Association rule mining and apriori algorithm has been selected for this research; other algorithms such as decision trees and Bayesian networks for classification, clustering can be used. Similarly the effects of public key cryptosystems on PPDM can also be studied. Another work suggested for future research is to combine the cryptographic and randomization techniques so as to improve efficiency without the loss of accuracy. The impacts of this combined technique will be a good research area for future work.

References

1. Evfimievski, A., Gehrke, J., Srikant, R.: Limiting privacy breaches in privacy preserving data mining. IBM Almaden Research Center (2004)
2. Yang, Z., Zhong, S., Wright, R.N.: Privacy-preserving classification of customer data without loss of accuracy. Computer Science Department, Stevens Institute of Technology, DIMACS Center, Rutgers University, Piscataway (2004)

3. Oliveira, S.M., Zaiane, O.R.: Algorithms for balancing privacy and knowledge discovery in association rule mining. Embrapa Information Technology Department of Computing Science (2003)
4. Pinkas, B.: Cryptographic techniques for privacy preserving data mining. HP Labs (2003)
5. Evfimievski, A.: Randomization in privacy preserving data mining. Cornell University Ithaca, NY 14853, USA (2002)
6. Agrawal, D., Agarwal, C.C.: On the design and quantification of privacy preserving data mining algorithms. In: Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 2001
7. Oliveira, S.R.: Data transformation for privacy-preserving data mining. University of Alberta (2005)
8. Lindell, Y., Pinkas, B.: Privacy preserving data mining. Department of Computer Science Weizmann Institute of Science Rehovot, Israel (2002)
9. Evfimievski, A., Srikant, R., Agrawal, R.: Privacy preserving mining of association rules. IBM Almaden Research Center, USA (2002)
10. Oliveira, S.R.M., Zaıane, O.R.: Protecting sensitive knowledge by data sanitization. University of Alberta, Edmonton, Canada (2003)
11. Oliveira, S.R.M., Zaıane, O.R.: Secure association rule sharing. University of Alberta, Edmonton, Canada (2003)
12. Oliveira, S.R.M.: Privacy preserving frequent itemset mining. University of Alberta (2002)
13. Natwichai, J., Maria, X., Orlowska, E.: A reconstruction-based algorithm for classification rules hiding. School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia (2006)
14. Ozel, S.A., Güvenir, H.A.: An algorithm for mining association rules using perfect hashing and database pruning. Bilkent University, Department of Computer Engineering, Ankara, Turkey (2000)
15. Aggarwal, C.C., Yu, P.S.: A general survey of privacy-preserving data mining models and algorithms, vol. 34, pp. 11–52 (2008)
16. Evfimievski, A., Grandison, T.: Privacy Preserving Data Mining. IBM Almaden Research Center, USA (2009)
17. Crawford, K., Schultz, J.: Big data and due process: towards a framework to redress predictive privacy harms (2014)
18. Lu, R., Zhu, H., Liu, X., Lio, J.K., Shao, J.: Towards efficient and privacy-preserving computing in big data. IEEE Network **28**(4), 46–50 (2014)
19. Richards, N.M., King, J.: Big data ethics. Wake Forest (2014)
20. Muhammed, N., Chen, R., Fung, B., Yu, P.S.: Differentially private data release for data mining (2011)
21. Dua, S., Du, X.: Data Mining and Machine Learning in Cyber Security. Taylor and Francis Group, Boca Raton (2016)

A BGN Type Outsourcing the Decryption of CP-ABE Ciphertexts

Li Zhenlin[1,2(✉)], Zhang Wei[2], Ding Yitao[2], and Bai Ping[2]

[1] Department of Electronic Technique, Engineering University of PAP,
Xi'an, Shaanxi, China
lizhenlin19921109@163.com

[2] Key Laboratory of Information Security, Engineering College of PAP,
Xi'an, Shaanxi, China
zhaangweei@yeah.net, 15803396982@163.com,
15502960211@163.com

Abstract. Cloud computing security is the key bottleneck that restricts its development, and access control on the result of cloud computing is a hot spot of current research. Based on the somewhat homomorphic encryption BGN and combined with Green's scheme that proposed outsourcing the decryption of CP-ABE (Ciphertext-Policy Attribute-Based Encryption) ciphertexts, we constructed a BGN type outsourcing the decryption of CP-ABE ciphertexts. In our construction, partial decryption of ciphertexts is outsourced to the cloud, and only users whose attribute meets the access policy will get the correct decryption. And the scheme supports arbitrary homomorphic additions and one homomorphic multiplication on ciphertexts. Finally, we prove its semantic security under the subgroup decision assumption and compare it with other schemes.

1 Introduction

With the emerge of cloud computing [1], the development of information industry is moving in the fast lane. Cloud computing provides users with massive storage services and powerful computing services, which remarkably makes a contribution to economy [2–5]. However, security issues associated with cloud computing have become increasingly prominent [6]. Kaufman [7] pointed out that the security issue of cloud services was not only one of the biggest challenge of difficulties it faced, but also the problem that should be solved as soon as possible.

If the users save their sensitive data to the cloud server in plaintext, then because the cloud may copy even distort the information, but users do not know such unauthorized behavior of the cloud, which may cause immeasurable loss, the cloud will not be unconditional trusted. In order to prevent malicious leakage and illegal access to sensitive data, users can outsource their data in the encrypted form.

The traditional encryption and decryption model of cloud computing cannot achieve fine-grained access control on the results of cloud computing. In reality, we do not need everyone to gain the final results. In 1984 Shamir [8] proposed Identity-Based Encryption (IBE), in which a user's public key was generated by a unique identifier

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_6

that was related to his/her identity, and the servers did not need query the user's public key certificate any more. Attribute-Based Encryption (ABE), proposed by Sahai and Waters [9], is seen as a promotion of IBE. In ABE system, the user's key and the ciphertexts are associated with attribute, and only when attribute meets the access policy, the user will get the correct decryption, which succeeds in fine-grained access control on the ciphertexts. Due to such good characteristics, ABE scheme has attracted great attention of cryptographers. A large number of relevant research on ABE have emerged in recent years [10–13], and it also has been widely applied to cloud computing security algorithm [14–16], which becomes an important tool for data protection in cloud computing.

In this paper, based on the classic somewhat homomorphic encryption scheme BGN [17], adopting the method of [13] in which we called it outsourcing the decryption of CP-ABE ciphertexts, we propose a BGN type outsourcing the decryption of CP-ABE ciphertexts. In our scheme, partial decryption of ciphertexts is outsourced to the cloud, which greatly reduces the computing overhead of users. The user's private key is associated with his/her attributes, and access control policy is embedded into the ciphertexts, and only the users whose attributes satisfy the access policy can decrypt the ciphertexts. Meanwhile, our scheme can operate on ciphertexts for arbitrary additions and one multiplication.

In Sect. 2, we give the preliminary knowledge of this paper. We present our construction of outsourcing and analyze the homomorphic properties of the scheme in Sect. 3. In Sects. 4 and 5, its security and performance analysis is described respectively. In the next chapter, we make a conclusion.

2 Preliminares

2.1 Bilinear Map

Let $\mathbb{G}$ and $\mathbb{G}_T$ be two multiplicative cyclic groups of prime order p. Let g be a generator of $\mathbb{G}$ and $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ be a bilinear map with the following properties:

1. Bilinearity: for all $u, k \in \mathbb{G}$ and $a, b \in Z_p$, then $e\left(u^a, k^b\right) = e(u, k)^{ab}$.
2. Non-degeneracy: $e(g, g) \neq 1$.
3. Computable: the bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ can be computed in polynomial time.

2.2 Access Structures

Definition 1 (Access Structure [18]). Let $\{P_1, P_2, \cdots, P_n\}$ be a set of participants and let $P = 2^{\{P_1, P_2, \cdots, P_n\}}$. And access structure Γ is a non-empty subset of $\{P_1, P_2, \cdots, P_n\}$. We define its monotone property as follows: If $A \in \Gamma$ and $A \subseteq B$, then $B \in \Gamma$. We call the sets in Γ the authorized sets, otherwise the unauthorized sets.

2.3 Linear Secret Sharing Schemes

Definition 2 (Linear Secret-Sharing Schemes (LSSS) [18]**).** A secret-sharing scheme Π over a set of participants P is called linear (over Z_p) if

1. The shares of the participants form a vector over Z_p.
2. There exists a $l \times n$ matrix $\boldsymbol{M}$ that is called the share-generating matrix for Π. We define a function ρ that maps every row of the share-generating matrix to a related participant, i.e., for $i = 1, 2, \cdots l$, the value $\rho(i)$ is the participant which is associated with row i. And we build a column vector $\boldsymbol{v} = (s, y_2, \cdots, y_n)$, in which $y_2, \cdots, y_n \in Z_p$ are chosen randomly, and $s \in Z_p$ is just the secret to be shared, then $\boldsymbol{Mv}$ is the vector of l shares of the secret s according to Π. The share $(\boldsymbol{Mv})_i$ belongs to participant $\rho(i)$.

Definition 3 (Linear Reconstruction [18]**).** Each linear secret sharing-scheme has the linear reconstruction property: Suppose that Π is an LSSS for the access structure Γ. Let $S \in \Gamma$ be an authorized set, and let $I \subseteq \{1, 2, \ldots, l\}$ and $I = \{i : \rho(i) \in S\}$. Then, if $\{\lambda_i\}$ are valid shares of any secret s according to Π, there must exist constants $\{w_i \in Z_p\}_{i \in I}$ such that $\sum_{i \in I} w_i \boldsymbol{M}_i \boldsymbol{v} = s$.

2.4 BGN Scheme

The BGN [17] is a classic somewhat homomorphic encryption that is proposed by Boneh, Goh and Nissim, and BGN scheme supports arbitrary homomorphic additions and one homomorphic multiplication. As all know, BGN is the first somewhat homomorphic encryption after the concept of homomorphic encryption was proposed in [19], and in 2010 Gentry [20] implemented BGN on lattice. The scheme is described as follows:

KeyGen(τ): Given a security parameter $\tau \in Z^+$, run $\mathcal{G}(\tau)$ to obtain a tuple $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$. Let $n = q_1 q_2$. Pick two generators $k, u \overset{R}{\longleftarrow} \mathbb{G}$ randomly and set $h = u^{q_2}$. Then h is a random generator of the subgroup of $\mathbb{G}$ of order q_1. The public key is $PK = (n, \mathbb{G}, \mathbb{G}_1, e, k, h)$ and the private key is $SK = q_1$.

Encrypt(PK, M): The message space is described as $m \in \{0, 1, \cdots, T\}$ with $T < q_2$. We use public key PK to encrypt a message m, pick a random $r \overset{R}{\longleftarrow} \{0, 1, \cdots n-1\}$ and compute $C = k^m h^r \in$ G. Output C as the ciphertext.

Decrypt(SK, C): We use the private key $SK = q_1$ to decrypt a ciphertext C, observe that $C^{q_1} = (k^m h^r)^{q_1} = (k^{q_1})^m$. To obtain m, we compute the discrete log of C^{q_1} base k^{q_1}. Since $0 \le m \le T$ this takes expected time $O(\sqrt{T})$ using Pollard's lambda [21] method.

Homomorphic properties: The BGN scheme is clearly additively homomorphic:

$$C = C_1 C_2 h^r = k^{m_1} h^{r_1} \cdot k^{m_2} h^{r_2} \cdot h^r = k^{m_1 + m_2} h^{r_1 + r_2 + r} \in \mathbb{G}$$

Multiplicatively homomorphic: Let $k_1 = e(k, k)$ and $h_1 = e(k, h)$, then k_1 is of order n and h_1 is of order q_1. There is some (unknown) $\beta \in Z$ such that $h = k^{\beta q_2}$. We have:

$$\begin{aligned} C &= e(C_1, C_2)h_1^r \\ &= e(k^{m_1}h^{r_1}, k^{m_2}h^{r_2})h_1^r \\ &= k_1^{m_1m_2}h_1^{m_1r_2+m_2r_1+\beta q_2r_1r_2+r} \\ &= k_1^{m_1m_2}h_1^{\tilde{r}} \in \mathbb{G}_1 \end{aligned}$$

Where $\tilde{r} = m_1r_2 + m_2r_1 + \beta q_2r_1r_2 + r$ is distributed uniformly in z_n. The new ciphertext $C \in \mathbb{G}_1$, because there is no efficient algorithm to make $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}$, the scheme can operate on ciphertexts for only one multiplication.

2.5 Outsourcing the Decryption of ABE Ciphertexts Model

Outsourcing the decryption of attribute-based encryption ciphertexts is proposed by Green, Hohenberger and Waters [13]. The difference from traditional attribute-based encryption is that a transformation algorithm is added in the new scheme, in which partial decryption is outsourced to the cloud, and clients only compute on a little data feedback by the cloud. This method that is called the outsourcing makes full use of the powerful computing ability of the cloud, which greatly improves the decryption efficiency of clients. The traditional attribute-based encryption model and outsourcing the decryption of attribute-based encryption ciphertexts model are shown in Figs. 1 and 2 respectively:

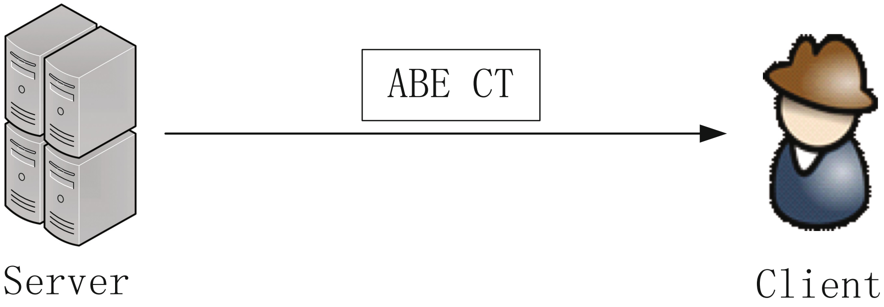


Fig. 1. Traditional ABE model

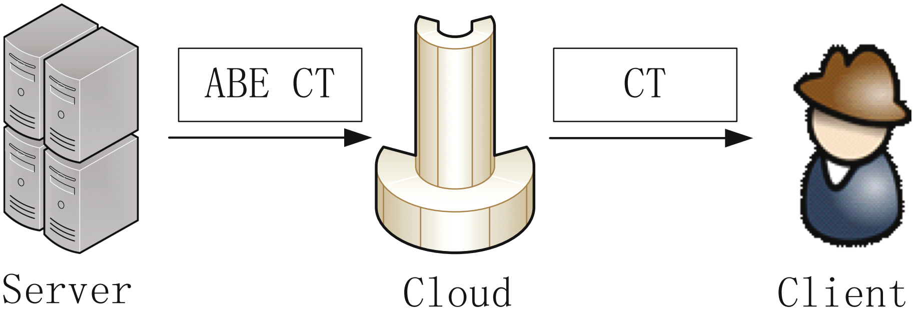


Fig. 2. Outsourcing the decryption of ABE ciphertexts model

In the traditional attribute-based encryption model, which is shown in Fig. 1, the clients must download all ABE ciphertexts to decrypt. Obviously the overhead of storage and computation is too much expensive. In order to solve such shortcomings, the outsourcing model shown in Fig. 2 is designed. The decryption of ABE ciphertexts will be outsourced to cloud which sends partial ciphertexts back, and the clients only need download a small amount of data and compute some simple operations, the storage and computation overhead of the procedure has remarkable reduction.

3 Our Construction

In this part, we construct a BGN type outsourcing the decryption of CP-ABE ciphertexts. Combining the BGN scheme with the idea of outsourcing decryption of attribute-based ciphertexts, we present our construction that can realize access control on the results of cloud outsourcing. Our scheme consists of the following five algorithms:

Setup(λ, U): The setup algorithm takes as input a security parameter λ and a universe description $U = \{0,1\}^*$. It runs $\mathcal{G}(\lambda)$ to obtain a tuple $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$, $\mathbb{G}, \mathbb{G}_1$ are two groups of order $n = q_1 q_2$ and $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_1$ be a bilinear map. It picks two generators $k, u \xleftarrow{R} \mathbb{G}$ randomly and set $h = u^{q_2}$, then h is a random generator of the subgroup of $\mathbb{G}$ of order q_1. It then chooses two group $\mathbb{G}', \mathbb{G}'_{\mathrm{T}}$ of prime order p and a hash function F that maps $\{0,1\}^*$ to $\mathbb{G}'$ and hash function H that maps $\mathbb{G}'_{\mathrm{T}}$ to $(0,1)$. Let g be a generator of $\mathbb{G}'$ and $e' : \mathbb{G}' \times \mathbb{G}' \to \mathbb{G}'_{\mathrm{T}}$ be a bilinear map. What's more, it chooses exponents $\alpha, a \in Z_p$ randomly. The algorithm sets MSK $= (g^{\alpha}, PK)$ as the master secret key. And the public parameters is PK $= (n, g, k, h, e, e'(g,g)^{\alpha}, g^a, F, H, \mathbb{G}, \mathbb{G}_1)$.

Encrypt(PK, m, ($\boldsymbol{M}$, ρ)): The encryption algorithm takes as input the public parameters PK and a message m to encrypt. In addition, it takes as input an LSSS access structure ($\boldsymbol{M}$, ρ). The function ρ associates rows of $\boldsymbol{M}$ to attributes. Let $\boldsymbol{M}$ be an $l \times n$ matrix. The algorithm first chooses a random vector $\boldsymbol{v} = (s, y_2, \cdots, y_n) \in Z_p^n$, and s is the secret to be shared. For $i = 1, 2, \cdots, l$, it computes $\lambda_i = \boldsymbol{M}_i \cdot \mathbf{v}$, in which $\boldsymbol{M}_i$ is the vector corresponding to the i th row of $\boldsymbol{M}$. In addition, the algorithm chooses random R, $r_1, \cdots, r_l \in Z_p$. Output the ciphertext $CT =$

$$c = k^{mH(e'(g,g)^{\alpha s})} h^R, C' = g^s$$

$$\left(C_1 = g^{a\lambda_1} \cdot F(\rho(1))^{-r_1}, D_1 = g^{r_1}\right)$$

$$\cdots\cdots$$

$$\left(C_l = g^{a\lambda_l} \cdot F(\rho(l))^{-r_l}, D_l = g^{r_l}\right)$$

KeyGen(MSK, S): The keygen algorithm chooses $t' \in Z_p$ randomly, then it takes as input MSK and an attribute set S to obtain $SK'\left(PK, K' = g^{\alpha} g^{at'}, L' = g^{t'}, \{K'_x = F(x)^{t'}\}_{x \in S}\right)$. It chooses a random value $z \in Z_p$. Let $t = t'/z$, it then published the transformation key TK as:

$$PK, K = K'^{1/z} = g^{\alpha/z} g^{at}, L = L'^{1/z} = g^{t}, \{K_x\}_{x \in S} = \left\{K_x'^{1/z}\right\}_{x \in S}$$

and the private key is $SK = (q_1, z, \mathrm{TK})$.

Transform$(\mathrm{TK}, \mathrm{CT})$: The transformation algorithm takes as input a transformation key $\mathrm{TK} = \left(\mathrm{PK}, \mathrm{K}, \mathrm{L}, \{K_x\}_{x \in S}\right)$ for a set S and a ciphertext $\mathrm{CT} = (c, C', C_1, \cdots, C_l)$ for access structure $(\boldsymbol{M}, \rho)$. If S does not satisfy the access structure, it outputs $\perp$. Suppose that S satisfies the access structure and let $I \subset \{1, 2, \cdots, l\}$ be defined as $I = \{i : \rho(i) \in S\}$. Then, let $\left\{\omega_i \in Z_p\right\}_{i \in I}$ be a set of constants such that if $\{\lambda_i\}$ are valid shares of any secret s according to $\boldsymbol{M}$, then $\sum_{i \in I} \omega_i \lambda_i = s$. The transformation algorithm calculates:

$$\begin{aligned}
Q &= e'(C', K) / \left(e'\left(\prod_{i \in I} C_i^{w_i}, L \right) \cdot \prod_{i \in I} e'\left(D_i^{w_i}, K_{\rho(i)}\right) \right) \\
&= e'(g, g)^{s\alpha/z} e'(g, g)^{sat} / \left(\left(\prod_{i \in I} e'(g, g)^{ta\lambda_i w_i} \right) \right) \\
&= e'(g, g)^{s\alpha/z}
\end{aligned}$$

It outputs the partially decrypted ciphertext $\mathrm{CT}' = (c, Q)$.

Decrypt(SK, CT): The decryption algorithm takes as input a private key $SK = (q_1, z, \mathrm{TK})$ and a ciphertext CT. If the ciphertext is not partially decrypted, then the algorithm first executes transformation algorithm. If the output is $\perp$, then this algorithm outputs $\perp$ as well. Otherwise, it uses (z, Q) to obtain $e'(g, g)^{s\alpha} = Q^z$, then decrypts c using the partial private key q_1, observe that $c^{q_1} = \left(k^{mH(e'(g,g)^{\alpha s})} h^R\right)^{q_1} = \left(k^{H(e'(g,g)^{\alpha s}) q_1}\right)^m$, and using Pollard's lambda method, we compute the discrete log of C_{q_1} base $k^{H(e'(g,g)^{\alpha s}) q_1}$ to cover m.

Our outsourcing construction is based on the BGN scheme, so it satisfies the properties of arbitrary additions and one multiplication.

1. Additively Homomorphic: For two ciphertexts $c_1 = k^{m_1 H(e'(g,g)^{\alpha s})} h^{R_1} \in \mathbb{G}$ and $c_2 = k^{m_2 H(e'(g,g)^{\alpha s})} h^{R_2} \in \mathbb{G}$, we have:

$$\begin{aligned}
c' &= c_1 c_2 h^R \\
&= \left(k^{m_1 H(e'(g,g)^{\alpha s})} h^{R_1}\right) \cdot \left(k^{m_2 H(e'(g,g)^{\alpha s})} h^{R_2}\right) h^R \\
&= k^{(m_1 + m_2) H(e'(g,g)^{\alpha s})} h^{R_1 + R_2 + R} \in \mathbb{G}
\end{aligned}$$

The legal decryptor whose attribute meets the access policy can gain the value of $e'(g, g)^{s\alpha}$, then he will decrypt the ciphertexts through decryption algorithm.

2. Multiplicatively Homomorphic: Let $k_1 = e(k,k)$ and $h_1 = e(k,h)$, then k_1 is of order n and h_1 is of order q_1. There is some (unknown) $\beta \in Z$ such that $h = k^{\beta q_2}$. We have:

$$\begin{aligned} c' &= e(c_1, c_2) h_1^R \\ &= e\left(k^{m_1 H(e'(g,g)^{\alpha s})} h^{R_1}, k^{m_2 H(e'(g,g)^{\alpha s})} h^{R_2}\right) h_1^R \\ &= k_1^{m_1 m_2 H(e'(g,g)^{\alpha s})^2} h_1^{R + (R_1 m_2 + R_2 m_1) H(e'(g,g)^{\alpha s}) + \beta q_2 R_1 R_2} \in \mathbb{G}_1 \end{aligned}$$

In the same way, the legal users can work out $m_1 m_2$. Since there is no efficient algorithm to make $e : \mathbb{G}_1 \times \mathbb{G}_1 \to \mathbb{G}$, so the scheme can operate on ciphertexts for only one multiplication.

4 Security

4.1 The Subgroup Decision Problem

We define an algorithm $\mathcal{G}$ such that given a parameter $\tau \in Z^+$, it outputs a tuple $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$ in which $\mathbb{G}, \mathbb{G}_1$ are groups of order $n = q_1 q_2$ and $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_1$ is a bilinear map. On input τ, the algorithm $\mathcal{G}$ will work as follows:

1. Generate randomly two τ-bit primes q_1, q_2 and set $n = q_1 q_2 \in Z$.
2. Generate a bilinear group $\mathbb{G}$ of order n as defined above. And let g be a generator of $\mathbb{G}$ and $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_1$ be the bilinear map.
3. Output $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$.

Obviously the group action in $\mathbb{G}, \mathbb{G}_1$ and the bilinear map are computable in polynomial time in τ. Let $\tau \in Z^+$ and let $(q_1, q_2, \mathbb{G}, \mathbb{G}_1, e)$ be a tuple produced by $\mathcal{G}(\tau)$ where $n = q_1 q_2$. Consider the following problem: given $(n, \mathbb{G}, \mathbb{G}_1, e)$ and an element $x \in \mathbb{G}$, output '1' if the order of x is q_1 and output '0' otherwise; i.e., decide if an element x is in a subgroup of $\mathbb{G}$, without knowing the factorization of n. We call it the subgroup decision problem and define the advantage of $\mathcal{A}$ in solving the subgroup decision problem $\mathit{SD\text{-}Adv}_{\mathcal{A}}(\tau)$ as:

$$\mathit{SD\text{-}Adv}_{\mathcal{A}}(\tau) = \left| \begin{array}{l} \Pr\left[\begin{array}{r} \mathcal{A}(n, \mathbb{G}, \mathbb{G}_1, e, x) = 1 : (q_1, q_2, \mathbb{G}, \mathbb{G}_1, e) \leftarrow \mathcal{G}(\tau), \\ n = q_1 q_2, x \leftarrow \mathbb{G} \end{array}\right] \\ -\Pr\left[\begin{array}{r} \mathcal{A}(n, \mathbb{G}, \mathbb{G}_1, e, x^{q_2}) = 1 : (q_1, q_2, \mathbb{G}, \mathbb{G}_1, e) \leftarrow \mathcal{G}(\tau), \\ n = q_1 q_2, x \leftarrow \mathbb{G} \end{array}\right] \end{array} \right|$$

Definition 4 We say that $\mathcal{G}$ satisfies the subgroup decision assumption if $\mathit{SD\text{-}Adv}_{\mathcal{A}}(\tau)$ is a negligible function in τ for any polynomial time algorithm $\mathcal{A}$.

Theorem 1 Our scheme is semantically secure assuming $\mathcal{G}$ satisfies the subgroup decision assumption.

4.2 Proof

Suppose that a polynomial time algorithm $\mathcal{B}$ breaks the semantic security of the system with advantage $\varepsilon(\tau)$. That's to say, there will exist an algorithm $\mathcal{A}$ that breaks the subgroup decision assumption with the same advantage. Detailed proof procedure is as follows:

1. Algorithm $\mathcal{A}$ chooses a generator $g \in \mathbb{G}$ randomly, sends the public key $(n, \mathbb{G}, \mathbb{G}_1, e, g, x)$ to algorithm $\mathcal{B}$.
2. Algorithm $\mathcal{B}$ outputs two messages $m_0, m_1 \in \{0, 1, \cdots T\}$ to algorithm $\mathcal{A}$, and algorithm $\mathcal{A}$ responds with the ciphertext $C = g^{m_b}x^r \in \mathbb{G}$ for a random $b \xleftarrow{R} \{0, 1\}$ and random $r \xleftarrow{R} \{0, 1, \cdots, n-1\}$ to algorithm $\mathcal{B}$.
3. Algorithm $\mathcal{B}$ outputs $b' \in \{0, 1\}$ for b as its guess. If $b = b'$ algorithm $\mathcal{A}$ outputs 1 (i.e., x is uniformly distributed in a subgroup of $\mathbb{G}$); otherwise $\mathcal{A}$ outputs 0 (i.e., x is uniformly distributed in $\mathbb{G}$).

It is apparent that when x is uniformly distributed in $\mathbb{G}$, the challenge ciphertext C is uniform in $\mathbb{G}$ and is independent of b. Thus, in this case $\Pr|b = b'| = 1/2$. But then, when x is uniformly distributed in q_1-subgroup of $\mathbb{G}$, the public key and challenge C given to $\mathcal{B}$ are as in a real semantic security game. In this case, it is obvious that $\Pr|b = b'| > 1/2 + \varepsilon(\tau)$ by the definition of $\mathcal{B}$. It now follows that $\mathcal{A}$ satisfies $SD\text{-}Adv_{\mathcal{A}}(\tau) > \varepsilon(\tau)$ and hence $\mathcal{A}$ breaks the subgroup decision assumption with advantage $\varepsilon(\tau)$ as required.

Therefore, we prove semantic security of the scheme under the subgroup decision assumption. What's more, it's explicit that the leakage of the attribute does not affect the security of the system. Because even if an attacker got the attribute and the random parameter z, i.e., he could gain the value of $e'(g, g)^{s\alpha}$, however he would fail in computing $c^{q_1} = \left(k^{mH(e'(g,g)^{\alpha s})}h^R\right)^{q_1} = \left(k^{H(e'(g,g)^{\alpha s})q_1}\right)^m$ to cover m without q_1. On the other hand, if the attacker got nothing but q_1, his attribute did not meet the access policy, i.e., he could not work out $e'(g, g)^{s\alpha}$, he cannot decrypt the ciphertexts as well. To sum up, only the legitimate users can cover m in our scheme.

5 Performance Analysis

Green *et al.* [13] presented the idea of outsourcing the decryption of attribute-based encryption ciphertexts, and Boneh *et al.* [17] proposed a classic somewhat homomorphic encryption. In this section, we compared our scheme with the literature [13, 17] in the following aspects: whether to support homomorphic operation, the effect of attributes leak on security, the size of ciphertext and the decryption ops. The results are shown in Tables 1 and 2.

Table 1. Comparison with Green scheme

Scheme	Homomorphic	Effect of attributes leak on security
Green [13]	No	Deadly
Ours	Yes	Hardly

From Table 1, it is distinct that compared with [13], ours do support homomorphic operation on ciphertexts outsourced to the cloud. Moreover, the security is not directly determined by attributes, which means that the malicious users cannot carry out collusion attacks, our system security is based on the subgroup decision assumption.

Table 2. Comparison with BGN scheme

Scheme	Access control	Ciphertext size	Decryption Ops
BGN [17]	No	$O(T)$	$O(\sqrt{T})$
Ours	Yes	$O(T)$	$O(\sqrt{T}) + O_P$

From Table 2, O_P stands for the time to compute hash function H, and compared with BGN scheme, although the decryption overhead increases, the ciphertext length is just the same. On the other hand, the BGN scheme fails in providing fine-grained access control, however ours achieves restricting who can get the results of homomorphic encryption through employing ABE.

6 Summary

In this article, we bring the thought of outsourcing the decryption of ABE ciphertexts into BGN scheme, and propose our BGN type outsourcing the decryption of CP-ABE ciphertexts, which is suitable for the cloud environment. By using the method of attribute-based encryption, we can solve the problem of access control on cloud computing results, and the users' computation overhead in decryption reduces remarkably, because the process of outsourcing improves users' decrypting efficiency. Further work is to explore the combination of outsourcing the decryption of ABE ciphertexts with the full homomorphic encryption, and to construct a more efficient and practical outsourcing scheme for the full homomorphic encryption based on the cloud.

References

1. Hand, E.: Head in the clouds. Nature **449**(7165), 963 (2007)
2. Alamareen, A., Al-Jarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web. Eng. **11**(3), 1–14 (2016)
3. Almiani, M., Razaque, A., Al-Dmour, A.: Privacy preserving framework to support mobile government services. Int. J. Inf. Technol. Web. Eng. **11**(3), 65–78 (2016)
4. Dam, H.K., Ghose, A., Qasim, M.: An agent-mediated platform for business processes. Int. J. Inf. Technol. Web. Eng. **10**(2), 43–61 (2015)
5. Mezghani, K., Ayadi, F.: Factors explaining IS managers attitudes toward cloud computing adoption. Int. J. Technol. Human Interact. **12**(1), 1–20 (2016)
6. Khan, N., Al-Yasiri, A.: Cloud security threats and techniques to strengthen cloud computing adoption framework. Int. J. Inf. Technol. Web. Eng. **11**(3), 50–64 (2016)

7. Kaufman, L.M.: Data security in the world of cloud computing. IEEE Secur. Priv. **7**(4), 61–64 (2009)
8. Shamir A.: Identity-based cryptosystems and signature schemes. In: Workshop on the Theory and Application of Cryptographic Techniques, pp. 47–53. Springer, Heidelberg (1984)
9. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 457–473. Springer, Heidelberg (2005)
10. Pirretti, M., Traynor, P., McDaniel, P., et al.: Secure attribute-based systems. J. Comput. Secur. **18**(5), 799–837 (2010)
11. Ning, J., Dong, X., Cao, Z., et al.: White-box traceable ciphertext-policy attribute-based encryption supporting flexible attributes. IEEE Trans. Inf. Forensics Secur. **10**(6), 1 (2015)
12. Zhang, K., Ma, J., Liu, J., et al.: Adaptively secure multi-authority attribute-based encryption with verifiable outsourced decryption. Sci. China Inf. Sci. (2016)
13. Green, M., Hohenberger, S., Waters, B.: Outsourcing the decryption of ABE ciphertexts. In: USENIX Security Symposium (2011)
14. Wan, Z., Liu, J., Deng, R.H.: HASBE: a hierarchical attribute-based solution for flexible and scalable access control in cloud computing. IEEE Trans. Inf. Forensics Secur. **7**(2), 743–754 (2012)
15. Yang, K., Jia, X., Ren, K., et al.: DAC-MACS: effective data access control for multi-authority cloud storage systems. IEEE Trans. Inf. Forensics Secur. **8**(11), 1790–1801 (2013)
16. Wang, S., Zhou, J., Liu, J., et al.: An efficient file hierarchy attribute-based encryption scheme in cloud computing. IEEE Trans. Inf. Forensics Secur. **11**(6), 1 (2016)
17. Boneh, D., Goh, E.J., Nissim, K.: Evaluating 2-DNF formulas on ciphertexts. In: Theory of Cryptography Conference, pp. 325–341. Springer, Heidelberg (2005)
18. Beimel A.: Secure schemes for secret sharing and key distribution. Int. J. Pure Appl. Math. (1996)
19. Rivest, R.L., Adleman, L., Dertouzos, M.L.: On data banks and privacy homomorphisms. Found. Secure Comput. **4**(11), 169–180 (1978)
20. Gentry, C., Halevi, S., Vaikuntanathan, V.: A simple BGN-type cryptosystem from LWE. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 506–522. Springer, Heidelberg (2010)
21. Menezes, A.J., Oorschot, P.V., Vanstone, S.A.: Handbook of Applied Cryptography, pp. 425–488. CRC Press, Boca Raton (1999)

Performance Evaluation of WMN-PSOHC and WMN-PSO Simulation Systems for Node Placement in Wireless Mesh Networks: A Comparison Study

Shinji Sakamoto[1](✉), Kosuke Ozera[1], Tetsuya Oda[2], Makoto Ikeda[2], and Leonard Barolli[2]

[1] Graduate School of Engineering, Fukuoka Institute of Technology (FIT),
3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
shinji.t.sakamoto@gmail.com, kosuke.o.fit@gmail.com

[2] Department of Information and Communication Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
oda.tetsuya.fit@gmail.com, makoto.ikd@acm.org, barolli@fit.ac.jp

Abstract. Wireless Mesh Networks (WMNs) have many advantages such as low cost and increased high speed wireless Internet connectivity, therefore WMNs are becoming an important networking infrastructure. In our previous work, we implemented a Particle Swarm Optimization (PSO) based simulation system for node placement in WMNs, called WMN-PSO. Also, we implemented a simulation system based on Hill Climbing (HC) for solving node placement problem in WMNs, called WMN-HC. In this paper, we implement a hybrid simulation system based on PSO and HC, called WMN-PSOHC. We compare WMN-PSO with WMN-PSOHC by conducting computer simulations. The simulation results show that the WMN-PSOHC has better performance than WMN-PSO.

1 Introduction

The wireless networks and devises are becoming increasingly popular and they provide users access to information and communication anytime and anywhere [11–16]. Wireless Mesh Networks (WMNs) are gaining a lot of attention because of their low cost nature that makes them attractive for providing wireless Internet connectivity. A WMN is dynamically self-organized and self-configured, with the nodes in the network automatically establishing and maintaining mesh connectivity among them-selves (creating, in effect, an ad hoc network). This feature brings many advantages to WMNs such as low up-front cost, easy network maintenance, robustness and reliable service coverage [1]. Moreover, such infrastructure can be used to deploy community networks, metropolitan area

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_7

networks, municipal and corporative networks, and to support applications for urban areas, medical, transport and surveillance systems.

Mesh node placement in WMN can be seen as a family of problems, which are shown (through graph theoretic approaches or placement problems, e.g. [8,19]) to be computationally hard to solve for most of the formulations [32]. In fact, the node placement problem considered here is even more challenging due to two additional characteristics:

(a) locations of mesh router nodes are not pre-determined, in other wards, any available position in the considered area can be used for deploying the mesh routers.
(b) routers are assumed to have their own radio coverage area.

Here, we consider the version of the mesh router nodes placement problem in which we are given a grid area where to deploy a number of mesh router nodes and a number of mesh client nodes of fixed positions (of an arbitrary distribution) in the grid area. The objective is to find a location assignment for the mesh routers to the cells of the grid area that maximizes the network connectivity and client coverage. Node placement problems are known to be computationally hard to solve [17,18,33]. In some previous works, intelligent algorithms have been recently investigated [2–4,6,7,9,10,20,22,24–26,34].

In our previous work, we implemented a Particle Swarm Optimization (PSO) based simulation system, called WMN-PSO [27]. Also, we implemented a simulation system based on Hill Climbing (HC) for solving node placement problem in WMNs, called WMN-HC [23].

In this paper, we implement a hybrid simulation system based on PSO and HC. We call this system WMN-PSOHC. We compare the performance of hybrid WMN-PSOHC system with WMN-PSO.

The rest of the paper is organized as follows. The mesh router nodes placement problem is defined in Sect. 2. We present our designed and implemented hybrid simulation system in Sect. 3. The simulation results are given in Sect. 4. Finally, we give conclusions and future work in Sect. 5.

2 Node Placement Problem in WMNs

For this problem, we have a grid area arranged in cells we want to find where to distribute a number of mesh router nodes and a number of mesh client nodes of fixed positions (of an arbitrary distribution) in the considered area. The objective is to find a location assignment for the mesh routers to the area that maximizes the network connectivity and client coverage. Network connectivity is measured by Size of Giant Component (SGC) of the resulting WMN graph, while the user coverage is simply the number of mesh client nodes that fall within the radio coverage of at least one mesh router node and is measured by Number of Covered Mesh Clients (NCMC).

An instance of the problem consists as follows.

- N mesh router nodes, each having its own radio coverage, defining thus a vector of routers.
- An area $W \times H$ where to distribute N mesh routers. Positions of mesh routers are not pre-determined and are to be computed.
- M client mesh nodes located in arbitrary points of the considered area, defining a matrix of clients.

It should be noted that network connectivity and user coverage are among most important metrics in WMNs and directly affect the network performance.

In this work, we have considered a bi-objective optimization in which we first maximize the network connectivity of the WMN (through the maximization of the SGC) and then, the maximization of the NCMC.

In fact, we can formalize an instance of the problem by constructing an adjacency matrix of the WMN graph, whose nodes are router nodes and client nodes and whose edges are links between nodes in the mesh network. Each mesh node in the graph is a triple $\boldsymbol{v} =< x, y, r >$ representing the 2D location point and r is the radius of the transmission range. There is an arc between two nodes $\boldsymbol{u}$ and $\boldsymbol{v}$, if $\boldsymbol{v}$ is within the transmission circular area of $\boldsymbol{u}$.

3 Proposed and Implemented Simulation System

3.1 PSO

In PSO a number of simple entities (the particles) are placed in the search space of some problem or function and each evaluates the objective function at its current location. The objective function is often minimized and the exploration of the search space is not through evolution [21]. However, following a widespread practice of borrowing from the evolutionary computation field, in this work, we consider the bi-objective function and fitness function interchangeably. Each particle then determines its movement through the search space by combining some aspect of the history of its own current and best (best-fitness) locations with those of one or more members of the swarm, with some random perturbations. The next iteration takes place after all particles have been moved. Eventually the swarm as a whole, like a flock of birds collectively foraging for food, is likely to move close to an optimum of the fitness function.

Each individual in the particle swarm is composed of three $\mathcal{D}$-dimensional vectors, where $\mathcal{D}$ is the dimensionality of the search space. These are the current position $\mathbf{x}_i$, the previous best position $\mathbf{p}_i$ and the velocity $\mathbf{v}_i$.

The particle swarm is more than just a collection of particles. A particle by itself has almost no power to solve any problem; progress occurs only when the particles interact. Problem solving is a population-wide phenomenon, emerging from the individual behaviors of the particles through their interactions. In any case, populations are organized according to some sort of communication structure or topology, often thought of as a social network. The topology typically consists of bidirectional edges connecting pairs of particles, so that if j is in i's neighborhood, i is also in j's. Each particle communicates with some other particles and is affected by the best point found by any member of its topological neighborhood. This is just the vector $\mathbf{p}_i$ for that best neighbor, which we will denote with $\mathbf{p}_g$. The potential kinds of population "social networks" are hugely varied, but in practice certain types have been used more frequently.

In the PSO process, the velocity of each particle is iteratively adjusted so that the particle stochastically oscillates around $\mathbf{p}_i$ and $\mathbf{p}_g$ locations.

Initialization. Our proposed system starts by generating an initial solution randomly, by *ad hoc* methods [34]. We decide the velocity of particles by a random process considering the area size. For instance, when the area size is $W \times H$, the velocity is decided randomly from $-\sqrt{W^2 + H^2}$ to $\sqrt{W^2 + H^2}$.

Particle-pattern. A particle is a mesh router. A fitness value of a particle-pattern is computed by combination of mesh routers and mesh clients positions. In other words, each particle-pattern is a solution as shown is Fig. 1. Therefore, the number of particle-patterns is a number of solutions.

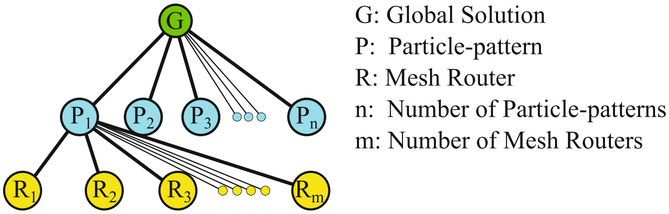


Fig. 1. Relationship among global solution, particle-patterns and mesh routers.

Fitness function. One of most important thing in PSO algorithm is to decide the determination of an appropriate objective function and its encoding. In our case, each particle-pattern has an own fitness value and compares other particle-pattern's fitness value in order to share information of global solution. The fitness function follows a hierarchical approach in which the main objective is to maximize the SGC in WMN. Thus, the fitness function of this scenario is defined as

$$\text{Fitness} = 0.7 \times \text{SGC}(\boldsymbol{x}_{ij}, \boldsymbol{y}_{ij}) + 0.3 \times \text{NCMC}(\boldsymbol{x}_{ij}, \boldsymbol{y}_{ij}).$$

Routers replacement method. A mesh router has x, y positions and velocity. Mesh routers are moved based on velocities. There are many moving methods in PSO field, such as:

Constriction Method (CM)
CM is a method which PSO parameters are set to a week stable region ($\omega = 0.729$, $C_1 = C2 = 1.4955$) based on analysis of PSO by M. Clerc et. al. [5,30].

Random Inertia Weight Method (RIWM)
In RIWM, the ω parameter is changing randomly from 0.5 to 1.0. The C_1 and C_2 are kept 2.0. The ω can be estimated by the week stable region. The average of ω is 0.75 [30].
Linearly Decreasing Inertia Weight Method (LDIWM)
In LDIWM, C_1 and C_2 are set to 2.0, constantly. On the other hand, the ω parameter is changed linearly from unstable region ($\omega = 0.9$) to stable region ($\omega = 0.4$) with increasing of iterations of computations [30,31].
Linearly Decreasing Vmax Method (LDVM)
In LDVM, PSO parameters are set to unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). A value of V_{max} which is maximum velocity of particles is considered. With increasing of iteration of computations, the V_{max} is kept decreasing linearly [29].
Rational Decrement of Vmax Method (RDVM)
In RDVM, PSO parameters are set to unstable region ($\omega = 0.9$, $C_1 = C_2 = 2.0$). The V_{max} is kept decreasing with the increasing of iterations as

$$V_{max}(x) = \sqrt{W^2 + H^2} \times \frac{T - x}{x}.$$

where, W and H are the width and the height of the considered area, respectively. Also, T and x are the total number of iterations and a current number of iteration, respectively [28].

3.2 HC Algorithm

HC is local search algorithm and is based on incremental improvements of solutions as follows: it starts with a solution (which may be randomly generated or ad hoc computed) considered as the current solution in the search space. The algorithm examines its neighboring solutions and if a neighbor is better than current solution then it can become the current solution; the algorithm keeps moving from one solution to another one in the search space until no further improvements are possible. There are several variants of the algorithm depending on whether a simple climbing, steepest ascent climbing or stochastic climbing is done:

- *Simple climbing*: the next neighbor solution is the first that improves current solution.
- *Steepest ascent climbing*: all neighbor solutions are examined and the best one is chosen as next solution.
- *Stochastic climbing*: a neighbor is selected at random, and according to yielded improvement of that neighbor is decided whether to choose it as next solution or to examine another neighbor. This kind of climbing has more general forms known as Metropolis and Simulated Annealing algorithms.

There are several versions of the algorithm are possible depending on the way a neighbor solution is selected. It should be noted that HC usually ends up in local optima, which can be overcome in some cases by adopting additional techniques such as:

Algorithm 1. Pseudo code of PSO.

```
/* Generate the initial solutions and parameters */
Computation maxtime:= $T_{max}$, $t = 0$;
Number of particle-patterns:= $m$, $2 \leq m \in \boldsymbol{N}^1$;
Particle-patterns initial solution:= $\boldsymbol{P}_i^0$;
Global initial solution:= $\boldsymbol{G}^0$;
Particle-patterns initial position:= $\boldsymbol{x}_{ij}^0$;
Particles initial velocity:= $\boldsymbol{v}_{ij}^0$;
PSO parameter:= $\omega$, $0 < \omega \in \boldsymbol{R}^1$;
PSO parameter:= $C_1$, $0 < C_1 \in \boldsymbol{R}^1$;
PSO parameter:= $C_2$, $0 < C_2 \in \boldsymbol{R}^1$;
/* Start PSO */
Evaluate($\boldsymbol{G}^0, \boldsymbol{P}^0$);
/* "Evaluate" does calculate present fitness value of each Particle-patterns. */
while $t < T_{max}$ do
   /* Update velocities and positions */
   $\boldsymbol{v}_{ij}^{t+1} = \omega \cdot \boldsymbol{v}_{ij}^t$
         $+C_1 \cdot \text{rand}() \cdot (best(P_{ij}^t) - x_{ij}^t)$
         $+C_2 \cdot \text{rand}() \cdot (best(G^t) - x_{ij}^t)$;
   $\boldsymbol{x}_{ij}^{t+1} = \boldsymbol{x}_{ij}^t + \boldsymbol{v}_{ij}^{t+1}$;
   Update_Solutions($\boldsymbol{G}^t, \boldsymbol{P}^t$);
   /* "Update_Solutions" compares and updates the Particle-pattern's best solutions
   and the global best solutions if their fitness value is better than previous. */
   Evaluate($\boldsymbol{G}^{(t+1)}, \boldsymbol{P}^{(t+1)}$);
   $t = t + 1$;
end while
return Best found pattern of particles as solution;
```

(a) getting back to a previous state and exploring another direction;
(b) jumping to a new solution, possibly "far away" from current solution;
(c) considering several search direction in solution space at the same time.

3.3 Implemented Simulation Systems

We present here the implementation of two simulation systems.

3.3.1 WMN-PSO

We show the PSO algorithm in Algorithm 1 for the mesh router node placement problem in WMNs. We implemented a simulator that uses PSO algorithm to solve the node placement problem in WMNs. We call this simulator WMN-PSO. Our system can generate instances of the problem using different iterations of clients and mesh routers.

3.3.2 WMN-PSOHC

We show the PSO-HC hybrid algorithm in Algorithm 2. WMN-PSO system is improved by HC mechanism. We call this simulator WMN-PSOHC.

Algorithm 2. Pseudo code of PSO-HC.

```
/* Generate the initial solutions and parameters */
Computation maxtime:= $T_{max}$, $t := 0$;
Number of particle-patterns:= $m$, $2 \leq m \in \boldsymbol{N}^1$;
Particle-patterns initial solution:= $\boldsymbol{P}_i^0$;
Global initial solution:= $\boldsymbol{G}^0$;
Particle-patterns initial position:= $\boldsymbol{x}_{ij}^0$;
Particles initial velocity:= $\boldsymbol{v}_{ij}^0$;
PSO parameter:= $\omega$, $0 < \omega \in \boldsymbol{R}^1$;
PSO parameter:= $C_1$, $0 < C_1 \in \boldsymbol{R}^1$;
PSO parameter:= $C_2$, $0 < C_2 \in \boldsymbol{R}^1$;
/* Start PSO-HC */
Evaluate($\boldsymbol{G}^0, \boldsymbol{P}^0$);
while $t < T_{max}$ do
  /* Update velocities and positions */
  $\boldsymbol{v}_{ij}^{t+1} = \omega \cdot \boldsymbol{v}_{ij}^t$
      $+C_1 \cdot \text{rand}() \cdot (best(P_{ij}^t) - x_{ij}^t)$
      $+C_2 \cdot \text{rand}() \cdot (best(G^t) - x_{ij}^t)$;
  $\boldsymbol{x}_{ij}^{t+1} = \boldsymbol{x}_{ij}^t + \boldsymbol{v}_{ij}^{t+1}$;
  /* if fitness value is increased, a new solution will be accepted. */
  if Evaluate($\boldsymbol{G}^{(t+1)}, \boldsymbol{P}^{(t+1)}$) ¿= Evaluate($\boldsymbol{G}^{(t)}, \boldsymbol{P}^{(t)}$) then
    Update_Solutions($\boldsymbol{G}^t, \boldsymbol{P}^t$);
    Evaluate($\boldsymbol{G}^{(t+1)}, \boldsymbol{P}^{(t+1)}$);
  else
    /* "Reupdate_Solutions" makes particle back to previous position */
    Reupdate_Solutions($\boldsymbol{G}^{(t+1)}, P^{(t+1)}$);
  end if
  $t = t + 1$;
end while
Update_Solutions($\boldsymbol{G}^t, \boldsymbol{P}^t$);
return Best found pattern of particles as solution;
```

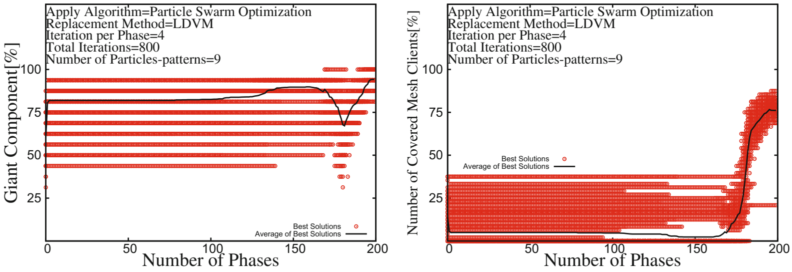
4 Simulation Results

In this section, we show simulation results using WMN-PSO and WMN-PSOHC systems. In this work, we set the same parameters in order to compare two simulation systems. We consider normal distribution of mesh clients. The number of mesh routers is considered 16 and the number of mesh clients 48. The total number of iterations is considered 800 and the iterations per phase is considered 4. We consider the number of particle-patterns 9. We conducted simulations 100 times, in order to avoid the effect of randomness and create a general view of results. We show the parameter setting for both WMN-PSO and WMN-PSOHC in Table 1.

We show the simulation results from Figs. 2, 3 and 4. In Fig. 2, we see that solutions converge after 170 phases and NCMC does not reach maximum (100%). We show simulations results for WMN-PSOHC in Fig. 3. Figure 3(a) shows that the SGC increases gradually and all solution reaches maximum value. Also, we

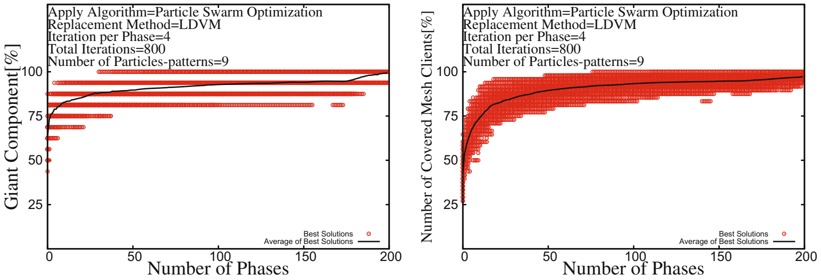
Table 1. WMN-PSO and WMN-PSOHC Parameters.

Parameters	Values
Clients distribution	Normal distribution
Area size	32.0 × 32.0
Number of mesh routers	16
Number of mesh clients	48
Total iterations	800
Iteration per phase	4
Number of particle-patterns	9
Radius of a mesh router	2.0
Replacement method	LDVM



(a) SGC (b) NCMC

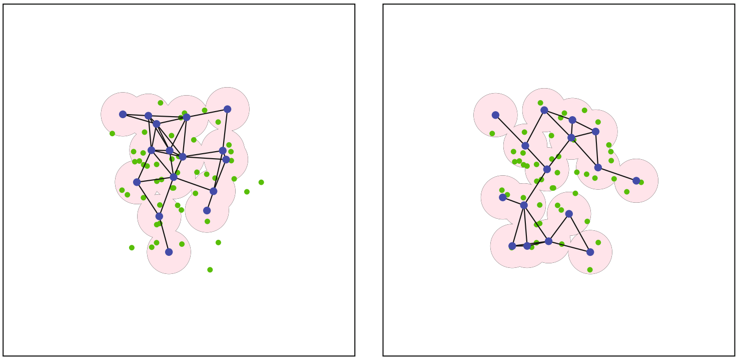
Fig. 2. Simulation results for WMN-PSO.



(a) SGC (b) NCMC

Fig. 3. Simulation results for WMN-PSOHC.

can see that WMN-PSOHC converges faster than WMN-PSO and NCMC reaches 100% as shown in Fig. 3(b). We show the visualized results for WMN-PSO and WMN-PSOHC in Fig. 4(a) and (b), respectively. We see that all mesh routers are connected for both systems. However, some mesh clients are not covered in WMN-PSO. On the other hand, all mesh clients are covered by mesh routers in



(a) WMN-PSO

(b) WMN-PSOHC

Fig. 4. Visualized image of simulation results.

WMN-PSOHC. Comparing WMN-PSO with WMN-PSOHC, WMN-PSOHC has better performance than WMN-PSO.

5 Conclusions

In this work, we compared hybrid simulation system based on PSO and HC (called WMN-PSOHC) with WMN-PSO.

From the simulation results, we conclude the WMN-PSOHC has better performance than WMN-PSO.

In our future work, we would like to evaluate the performance of the proposed system for different parameters and patterns.

Acknowledgement. This work is supported by a Grant-in-Aid for Scientific Research from Japanese Society for the Promotion of Science (JSPS KAKENHI Grant Number 15J12086). The authors would like to thank JSPS for the financial support.

References

1. Akyildiz, I.F., Wang, X., Wang, W.: Wireless mesh networks: a survey. Comput. Netw. **47**(4), 445–487 (2005)
2. Amaldi, E., Capone, A., Cesana, M., Filippini, I., Malucelli, F.: Optimization models and methods for planning wireless mesh networks. Comput. Netw. **52**(11), 2159–2171 (2008)
3. Barolli, A., Spaho, E., Barolli, L., Xhafa, F., Takizawa, M.: QoS routing in ad-hoc networks using GA and multi-objective optimization. Mob. Inf. Syst. **7**(3), 169–188 (2011)
4. Behnamian, J., Ghomi, S.F.: Development of a PSO-SA hybrid metaheuristic for a new comprehensive regression model to time-series forecasting. Expert Syst. Appl. **37**(2), 974–984 (2010)

5. Clerc, M., Kennedy, J.: The particle swarm-explosion, stability, and convergence in a multidimensional complex space. IEEE Trans. Evol. Comput. **6**(1), 58–73 (2002)
6. Cunha, M.C., Sousa, J.: Water distribution network design optimization: simulated annealing approach. J. Water Resour. Plan. Manage. **125**(4), 215–221 (1999)
7. Del Valle, Y., Venayagamoorthy, G.K., Mohagheghi, S., Hernandez, J.C., Harley, R.G.: Particle swarm optimization: basic concepts, variants and applications in power systems. IEEE Trans. Evol. Comput. **12**(2), 171–195 (2008)
8. Franklin, A.A., Murthy, C.S.R.: Node placement algorithm for deployment of two-tier wireless mesh networks. In: Proceedings of Global Telecommunications Conference, pp. 4823–4827 (2007)
9. Ge, H., Du, W., Qian, F.: A hybrid algorithm based on particle swarm optimization and simulated annealing for job shop scheduling. In: Third International Conference on Natural Computation (ICNC-2007), vol. 3, pp. 715–719 (2007)
10. Girgis, M.R., Mahmoud, T.M., Abdullatif, B.A., Rabie, A.M.: Solving the wireless mesh network design problem using genetic algorithm and simulated annealing optimization methods. Int. J. Comput. Appl. **96**(11), 1–10 (2014)
11. Goto, K., Sasaki, Y., Hara, T., Nishio, S.: Data gathering using mobile agents for reducing traffic in dense mobile wireless sensor networks. Mob. Inf. Syst. **9**(4), 295–314 (2013)
12. Hiyama, M., Sakamoto, S., Kulla, E., Ikeda, M., Barolli, L.: Experimental results of a MANET testbed for different settings of HELLO packets of OLSR protocol. J. Mob. Multimedia **9**(1–2), 27–38 (2013)
13. Inaba, T., Sakamoto, S., Kulla, E., Caballe, S., Ikeda, M., Barolli, L.: An integrated system for wireless cellular and ad-hoc networks using fuzzy logic. In: International Conference on Intelligent Networking and Collaborative Systems (INCoS-2014), pp. 157–162 (2014)
14. Inaba, T., Elmazi, D., Liu, Y., Sakamoto, S., Barolli, L., Uchida, K.: Integrating wireless cellular and ad-hoc networks using fuzzy logic considering node mobility and security. In: The 29th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA-2015), pp. 54–60 (2015). doi:10.1109/WAINA.2015.116
15. Inaba, T., Elmazi, D., Sakamoto, S., Oda, T., Ikeda, M., Barolli, L.: A secure-aware call admission control scheme for wireless cellular networks using fuzzy logic and its performance evaluation. J. Mob. Multimedia **11**(3&4), 213–222 (2015)
16. Inaba, T., Sakamoto, S., Oda, T., Ikeda, M., Barolli, L.: A testbed for admission control in WLAN: a fuzzy approach and its performance evaluation. In: International Conference on Broadband and Wireless Computing, pp. 559–571. Springer (2016)
17. Lim, A., Rodrigues, B., Wang, F., Xu, Z.: *k*-center problems with minimum coverage. In: Computing and Combinatorics, pp. 349–359 (2004)
18. Maolin, T., et al.: Gateways placement in backbone wireless mesh networks. Int. J. Commun. Netw. Syst. Sci. **2**(1), 44 (2009)
19. Muthaiah, S.N., Rosenberg, C.P.: Single gateway placement in wireless mesh networks. In: Proceedings of 8th International IEEE Symposium on Computer Networks, pp. 4754–4759 (2008)
20. Naka, S., Genji, T., Yura, T., Fukuyama, Y.: A hybrid particle swarm optimization for distribution state estimation. IEEE Trans. Power Syst. **18**(1), 60–68 (2003)
21. Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intell. **1**(1), 33–57 (2007)

22. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A comparison study of simulated annealing and genetic algorithm for node placement problem in wireless mesh networks. J. Mob. Multimedia **9**(1–2), 101–110 (2013)
23. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A comparison study of hill climbing, simulated annealing and genetic algorithm for node placement problem in WMNs. J. High Speed Netw. **20**(1), 55–66 (2014)
24. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Performance evaluation considering iterations per phase and SA temperature in WMN-SA system. Mob. Inf. Syst. **10**(3), 321–330 (2014)
25. Sakamoto, S., Lala, A., Oda, T., Kolici, V., Barolli, L., Xhafa, F.: Application of WMN-SA simulation system for node placement in wireless mesh networks: a case study for a realistic scenario. Int. J. Mob. Comput. Multimedia Commun. (IJMCMC) **6**(2), 13–21 (2014)
26. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: An integrated simulation system considering WMN-PSO simulation system and network simulator 3. In: International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 187–198. Springer (2016)
27. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation and evaluation of a simulation system based on particle swarm optimisation for node placement problem in wireless mesh networks. Int. J. Commun. Netw. Distrib. Syst. **17**(1), 1–13 (2016)
28. Sakamoto, S., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: Implementation of a new replacement method in WMN-PSO simulation system and its performance evaluation. In: The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016), pp. 206–211 (2016) doi:10.1109/AINA.2016.42
29. Schutte, J.F., Groenwold, A.A.: A study of global optimization using particle swarms. J. Global Optim. **31**(1), 93–108 (2005)
30. Shi, Y.: Particle swarm optimization. IEEE Connections **2**(1), 8–13 (2004)
31. Shi, Y., Eberhart, R.C.: Parameter selection in particle swarm optimization. In: Evolutionary Programming VII, pp. 591–600 (1998)
32. Vanhatupa, T., Hannikainen, M., Hamalainen, T.: Genetic algorithm to optimize node placement and configuration for WLAN planning. In: Proceedings of 4th IEEE International Symposium on Wireless Communication Systems, pp. 612–616 (2007)
33. Wang, J., Xie, B., Cai, K., Agrawal, D.P.: Efficient mesh router placement in wireless mesh networks. In: Proceedings of IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems (MASS-2007), pp. 1–9 (2007)
34. Xhafa, F., Sanchez, C., Barolli, L.: Ad hoc and neighborhood search methods for placement of mesh routers in wireless mesh networks. In: Proceedings of 29th IEEE International Conference on Distributed Computing Systems Workshops (ICDCS-2009), pp. 400–405 (2009)

Effects of Number of Activities the Member Failures on Qualified Voting in P2P Mobile Collaborative Team: A Comparison Study for Two Fuzzy-Based Systems

Yi Liu[1](✉), Keita Matsuo[2], Makoto Ikeda[2], and Leonard Barolli[2]

[1] Graduate School of Engineering, Fukuoka Institute of Technology (FIT),
3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
ryuui1010@gmail.com

[2] Department of Information and Communication Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811-0295, Japan
{kt-matsuo,barolli}@fit.ac.jp, makoto.ikd@acm.org

Abstract. Mobile computing has many application domains. One important domain is that of mobile applications supporting collaborative work. In a collaborative work, the members of the team has to take decision or solve conflicts in project development (such as delays, changes in project schedule, task assignment, etc.) and therefore members have to vote. Voting can be done in many ways, and in most works in the literature consider majority voting, in which every member of the team accounts on for a vote. In this work, we consider a more realistic case where a vote does not account equal for every member, but accounts on according to member's active involvement and reliability in the groupwork. We present a voting model, that we call qualified voting, in which every member has a voting score according to four parameters: Number of Activities the Member Participates (NAMP), Number of Activities the Member has Successfully Finished (NAMSF), Number of Online Discussions the Member has Participated (NODMP), Number of Activities the Member Failures (NAMF). Then, we use fuzzy based model to compute a voting score for the member. In this paper, we present two fuzzy-based voting systems (calles FVS1 and FVS2). We make a comparison study between FVS1 and FVS2. The simulation results show that with increasing of the number of activities the member failures, the VS is decreased. When NAMP, NAMSF and NODMP are high, the voting sore is high. The proposed system can choose peers with good voting score in P2P mobile collaborative team. Comparing the complexity of FVS1 and FVS2, the FVS2 is more complex than FVS1. However, it considers also the number of activities the member failures which makes the voting process better.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_8

1 Introduction

Peer to Peer technologies has been among most disruptive technologies after Internet. Indeed, the emergence of the P2P technologies changed drastically the concepts, paradigms and protocols of sharing and communication in large scale distributed systems. As pointed out since early 2000 years [1–5], the nature of the sharing and the direct communication among peers in the system, being these machines or people, makes possible to overcome the limitations of the flat communications through email, newsgroups and other forum-based communication forms.

The usefulness of P2P technologies on one hand has been shown for the development of stand alone applications. On the other hand, P2P technologies, paradigms and protocols have penetrated other large scale distributed systems such as Mobile Ad hoc Networks (MANETs), Groupware systems, Mobile Systems to achieve efficient sharing, communication, coordination, replication, awareness and synchronization. In fact, for every new form of Internet-based distributed systems, we are seeing how P2P concepts and paradigms again play an important role to enhance the efficiency and effectiveness of such systems or to enhance information sharing and online collaborative activities of groups of people. We briefly introduce below some common application scenarios that can benefit from P2P communications.

With the fast development in mobile technologies we are witnessing how the mobile devices are widely used for supporting collaborative team work. Indeed, by using mobile devices (such as PDAs, smartphones, etc.) members of a team can not only be geographically distributed, they can also be supported on the move, when network connection can change over time. In this paper, we propose a fuzzy-based system for qualified voting in P2P mobile collaborative team.

Fuzzy Logic (FL) is the logic underlying modes of reasoning which are approximate rather then exact. The importance of FL derives from the fact that most modes of human reasoning and especially common sense reasoning are approximate in nature. FL uses linguistic variables to describe the control parameters. By using relatively simple linguistic expressions it is possible to describe and grasp very complex problems. A very important property of the linguistic variables is the capability of describing imprecise parameters.

The concept of a fuzzy set deals with the representation of classes whose boundaries are not determined. It uses a characteristic function, taking values usually in the interval [0, 1]. The fuzzy sets are used for representing linguistic labels. This can be viewed as expressing an uncertainty about the clear-cut meaning of the label. But important point is that the valuation set is supposed to be common to the various linguistic labels that are involved in the given problem.

The fuzzy set theory uses the membership function to encode a preference among the possible interpretations of the corresponding label. A fuzzy set can be defined by examplification, ranking elements according to their typicality with respect to the concept underlying the fuzzy set [6].

In this paper, we present a fuzzy-based peer voting score system (FVS1) considering three parameters: Number of Activities the Member Participates (NAMP), Number of Activities the Member has Successfully Finished (NAMSF), Number of Online Discussions the Member has Participated (NODMP) to decide the Voting Score (VS). We also implement another FVS (FVS2) considering four parameters: NAMP, NAMSF, NODMP and Number of Activities the Member Failures (NAMF) as a new parameter. We evaluated the proposed systems by simulations. The simulation results show that with increasing of NAMP, NAMSF, and NODMP, the VS is increasing, but with increasing of NAMF, the VS is decreased. The proposed systems can choose peers with a good voting score in P2P mobile collaborative team.

The structure of this paper is as follows. In Sect. 2,we introduce the scenarios of collaborative teamwork. In Sect. 3, we introduce the vote weights and voting score. In Sect. 4, we introduce FL used for control. In Sect. 5, we present the proposed fuzzy-based system. In Sect. 6, we discuss the simulation results. Finally, conclusions and future work are given in Sect. 7.

2 Scenarios of Collaborative Teamwork

In this section, we describe and analyse some main scenarios of collaborative teamwork for which P2P technologies can support efficient system design.

2.1 Collaborative Teamwork and Virtual Campuses

Collaborative work through virtual teams is a significant way of collaborating in modern businesses, online learning, etc. Collaboration in virtual teams requires efficient sharing of information (both data sharing among the group members as well as sharing of group processes) and efficient communication among members of the team. Additionally, coordination and interaction are crucial for accomplishing common tasks through a shared workspace environment. P2P systems can enable fully decentralized collaborative systems by efficiently supporting different forms of collaboration [7]. One such form is using P2P networks, with super-peer structure as show in Fig. 1.

During the last two decades, online learning has become very popular and there is a widespread of virtual campuses or combinations of face-to-face with

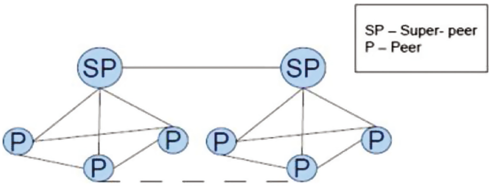


Fig. 1. Super-peer P2P group netwok.

semi-open teaching and learning. Virtual campuses are now looking at ways to effectively support learners, especially for online courses implemented as PBL-Project Based Learning or SBL Scenario Based Learning there is an increasing need to develop mobile applications that support these online groupwork learning paradigms [8]. In such setting, P2P technologies offer interesting solutions for (a) decentralizing the virtual campuses, which tend to grow and get further centralized with the increase of number of students enrolled, new degrees, and increase in academic activity; (b) in taking advantage of resources of students and developing volunteer based computing systems as part of virtual campuses and (c) alleviating the communication burden for efficient collaborative teamwork. The use of P2P libraries such as JXTA have been investigated to design P2P middleware for P2P eLearning applications. Also, the use of P2P technologies in such setting is used for P2P video synchronization in a collaborative virtual environment [9]. Recently, virtual campuses are also introducing social networking among their students to enhance the learning activities through social support and scaffolding. Again the P2P solutions are sought in this context [10] in combination with social networking features to enhance especially the interaction among learners sharing similar objectives and interest or accomplishing a common project.

2.2 Mobile Ad Hoc Networks (MANETs)

Mobile ad-hoc networks are among most interesting infrastructureless network of mobile devices connected by wireless having self-configuring properties. The lack of fixed infrastructure and of a centralized administration makes the building and operation in MANETS challenging. P2P networks and mobile ad hoc networks (MANETs) follow the same idea of creating a network without a central entity. All nodes (peers) must collaborate together to make possible the proper functioning of the network by forwarding information on behalf of others in the network [11]. P2P and MANETs share many key characteristics such as self-organization and decentralization due to the common nature of their distributed components. Both MANETs and P2P networks follow a P2P paradigm characterized by the lack of a central node or peer acting as a managing server, all participants having therefore to collaborate in order for the whole system to work. A key issue in both networks is the process of discovering the requested data or route efficiently in a decentralized manner. Recently, new P2P applications which uses wireless communication and integrates mobile devices such as PDA and mobile phones is emerging. Several P2P-based protocols can be used for MANETs such as Mobile P2P Protocol (MPP), which is based on Dynamic Source Routing (DSR), JXTA protocols, and MANET Anonymous Peer-to-peer Communication Protocol (MAPCP), which serves as an efficient anonymous communication protocol for P2P applications over MANET.

3 Vote Weights

3.1 Votes with Embedded Weight

The weights can be included in voting bulletins distributed to voters, which would then be copied into the votes sent to Counters. But this approach requires a strong assumption: the voters' application must be trusted not to forge weights. Since the voters' application may be tampered in some scenarios, namely when "voting anywhere" is considered, the voters' side cannot be trusted to give the correct input for the system when weights are considered.

The simple copy/paste of weights could be strengthened by adding a cleartext value of the weight when submitting a blinded vote digest for getting a signature from an Administrator. Then, the weight, checked and signed by all the required Administrators, could be added to the final vote submitted to Counters. A bit commitment value should also be added to the weight to prevent stolen, signed weights, to be used by other voters. The drawback of this approach is that protocol messages from voters to Administrators and from voters to Counters would increase in size, namely would double in size. This collides with the requirement of keeping the performance of system close to the performance of the initial version of REVS (Robust Electronic Voting System [12]).

3.2 Voting Score

Score voting (sometimes called range voting) is a single-winner voting system where voters rate candidates on a scale. The candidate with the highest rating wins. For comparison, consider ratings systems from site like: Internet Movie Database, Amazon, Yelp, and Hot or Not. Variations of score voting can use a score-style ballot to elect multiple candidates simultaneously.

Simplified forms of score voting automatically give skipped candidates the lowest possible score for the ballot they were skipped. Other forms have those ballots not affect the candidate's rating at all. Those forms not affecting the candidates rating frequently make use of quotas. Quotas demand a minimum proportion of voters rate that candidate in some way before that candidate is eligible to win [13].

4 Application of Fuzzy Logic for Control

The ability of fuzzy sets and possibility theory to model gradual properties or soft constraints whose satisfaction is matter of degree, as well as information pervaded with imprecision and uncertainty, makes them useful in a great variety of applications.

The most popular area of application is Fuzzy Control (FC), since the appearance, especially in Japan, of industrial applications in domestic appliances, process control, and automotive systems, among many other fields.

4.1 FC

In the FC systems, expert knowledge is encoded in the form of fuzzy rules, which describe recommended actions for different classes of situations represented by fuzzy sets.

In fact, any kind of control law can be modeled by the FC methodology, provided that this law is expressible in terms of "if ... then ..." rules, just like in the case of expert systems. However, FL diverges from the standard expert system approach by providing an interpolation mechanism from several rules. In the contents of complex processes, it may turn out to be more practical to get knowledge from an expert operator than to calculate an optimal control, due to modeling costs or because a model is out of reach.

4.2 Linguistic Variables

A concept that plays a central role in the application of FL is that of a linguistic variable. The linguistic variables may be viewed as a form of data compression. One linguistic variable may represent many numerical variables. It is suggestive to refer to this form of data compression as granulation [14].

The same effect can be achieved by conventional quantization, but in the case of quantization, the values are intervals, whereas in the case of granulation the values are overlapping fuzzy sets. The advantages of granulation over quantization are as follows:

- it is more general;
- it mimics the way in which humans interpret linguistic values;
- the transition from one linguistic value to a contiguous linguistic value is gradual rather than abrupt, resulting in continuity and robustness.

4.3 FC Rules

FC describes the algorithm for process control as a fuzzy relation between information about the conditions of the process to be controlled, x and y, and the output for the process z. The control algorithm is given in "if ... then ..." expression, such as:

If x is small and y is big, then z is medium;
If x is big and y is medium, then z is big.

These rules are called *FC rules*. The "if" clause of the rules is called the antecedent and the "then" clause is called consequent. In general, variables x and y are called the input and z the output. The "small" and "big" are fuzzy values for x and y, and they are expressed by fuzzy sets.

Fuzzy controllers are constructed of groups of these FC rules, and when an actual input is given, the output is calculated by means of fuzzy inference.

4.4 Control Knowledge Base

There are two main tasks in designing the control knowledge base. First, a set of linguistic variables must be selected which describe the values of the main control parameters of the process. Both the input and output parameters must be linguistically defined in this stage using proper term sets. The selection of the level of granularity of a term set for an input variable or an output variable plays an important role in the smoothness of control. Second, a control knowledge base must be developed which uses the above linguistic description of the input and output parameters. Four methods [15–18] have been suggested for doing this:

- expert's experience and knowledge;
- modelling the operator's control action;
- modelling a process;
- self organization.

Among the above methods, the first one is the most widely used. In the modeling of the human expert operator's knowledge, fuzzy rules of the form "If Error is small and Change-in-error is small then the Force is small" have been used in several studies [19,20]. This method is effective when expert human operators can express the heuristics or the knowledge that they use in controlling a process in terms of rules of the above form.

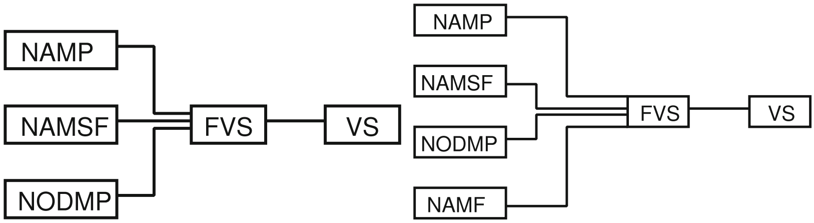
4.5 Defuzzification Methods

The defuzzification operation produces a non-FC action that best represent the membership function of an inferred FC action. Several defuzzification methods have been suggested in literature. Among them, four methods which have been applied most often are:

- Tsukamoto's Defuzzification Method;
- The Center of Area (COA) Method;
- The Mean of Maximum (MOM) Method;
- Defuzzification when Output of Rules are Function of Their Inputs.

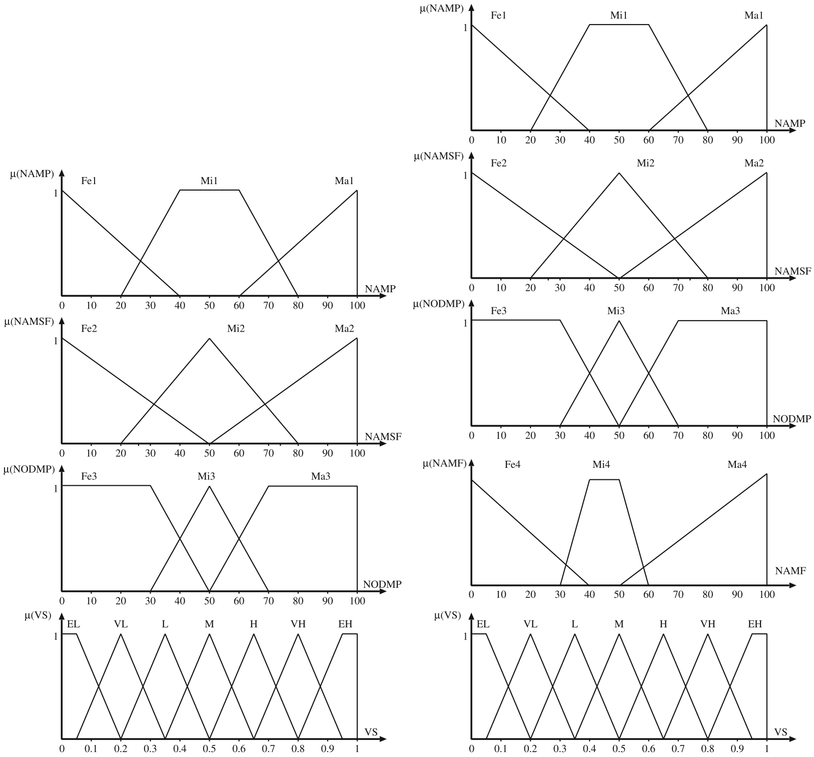
5 Proposed Fuzzy-Based Peer Voting Score System

To complete a certain task in P2P mobile collaborative team work, peers often have to in teract with unknow peers. Thus, it is important that group members must slect reliable peers to interact.



(a) Structure of FVS1 (b) Structure of FVS2

Fig. 2. Structure of FVS.



(a) Membership functions of FVS1 (b) Membership functions of FVS2

Fig. 3. Membership functions of FVS.

The structure of FVS1 is shown in Fig. 2(a) and the membership functions for FVS1 are shown in Fig. 3(a). The Fuzzy Rule Base (FRB) of FVS1 is shown in

Table 1. FRB.

Rule	NAMP	NAMF	NODMP	VS
1	Fe1	Fe2	Fe3	EL
2	Fe1	Fe2	Mi3	EL
3	Fe1	Fe2	Ma3	L
4	Fe1	Mi2	Fe3	EL
5	Fe1	Mi2	Mi3	VL
6	Fe1	Mi2	Ma3	M
7	Fe1	Ma2	Fe3	VL
8	Fe1	Ma2	Mi3	L
9	Fe1	Ma2	Ma3	H
10	Mi1	Fe2	Fe3	EL
11	Mi1	Fe2	Mi3	L
12	Mi1	Fe2	Ma3	M
13	Mi1	Mi2	Fe3	VL
14	Mi1	Mi2	Mi3	M
15	Mi1	Mi2	Ma3	H
16	Mi1	Ma2	Fe3	L
17	Mi1	Ma2	Mi3	H
18	Mi1	Ma2	Ma3	VH
19	Ma1	Fe2	Fe3	VL
20	Ma1	Fe2	Mi3	M
21	Ma1	Fe2	Ma3	VH
22	Ma1	Mi2	Fe3	L
23	Ma1	Mi2	Mi3	H
24	Ma1	Mi2	Ma3	VH
25	Ma1	Ma2	Fe3	M
26	Ma1	Ma2	Mi3	VH
27	Ma1	Ma2	Ma3	VVH

Table 1 and consists of 27 rules. In this work, we consider the NAMF as a new parameter together with three parameters to decide the VS. We call this system FVS2. The structure of FVS2 and membership functions are shown in Figs. 2(b) and 3(b), respectively. In Table 2, we show the FRB of FVS2, which consists of 81 rules.

Table 2. FRB.

Rule	NAMP	NAMSF	NODMP	NAMF	VS	Rule	NAMP	NAMSF	NODMP	NAMF	VS
1	Fe1	Fe2	Fe3	Fe4	VL	41	Mi1	Mi2	Mi3	Mi4	M
2	Fe1	Fe2	Fe3	Mi4	EL	42	Mi1	Mi2	Mi3	Ma4	VL
3	Fe1	Fe2	Fe3	Ma4	EL	43	Mi1	Mi2	Ma3	Fe4	EH
4	Fe1	Fe2	Mi3	Fe4	L	44	Mi1	Mi2	Ma3	Mi4	VH
5	Fe1	Fe2	Mi3	Mi4	VL	45	Mi1	Mi2	Ma3	Ma4	L
6	Fe1	Fe2	Mi3	Ma4	EL	46	Mi1	Ma2	Fe3	Fe4	VH
7	Fe1	Fe2	Ma3	Fe4	H	47	Mi1	Ma2	Fe3	Mi4	L
8	Fe1	Fe2	Ma3	Mi4	L	48	Mi1	Ma2	Fe3	Ma4	VL
9	Fe1	Fe2	Ma3	Ma4	EL	49	Mi1	Ma2	Mi3	Fe4	EH
10	Fe1	Mi2	Fe3	Fe4	L	50	Mi1	Ma2	Mi3	Mi4	H
11	Fe1	Mi2	Fe3	Mi4	EL	51	Mi1	Ma2	Mi3	Ma4	L
12	Fe1	Mi2	Fe3	Ma4	EL	52	Mi1	Ma2	Ma3	Fe4	EH
13	Fe1	Mi2	Mi3	Fe4	M	53	Mi1	Ma2	Ma3	Mi4	VH
14	Fe1	Mi2	Mi3	Mi4	VL	54	Mi1	Ma2	Ma3	Ma4	H
15	Fe1	Mi2	Mi3	Ma4	EL	55	Ma1	Fe2	Fe3	Fe4	H
16	Fe1	Mi2	Ma3	Fe4	VH	56	Ma1	Fe2	Fe3	Mi4	L
17	Fe1	Mi2	Ma3	Mi4	M	57	Ma1	Fe2	Fe3	Ma4	EL
18	Fe1	Mi2	Ma3	Ma4	VL	58	Ma1	Fe2	Mi3	Fe4	VH
19	Fe1	Ma2	Fe3	Fe4	M	59	Ma1	Fe2	Mi3	Mi4	M
20	Fe1	Ma2	Fe3	Mi4	VL	60	Ma1	Fe2	Mi3	Ma4	VL
21	Fe1	Ma2	Fe3	Ma4	EL	61	Ma1	Fe2	Ma3	Fe4	EH
22	Fe1	Ma2	Mi3	Fe4	VH	62	Ma1	Fe2	Ma3	Mi4	VH
23	Fe1	Ma2	Mi3	Mi4	L	63	Ma1	Fe2	Ma3	Ma4	M
24	Fe1	Ma2	Mi3	Ma4	VL	64	Ma1	Mi2	Fe3	Fe4	VH
25	Fe1	Ma2	Ma3	Fe4	EH	65	Ma1	Mi2	Fe3	Mi4	M
26	Fe1	Ma2	Ma3	Mi4	H	66	Ma1	Mi2	Fe3	Ma4	VL
27	Fe1	Ma2	Ma3	Ma4	L	67	Ma1	Mi2	Mi3	Fe4	EH
28	Mi1	Fe2	Fe3	Fe4	L	68	Ma1	Mi2	Mi3	Mi4	VH
29	Mi1	Fe2	Fe3	Mi4	VL	69	Ma1	Mi2	Mi3	Ma4	L
30	Mi1	Fe2	Fe3	Ma4	EL	70	Ma1	Mi2	Ma3	Fe4	EH
31	Mi1	Fe2	Mi3	Fe4	H	71	Ma1	Mi2	Ma3	Mi4	EH
32	Mi1	Fe2	Mi3	Mi4	L	72	Ma1	Mi2	Ma3	Ma4	H
33	Mi1	Fe2	Mi3	Ma4	EL	73	Ma1	Ma2	Fe3	Fe4	EH
34	Mi1	Fe2	Ma3	Fe4	VH	74	Ma1	Ma2	Fe3	Mi4	H
35	Mi1	Fe2	Ma3	Mi4	M	75	Ma1	Ma2	Fe3	Ma4	L
36	Mi1	Fe2	Ma3	Ma4	VL	76	Ma1	Ma2	Mi3	Fe4	EH
37	Mi1	Mi2	Fe3	Fe4	M	77	Ma1	Ma2	Mi3	Mi4	VH
38	Mi1	Mi2	Fe3	Mi4	VL	78	Ma1	Ma2	Mi3	Ma4	H
39	Mi1	Mi2	Fe3	Ma4	EL	79	Ma1	Ma2	Ma3	Fe4	EH
40	Mi1	Mi2	Mi3	Fe4	VH	80	Ma1	Ma2	Ma3	Mi4	EH
						81	Ma1	Ma2	Ma3	Ma4	VH

The input parameters for FVS1 and FVS2 are: NAMP, NAMSF, NODMP, NAMF and the output linguistic parameter is VS. The term sets of *NAMP*, *NAMSF*, *NODMP* and *NAMF* are defined respectively as:

$$\begin{aligned} NAMP &= \{Few1,\ Middle1,\ Many1\} \\ &= \{Fe1,\ Mi1,\ Ma1\}; \\ NAMSF &= \{Few2,\ Middle2,\ Many2\} \\ &= \{Fe2,\ Mi2,\ Ma2\}; \\ NODMP &= \{Few3,\ Middle3,\ Many3\} \\ &= \{Fe3,\ Mi3,\ Ma3\}; \\ NAMF &= \{Few4,\ Middle4,\ Many4\} \\ &= \{Fe4,\ Mi4,\ Ma4\}. \end{aligned}$$

and the term set for the output *VS* is defined as:

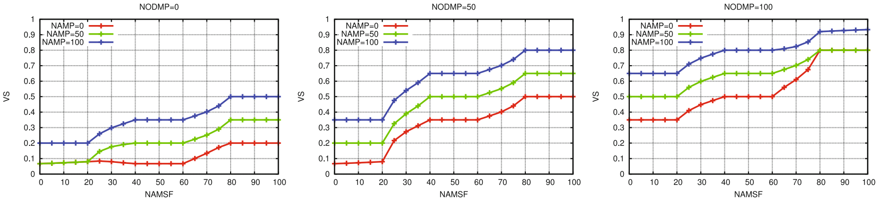
$$VS = \begin{pmatrix} Extremely\ Low \\ Very\ Low \\ Low \\ Middle \\ High \\ Very\ High \\ Extremely\ High \end{pmatrix} = \begin{pmatrix} EL \\ VL \\ L \\ M \\ H \\ VH \\ EH \end{pmatrix}$$

6 Simulation Results

In this section, we present the simulation results for our proposed system. In our system, we decided the number of term sets by carrying out many simulations. These simulation results were carried out in MATLAB.

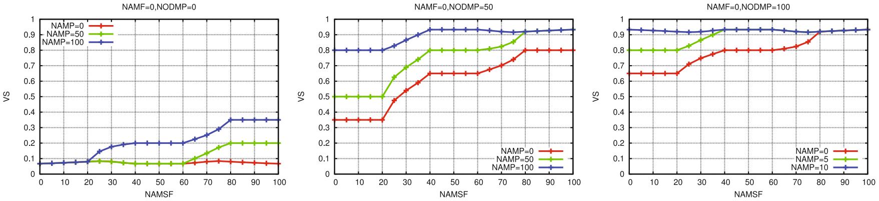
For FVS1, we show the relation of NAMP, NAMF, NODMP and VS in Fig. 4. In this simulation, we consider the NODMP as a constant parameter. From the simulations results, we conclude that with increasing of NAMP, NAMSF, and NODMP, the VS is increasing. Thus, the proposed system can choose peers with a good voting score in P2P mobile collaborative team.

For FVS2, in Fig. 5, we show the relation between NAMP, NAMSF, NODMP, NAMF and VS. When NODMP and NAMF are considered as constant parameters. In Figs. 6 and 7, we increase the NAMF value to 50 and 100 percent, respectively. With the increase of the NAMF, the VS is decreased. When the NAMF is high, the VS values of FVS2 are lower than FVS1. This shows that the NAMF has a great effect on the voting score of proposed system.



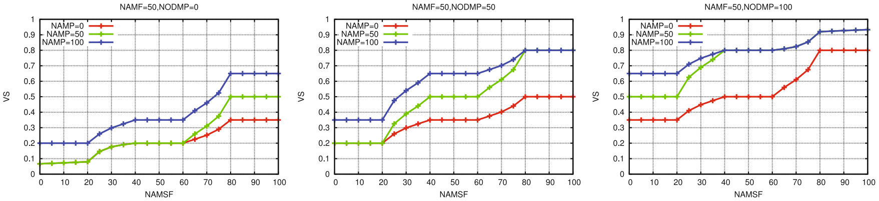
(a) NODMP=0 (b) NODMP=50 (c) NODMP=100

Fig. 4. Voting score for different NODMP (FVS1).



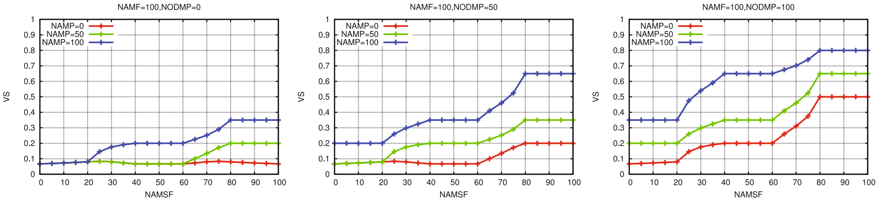
(a) NODMP=0 (b) NODMP=50 (c) NODMP=100

Fig. 5. Voting score for different NODMP when the NAMF=0 (FVS2).



(a) NODMP=0 (b) NODMP=50 (c) NODMP=100

Fig. 6. Voting score for different NODMP when the NAMF=50 (FVS2).



(a) NODMP=0 (b) NODMP=50 (c) NODMP=100

Fig. 7. Voting score for different NODMP when the NAMF=100 (FVS2).

7 Conclusions and Future Work

In this paper, we proposed two fuzzy-based systems to decide the VS. We took into consideration four parameters: NAMP, NAMSF, NODMP and NAMF. We evaluated the performance of proposed systems by computer simulations. From the simulations results, we conclude as follows.

- With increasing of the NAMF parameter, the VS is decreased.
- When NAMP, NAMSF and NODMP are high, the VS is high.
- The proposed system can choose peers with a good voting score in P2P mobile collaborative team.
- Comparing the complexity, the FVS2 is more complex than FVS1. However, FVS2 considers also the NAMF, which makes the voting process better.

In the future, we would like to make extensive simulations to evaluate the proposed systems and compare the performance with other systems.

References

1. Oram, A. (ed.): Peer-to-Peer: Harnessing the Power of Disruptive Technologies. O'Reilly and Associates (2001)
2. Sula, A., Spaho, E., Matsuo, K., Barolli, L., Xhafa, F., Miho, R.: A new system for supporting children with autism spectrum disorder based on IoT and P2P technology. Int. J. Space-Based Situated Comput. **4**(1), 55–64 (2014). doi:10.1504/IJSSC.2014.060688
3. Di Stefano, A., Morana, G., Zito, D.: QoS-aware services composition in P2PGrid environments. Int. J. Grid Util. Comput. **2**(2), 139–147 (2011). doi:10.1504/IJGUC.2011.040601
4. Sawamura, S., Barolli, A., Aikebaier, A., Takizawa, M., Enokido, T.: Design and evaluation of algorithms for obtaining objective trustworthiness on acquaintances in P2P overlay networks. Int. J. Grid Util. Comput. **2**(3), 196–203 (2011). doi:10.1504/IJGUC.2011.042042
5. Higashino, M., Hayakawa, T., Takahashi, K., Kawamura, T., Sugahara, K.: Management of streaming multimedia content using mobile agent technology on pure P2P-based distributed e-learning system. Int. J. Grid Util. Comput. **5**(3), 198–204 (2014). doi:10.1504/IJGUC.2014.062928
6. Terano, T., Asai, K., Sugeno, M.: Fuzzy Systems Theory and Its Applications. Academic Press, INC. Harcourt Brace Jovanovich Publishers (1992)
7. Xhafa, F., Poulovassilis, A.: Requirements for distributed event-based awareness in P2P groupware systems. In: Proceedings of AINA 2010, pp. 220–225, April 2010
8. Xhafa, F., Barolli, L., Caballé, S., Fernandez, R.: Supporting scenario-based online learning with P2P group-based systems. In: Proceedings of NBiS 2010, pp. 173–180, September 2010
9. Gupta, S., Kaiser, G.: P2P video synchronization in a collaborative virtual environment. In: Proceedings of the 4th International Conference on Advances in Web-Based Learning (ICWL05), pp. 86–98 (2005)
10. Martnez-Alemn, A.M., Wartman, K.L.: Online Social Networking on Campus Understanding What Matters in Student Culture. Taylor and Francis, Routledge (2008)

11. Spaho, E., Kulla, E., Xhafa, F., Barolli, L.: P2P Solutions to Efficient Mobile Peer Collaboration in MANETs. In: Proceedings of 3PGCIC 2012, pp. 379–383, November 2012
12. Joaquim, R., Zùquete, A., Ferreira, P.: REVS – a robust electronic voting system. IADIS Int. J. WWW/Internet 1(2) (2003)
13. https://electology.org/score-voting
14. Kandel, A.: Fuzzy Expert Systems. CRC Press, Boca Raton (1992)
15. Zimmermann, H.J.: Fuzzy Set Theory and Its Applications, Revised 2 edn. Kluwer Academic Publishers, Dordrecht (1991)
16. McNeill, F.M., Thro, E.: Fuzzy Logic a Practical Approach. Academic Press Inc., Boston (1994)
17. Zadeh, L.A., Kacprzyk, J.: Fuzzy Logic for the Management of Uncertainty. Wiley, Hoboken (1992)
18. Procyk, T.J., Mamdani, E.H.: A linguistic self-organizing process controller. Automatica **15**(1), 15–30 (1979)
19. Klir, G.J., Folger, T.A.: Fuzzy Sets, Uncertainty, and Information. Prentice Hall, Englewood Cliffs (1988)
20. Munakata, T., Jani, Y.: Fuzzy systems: an overview. Commun. ACM **37**(3), 69–76 (1994)

A User Prediction and Identification System for Tor Networks Using ARIMA Model

Tetsuya Oda[1](✉), Miralda Cuka[3], Ryoichiro Obukata[3], Makoto Ikeda[2], and Leonard Barolli[2]

[1] Department of Information and Computer Engineering, Okayama University of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700–0005, Japan
oda.tetsuya.fit@gmail.com

[2] Department of Information and Communication Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811–0295, Japan
makoto.ikd@acm.org, barolli@fit.ac.jp

[3] Graduate School of Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811–0295, Japan
mcuka91@gmail.com, obukenkyuu@gmail.com

Abstract. Due to the amount of anonymity afforded to users of the Tor infrastructure, Tor has become a useful tool for malicious users. With Tor, the users are able to compromise the non-repudiation principle of computer security. Also, the potentially hackers may launch attacks such as DDoS or identity theft behind Tor. For this reason, there are needed new systems and models to detect the intrusion in Tor networks. In this paper, we present the application of Autoregression Integrated Moving Average (ARIMA) for prediction of user behavior in Tor networks. We constructed a Tor server and a Deep Web browser (Tor client) in our laboratory. Then, the client sends the data browsing to the Tor server using the Tor network. We used Wireshark Network Analyzer to get the data and then used the ARIMA model to make the prediction. The simulation results show that proposed system has a good prediction of user behavior in Tor networks.

1 Introduction

The Onion Router (Tor) [1,2] is an implementation of an Onion Routing network, where users expect a large degree of privacy. This privacy is reflected in the perfect forward secrecy exhibited by Tor connections such that traffic captured at any single network location during transit only uncovers the previous and next waypoints [3]. Due to the amount of anonymity afforded to users of the Tor infrastructure, Tor has become a useful tool for malicious users. With Tor, the users are able to compromise the non-repudiation principle of computer security. Also, the potentially hackers may launch attacks such as DDoS or identity theft behind Tor.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_9

The Tor has been designed to make it possible for users to surf the Internet anonymously, so their activities and location cannot be discovered by government agencies, corporations, or anyone else. Compared with other anonymizers Tor is more popular and has more visibility in the academic and hacker communities. Tor is a low-latency, circuit-based and privacy-preserving anonymizing platform and network. It is one of several systems that have been developed to provide Internet users with a high level of privacy and anonymity in order to cope with the censorship measures taken by authorities and to protect against the constantly increasing threats to these two key security properties.

There are two main approaches to the design of Intrusion Detection Systems (IDSs). In a misuse detection based IDS, intrusions are detected by looking for activities that correspond to known signatures of intrusion or vulnerabilities. On the other hand, anomaly detection based IDS detects intrusions by searching for abnormal network traffic. The abnormal traffic pattern can be defined either as the violation of accepted thresholds for the legitimate profile developed for the normal behavior.

In [4], the authors designed and implemented TorWard, which integrates an Intrusion Detection System (IDS) at Tor exit routers for Tor malicious traffic discovery and classification. The system can avoid legal and administrative complaints and allows the investigation to be performed in a sensitive environment such as a university campus. An IDS is used to discover and classify malicious traffic. The authors performed comprehensive analysis and extensive real-world experiments to validate the feasibility and effectiveness of TorWard.

One of the most commonly used approaches in expert system based on intrusion detection is a rule-based analysis using soft computing techniques such Fuzzy Logic (FL), Artificial Neural Networks (ANNs), Probabilistic Reasoning (PR), and Genetic Algorithms (GAs). They are good approaches capable of finding patterns for abnormal and normal behavior. In some studies, the NNs have been implemented with the capability to detect normal and attack connections [5].

In [6], a specific combination of two NN learning algorithms, the Error Backpropagation and the Levenberg-Marquardt algorithm, is used to train an artificial NN to model the boundaries of the clusters of recorded normal behavior. It is shown that the training dataset, consisting of a combination of recorded normal instances and artificially generated intrusion instances, successfully guides the NN towards learning the complex and irregular cluster boundary in a multidimensional space. The performance of the system is tested on unseen network data containing various intrusion attacks [6].

In [7] is presented a NN-based intrusion detection method for the internet-based attacks on a computer network. The IDSs have been created to predict and thwart current and future attacks. The NNs are used to identify and predict unusual activities in the system. In particular, feed-forward NNs with the Back-propagation training algorithm were employed and the training and testing data were obtained from the Defense Advanced Research Projects Agency (DARPA) intrusion detection evaluation data sets. The experimental results on real-data showed promising results on detection intrusion systems using NNs.

In [8], the authors deal with packet behavior as parameters in anomaly intrusion detection. The proposed IDS uses a Back-propagation Artificial Neural Network (ANN) to learn system's behavior. The authors used the KDD'99 data set for experiments and the obtained satisfying results.

In [9] is presented a deep learning approach for network intrusion detection system. The proposed system use self-taught learning, a deep learning technique based on sparse auto-encoder and soft-max regression, to develop an NIDS. The experimental results on the test data showed promising results on detection intrusion systems using NIDS.

In this paper, we present the application of Autoregression Integrated Moving Average (ARIMA) model for user behavior in Tor networks. We used the ARIMA and constructed a Tor server and a Deep Web browser (client) in our laboratory. Then, the client sends the data browsing to the Tor server using the Tor network. We used Wireshark Network Analyzer to get the data and then use the ARIMA to make the prediction. For evaluation we considered Number of Packets (NoP) metric. We present some simulation results considering Tor client.

The structure of the paper is as follows. In Sect. 2, we present a short description of Deep Web and Tor. In Sect. 3, we give an overview of ARIMA. In Sect. 4, we present an overview of R. In Sect. 5, we present the proposed model. In Sect. 6, we discuss the simulation results. Finally, conclusions and future work are given in Sect. 7.

2 Deep Web and Tor Overview

2.1 Deep Web

The Deep Web (also called the Deepnet, Invisible Web or Hidden Web) is the portion of World Wide Web content that is not indexed by standard search engines [10,11]. Most of the Web's information is far from the search sites and standard search engines do not find it. Traditional search engines cannot see or retrieve content in the Deep Web. The portion of the Web that is indexed by standard search engines is known as the Surface Web. Now, the Deep Web is several orders of magnitude larger than the Surface Web. The most famous of the deep web browsers is called Tor.

The Deep Web is both surprising and sinister and accounts for in excess of 90% of the overall Internet [12]. The Google and other search engines deal only with the indexed surface web. The deep-dark web hosts illegal markets, such as the Silk Road, malware emporiums, illegal pornography, and covert meeting places and messaging services. The pervasiveness of the Internet provides easy access to darkweb sites from anywhere in the world. The growth of the dark web has been paralleled by an increasing number of anonymity web-overlay services, such as Tor, which allow criminals, terrorists, hackers, paedophiles and the like to shop and communicate with impunity. Law enforcement and security agencies have had only very limited success in combating and containing this dark menace.

2.2 Tor

Tor is a low-latency, circuit-based and privacy-preserving anonymizing platform and network. It is one of several systems that have been developed to provide Internet users with a high level of privacy and anonymity in order to cope with the censorship measures taken by authorities and to protect against the constantly increasing threats to these two key security properties [13–16].

The Tor main design goals are to prevent attackers from linking communication partners, or from linking multiple communications to or from a single user. Tor relies on a distributed overlay network and onion routing to anonymize TCP-based applications like web browsing, secure shell, or peer-to-peer communications.

The Tor network is composed of the Tor client, an entry/guard node, several relays and the exit node. The Tor client is a software, installed on each Tor user's device. It enables user to create a Tor anonymizing circuit and to handle all the cryptographic keys, needed to communicate with all nodes within the circuit. The Entry Node is the first node in the circuit that receives the client request and forwards it to the second relay in the network. The Exit Node is the last Tor-relay in the circuit. Once the connection request leaves the entry node, it will be forwarded, through relays in the circuit, all the way to the exit node. The latter receives the request and relays it to the final destination.

When a client wants to communicate with a server via Tor, he selects n nodes of the Tor system (where n is typically 3) and builds a circuit using those selected nodes. Messages are then encrypted n times using the following onion encryption scheme. The messages are first encrypted with the key shared with the last node (called the exit node of the circuit) and subsequently with the shared keys of the intermediate node. As a result of this onion routing, each intermediate node only knows its predecessor and successor, but no other nodes of the circuit. In addition, the onion encryption ensures that only the last node is able to recover the original message.

A Tor client typically uses multiple simultaneous circuits. As a result, all streams of a user are multiplexed over these circuits. For example, a BitTorrent user can use one of the circuits for his connections to the tracker and other circuits for his connections to the peers.

3 ARIMA

The basic models of time series proposed by Box and Jenkins include Auto-regression Model, Moving Average Model, Auto-regression Moving Average Model and Autoregression Integrated Moving Average Model. The Autoregressive Moving Average Model (ARMA) is a relatively mature model which contains Autoregressive (AR) Model, Moving Average (MA) Model and ARMA Model. ARMA (p, q) Model is expressed as follows:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} \cdots - \theta_\eta \epsilon_{t-\eta} \qquad (1)$$

where y_t is forecasting value of time, $t, \phi_i, \theta_j = (i = 1, 2, \cdots, p; j = 1, 2, \cdots, q)$ are model parameters, ϵ_t is the random process of white noise whose mean value is zero, p and q are orders of the model [17]. The ARIMA model is the extension of ARMA model. The ARMA model is usually used to dispose stable time series. If the series are non-stationary, it can be transfonned into a stationary time series using the d-th difference process, d is usually zero, one or two. Then, the series after difference is modeled by ARMA. The whole above process is called ARIMA. The prediction process of ARIMA model is as follows.

- Stationary Identification of Sequence: Check series' stationary with test methods of ADF root of unity.
- Series' Stationary Processing: If data series are unstable, we need to conduct difference processing until the data after disposed meet stationary condition. The order of difference is d when time series are stable.
- The Estimation of Parameters: We need check whether it has statistical significance.
- Conduct Hypothesis Testing: We need to judge whether residual sequence of the model is white noise.
- Conduct Forecasting Analysis: The forecasting analysis are conducted by using models that has been checked to be qualified.

4 The R Environment

The R is an integrated suite of software facilities for data manipulation, calculation and graphical display [18]. Among other things it has the following features.

- An effective data handling and storage facility.
- A suite of operators for calculations on arrays, in particular matrices.
- A large, coherent, integrated collection of intermediate tools for data analysis.
- Graphical facilities for data analysis and display either directly at the computer or on hardcopy.
- A well developed, simple and effective programming language (called "S") which includes conditionals, loops, user defined recursive functions and input and output facilities. Indeed most of the system supplied functions are themselves written in the S language.

The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software. R is very much a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages. However, most programs written in R are essentially ephemeral, written for a single piece of data analysis.

Many people use R as a statistics system. The R is an environment within which many classical and modern statistical techniques have been implemented. A few of these are built into the base R environment, but many are supplied as packages.

5 Proposed Intrusion Detection Model for Tor Networks

The proposed system model is shown in Fig. 1. We call this system: User Behavior Prediction System using ARIMA (UBPS-ARIMA). We used UBPS-ARIMA and constructed a Tor server and a Deep Web browser (Tor client) in our laboratory. Then, the client sends the data browsing to the Tor server using the Tor network. We used Wireshark Network Analyzer [19] to get the data. The data are stored in the log files. The system runs until the number of loops is achieved.

The data in the log files are considered as old data and the current data are the input of ARIMA model. The UBPS-ARIMA can predict the user behavior using these data.

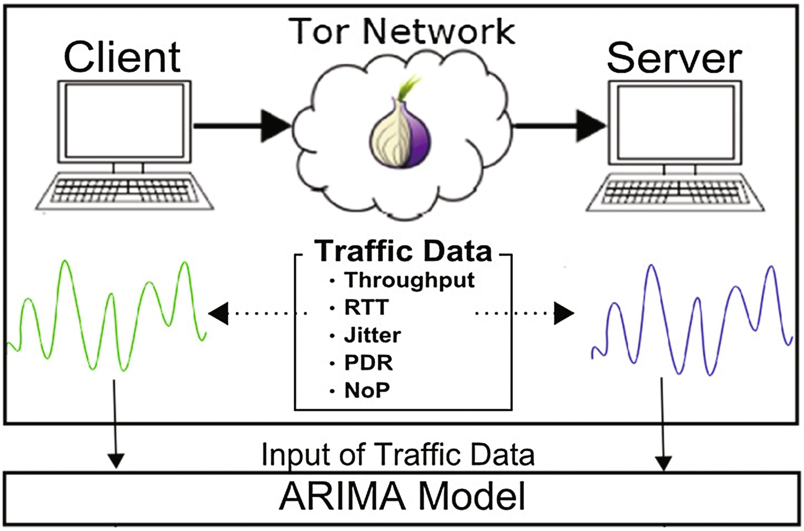


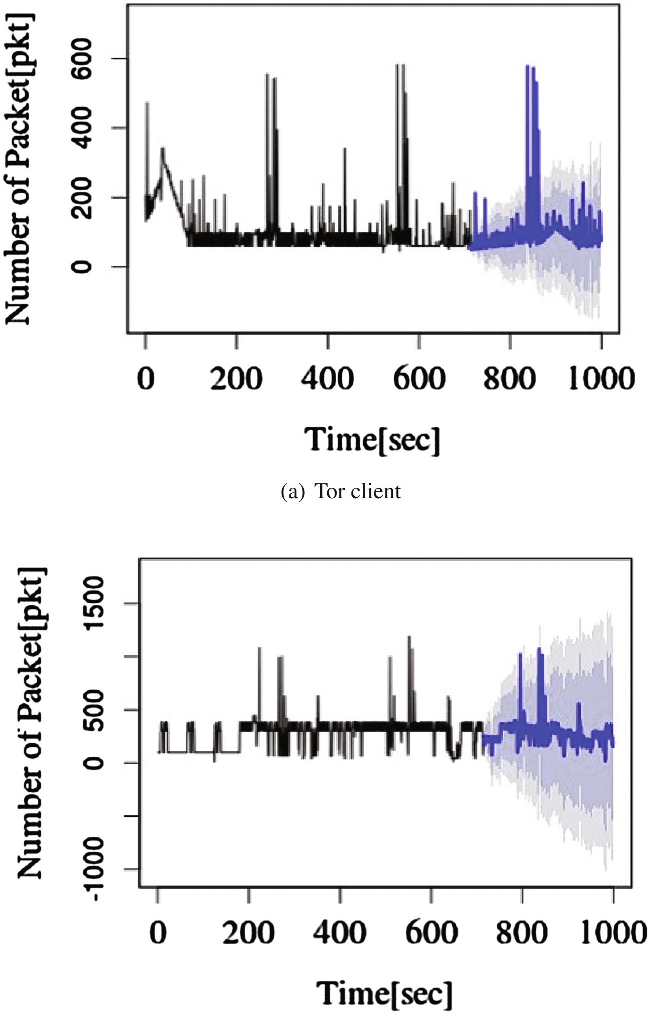
Fig. 1. Proposed system model.

6 Simulation Results

We carried out some simulations using the UBPS-ARIMA. In Fig. 2, we show the Number of Packet (NoP) parameter for Tor client and Tor server. The simulation parameters are shown in Table 1. We ran the simulation 10000 times. In Fig. 2(a) are shown the data for Tor client and in Fig. 2(b) for Tor server. As evaluation parameter, we used the NoP. From 0 [sec] to 700 [sec] are shown the training data. Then, from 701 [sec] to 1000 [sec] are shown the predicted data. The simulation results show that UBPS-ARIMA model has a good prediction.

Table 1. Simulation parameters.

Parameters	Values
Number of training data	1000
Number of measurements	400
Predictive model	ARIMA model



(a) Tor client

(b) Tor server

Fig. 2. Simulation results.

7 Conclusions

The Tor network has become a useful tool for malicious users. With Tor, the users are able to compromise the non-repudiation principle of computer security. Also, the potentially hackers may launch attacks such as DDoS or identity theft behind Tor. For this reason, there are needed new systems and models to detect the intrusion in Tor networks.

In this paper, we presented the application of ARIMA for prediction of user behavior in Tor networks. We used ARIMA model and constructed a Tor server and a Deep Web browser. Then, the client sent the data browsing to the Tor server using the Tor network. We used Wireshark Network Analyzer to get the data and then used the ARIMA to make prediction of user behavior. The simulation results show that UBPS-ARIMA has a good prediction of user behavior.

References

1. Tor Project Web Site. http://www.torproject.org/
2. Dingledine, R., Mathewson, N., Syverson, P.: Deploying low-latency anonymity: design challenges and social factors. IEEE Secur. Priv. **5**(5), 83–87 (2007)
3. Dingledine, R., Mathewson, N., Syverson, P.: Tor: the second-generation Onion Router. In: Proceedings of the 13th Conference on USENIX Security Symposium (SSYM-2004), vol. 13, p. 21 (2004)
4. Ling, Z., Luo, J., Wu, K., Yu, W., Fu, X.: TorWard: discovery of malicious traffic over Tor. In: Proceedings of IEEE INFOCOM 2014, pp. 1402–1410, April 2014
5. Reddy, E.K.: Neural networks for intrusion detection and its applications. In: Proceedings of the World Congress on Engineering 2013 Vol. II, WCE-2013, July 2013
6. Linda, O., Vollmer, T., Manic, M.: Neural network based intrusion detection system for critical infrastructures. In: Proceedings of International Joint Conference on Neural Networks (IJCNN-2009), pp. 1827–1834, June 2009
7. Shum, J., Malki, H.A.: Network intrusion detection system using neural networks. In: Proceedings of Fourth International Conference on Natural Computation (ICNC-2008), pp. 242–246, October 2008
8. Al-Janabi, S.T.F., Saeed, H.A.: A neural network based anomaly intrusion detection system. In: Developments in E-systems Engineering (DeSE), pp. 221–226, December 2011
9. Niyaz, Q., Sun, W., Javaid, A.Y., Alam, M.: A deep learning approach for network intrusion detection system. In: Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (Formerly BIONETICS), BICT-15, vol. 15, pp. 21–26 (2015)
10. Lang Hong, J.: Deep web data extraction. In: Proceedings of IEEE International Conference on Systems Man and Cybernetics (SMC-2010), pp. 3420–3427, October 2010
11. Singh, M.P.: Deep web structure. IEEE Internet Comput. **6**(5), 4–5 (2002)
12. Stupples, D.: Security challenge of Tor and the deep web. In: 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), December 2013
13. Biryukov, A.: Trawling for Tor hidden services: detection, measurement, deanonymization. In: Proceedings of IEEE Symposium on Security and Privacy (SP-2013), pp. 80–94, November 2013

14. Dhungel, P., Steiner, M., Rimac, I., Hilt, V., Ross, K.W.: Waiting for anonymity: understanding delays in the Tor overlay. In: Proceedings of IEEE Tenth International Conference on Peer-to-Peer Computing (P2P-2010), pp. 1–4, August 2010
15. Xin, L., Neng, W.: Design improvement for Tor against low-cost traffic attack and low-resource routing attack. In: Proceedings of WRI International Conference on Communications and Mobile Computing (CMC-2009), pp. 549–554, January 2009
16. Syverson, P.: A peel of onion. In: Proceedings of ACSAC-2011, pp. 123–135, December 2011
17. Min, Y., Bin, W., Liang-Ii, Z., Xi, C.: Wind speed forecasting based on EEMD and ARIMA. In: Chinese Automation Congress (CAC-2015), pp. 1299–1302 (2015)
18. The R Project for Statistical Computing. http://www.r-project.org/
19. WireShark Web Site. http://www.wireshark.org/

Implementation of an Actor Node for an Ambient Intelligence Testbed: Evaluation and Effects of Actor Node on Human Sleeping Condition

Ryoichiro Obukata[1], Miralda Cuka[1], Donald Elmazi[1], Tetsuya Oda[2](✉),
Keita Matsuo[3], and Leonard Barolli[3]

[1] Graduate School of Engineering, Fukuoka Institute of Technology (FIT),
3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811–0295, Japan
obukenkyuu@gmail.com, mcuka91@gmail.com, donald.elmazi@gmail.com

[2] Department of Information and Computer Engineering, Okayama University
of Science (OUS), 1-1 Ridaicho, Kita-ku, Okayama 700–0005, Japan
oda.tetsuya.fit@gmail.com

[3] Department of Information and Communication Engineering, Fukuoka Institute
of Technology (FIT), 3-30-1 Wajiro-Higashi, Higashi-Ku, Fukuoka 811–0295, Japan
{kt-matsuo,barolli}@fit.ac.jp

Abstract. Ambient intelligence (AmI) deals with a new world of ubiquitous computing devices, where physical environments interact intelligently and unobtrusively with people. AmI environments can be diverse, such as homes, offices, meeting rooms, schools, hospitals, control centers, vehicles, tourist attractions, stores, sports facilities, and music devices. In this paper, we present the implementation and evaluation of actor node an AmI testbed using Raspberry Pi mounted on Raspbian OS. For evaluation, we considered respiratory rate and heart rate metrics. We carried out an experiments and clustered sensed data by k-means clustering algorithm. From experimental results, we found that the implemented AmI testbed gives a good effect to human during sleeping.

1 Introduction

Ambient Intelligence (AmI) is the vision that technology will become invisible, embedded in our natural surroundings, present whenever we need it, enabled by simple and effortless interactions, attuned to all our senses, adaptive to users and context and autonomously acting [1]. High quality information and content must be available to any user, anywhere, at any time, and on any device.

In order that AmI becomes a reality, it should completely envelope humans, without constraining them. Distributed embedded systems for AmI are going to change the way we design embedded systems, in general, as well as the way we think about such systems. But, more importantly, they will have a great

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_10

impact on the way we live. Applications ranging from safe driving systems, smart buildings and home security, smart fabrics or e-textiles, to manufacturing systems and rescue and recovery operations in hostile environments, are poised to become part of society and human lives.

There are a lot of works done on testbed for AmI. In [2], the authors present a simulation environment that offers a library of Networked Control Systems (NCS) blocks. Thus, the constraints can be considered and integrated in the design process. They describe a real process, an inverted pendulum, which is automated based on Mica nodes. These nodes were designed especially for AmI purposes. This real NCS serves as a challenging benchmark for proving the AmI suitability of the controllers.

In [3], the authors present the development of an adaptive embedded agent, based on a hybrid PCA-NFS scheme, able to perform true real-time control of AmI environments in the long term. The proposed architecture is a single-chip HW/SW architecture. It consists of a soft processor core (SW partition), a set of NFS cores (HW partition), the HW/SW interface, and input/output (I/O) peripherals. An application example based on data obtained in an ubiquitous computing environment has been successfully implemented using an FPGA of Xilinx's Virtex 5 family [4].

In [5], the authors describe a framework to Context Acquisition Services and Reasoning Algorithms (CASanDRA) to be directly consumed by any type of application needing to handle context information. CASanDRA decouples the acquisition and inference tasks from the application development by offering a set of interfaces for information retrieval. The framework design is based on a data fusion-oriented architecture. CASanDRA has been designed to be easily scalable; it simplifies the integration of both new sensor access interfaces and fusion algorithms deployment, as it also aims at serving as a testbed for research.

In this work, we implement an actor node for an AmI testbed. We investigate the effects of the actor node on human sleeping condition by using the k-means clustering algorithm. As evaluation metrics, we considered respiratory rate and heart rate.

The structure of the paper is as follows. In Sect. 2, we present a short description of AmI. In Sect. 3, we give a brief introduction of k-means clustering algorithm. In Sect. 4, we show the description and design of the testbed. In Sect. 5, we discuss the experimental results. Finally, conclusions and future work are given in Sect. 6.

2 Ambient Intelligence (AmI)

In the future, small devices will monitor the health status in a continuous manner, diagnose any possible health conditions, have conversation with people to persuade them to change the lifestyle for maintaining better health, and communicates with the doctor, if needed [6]. The device might even be embedded into the regular clothing fibers in the form of very tiny sensors and it might communicate with other devices including the variety of sensors embedded into the home to monitor the lifestyle. For example, people might be alarmed about the lack of a healthy diet based on the items present in the fridge and based on what they are eating outside regularly.

The AmI paradigm represents the future vision of intelligent computing where environments support the people inhabiting them [7–9]. In this new computing paradigm, the conventional input and output media no longer exist, rather the sensors and processors will be integrated into everyday objects, working together in harmony in order to support the inhabitants [10]. By relying on various artificial intelligence techniques, AmI promises the successful interpretation of the wealth of contextual information obtained from such embedded sensors, and will adapt the environment to the user needs in a transparent and anticipatory manner.

3 The *k*-means Algorithm

Here, we briefly describes the standard k-means algorithm [11]. The k-means is a typical clustering algorithm in data mining and which is widely used for clustering large set of data. The k-means algorithm is one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-known cluster [12]. It is a partitioning clustering algorithm, this method is to classify the given date objects into k different clusters through the iterative, converging to a local minimum. So the results of generated clusters are compact and independent. The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance. The next phase is to take each data object to the nearest center [13]. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum. Supposing that the target object is x, x_i indicates the average of cluster C_i, criterion function is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - x_i|^2 , \tag{1}$$

where E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data object and cluster center. The Euclidean distance between one vector $x = (x_1, x_2, \ldots, x_n)$ and another vector $y = (y_1, y_2, \ldots, y_n)$. The Euclidean distance $d(x_i, y_i)$ can be obtained as follow:

$$d(x_i, y_i) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}} . \tag{2}$$

The process of k-means algorithm in Algorithm 1. The k-means clustering algorithm always converges to local minimum. Before the k-means algorithm

Algorithm 1. The process of k-means algorithm.

```
Input: Number of desired clusters, $k$, and a database $D = d_1, d_2, \ldots, d_n$ containing $n$ data objects;
Output: A set of $k$ clusters;
Randomly select $k$ data objects from dataset $D$ as initial cluster centers;
Calculate the distance between each data object $d_i$ $(1 \leq i \leq n)$ and all $k$ cluster centers $c_j$ $(1 \leq j \leq k)$ and assign data object $d_i$ to the nearest cluster;
For each cluster $j$ $(1 \leq j \leq k)$, recalculate the cluster center;
Until no changing in the center of clusters;
```

converges, calculations of distance and cluster centers are done while loops are executed a number of times, where the positive integer t is known as the number of k-means iterations. The precise value of t varies depending on the initial starting cluster centers [14]. The distribution of data points has a relationship with the new clustering center, so the computational time complexity of the k-means algorithm is $O(nkt)$. The n is the number of all data objects, k is the number of clusters, t is the iterations of algorithm. Usually requiring $k \ll n$ and $t \ll n$.

4 Testbed Description

In Fig. 1 is shown the structure of AmI testbed. Our testbed is composed of five Raspberry Pi 3 Model B [15–18]. The Raspberry Pi is a credit card-sized single-board computer developed by the Raspberry Pi Foundation [19]. The operating systems mounted on these machines are Raspbian version Debian 7.8 with kernel 3.18.11 [20].

We use Microwave Sensor Module (MSM) called DC6M4JN3000, which emits microwaves in the direction of a human or animal subject [21]. These microwaves

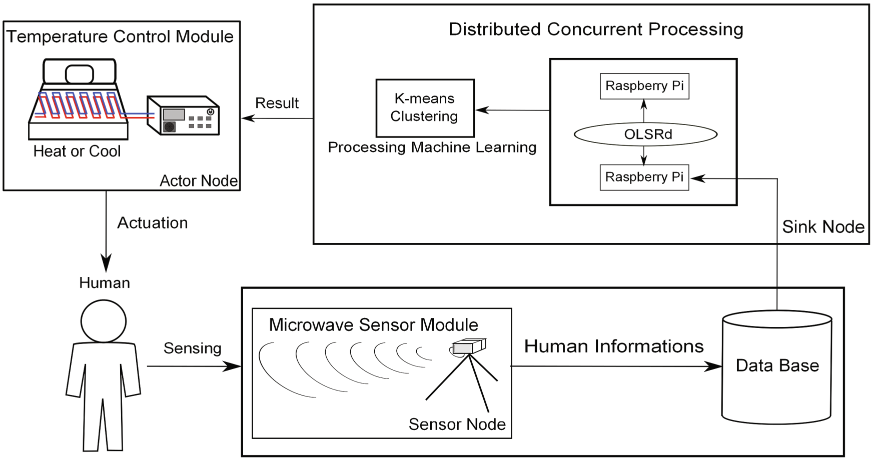


Fig. 1. Structure of AmI testbed.

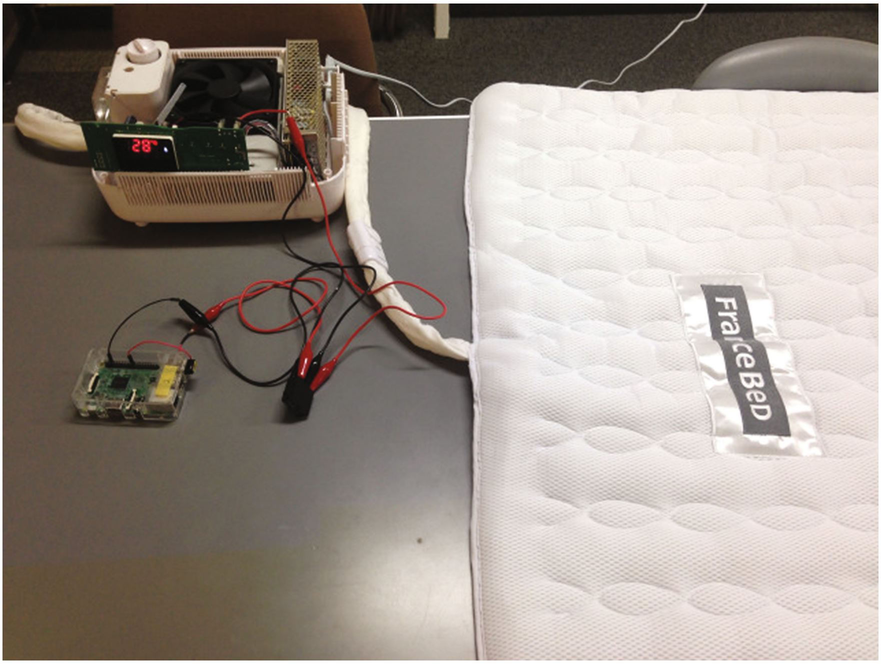


Fig. 2. Snapshot of AmI testbed actor node.

reflect back off the surface of the subject and change slightly in accordance with movements of the subject's heart and lungs. From these changes, the DC6M4JN3000 measures biological information such as heart and respiratory rates.

The DC6M4JN3000 is capable of measuring heart rate within a margin of error of ±10 [%] when placed roughly three meters away from the target subject. The unit uses microwaves, so it can detect targets located behind obstacles such as mattresses, doors, and walls. This makes it possible to measure biological information even when the target is asleep or in situations where the targets privacy must be maintained (such as in the washroom or bathroom), thereby enabling this sensor module to boost the level of service given in elderly care or nursing care.

For actor node, we use Reidan Shiki PAD (see Fig. 2), which can be used for cooling and heating the bed [22]. The Reidan Shiki PAD can adjust the temperature between 15 and 48 [°C].

As shown in Fig. 1, by using the MSM the system get the respiratory rate and heart rate data. Which are sent to sink node equipped with five Raspberry Pi 3 Model B computers [23,24]. In the sink is carried out the distributed concurrent processing for k-means algorithm. Then, the results is sent to temperature control module to control the Reidan Shiki PAD temperature.

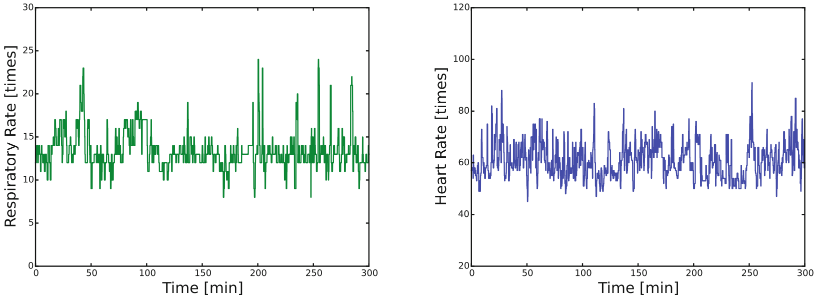
5 Experimental Results

We carried out experiments with a student of our laboratory in the cases when we use or do not use the implemented testbed. The experimental parameters are shown in Table 1. We collected data for respiratory rate and heart rate.

Table 1. Simulation parameters.

Parameters	Values
Number of clusters	3
Initial centroids	Random
Precompute distance	True

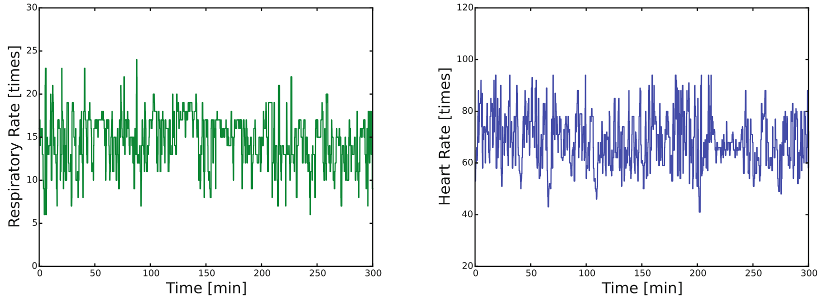
In Figs. 3 and 4, we show the respiratory rate and heart rate in two cases: not using AmI testbed and using AmI testbed, respectively. In Fig. 5, we present the result of clustered data using k-means algorithm for these two cases. We can see 3 regions of clustering: Rapid Eye Movement (REM) sleep, light non-REM sleep, deep non-REM sleep. The blue cluster shows deep non-REM sleep, the red cluster shows light non-REM sleep and the green cluster shows REM sleep. The "+" mark shows the center of gravity of each cluster. Comparing Fig. 5(a) and (b), we can see that the actor node gives good effect to human during sleeping, because in the case of using AmI tesbed there are more blue cluster points compared with case of not using AmI tesbed. This shows that using AmI testbed the human has a better sleeping condition.



(a) Respiratory rate

(b) Heart rate

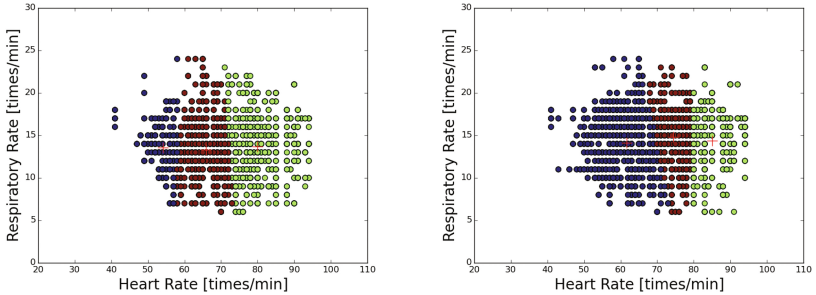
Fig. 3. Sensing data during sleeping (not using AmI testbed).



(a) Respiratory rate

(b) Heart rate

Fig. 4. Sensing data during sleeping (using AmI testbed).



(a) Clustering of not using AmI testbed.

(b) Clustering of using AmI testbed.

Fig. 5. Experimental results using k-`means` clustering algorithm.

6 Conclusions

In this paper, we presented the implementation and evaluation of an actor node for AmI testbed. We carried out an experiments and clustered sensed data by k-means clustering algorithm. From experimental results, we found that the implemented AmI testbed gives a good effect to human during sleeping.

In the future, we would like to make extensive experiments using the implemented AmI testbed.

References

1. Lindwer, M., Marculescu, D., Basten, T., Zimmermann, R., Marculescu, R., Jung, S., Cantatore, E.: Ambient intelligence visions and achievements: linking abstract ideas to real-world concepts. In: Design, Automation and Test in Europe Conference and Exhibition, pp. 10–15 (2003)
2. Gabel, O., Litz, L., Reif, M.: NCS testbed for ambient intelligence. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 1, pp. 115–120 (2005)

3. del Campo, I., Martinez, M.V., Echanobe, J., Basterretxea, K.: A hardware/software embedded agent for realtime control of ambient-intelligence environments. In: IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2012)
4. Virtex 5 Family Overview, Xilinx Inc., San Jose, CA (2009)
5. Bernardos, A.M., Tarrio, P., Casar, J.R.: CASanDRA: a framework to provide context acquisition services and reasoning algorithms for ambient intelligence applications. In: International Conference on Parallel and Distributed Computing, Applications and Technologies, pp. 372–377 (2009)
6. Acampora, G., Cook, D., Rashidi, P., Vasilakos, A.V.: A survey on ambient intelligence in health care. Proc. IEEE **101**(12), 2470–2494 (2013)
7. Aarts, E., Wichert, R.: Ambient intelligence. In: Bullinger, H.-S. (ed.) Technology Guide, pp. 244–249. Springer, Heidelberg (2009)
8. Aarts, E., de Ruyter, B.: New research perspectives on ambient intelligence. J. Ambient Intell. Smart Environ. **1**(1), 5–14 (2009)
9. Vasilakos, A., Pedrycz, W.: Ambient Intelligence, Wireless Networking, and Ubiquitous Computing. Artech House Inc., Norwood (2006)
10. Sadri, F.: Ambient intelligence: a survey. ACM Comput. Surv. **43**(4), 36:1–36:66 (2011)
11. Na, S., Xumin, L., Yong, G.: Research on k-means clustering algorithm: an improved k-means clustering algorithm. In: Third International Symposium on Intelligent Information Technology and Security Informatics (IITSI), pp. 63–67 (2010)
12. Jigui, S., Jie, L., Lianyu, Z.: Clustering algorithms research. J. Softw. **19**(1), 48–61 (2008)
13. Fahim, A.M., Salem, A.M., Torkey, F.A.: An efficient enhanced k-means clustering algorithm. J. Zhejiang Univ. Sci. A **10**, 1626–1633 (2006)
14. Abdul Nazeer, K.A., Sebastian, M.P.: Improving the accuracy and efficiency of the k-means clustering algorithm. In: Proceedings of the World Congress on Engineering, vol. 1 (2009)
15. Oda, T., Barolli, A., Sakamoto, S., Barolli, L., Ikeda, M., Uchida, K.: Implementation and experimental results of a WMN testbed in indoor environment considering LoS scenario. In: The 29-th IEEE International Conference on Advanced Information Networking and Applications (AINA-2015), pp. 37–42 (2015)
16. Oda, T., Barolli, L.: Experimental results of a Raspberry Pi based WMN testbed considering CPU frequency. In: The 30-th IEEE International Conference on Advanced Information Networking and Applications (IEEE AINA-2016), pp. 981–986 (2016)
17. Oda, T., Elmazi, D., Yamada, M., Obukata, R., Barolli, L., Takizawa, M.: Experimental results of a Raspberry Pi based WMN testbed in indoor environment: a comparison study of LoS and NLoS scenarios. In: The 19-th International Conference on Network-Based Information Systems (NBiS-2016), pp. 9–14 (2016)
18. Obukata, R., Oda, T., Barolli, L.: Design of an ambient intelligence Testbed for improving quality of life. In: The 9-th International Symposium on Mining and Web (MAW-2016), pp. 714–719 (2016)
19. Raspberry Pi Foundation. http://www.raspberrypi.org/
20. Raspbian: FrontPage. https://www.raspbian.org/
21. Sharp to Release Microwave Sensor Module. http://www.sharp-world.com/corporate/news/150625.html/
22. Corporate History - FRANCEBED. http://www.francebed.co.jp/eng/about/history/index.html

23. Oda, T., Elmazi, D., Ishitaki, T., Barolli, A., Matsuo, K., Barolli, K.: Experimental results of a Raspberry Pi based WMN testbed for multiple flows and distributed concurrent processing. In: The 10-th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA-2015), pp. 201–206 (2015)
24. Obukata, R., Oda, T., Elmazi, D., Ikeda, M., Matsuo, K., Barolli, L.: Performance evaluation of an AmI testbed for improving QoL: evaluation using clustering approach considering parallel processing. In: Proceedings of the 11-th International Conference on Broad-Band Wireless Computing, Communication and Applications (BWCCA-2016). Lecture Notes on Data Engineering and Communications Technologies, vol. 2, pp. 623–630 (2016)

Designing the Light Weight Rotation Boolean Permutation on Internet of Things

Yu Zhou(✉)

Science and Technology on Communication Security Laboratory,
Chengdu 610041, China
zhouyu.zhy@tom.com

Abstract. Encryption algorithms in Internet of Things are a piece of small area with small-scale, it need some light weight encryption algorithms. This paper focuses on the component of encryption algorithms, some light weight of the rotation boolean permutations are perfectly characterized by the matrix of linear expressions. Three methods of rotation nonlinear boolean permutations are constructed. The subfunctions of the three permutations have three monomials, hight degree, 2-algebra immunity. All three classes of rotation nonlinear boolean permutations are fully determination by the first component Boolean function, respectively.

1 Introduction

Internet of Things is an up-and-coming information and technology industry, the project of Internet of Things which is a piece of small area with small-scale and self-system obtain gratifying achievement and bright future. But there are some serious hidden danger and potential crisis problems [1,2]. For example, security issues. Wireless sensor network's characteristics present new challenges in information security area. Along with one-time, unattended, wireless communications, low-cost and resource-constrained, sensors easily appear to abnormalities, physical attacks by attackers, Trojan attacks, virus damage, keys decryption, DOS, eavesdropping and traffic analysis are really threats. The trouble is a challenge that design of key storage, distribution, encryption and decryption mechanism caused by wireless sensor network's large and resource constraints.

In order to design encryption algorithms using in resource constraints, we need design some basic components of encryption algorithms. In collaborative networks, there are some symmetric algorithms which ensure the security of many data. Stream cipher is an important class in symmetric cryptosystem. It is because a good Stream cipher is faster in implementation, and it can produce sequences with large period and good statistical properties. Thus, in order to design a good Stream cipher, one should design some good components, for example Linear feedback shift registers (LFSR), S-box, Maximum Distance Separable (MDS) and so on.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_11

In this paper, we study rotation-invariant n-bit invertible (bijective) functions, this component was firstly introduced in Daemen's 1995 PHD Thesis [5]. The defining property of shift-invariant transformations is the commutativity with translation. Shift-invariant transformations on binary vectors have a number of properties that make them suitable components for the state updating transformation of cryptographic finite state machines.

For hardware, these transformations can be implemented as an interconnected array of identical 1-bit output processors. The shift-invariance ensures that the computational load is optimally distributed.

For software, their regularity allows efficient implementations by employing bitwise logical operations. Moreover, binary shift-invariant transformations can be specified by a single Boolean function.

In 2006, SMS4 [12] was used for WAPI (Wireless LAN Authentication and Privacy Infrastructure) in China, this block cipher used a binary shift-invariant transformation: $C(x) = x \oplus (x <<< 2) \oplus (x <<< 10) \oplus (x <<< 18) \oplus (x <<< 24)$, where $x \in \mathbb{F}_2^{32}$, it had good cryptographic properties, for example the differential branch number and the linear branch number of this transformation was five, this is one of the best transformations of linear functions. And in 2015, Markku Juhani O. Saarinen [11] submit to the CBEAMr1 authenticated encryption algorithm for the first round CAESAR Competition. CBEAMr1 uses a slightly different notation from Daemen who used ϕ to denote non-invertible as well as invertible rotation-invariant functions.

Note that the permutation $(\underbrace{f, \cdots, f}_{n})$ fall into the categories linear (with respect to bitwise addition) and nonlinear. In the nonlinear case, [5] obtained a distinction is made between transformations with finite and those with infinite neighborhood. And dedicated to the study of the propagation and correlation properties of binary shift-invariant permutations with finite neighborhood. [11] used the rotation boolean function $(f(x_0, x_1, x_2, x_3, x_4) = x_0x_1x_3x_4 \oplus x_0x_2x_3 \oplus x_0x_1x_4 \oplus x_1x_2x_3 \oplus x_2x_3x_4 \oplus x_0x_3 \oplus x_1x_3 \oplus x_2x_3 \oplus x_2x_4 \oplus x_3x_4 \oplus x_1 \oplus x_3 \oplus x_4), x_i \in \mathbb{F}_2, 0 \leq i \leq 4)$ for CBEAMr1 authenticated encryption, but this function is very complexity for hardware, since this function can not be implemented with less than eight logical instructions, so this encryption defined a new data type in order to fit into the register sets of various CPU architectures.

By [5,11], we find that they did not give how to construct this permutation (named by rotation boolean permutation) as simply as possible from cryptographical security. Based on the above consideration, we study the following questions:

(1) What is the property of the rotation linear boolean permutations?
(2) How to construct the rotation nonlinear boolean permutations.

The organization of this paper is as follows. In Sect. 2, the basic concepts and notions are presented. In Sect. 3, rotation linear boolean permutations is perfectly characterization. Three method to construct rotation nonlinear boolean permutation are presented, and its hardware implementation consumption can be analysis in Sect. 4. Finally, Sect. 5 concludes this paper.

2 Preliminaries

Let $\mathbb{B}_n$ denote the set of n variables Boolean functions. We denote by $\oplus$ the additions in $\mathbb{F}_2$, in $\mathbb{F}_2^n$ and in $\mathbb{B}_n$. Every Boolean function $f(x) \in \mathbb{B}_n$ admits a unique representation called its algebraic normal form (ANF) as a polynomial over $\mathbb{F}_2$:

$$f(x_1, \cdots, x_n) = a_0 \oplus \bigoplus_{1\leq i\leq n} a_i x_i \oplus \bigoplus_{1\leq i<j\leq n} a_{i,j} x_i x_j \oplus \cdots \oplus a_{1,\cdots,n} x_1 x_2 \cdots x_n$$

where the coefficients $a_0, a_i, a_{i,j}, \cdots, a_{1,\cdots,n} \in \mathbb{F}_2$. The algebraic degree, $deg(f)$, is the number of variables in the highest order term with non-zero coefficient. The support of a Boolean function $f(x) \in \mathbb{B}_n$ is defined as $Supp(f) = \{(x_1, \cdots, x_n) \mid f(x_1, \cdots, x_n) = 1\}$. We say that a Boolean function $f(x)$ is balanced if its truth table contains an equal number of ones and zeros, i.e., if its Hamming weight equals 2^{n-1}. A Boolean function is affine if there exists no term of degree > 1 in the ANF and the set of all affine functions is denoted by $\mathbb{A}_n$. An affine function with constant term equal to zero is called a linear function.

Definition 1. The Walsh spectrum of $f(x) \in \mathbb{B}_n$ is defined as

$$\mathscr{F}(f \oplus \varphi_\alpha) = \sum_{x\in\mathbb{F}_2^n} (-1)^{f(x)\oplus\alpha x},$$

where $\varphi_\alpha = \alpha x = \alpha_1 x_1 \oplus \alpha_2 x_2 \oplus \cdots \oplus \alpha_n x_n$.

Definition 2. The cross-correlation function between $f(x), g(x) \in \mathbb{B}_n$ is defined as

$$\triangle_{f,g}(\alpha) = \sum_{x\in\mathbb{F}_2^n} (-1)^{f(x)\oplus g(x\oplus\alpha)}, \alpha \in \mathbb{F}_2^n.$$

If $f(x) = g(x)$, then $\triangle_f(\alpha) = \sum_{x\in\mathbb{F}_2^n} (-1)^{f(x)\oplus f(x\oplus\alpha)}$.

Denoted $\triangle_{min} = \min\{|\triangle_f(\alpha)| \alpha \in \mathbb{F}_2^n, \alpha \neq 0^n\}$.

Two n-variable Boolean functions $f(x), g(x)$ are called to be perfectly uncorrelated if $\triangle_{f,g}(\alpha) = 0$ for all $\alpha \in \mathbb{F}_2^n$, and are called to be uncorrelated of degree k if $\triangle_{f,g}(\alpha) = 0$ for all $\alpha \in \mathbb{F}_2^n$ such that $0 \leq wt(\alpha) \leq k$.

The two indicators are called the global avalanche characteristics of Boolean functions(GAC [6]): $\sigma_f = \sum_{\alpha\in\mathbb{F}_2^n} \triangle_f^2(\alpha), \triangle_f = \max_{\alpha\in\mathbb{F}_2^n, wt(\alpha)\neq\mathbf{0}} \mid \triangle_f(\alpha) \mid$.

In order to study cross-correlation distributions between any two Boolean functions, we need the following definition:

Definition 3. [14] Let $f(x), g(x) \in \mathbb{B}_n$. If $D_a(f,g) : x \mapsto f(x) \oplus g(x \oplus a)$ is constant, a is said to be a *linear structure* of f and g. For convenience, let

$U_{f,g}^0 = \{a \in \mathbb{F}_2^n \mid f(x) \oplus g(x \oplus a) = 0, \forall x \in \mathbb{F}_2^n\}$;

$U_{f,g}^1 = \{a \in \mathbb{F}_2^n \mid f(x) \oplus g(x \oplus a) = 1, \forall x \in \mathbb{F}_2^n\}$;

If $\mathbf{0}^n \in U_{f,g}$, it is easy to know that $U_{f,g}^0$ and $U_{f,g} = U_{f,g}^0 \cup U_{f,g}^1$ are linear subspaces of $\mathbb{F}_2^n$.

In Definition 3, if $f(x) = g(x)$, then $U_f^0 = \{a \in \mathbb{F}_2^n \mid f(x) \oplus f(x \oplus a) = 0, \forall x \in \mathbb{F}_2^n\}$; $U_f^1 = \{a \in \mathbb{F}_2^n \mid f(x) \oplus f(x \oplus a) = 1, \forall x \in \mathbb{F}_2^n\}$. U_f^0 and $U_f = U_f^0 \cup U_f^1$ are linear subspaces of $\mathbb{F}_2^n$.

For $f(x) \in \mathbb{B}_n$, the annihilators of f is the set $Ann(f) = g \in \mathbb{B}_n : f \cdot g = 0$. The algebraic immunity $AI(f)$ is the minimum degree of nonzero functions $g \in \mathbb{B}_n$ such that $gf = 0$ or $g(1 \oplus f) = 0$. Namely, $AI(f) = \min\{deg(g) : 0 \neq g \in ANN(f) \cup Ann(1 \oplus f) = (f \oplus 1) \cup (f)\}$.

Definition 4. Let $F(x) = (f_1(x), f_2(x), ..., f_n(x)) \in \mathbb{F}_2^n$ and $f_i(x) \in \mathbb{B}_n$, $x \in \mathbb{F}_2^n$. $F(x)$ is called to a boolean permutation, if $F(x)$ is an one to one mapping from $\mathbb{F}_2^n$ to $\mathbb{F}_2^n$.

Lemma 1. *Let $F(x) = (f_1(x), f_2(x), ..., f_n(x)) \in \mathbb{F}_2^n$ and $f_i(x) \in \mathbb{B}_n$, $x \in \mathbb{F}_2^n$. $F(x)$ is a boolean permutation if and only if $\bigoplus_{i=1}^n c_i f_i(x)$ is a balanced function, where $(0, 0, ..., 0) \neq (c_1, c_2, ..., c_n) \in \mathbb{F}_2^n$.*

In this paper, we will study a specially permutation, named the rotation boolean permutation.

Definition 5. Let $f(x_1, x_2, ..., x_{n-1}, x_n) \in \mathbb{F}_2^n$. $F(x)$ is called a rotation boolean permutation(denoted by RBP), if $F(x) = (f^0(x), f^1(x), ..., f^{n-1}(x))$ is a boolean permutation, where

$$\begin{aligned}
&f^0(x) = f(x_1, x_2, ..., x_{n-1}, x_n),\\
&f^1(x) = f(x_2, x_3, ..., x_n, x_1),\\
&...\\
&f^{n-1}(x) = f(x_n, x_1, ..., x_{n-2}, x_{n-1}).
\end{aligned}$$

For example, if $f(x_1, x_2, x_3) = x_1x_2 \oplus x_3$, then $f^1 = x_2x_3 \oplus x_1$, $f^2 = x_3x_1 \oplus x_2$. It is easy to know $f^n = f^0 = f(x)$.

In term of Definition 5, we know that $F(x)$ is fully determined by $f(x_0\ , x_1\ , \cdots,\ x_{n-1})$. So, we called $f(x_0, x_1, \cdots, x_{n-1})$ a basic function of a rotation boolean permutation $F(x)$. Thus, the set of n-bit rotation boolean function can be partitioned into 4 subsets:

(1) Basic function: $f(x_0, x_1, \cdots, x_{n-1})$;
(2) Reverse of basic function: $f_r(x_0, x_1, \cdots, x_{n-1}) = f(x_{n-1}, x_{n-2}, .., x_1, x_0)$;
(3) Complement of basic function: $f_c(x_0, x_1, \cdots, x_{n-1}) = 1 \oplus f(x_0, x_1, x_2, ..., x_{n-2}, x_{n-1})$;
(4) Reverse complement of basic function: $f_{rc}(x_0, x_1, \cdots, x_{n-1}) = 1 \oplus f(x_{n-1}, x_{n-2}, ..., x_1, x_0)$.

That means, if we find a basic function of a rotation boolean permutation, then we can obtain three classed permutation: reverse, complement and reverse complement rotation boolean permutations. Thus, how to find a basic rotation boolean permutation is important.

3 Rotation Linear Boolean Permutation

At first, we give a result about rotation linear boolean permutation.

Theorem 1. *Let* $f(x_1, x_2, ..., x_{n-1}, x_n) = a_1x_1 \oplus a_2x_2 \oplus \cdots c_nx_n \oplus a_0 \in \mathbb{B}_n$, $a)i \in \mathbb{F}_2^n (0 \leq i \leq n-1)$. *Then* $F(x) = (f^0(x), f^1(x), ..., f^{n-1}(x))$ *is a rotation boolean permutation if and only if*

$$A = \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_{n-1} & a_n \\ a_n & a_1 & a_2 & \cdots & a_{n-2} & a_{n-1} \\ a_{n-1} & a_n & a_1 & \cdots & a_{n-3} & a_{n-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_2 & a_3 & a_4 & \cdots & a_n & a_1 \end{pmatrix}$$

is a reversible matrix on $\mathbb{F}_2$.

Proof. It is easy to proof. □

For a rotation linear boolean permutation, we know that this rotation boolean permutations are fully determined by the reversible matrix. So the number of rotation linear permutations is at most the number of the reversible matrix.

4 Rotation Nonlinear Boolean Permutation

In this section, we will analyze rotation nonlinear boolean permutation, and give three constructions at first, then analysis its hardware implementation consumption.

4.1 The First Construction

Construction 1. *Let*

$$f(x_1, x_2, ..., x_{n-1}, x_n) = x_1 \oplus x_2x_3 \cdots x_{n-1} \oplus x_2x_3 \cdots x_{n-1}x_n$$

be a Boolean function with $n (n \geq 4)$*-variable. Then* $F(x) = (f^0, f^1, f^2, \cdots, f^{n-1})$ *is a rotation boolean permutation.*

Proof. According to the ANF of $f(x)$, then $f^0 = x_1 \oplus x_2x_3 \cdots x_{n-1}(1 \oplus x_n), f^1 = x_2 \oplus x_3x_4 \cdots x_n(1 \oplus x_1), f^2 = x_3 \oplus x_4x_5 \cdots x_1(1 \oplus x_2), \cdots, f^{n-1} = x_n \oplus x_1x_2 \cdots x_{n-2}(1 \oplus x_{n-1})$. There are four cases:

(1) When $wt(x) < n-2$. Then $x_2x_3 \cdots x_{n-1}$=$x_3x_4 \cdots x_n$=$\cdots$=$x_1x_2 \cdots x_{n-2}$=0, that is, $F(\alpha) \neq F(\beta)$ for any $\alpha, \beta \in \mathbb{F}_2^n$ satisfying $0 \leq wt(\alpha), wt(\beta) < n-2$ and $\alpha \neq \beta$.

The number (denoted byT_1) of $\alpha \in \mathbb{F}_2^n (wt(\alpha) < n-2)$ in this case is $T_1 = \sum_{i=0}^{n-3} \binom{n}{i}$.

(2) When $wt(x) = n-2$. Then only one of $x_2x_3 \cdots x_{n-1}$, $x_3x_4 \cdots x_n$, $\cdots$, $x_1x_2 \cdots x_{n-2}$ is equal to 1. For simplicity, let $x_2x_3 \cdots x_{n-1} = 1$, we have

$x_2 = x_3 = \cdots = x_{n-1} = 1$ and $x_1 = x_n = 0$. Thus if $\alpha = (0, \underbrace{1, 1, \cdots, 1}_{n-2}, 0)$, then $F(\alpha) = (\underbrace{1, 1, \cdots, 1}_{n-1}, 0)$. Denoted by the set $A = \{(0, \underbrace{1, 1, \cdots, 1}_{n-2}, 0), (0, 0, \underbrace{1, 1, \cdots, 1}_{n-2}), \cdots, (\underbrace{1, 1, \cdots, 1}_{n-2}, 0, 0)\}$. It is easy verifies that $F(\alpha) \neq F(\beta)$ if $\alpha, \beta \in A$ and $\alpha \neq \beta$.

Then let $B = \{\alpha \in \mathbb{F}_2^n \mid wt(\alpha) = n-2, \alpha \notin A\}$. $\mid B \mid = \binom{n}{n-2} - \mid A \mid = \binom{n}{n-2} - n$. Note that $x_2 x_3 \cdots x_{n-1} = x_3 x_4 \cdots x_n = \cdots = x_1 x_2 \cdots x_{n-2} = 0$, if $x \in B$. This means $\alpha = F(\alpha) \neq F(\beta) = \beta$ if $\alpha, \beta \in B$ and $\alpha \neq \beta$.

The number (denoted by T_2) of $\alpha \in \mathbb{F}_2^n (wt(\alpha) = n-2)$ in this case is $T_2 = \binom{n}{n-2}$.

(3) When $wt(x) = n-1$. That is, let $x_i = 0$ and $x_j = 1$ for $1 \leq i \neq j \leq n$. So, $\alpha = (1, 1, \cdots, 1, \underbrace{0}_{i}, 1, \cdots, 1) \in \mathbb{F}_2^n$, $F(\alpha) = (\underbrace{1, 1, \cdots, 1}_{i-1}, 0, 0, \underbrace{1, \cdots, 1}_{n-i-1})$.

The number(denoted by T_3) of $\alpha \in \mathbb{F}_2^n (wt(\alpha) = n-1, n)$ in this case is $T_3 = \binom{n}{n-1} + 1 = n + 1$.

(4) $wt(x) = n$, that is, if $wt(\alpha) = n$, then $F(\alpha) = (1, 1, \cdots, 1, 1)$.

Combining the above four cases, we know that the number (denoted by T) of value with $F(x)$ is $T = T_1 + T_2 + T_3 = \sum_{i=0}^{n-3} \binom{n}{i} + \binom{n}{n-2} + n + 1 = 2^n$.

Thus, $F(x)$ is a rotation boolean permutation on $\mathbb{F}_2^n$. □

Remark 1. In **Construction 1**.

(1) We call $f(x_1, x_2, ..., x_{n-1}, x_n) = x_1 \oplus x_2 x_3 \cdots x_{n-1} \oplus x_2 x_3 \cdots x_{n-1} x_n$ a basic function, denoted by f_{bf}^0.

(2) It is easy to find that this boolean permutation $F(x) = (f^0, f^1, \cdots, f^{n-1})$ has some fixedly points, that is, $F(x) = x$. For example $x = (0, 0, \cdots, 0)$. In order to eliminate these fixedly points, we can change $F(x) = (f^0, f^1, \cdots, f^{n-1})$ by $G(x) = (f^0 \oplus 1, f^1, \cdots, f^{n-1})$, or by $G(x) = (f^0, f^1 \oplus 1, \cdots, f^{n-1})$, etc.

Lemma 2. *Let $f(x_1, x_2, ..., x_{n-1}, x_n) = x_1 x_2 x_3 \cdots x_{n-1} x_n$ be a Boolean function with n-variable. Then the Walsh spectrum is three values for any $\alpha \in \mathbb{F}_2^n$:*

$$\mathscr{F}(f \oplus \varphi_\alpha) = \begin{cases} 2, & wt(\alpha) \equiv 0 \mod 2, wt(\alpha) > 0; \\ -2, & wt(\alpha) \equiv 1 \mod 2; \\ 2^n - 2, & wt(\alpha) = 0. \end{cases}.$$

Theorem 2. *Let $f(x_1, x_2, ..., x_{n-1}, x_n) = x_1 \oplus x_2 x_3 \cdots x_{n-1} \oplus x_2 x_3 \cdots x_{n-1} x_n$ be a Boolean function with n-variable. Then f satisfies the following properties:*

1. *balanced;*
2. *$deg(f) = n - 1$;*
3. *$AI(f) = 2$;*
4. *$N_f = 2$;*
5. *The Walsh spectrum is four values: $\{0, \pm 4, 2^n - 4\}$.*

Proof. According to the definition of Walsh spectrum and Lemma 2, we have

$$\begin{aligned}
\mathscr{F}(f\oplus\alpha) &= \sum_{x\in\mathbb{F}_2^n}(-1)^{f(x)\oplus\alpha\cdot x}\\
&= \sum_{x\in\mathbb{F}_2^n}(-1)^{x_1\oplus x_2x_3\cdots x_{n-1}\oplus x_2x_3\cdots x_{n-1}x_n\oplus\alpha_1x_1\oplus\alpha_2x_2\cdots\oplus\alpha_nx_n}\\
&= (1+(-1)^{1\oplus\alpha_1})\sum_{(x_2,x_3,\cdots,x_n)\in\mathbb{F}_2^{n-1}}(-1)^{x_2x_3\cdots x_{n-1}\oplus x_2x_3\cdots x_{n-1}x_n\oplus\alpha_2x_2\cdots\oplus\alpha_nx_n}\\
&= (1-(-1)^{\alpha_1})[\sum_{(x_2,\cdots,x_{n-1})\in\mathbb{F}_2^{n-2}}(-1)^{x_2x_3\cdots x_{n-1}\oplus\alpha_2x_2\oplus\cdots\oplus\alpha_{n-1}x_{n-1}}+\\
&\quad (-1)^{\alpha_n}\sum_{(x_2,\cdots,x_{n-1})\in\mathbb{F}_2^{n-2}}(-1)^{\alpha_2x_2\oplus\cdots\oplus\alpha_{n-1}x_{n-1}}]\\
&= \begin{cases}0, & \alpha_1=0,\alpha_i\in\mathbb{F}_2, 2\le i\le n;\\ -4, & \alpha_1=1, wt((\alpha_2,\cdots,\alpha_{n-1}))\equiv 1 \mod 2, \alpha_n\in\mathbb{F}_2;\\ 4, & \alpha_1=1, wt((\alpha_2,\cdots,\alpha_{n-1}))\equiv 0 \mod 2, \alpha_n\in\mathbb{F}_2;\\ 2^n-4, & \alpha_1=1, wt(\alpha_2,\cdots,\alpha_{n-1},\alpha_n))=0.\end{cases}
\end{aligned}$$

Based on the distribution of Walsh spectrum, 1,4 and 5 are easy to be proved.

It is easy to find the annihilator of $f(x)$ is $(1\oplus x_1)x_n$. Thus, $AI(f)=2$. □

Example 1. (1) For $n=4$, then the truth table of this function in Theorem 2 is $\overline{f}=(0x02, 0xfd)$ (in hexadecimal). The Walsh spectrum is $\overline{\mathscr{F}}=(0,0,0,0,0,0,0,0,12,-4,4,4,4,4,-4,-4)$;

(2) For $n=5$, then the truth table of this function in Theorem 2 is $\overline{f}=(0x00, 0x02, 0xff, 0xfd)$. The Walsh spectrum is $\overline{\mathscr{F}}=(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,28,-4,4,4,4,4,-4,-4,4,4,-4,-4,-4,-4,4,4)$.

(3) For n, then the truth table of this function in Theorem 2 is $\overline{f}=(\underbrace{0x00,\cdots,0x00}_{2^{n-4}-1}, 0x02, \underbrace{0xff,\cdots,0xff}_{2^{n-4}-1}, 0xfd)$.

4.2 The Second Construction

Construction 2. *Let*

$$f(x_1,x_2,...,x_{n-1},x_n)=x_n\oplus x_{n-1}x_{n-2}\cdots x_5x_3x_2x_1\oplus x_nx_{n-1}\cdots x_5x_3x_2$$

be a Boolean function with $n(n\ge 5)$-variable. Then $F(x)=(f^0,f^1,f^2,\cdots,f^{n-1})$ is a rotation boolean permutation.

Proof. According to the ANF of $f(x)$, then $f^0=x_n\oplus x_{n-1}x_{n-2}\cdots x_5x_3x_2(x_1\oplus x_n)$, $f^1=x_1\oplus x_nx_{n-1}\cdots x_6x_4x_3(x_2\oplus x_1)$, $f^2=x_2\oplus x_1x_n\cdots x_7x_5x_4(x_3\oplus x_2)$, $\cdots f^{n-1}=x_{n-1}\oplus x_{n-2}x_{n-3}\cdots x_4x_2x_1(x_n\oplus x_{n-1})$. There are five cases:

(1) When $wt(x) < n-3$. Then $x_{n-1}x_{n-2}\cdots x_5x_3x_2 = x_nx_{n-1}\cdots x_6x_4x_3 = \cdots = x_{n-2}x_{n-3}\cdots x_4x_2x_1 = 0$, that is, if $\alpha,\beta \in \mathbb{F}_2^n$ satisfying $0 \le wt(\alpha), wt(\beta) < n-3$ and $\alpha \neq \beta$, then $F(\alpha) \neq F(\beta)$.

The number(denoted byT_1) of $\alpha \in \mathbb{F}_2^n(wt(\alpha) < n-3)$ in this case is $T_1 = \sum_{i=0}^{n-4}\binom{n}{i}$.

(2) When $wt(x) = n-3$. Then only one of $x_{n-1}x_{n-2}\cdots x_5x_3x_2$, $x_nx_{n-1}\cdots x_6x_4x_3$, $\cdots$, $x_{n-2}x_{n-3}\cdots x_4x_2x_1$ is equal to 1. For simplicity, let $x_{n-1}x_{n-2}\cdots x_5x_3x_2 = 1$, we have $x_2 = x_3 = x_5 = \cdots = x_{n-2} = x_{n-1} = 1$ and $x_1 = x_4 = x_n = 0$. Thus if $\alpha = (0,1,1,0,\underbrace{1,1,\cdots,1}_{n-5},0)$, then $F(\alpha) = (0,0,1,1,0,\underbrace{1,1,\cdots,1}_{n-5})$.

By the same method of Construction 1, denoted by the set $A = \{(0,1,1,0,\underbrace{1,1,\cdots,1}_{n-5},0), (0,0,1,1,0,\underbrace{1,1,\cdots,1}_{n-5}), \cdots, (1,1,0,\underbrace{1,1,\cdots,1}_{n-5},0,0)\}$. It is easy verifies that $F(\alpha) \neq F(\beta)$ if $\alpha,\beta \in A$ and $\alpha \neq \beta$.

Then let $B = \{\alpha \in \mathbb{F}_2^n \mid wt(\alpha) = n-3, \alpha \notin A\}$. $\mid B \mid = \binom{n}{n-3} - \mid A \mid = \binom{n}{n-3} - n$. Note that $x_{n-1}x_{n-2}\cdots x_5x_3x_2 = x_nx_{n-1}\cdots x_6x_4x_3 = \cdots = x_{n-2}x_{n-3}\cdots x_4x_2x_1 = 0$, if $x \in B$. This means $F(\alpha) \neq F(\beta)$ if $\alpha,\beta \in B$ and $\alpha \neq \beta$.

The number(denoted by T_2) of $\alpha \in \mathbb{F}_2^n(wt(\alpha) = n-3)$ in this case is $T_2 = \binom{n}{n-3}$.

(3) When $wt(x) = n-2$. Suppose $x_1 = \cdots = x_{n-2} = 1$ and $x_{n-1} = x_n = 0$. Then $F(x) = (0,\underbrace{1,1,\cdots,1}_{n-2},0)$. It is easy to verify that $wt(F(x)) = n-2$ for any $wt(x) = n-2$, and $F(\alpha) \neq F(\beta)$ for $wt(\alpha) = wt(\beta) = n-2$ and $\alpha \neq \beta$.

The number(T_3) of of $\alpha \in \mathbb{F}_2^n(wt(\alpha) = n-2)$ in this case is $T_3 = \binom{n}{n-2}$.

(4) When $wt(x) = n-1$. Suppose $x_1 = \cdots = x_{n-1} = 1$ and $x_n = 0$. Then $F(x) = (\underbrace{1,1,\cdots,1}_{n-1},0)$. It is easy to verify that $F(x) = x$ for any $wt(x) = n-1$.

The number(T_4) of $\alpha \in \mathbb{F}_2^n(wt(\alpha) = n-1)$ in this case is $T_4 = n$.

(5) When $wt(x) = n$. Then $F(\underbrace{1,1,\cdots,1}_{}) = (\underbrace{1,1,\cdots,1}_{})$. The number($T_5$) In this case is $T_5 = n$.

Combining the above five cases, we know that the number(denoted by T) of value with $F(x)$ is $T = T_1 + T_2 + T_3 + T_4 + T_5 = \sum_{i=0}^{n-4}\binom{n}{i} + \sum_{i=0}^{n-3}\binom{n}{i} + \binom{n}{n-2} + n + 1 = 2^n$.

Thus, $F(x)$ is a rotation boolean permutation on $\mathbb{F}_2^n$. □

Remark 2. In **Construction 2.**

(1) We call $f(x_1, x_2, ..., x_{n-1}, x_n) = x_n \oplus x_{n-1}x_{n-2}\cdots x_5x_3x_2x_1 \oplus x_nx_{n-1}\cdots x_5x_3x_2$ a 2-nd basic function, denoted by f_{bf}^1.

(2) It is easy to find that this boolean permutation $F(x) = (f^0, f^1, \cdots, f^{n-1})$ has some fixedly points, that is, $F(x) = x$, for example $(0,0,\cdots,0)$. In order to eliminate these fixedly points, we can change $F(x) = (f^0, f^1, \cdots, f^{n-1})$ by $G(x) = (f^0 \oplus 1, f^1, \cdots, f^{n-1})$, or by $G(x) = (f^0, f^1 \oplus 1, \cdots, f^{n-1})$, etc.

Example 2. The truth table of $x_n \oplus x_{n-1}x_{n-2}\cdots x_5x_3x_2x_1 \oplus x_nx_{n-1}\cdots x_5x_3x_2 (n \geq 5)$ has the following property:

(1) For n, the truth table is $(\underbrace{0x00, \cdots, 0x00}_{2^{n-4}-2}, 0x02,\ 0x02,\ \underbrace{0xff, \cdots, 0xff}_{2^{n-4}-2},$ $0xfd,\ 0xfd)$;

When $n = 5$, the truth table is $(0x02, 0x02, 0xfd, 0xfd)$;

When $n = 6$, the truth table is $(0x00, 0x00, 0x02, 0x02, 0xff, 0xff, 0xfd, 0xfd)$.

(2) This Boolean function satisfies $f(a) \oplus f(a \oplus 1) = 1$ for any $a \in \mathbb{F}_2^n$.

Theorem 3. *Let* $f(x_1, x_2, ..., x_{n-1}, x_n) = x_n \oplus x_{n-1}x_{n-2}\cdots x_5x_3x_2x_1 \oplus x_nx_{n-1} \cdots x_5x_3x_2$ *be a Boolean function with n-variable. Then* f *satisfies the following properties:*

1. *balanced;*
2. $deg(f) = n - 2$*;*
3. $AI(f) = 2$*;*
4. $N_f = 4$*;*
5. *The Walsh spectrum is four values:* $\{0, \pm 8, 2^n - 8\}$.

Proof. It is easy to proof. □

4.3 The Third Construction

Construction 3. *Let*

$$f(x_1, x_2, ..., x_{n-1}, x_n) = x_n \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2x_1$$

be a Boolean function with $n(n \geq 6)$*-variable. Then* $F(x) = (f^0, f^1, f^2, \cdots, f^{n-1})$ *is a rotation boolean permutation.*

Proof. According to the ANF of $f(x)$, then $f^0 = x_n \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2(x_1 \oplus 1), f^1 = x_1 \oplus x_nx_{n-1}\cdots x_8x_7x_4x_3(x_2 \oplus 1), f^2 = x_2 \oplus x_1x_n \cdots x_9x_8x_5x_4(x_3 \oplus 1), \cdots f^{n-1} = x_{n-1} \oplus x_{n-2}x_{n-3}\cdots x_6x_5x_2x_1(x_n \oplus 1)$. There are six cases:

(1) When $wt(x) < n - 4$. Then $x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 = x_nx_{n-1}\cdots x_8x_7 x_4x_3 = x_1x_n \cdots x_9x_8x_5x_4 = \cdots x_{n-2}x_{n-3}\cdots x_6x_5x_2x_1 = 0$, that is, if $\alpha, \beta \in \mathbb{F}_2^n$ satisfying $0 \leq wt(\alpha), wt(\beta) < n - 3$ and $\alpha \neq \beta$, then $F(\alpha) \neq F(\beta)$.

The number(denoted by T_1) of $\alpha \in \mathbb{F}_2^n (wt(\alpha) < n - 4)$ in this case is $T_1 = \sum_{i=0}^{n-5} \binom{n}{i}$.

(2) When $wt(x) = n-4$. Then only one of $x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2, x_nx_{n-1}\cdots x_8x_7x_4x_3, x_1x_n \cdots x_9x_8x_5x_4, \cdots, x_{n-2}x_{n-3}\cdots x_6x_5x_2x_1$ is equal to 1. For simplicity, let $x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 = 1$, we have $x_{n-1} = x_{n-2} = \cdots = x_7 = x_6 = x_3 = x_2 = 1$ and $x_1 = x_4 = x_5 = x_n = 0$. Thus, if $\alpha = (0, 1, 1, 0, 0, \underbrace{1, 1, \cdots, 1}_{n-6}, 0)$, then $F(\alpha) = (1, 0, 1, 1, 0, 0, \underbrace{1, 1, \cdots, 1}_{n-6})$.

By the same method of Construction 2, denoted by the set $A = \{(0,1,1,0,0,\underbrace{1,1,\cdots,1}_{n-6},0), \quad (0,0,1,1,0,0,\underbrace{1,1,\cdots,1}_{n-6}), \quad (1,0,0,1,1,0,0,\underbrace{1,1,\cdots,1}_{n-6}), \quad \cdots, (1,1,0,0,\ \underbrace{1,1,\cdots,1}_{n-6},0,0)\}$. It is easy verifies that $F(\alpha) \neq F(\beta)$ if $\alpha, \beta \in A$ and $\alpha \neq \beta$.

Then let $B = \{\alpha \in \mathbb{F}_2^n \mid wt(\alpha) = n-4, \alpha \notin A\}$. $\mid B \mid = \binom{n}{n-4} - \mid A \mid = \binom{n}{n-4} - n$. Note that $x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 = x_nx_{n-1}\cdots x_8x_7x_4x_3 = x_1x_n\cdots x_9x_8x_5x_4 = \cdots\ x_{n-2}x_{n-3}\cdots\ x_6x_5x_2x_1 = 0$, if $x \in B$. This means $F(\alpha) \neq F(\beta)$ if $\alpha, \beta \in B$ and $\alpha \neq \beta$.

The number(denoted by T_2) of $\alpha \in \mathbb{F}_2^n(wt(\alpha) = n-4)$ in this case is $T_2 = \binom{n}{n-4}$.

(3) When $wt(x) = n-3$. Suppose $x_1 = \cdots = x_{n-3} = 1$ and $x_{n-2} = x_{n-1} = x_n = 0$. Then $F(x) = (0,\underbrace{1,1,\cdots,1}_{n-3},0,0)$. It is easy to verify that $wt(F(x)) = n-3$ for any $wt(x) = n-3$, and $F(\alpha) \neq F(\beta)$ for $wt(\alpha) = wt(\beta) = n-3$ and $\alpha \neq \beta$.

The number(T_3) of of $\alpha \in \mathbb{F}_2^n(wt(\alpha) = n-3)$ in this case is $T_2 = \binom{n}{n-3}$.

(4) When $wt(x) = n-2$. Suppose $x_1 = \cdots = x_{n-2} = 1$ and $x_{n-1} = x_n = 0$. Then $wt(F(x)) = wt((0,\underbrace{1,1,\cdots,1}_{n-1})) = n-1$. Meanwhile, suppose $x_2 = x_4 = \cdots = x_n = 1$ and $x_1 = x_3 = 0$. Then $wt(F(x)) = wt((1,0,1,0,\underbrace{1,1,\cdots,1}_{n-4})) = n-2$. It is easy to verify that there are two cases:

(1) When $wt(x) = n-2$ and there are two consecutive locations in x are equal to 0, that is, $x_i = x_{i+1} = 0 (1 \leq i \leq n)$. Then $wt(F(x)) = n-1$, and $F(\alpha) \neq F(\beta)$ if $\alpha, \beta (\alpha \neq \beta)$ are in this case. The number of x in this case is n.

(2) When $wt(x) = n-2$ and there are two discontinuousness locations in x are equal to 0, that is, $x_i = x_j = 0 (1 \leq i < j \leq n)$. Then $wt(F(x)) = n-2$, and $F(\alpha) \neq F(\beta)$ if $\alpha, \beta (\alpha \neq \beta)$ are in this case. The number of x in this case is $\binom{n}{2} - n$.

The number(T_4) of $\alpha \in \mathbb{F}_2^n(wt(\alpha) = n-2)$ in two cases is $T_4 = n + \binom{n}{2} - n = \binom{n}{2}$.

(5) When $wt(x) = n-1$. Suppose $x_1 = \cdots = x_{n-1} = 1$ and $x_n = 0$. Then $F(x) = (0,\underbrace{1,1,\cdots,1}_{n-2},0)$. It is easy to verify that $wt(F(x)) = n-2$ for any $wt(x) = n-1$, and $F(\alpha) \neq F(\beta)$ for $wt(\alpha) = wt(\beta) = n-2$ and $\alpha \neq \beta$. The number(T_5) in this case is $T_5 = n$.

(6) When $wt(x) = n$. Then $F(\underbrace{1,1,\cdots,1}_{n}) = (\underbrace{1,1,\cdots,1}_{n})$.

Combining the above five cases, we know that the number(denoted by T) of value with $F(x)$ is $T = T_1 + T_2 + T_3 + T_4 + T_5 + 1 = \sum_{i=0}^{n-5} \binom{n}{i} + \binom{n}{n-4} + \binom{n}{n-3} + \binom{n}{2} + n + 1 = 2^n$.

Thus, $F(x)$ is a rotation boolean permutation on $\mathbb{F}_2^n$. □

Remark 3. In **Construction 3.**

(1) We call $f(x_1, x_2, ..., x_{n-1}, x_n)$=$x_n \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 \oplus x_{n-1}x_{n-2} \cdots x_7x_6x_3x_2x_1$ a 3-th basic function, denoted by f_{bf}^2.

(2) It is easy to find that this boolean permutation $F(x) = (f^0, f^1, \cdots, f^{n-1})$ has some fixedly points, that is, $F(x) = x$, for example $(0, 0, \cdots, 0)$. In order to eliminate these fixedly points, we can change $F(x) = (f^0, f^1, \cdots, f^{n-1})$ by $G(x) = (f^0 \oplus 1, f^1, \cdots, f^{n-1})$, or by $G(x) = (f^0, f^1 \oplus 1, \cdots, f^{n-1})$, etc.

Example 3. The truth table of $f(x_1, x_2, ..., x_{n-1}, x_n) = x_n \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2x_1 (n \geq 6)$ has the following preposition:

(1) For n, the truth table is $(\underbrace{0x00, \cdots, 0x00}_{2^{n-4}-4}, \underbrace{0x02, \cdots, 0x02}_{4}, \underbrace{0xff, \cdots, 0xff}_{2^{n-4}-4}, \underbrace{0xfd, \cdots, 0xfd}_{4})$;

When $n = 6$, the truth table is $(0x02, 0x02, 0x02, 0x02, 0xfd, 0xfd, 0xfd, 0xfd)$; When $n = 7$, the truth table is $(0x00, 0x00, 0x00, 0x00, 0x02, 0x02, 0x02, 0x02, 0xff, 0xff, 0xff, 0xff, 0xfd, 0xfd, 0xfd, 0xfd)$.

(2) This Boolean function satisfies $f(a) \oplus f(a \oplus 1) = 1$ for any $a \in \mathbb{F}_2^n$.

Theorem 4. *Let* $f(x_1, x_2, ..., x_{n-1}, x_n) = x_n \oplus x_{n-1}x_{n-2}\cdots x_7x_6x_3x_2 \oplus x_{n-1} \cdots x_7x_6x_3x_2x_1$ *be a Boolean function with n-variable. Then f satisfies the following properties:*

1. *balanced;*
2. *deg*$(f) = n - 3$*;*
3. *AI*$(f) = 2$*;*
4. $N_f = 8$*;*
5. *The Walsh spectrum is four values:* $\{0, \pm 16, 2^n - 16\}$.

Proof. It is easy to proof. □

In hardware implementation, we find some good properties:

(1) The three classes of rotation boolean permutations are fully determined by the basic Boolean function, respectively. This means, we need only store one Boolean function in a permutation with n-input, but not n Boolean functions.

(2) The truth of the three classes of rotation boolean permutations are 4-value $\{0x00$, $0x02$, $0xff$, $0xfd\}$, it consumes very little storage space.

(3) The *ANF* of the three classes of rotation boolean permutations has 3 monomial forms, it consumes a small number of gates.

From here we see that the three classes of rotation boolean permutations can be used in encryption algorithm with Wireless sensor network and Internet of Things.

5 Conclusions

In this paper, we gave some light weight of the rotation boolean permutation are perfectly characterized by the matrix of linear expressions. Three methods of rotation nonlinear boolean permutations are constructed. The sub-functions of the three permutations have three monomials, high degree, 2-algebra immunity. All three classes of rotation nonlinear boolean permutations are fully determination by the first component Boolean function, respectively.

References

1. Evans, D.: The Internet of Things–How the next Evolution of the Internet is Changing Everything. Cisco Internet Business Solutions Group (IBSG), April 2011. www.flickr.com/photos.ciscoibsg/sets/72157626611102387
2. Brockmeier, K.: Gartner Adds Big Date, Gamification, and Internet of Things to Its Hype Cycle, Read Write Enterprise, Trend Analysis, 11 August 2011. www.readwriteweb.com/enterprise/2011/08/gartner-adds-big-data-gamifica.php
3. Adams, C.M., Tavares, S.E.: Generating and counting binary bent sequences. IEEE Trans. Inf. Theor. **36**(5), 1170–1173 (1990)
4. Webster, A.F.: Plaintext/ciphertext bit dependencies in cryptographic system. Master's Thesis, Department of Electrical Engineering, Queen's University, Ontario, Cannada (1985)
5. Daemen, J.: Cipher and Hash Function Design Strategies based on linear and differential cryptananlysis. PhD thesis, K.U.Leuven, March 1995
6. Zhang, X.M., Zheng, Y.L.: GAC- the criterion for global avalanche characteristics of cryptographic functions. J. Univ. Comput. Sci. **1**(5), 316–333 (1995)
7. Sarkar, P., Maitra, S.: Cross-correlation analysis of cryptographically useful boolean functions and S-boxes. Theor. Comput. Syst. **35**, 39–57 (2002)
8. Charpin, P., Pasalic, E.: On propagation characteristics of resilient functions. In: SAC 2002. LNCS, vol. 2595, pp. 175–195. Springer (2002)
9. Meier, W., Pasalic, E., Carlet, C.: Algebraic attacks and decomposition of Boolean functions. In: Advances in Cryptology-Eurocrypt. LNCS, vol. 3027, pp. 474–491. Springer, Heidelberg (2004)
10. Zhang, W.G., Xiao, G.Z.: Constructions of almost optimal resilient Boolean functions on large even number of variables. IEEE Trans. Inf. Theory **55**(12), 5822–5831 (2009)
11. Saarinen, M.-J.O.: The CBEAMr1 Authenticated Encryption Algorithm. http://www.cbeam.mx
12. http://www.oscca.gov.cn/Doc/6/News_1106.htm
13. Zhou, Y., Xie, M., Xiao, G.: On the global avalanche characteristics of two Boolean functions and the higher order nonlinearity. Inf. Sci. **180**, 256–265 (2010)
14. Zhou, Y.: On the distribution of auto-correlation value of balanced Boolean functions. Adv. Math. Commun. **7**(3), 335–347 (2013)
15. Zhou, Y., Wang, L., Wang, W., Xiaoni, D.: One sufficient and necessary condition on balanced Boolean functions with $\sigma_f = 2^{2n} + 2^{n+3} (n \geq 3)$. Int. J. Found. Comput. Sci. **25**(3), 343–353 (2014)

The Construction Method of Clue Words Thesaurus in Chinese Patents Based on Iteration and Self-filtering

Na Deng[1(✉)], Xu Chen[2], Ou Ruan[1], Chunzhi Wang[1], Zhiwei Ye[1],
and Jingbai Tian[1]

[1] School of Computer, Hubei University of Technology, Wuhan, China
iamdengna@163.com
[2] School of Information and Safety Engineering,
Zhongnan University of Economics and Law, Wuhan, China
chenxu@zuel.edu.cn

Abstract. Patent analysis and mining can excavate valuable information hidden in patent texts, and help enterprises to make correct decisions. As an important step in patent mining, whether patent semantic annotation is correct or not directly affects the results of mining. During the annotation of effect statements, whether manual or automatic, we need to use clue words to judge. This paper presents a construction method of clue words thesaurus in Chinese patents based on iteration and self-filtering, in order to improve the accuracy of effect statements' annotation.

1 Introduction

In the era of big data, all walks of life have accumulated a large amount of data. Mining the hidden information in these data can help management to make decisions, thereby improves the quality and efficiency of production and life. As a carrier of human intelligence and innovation, patents have become a major source of data mining and analysis. Patents are rich in technical, economic and legal information, and the effective use of which can provide important support for enterprises in risks avoidance, patent acquisition, maintenance of interests, technological innovation and so on. Patent technology/effect matrix is a tool often used in patent analysis and mining, and displays technologies and effects of multiple patents in the form of matrix, to help applications discover patent minefields and blank areas. In the process of constructing patent technology/effect matrix, the key step is to annotate patent effect statements. Through the observation of a large number of patents, most effect statements contain some specific clue words, such as: is used to, application, have and so on. The recognition of these words plays an important role in the annotation of effect statements, whether in manual or automatic annotation. The accuracy of clue words' recognition directly influences on the accuracy of effect statements' annotation.

Patent abstract is a microcosm of the background, purpose, method and function descriptions of a patent, thus it is an important source of patent analysis and mining. In the process of manual annotation of effect statements, annotators read patent abstract to judge which clause is an effect statement according to the semantics contained in it.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_12

In the authors' previous research, we make use of the distribution and morphological characteristics of patent effect statements to extract effect statements automatically. Regardless of manual or automatic annotation, we need to determine whether the clause contains any word about patent function or the application scope of the patent. In order to make the annotation more accurate, clue words thesaurus should be as accurate and complete as possible. Based on the idea of iteration, we make the recognition of clue words and the extraction of effect statements carry on alternatively; in addition, we utilize the difference between effect statements and non-effect statements to filter clue words. This paper aims to present a construction method of clue words thesaurus in Chinese patents based on iteration and self-filtering.

2 Related Work

Recently, there are some researches about patent annotation at home and abroad. Through the expression of TRIZ in the theory of scientific effect knowledge research and using Natural Language Processing method, [1] automatically annotates the patent design goals and the patent realization principle. [2] proposes a novel ontology-based automatic semantic annotation approach based on the thorough analysis of patent documents, which combines both structure and content characteristics, and integrates multiple techniques from various aspects. [3] aims to automatically annotate four types of bibliographical references in Chinese patent documents, such as patents, standards, papers, and other monographs public documents. [4] annotates key phrases for two semantic categories: PROBLEMS and SOLUTIONS.

As far as technology/effect matrix concerned, in order to mine implicit semantic information in patents, [5] applies semantic role labeling to create technology-effect matrix. [6] presents a method for matrix structure construction based on feature degree and lexical model.

In the authors' previous work about patent analysis and mining [7–10], we mainly focused on the removal of stop words in patents, intelligent recommendation of the traditional Chinese medicine patents and effect annotation. In [7, 8], we adopt a semi-supervised machine learning method named co-training, which cooperates keyword extraction with list extraction, and incrementally annotates functional clauses in patent abstract. The clue word concerned about in this paper is a refinement of keyword in [7, 8], and we bring the idea of list extraction into the location of candidate effect clauses.

The rest of paper is organized as follows: Sect. 3 explained different categories of clue words and an obvious feature of clue words. Section 4 described the algorithm in detail. Section 5 analyzed the experimental results. Section 6 concluded the paper and prospected the future work.

3 Clue Words

Clue words refer to the words which explain the function or application scope of invention in patent abstract text. According to different situations, we divide clue words into the following seven categories.

(1) leading word: a word used to guide the emergence of effect clause. For example: “have”, “can”, “apply to”, “used to”, “make”, etc.
(2) facet word: a word used to indicate which aspects have changed brought by a patent invention. For example: “cost”, “performance”, “quality”, “efficiency”, etc.
(3) changing word: a word reveals what changes have been made by a patent invention. For example: “improve”, “simple”, “lower”, “avoid” and so on.
(4) degree word: a word used to indicate the extent to which a patent invention have changed. For example: “significant”, “obvious”, etc.
(5) application scope word: a word used to explain the application scope of invention. For example: “widely”, “scope”, “market”, etc.
(6) facet changing word: While abstract text is segmented into words, some facet word and changing word will be divided into a word in Chinese. For example: “high efficiency”, “high performance”, “effort”, “labor saving” etc.
(7) summary word: a word used to describe the effect or merit of a patented invention. For example: “advantage”, “characteristic”, “effect” etc.

One of obvious features of clue word is that it appears frequently in effect statements but rarely occurs in non-effect statements. Table 1 exhibits part of the clue words of each category.

Table 1. Part of the clue words of each category

(1)	have, be used to, can, make, achieve, support, thus
(2)	price, operation, time, speed, cost, depth, intensity
(3)	simple, avoid, reduce, increase, keep, compensate, prevent
(4)	obvious, remarkable, substantially, significant, thorough, enormously, greatly
(5)	widely, domain, spread, prospect, generalization, application, significance
(6)	high efficiency, high performance, labor saving, timesaving, anti-interference
(7)	advantage, effect, characteristic, function, improvement, help, curative effect

4 Algorithm

Making use of the feature of clue words described above, we use a method to filter clue words through words frequency statistics and self-filtering. First of all, annotate effect statements of some patents manually, find out all high frequency words and remove those that often appear in non-effect statements. After artificial selection, we get the initial high quality clue words set. Then, utilize the characteristic that effect clauses often occur successively, clue words are used to locate effect clauses in more patent to extract new clue words. The recognition of clue words and the extraction of effect statements carry on alternatively and gradually enriches clues words thesaurus. The algorithm stops until the clue thesaurus achieves a relatively stable state. The flow chart of the algorithm is shown in Fig. 1.

We focus on two key steps in Fig. 1: self-filtering and locating candidate effect statements in more patents.

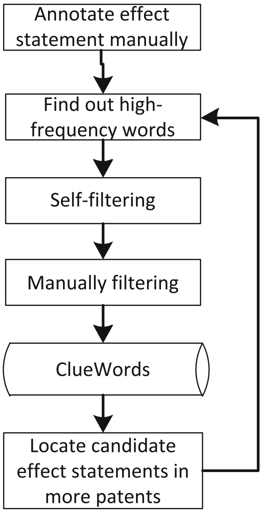

Annotate effect statement manually
Find out high-frequency words
Self-filtering
Manually filtering
ClueWords
Locate candidate effect statements in more patents

Fig. 1. The flow chart of the algorithm

4.1 Self-filtering

After finding out high frequency words of effect statements, it is necessary to filter these words in order to get high quality clue words. These high frequency words before filtration include not only the clue words, but also some common stop words. We use the feature of clue words to remove the words that are often used in the non-effect statements. In this paper, we call this kind of filtering method as self-filtering. We compare the high frequency words in effect statements with the high frequency words in patent abstracts without using any other filters or artificial screening. Self-filtering algorithm is as follows.

```
Algorithm: Self-filtering
Input: a set of Patent abstract texts, denoted as T;
annotated effect clauses, denoted as C; threshold th1;
threshold th2
Output: a set of candidate Clue Words, denoted as CWs
Begin
   $\{ClueWords\} = \emptyset$
   HashMap<String, Integer> HM1=segment_count(T);
   HashMap<String, Integer> HM2=segment_count(C);
   For each key $k$ in HM2:
      If ($HM2(k) > th1$ and $HM1(k) - HM2(k) < th2$)
         $\{CWs\} = \{CWs\} \cup \{k\}$
    End For
  End
```

Segments words for all abstract texts, calculated these words’ frequencies, and store them in a hashmap named HM1; Segments words for all annotated effect clauses, calculated these words’ frequencies, and store them in another hashmap named HM2; if a word is frequent in effect clauses, but not common in non-effect clauses, we regard it as candidate clue words.

4.2 Locating Candidate Effect Statements

This kind of idea is similar to chain extraction in the authors’ previous research. But there are some differences. In chain extraction, if two nonadjacent clauses are both effect clauses, then all clauses between them were also determined to be effect clauses. This approach can extract multiple effect statements one-time, but not suitable for our situation. Since the initial clue words set is very small, if we use this set to match in the abstract, it is probably there is only one or even no clause matches. In this paper, the aim we use clue words to match in new patent abstracts is to find more clue words. Here we use a simple and direct way: in a new patent abstract text, the clauses containing clue words and the clauses before and after them are extracted, which are seemed as candidate effect clauses. The algorithm of locating candidate effect statements is as follows.

```
Algorithm: locating candidate effect statements in a new
patent
Input:  a set of Clue Words, denoted as CWs; a new patent
abstract text, denoted as AT;
Output: candidate effect clauses, denoted as CECs
Begin
  String regex=”，|。|；|,|\\.|;|：|:|\\s+|\\?”;
  ArrayList<String> clauses=seperate2Clauses(AT, regex);
  For each clause c in clauses:
    If(c contains any word in CWs)
      $CECs = CECs \cup \{c\}$
    If(c is not the first clause in AT)
      $CECs = CECs \cup \{c_{-1}\}$
    If(c is not the last clause in AT)
      $CECs = CECs \cup \{c_{+1}\}$
  End For
End
```

In the algorithm, the abstract text is divided into clauses with punctuation. Those clauses containing clue words and the clauses before and behind them are extracted as candidate effect clauses. c_{-1} refers to the clause before c, and c_{+1} refers to the clause behind c.

5 Experiments

We use 40,000 patent abstract texts from Chinese universities and research institutions as the data source, and implement the algorithm to collect clue words.

5.1 Collection of Initial Clue Words

First of all, we manually annotate effect clauses of 300 patent abstracts, and use self-filtering to find the initial high-quality clue words. Table 2 shows the influence of different input parameters on the number and accuracy of clue words. As can be seen from Table 2, th1 and th2 jointly determine the number of clue words after filtering. When th2 is fixed, the smaller th1 is, the more the number of clue words will be returned after filtering. When th1 = 20 and th2 = 35, the accuracy of clue words is the highest. When th1 = 15 and th2 = 35, the number of clue words after manual screening is the highest. We selected these 30 clue words as the initial clue words set.

Table 2. The influence of different input parameters on the number and accuracy of clue words

th1	th2	Number of clue words after self-filtering	Number of clue words after manual screening	Precision
25	40	19	15	78.9%
30	40	14	11	78.6%
20	35	26	21	80.8%
20	30	20	16	80.0%
15	35	43	30	69.8%

5.2 Iteration

We use the 30 clue words gotten in Sect. 5.1 to locate candidate effect statements in more patent abstract texts. After several rounds of iteration, clue words thesaurus is constantly enriched. As shown in Table 3.

Table 3. Results of each iteration

	Number of patents	Size of clue words thesaurus
Initial	300	30
Round 1	1000	37
Round 2	2000	61
Round 3	3000	108
Round 4	4000	134
Round 5	5000	155

6 Conclusion and Future Work

In order to make the annotation of effect statements in Chinese patents more accurate, this paper proposes a construction method of clue words thesaurus based on iteration and self-filtering. Making use of frequency differences between clues words in effect statements and non-effect statements, we achieve clue words' filtration. Making use of the characteristic that effect statements often appear continuously, we locate three candidate effect clauses through a clue word. The recognition of clue words and the extraction of effect statements carry on alternatively and gradually enriches clues words thesaurus. The method of this paper is based on the characteristics of abstract text and clue word. It is simple, easy to implement and has satisfying experiment results. Future research may focus on the extraction of technical words from patents.

Acknowledgments. This paper was supported by Research Foundation for Advanced Talents of Hubei University of Technology (No. BSQD12131) and the Fundamental Research Funds for the Young Teachers' Innovation project of Zhongnan University of Economics and Law (No. 2014147).

References

1. Zhao, F., Jianhong, M.A.: Method of automatic annotation information for patents. J. Chongqing Univ. Posts Telecommun. (2015)
2. Wang, F., Lin, L.F., Yang, Z.: An ontology-based automatic semantic annotation approach for patent document retrieval in product innovation design. Appl. Mech. Mater. **446–447**, 1581–1590 (2013)
3. Jiang, C.: Automatic annotation of bibliographical references in Chinese patent documents. New Technol. Libr. Inf. Serv. (2015)
4. Kim, Y., Ryu, J., Myaeng, S.H.: A patent retrieval method using semantic annotations. In: Kdir 2009 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, Funchal, Madeira, Portugal, October, pp. 211–218. DBLP (2009)
5. He, Y., Li, Y., Meng, L.: A new method of creating patent technology-effect matrix based on semantic role labeling. In: International Conference on Identification, Information, and Knowledge in the Internet of Things, pp. 58–61. IEEE (2015)
6. Chen, Y.: Research of patent technology-effect matrix construction based on feature degree and lexical model. New Technol. Libr. Inf. Serv. (2012)
7. Chen, X., Deng, N.: A semi-supervised machine learning method for Chinese patent effect annotation. In: International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 243–250. IEEE Computer Society (2015)
8. Chen, X., Peng, Z., Zeng, C.: A co-training based method for Chinese patent semantic annotation. In: ACM International Conference on Information and Knowledge Management, pp. 2379–2382. ACM (2012)
9. Deng, N., Chen, X.: Automatically generation and evaluation of Stop words list for Chinese Patents. Telkomnika **13**(4), 1414 (2015)
10. Deng, N., Chen, X., Li, D.: Intelligent recommendation of Chinese traditional medicine patents supporting new medicine's R&D. J. Comput. Theor. Nanosci. **13**, 5907–5913 (2016)

Numerical Simulation for the Nonlinear Elliptic Problem

Qingli Zhao[1(✉)], Jin Li[1], and Lei Yang[2]

[1] School of Science, Shandong Jianzhu University, Jinan, China
shizilu@126.com, lijin@lsec.cc.ac.cn
[2] College of Science, China University of Petroleum, Qingdao, China
yanglei1021@upc.edu.cn

Abstract. With the rapid development of computing power, numerical simulation has become useful and powerful tool. Expanded mixed finite element method is introduced to solve the nonlinear elliptic problem in divergence form. Existence and uniqueness of the discrete problem are demonstrated. Optimal L^2-error estimates for three variables are got. Numerical examples are provided to validate the theoretical analysis.

1 Introduction

High-performance computing is developing quickly recently. High-performance computing is an important field with two branches: numerical simulations and big data analysis. It is well known that modeling and simulation technology is a method which is often used to analyze complex problems. Numerical simulation based on finite element methods is a major applications requiring high performance computing with floating-point operations, which includes fluid dynamics problems, electromagnetic problems and so on. These simulations methods solve the differential equations models of complex physics or mechanics problems, which are approximated by applying the numerical schemes with discrete values defined at grid points. Advanced modeling and numerical simulation technology [19] has become useful and powerful tool. In the simulation of complex physical and mechanics phenomenon, differential equation models are widely used. In this article, we consider the following nonlinear elliptic problem

$$\begin{cases} -\nabla \cdot a(u, \nabla u) = f(x), & x \in \Omega, \\ u = -g, & x \in \partial\Omega \end{cases} \tag{1}$$

in a bounded domain $\Omega \subset R^2$ with sufficiently smooth boundary $\partial\Omega$. The function $a = (a_1, a_2)^T : R \times R^2 \to R^2$ is twice continuously differentiable with bounded derivatives through the second order. It is known that the standard mixed finite element method computes both the scalar unknown and vector variables at the same time. Mixed method for linear and semi-linear second-order elliptic problem has received considerable attention [11,18]. In [13], Milner has studied the mixed method for quasilinear second-order elliptic problem.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_13

The mixed method of nonlinear elliptic problems has been investigated in [12,14,16]. The case of a only depend on ∇u which based on BDM element has been studied in [6]. The expanded mixed finite element method (EMFEM) is constructed by introducing three (or more) unknown variables. Chen [7,8] has studied the expanded mixed finite element method for linear and quasilinear second-order elliptic problems. Similar technique was presented by Arbogast [2,3] to develop the cell-centered finite differences scheme. We aim to present expanded mixed element method for the nonlinear elliptic problem in divergence form.

The paper is organized as follows. In Sect. 2, notation and weak form are given. In Sects. 3 and 4, we analysis the existence and uniqueness. In Sect. 5, optimal L^2 error estimates are established. In Sect. 6, numerical examples are carried out. Throughout this paper, C will denote a generic positive constant.

2 Notation, Weak Formulation and Approximation

Assume that the nonlinear operator associated with (1) is elliptic in the sense that matrix $A_{2\times 2} = [a_{ij}(u, \mathbf{z})]_{i,j=1,2} = [\partial a_i/\partial z_j]_{i,j=1,2}$ is symmetric and positive definite. If $\lambda_{min}, \lambda_{max}$ denote, respectively, the minimum and maximum eigenvalue of A, then, for all $\zeta = (\zeta_1, \zeta_2)^T \neq 0$ and $(u, \mathbf{z}) \in R \times R^2$,

$$0 < \lambda_{min}(u, \mathbf{z})|\zeta|^2 \leq \zeta^T A \zeta \leq \lambda_{max}(u, \mathbf{z})|\zeta|^2. \tag{2}$$

For $1 \leq q < \infty$ and k any nonnegative integer, let

$$W^{k,q}(\Omega) = \{f \in L^q(\Omega) \mid D^\alpha f \in L^q(\Omega),\ |\alpha| \leq k\}$$

denote the Sobolev spaces [1] endowed with the norm

$$\|f\|_{k,q,\Omega} = \left(\sum_{|\alpha|\leq k} \|D^\alpha f\|^q_{L^q(\Omega)} \right)^{1/q}.$$

The subscript Ω will always be omitted. Let $H^k(\Omega) = W^{k,2}(\Omega)$ with the norm $\|\cdot\|_k = \|\cdot\|_{k,2}$. The notation $\|\cdot\|$ mean $\|\cdot\|_{L^2(\Omega)}$ or $\|\cdot\|_{L^2(\Omega)^2}$. For $0 \leq s < +\infty$, let $W^{s,q}(\Omega)$, $W^{s,q}(\partial\Omega)$, $H^s(\Omega)$ and $H^s(\partial\Omega)$ denote the fractional-order Sobolev spaces with the norms $\|\cdot\|_{s,q,\Omega}$, $\|\cdot\|_{s,q,\partial\Omega}$, $\|\cdot\|_{s,\Omega}$ and $\|\cdot\|_{s,\partial\Omega}$. Denote by $(,)$ the inner product in either $L^2(\Omega)$ or $(L^2(\Omega))^2$, that is $(\xi, \eta) = \int_\Omega \xi\eta dx$, $\quad (\xi, \eta) = \int_\Omega \xi \cdot \eta dx$. The notation $\langle , \rangle$ means the inner product on the boundary: $\langle \xi, \eta \rangle = \int_{\partial\Omega} \xi\eta ds$. Let

$$\begin{aligned} &W = L^2(\Omega), \qquad \Lambda = (L^2(\Omega))^2, \\ &V = \mathbf{H}(div; \Omega) = \{v = (v_1, v_2)^T \in (L^2(\Omega))^2 \mid div\ v \in L^2(\Omega)\}, \\ &\mathbf{H}^s(div; \Omega) = \{v = (v_1, v_2)^T \in (L^2(\Omega))^2 \mid div\ v \in H^s(\Omega)\}. \end{aligned} \tag{3}$$

Define $\|v\|_{\mathbf{V}} = \|v\| + \|\nabla \cdot v\|$, $\|v\|_{\mathbf{H}^s(div;\Omega)} = \|v\| + \|\nabla \cdot v\|_s$. Let us introduce two variables $\lambda = \nabla u$, $\sigma = -a(u, \nabla u) = -a(u, \lambda)$, then we get the weak form: find $(u, \lambda, \sigma) \in W \times \Lambda \times V$, such that

$$(\sigma, \mu) + (a(u, \lambda), \mu) = 0, \quad \forall \mu \in \Lambda, \tag{4}$$

$$(\lambda, v) + (u, div\, v) = < g, v \cdot \gamma >, \quad \forall v \in V, \tag{5}$$

$$(div\ \sigma, \omega) = (f, \omega), \quad \forall \omega \in W. \tag{6}$$

Let $W_h(E) \times \Lambda_h(E) \times V_h(E)$ denote the mixed finite element spaces introduced in [4,5,10,15,17]. Some 2D RT type mixed elements are listed in Table 1.

Table 1. Some RT type mixed elements

Dim	Element	$V_h(E)$	$W_h(E)$
$2D$	Triangle	$RT_k(T) = P_k(E)^2 \oplus xP_k(E)$	$P_k(E)$
$2D$	Rectangle	$RT_{[k]}(T) = Q_{k+1,k}(E) \oplus Q_{k,k+1}(E)$	$Q_{k,k}(E)$

Choose

$$\begin{aligned} W_h(E) &= \{\omega \in W : \boldsymbol{\omega}|_E \in W_h(E), \quad \text{for each } E\}, \\ \Lambda_h(E) &= \{\mu \in \Lambda : \boldsymbol{\mu}|_E \in V_h(E), \quad \text{for each } E\}, \\ V_h(E) &= \{v \in V : v|_E \in V_h(E), \quad \text{for each } E\}. \end{aligned}$$

The discrete formulation of (4)–(6) is to find $(u_h, \lambda_h, \sigma_h) \in W_h \times \Lambda_h \times V_h$ such that

$$(\sigma_h, \mu) + (a(u_h, \lambda_h), \mu) = 0, \quad \forall \mu \in \Lambda_h, \tag{7}$$

$$(\lambda_h, v) + (u_h, div\ v) =< g, v \cdot \gamma >, \quad \forall v \in V_h, \tag{8}$$

$$(div\ \sigma_h, \omega) = (f, \omega), \quad \forall \omega \in W_h. \tag{9}$$

The error analysis below will make use of three projections as follows:

(I). The Raviart-Thomas-Nedelec projection $\pi_h : (H^r(\Omega))^2 \rightarrow V_h$ satisfies

$$(\nabla \cdot (v - \pi_h v), \omega)) = 0, \quad \forall \omega \in W_h,$$

Operator π_h has the approximation properties:

$$\begin{aligned} \|v - \pi_h v\|_r &\leq Ch^r \|v\|_r, \quad 1 \leq r \leq k+1, \\ \|\nabla \cdot (v - \pi_h v)\|_r &\leq Ch^r \|\nabla \cdot v\|_r, \quad 0 \leq r \leq k+1. \end{aligned}$$

(II). The standard L^2 projection P_h and R_h onto W_h and Λ_h, respectively

$$\begin{aligned} (\boldsymbol{\omega} - P_h \boldsymbol{\omega}, \nabla \cdot v) &= 0, \quad \forall \boldsymbol{\omega} \in W, \quad v \in V_h, \\ (\boldsymbol{\mu} - R_h \mu, \tau) &= 0, \quad \forall \boldsymbol{\mu} \in \Lambda, \quad \tau \in \Lambda_h. \end{aligned}$$

They have the approximation properties:

$$\|\boldsymbol{\omega} - P_h \boldsymbol{\omega}\|_{-s} \leq Ch^{r+s} \|\omega\|_r, \|\mu - R_h \mu\|_{-s} \leq Ch^{r+s} \|\mu\|_r, \ 0 \leq r, s \leq k+1.$$

3 Analysis of the Linearized Problem

We derive from (4)–(9) the following error equations

$$(a(u,\lambda)-a(u_h,\lambda_h),\mu)+(\sigma-\sigma_h,\mu)=0,\quad \forall\mu\in\Lambda_h, \tag{10}$$

$$(\lambda-\lambda_h,v)+(u-u_h,div\ v)=0,\quad \forall v\in V_h, \tag{11}$$

$$(div\ (\sigma-\sigma_h),\omega)=0,\quad \forall\omega\in W_h. \tag{12}$$

For $a=(a_1,a_2)^T$, we adopt the following integral form of Taylor's expansion

$$\begin{aligned} a(u_2,\lambda_2)-a(u,\lambda) &= -a_u(u,\lambda)(u-u_2)-a_\lambda(u,\lambda)(\lambda-\lambda_2)\\ &\quad +Q\left(u-u_2,\lambda-\lambda_2\right)\\ &= -\tilde{a}_u(u,\lambda)(u-u_2)-\tilde{a}_\lambda(u,\lambda)(\lambda-\lambda_2). \end{aligned} \tag{13}$$

Here we have used the following notations

$$\tilde{a}_u(u_2,\lambda_2)=\int_0^1 a_u(u_2^t,\lambda_2^t)dt,\quad \tilde{a}_\lambda(u_2,\lambda_2)=\int_0^1 a_\lambda(u_2^t,\lambda_2^t)dt,$$
$$Q(u-u_2,\lambda-\lambda_2)=\left\{Q_{a_1}\left(u-u_2,\lambda-\lambda_2\right),Q_{a_2}\left(u-u_2,\lambda-\lambda_2\right)\right\}^T\in R^2.$$

with $u_2^t=u+t(u_2-u),\quad \lambda_2^t=\lambda+t(\lambda_2-\lambda),\ 0\le t\le 1$. Note that $a_u\in(L^\infty(\Omega))^2,\quad a_\lambda\in(L^\infty(\Omega))^{2\times 2}$, combining (10)–(13), we obtain

$$(a(u,\lambda)-a(u_h,\lambda_h),\mu)+(\sigma-\sigma_h,\mu)=0,\quad \forall\mu\in\Lambda_h, \tag{14}$$

$$(\lambda-\lambda_h,v)+(u-u_h,div\ v)=0,\quad \forall v\in V_h, \tag{15}$$

$$(div\ (\sigma-\sigma_h),\omega)=0,\quad \forall\omega\in W_h. \tag{16}$$

Let $T_1=a_u(u,\lambda)\in R^2,\ T_2=a_\lambda(u,\lambda)\in R^{2\times 2}$, then (14) can be rewritten as

$$(T_1(u-u_h),\mu)+(T_2(\lambda-\lambda_h),\mu)+(\sigma-\sigma_h,\mu)=\left(Q(u-u_h,\lambda-\lambda_h),\mu\right). \tag{17}$$

Let $\Phi: W_h\times\Lambda_h\to W_h\times\Lambda_h$ be given by $\Phi(\bar{x},\bar{y})=(\overline{xx},\overline{yy})$, here $(\overline{xx},\overline{yy})$ is the solution of the following equations

$$(T_1(P_hu-\bar{x}),\mu)+(T_2(R_h\lambda-\bar{y}),\mu)+(\pi_h\sigma-\sigma_h),\mu)=(l,\mu),\ \forall\mu\in\Lambda_h, \tag{18}$$

$$(\lambda-\bar{y},v)+(u-\bar{x},div\ v)=0,\quad \forall v\in V_h, \tag{19}$$

$$(div(\sigma-\sigma_h),\omega)=0,\quad \forall\omega\in W_h, \tag{20}$$

where $l=T_1(P_hu-u)+T_2(R_h\lambda-\lambda)+Q(u-\overline{xx},\lambda-\overline{yy})+(\pi_h\sigma-\sigma)$. Define $\xi=P_hu-\bar{x},\quad \eta=R_h\lambda-\bar{y},\quad \phi=\pi_h\sigma-\sigma_h$, then (18)–(20) can be rewritten as below

$$(T_1\xi,\mu)+(T_2\eta,\mu)+(\phi,\mu)=(l,\mu),\quad \forall\mu\in\Lambda_h, \tag{21}$$

$$(\eta,v)+(\xi,div\ v)=0,\quad \forall v\in V_h, \tag{22}$$

$$(div\ \phi,\omega)=0,\quad \forall\omega\in W_h. \tag{23}$$

Lemma 1. *Let $\phi \in V,\ \eta,\ l \in (L^2(\Omega))^2$, if $\xi \in W_h$, then there exists a constant $C > 0$ such that*

$$\|\xi\| \leq C\left\{h(\|\eta\| + \|\phi\|) + h^2\|div\ \phi\| + \|l\|\right\}. \tag{24}$$

Proof. Now define M^* by $M^*\varphi = -div(T_2\nabla\varphi) + T_1 \cdot \nabla\varphi = \psi,\ \ in\ \Omega,\ \ \varphi|_{\partial\Omega} = 0$. It follows from [11] that the restrictions of the operator M^* to $H^2(\Omega) \cap H_0^1(\Omega)$ have bounded inverses; that is, for any $\psi \in L^2(\Omega)$, there is a unique $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$ such that $M^*\varphi = \psi$ and $\|\varphi\|_2 \leq C\|\psi\|$. Recall that $\|q\| = \sup\limits_{\psi \in L^2(\Omega), \psi \neq 0} \dfrac{(q,\psi)}{\|\psi\|}$. Employ a duality argument, we have

$$\begin{aligned}(\xi,\psi) = (\xi, M^*\varphi) &= (\xi, -\nabla\cdot(T_2\nabla\varphi) + T_1\nabla\varphi)\\ &= (\eta, \pi_h(T_2\nabla\varphi)) + (T_1\xi, \nabla\varphi - R_h\nabla\varphi) + (T_1\xi, R_h\nabla\varphi)\\ &= \Sigma_{j=1}^4 I_j.\end{aligned} \tag{25}$$

Using the properties of three projections, we get

$$\begin{aligned}I_1 &= (\eta, \pi_h(T_2\nabla\varphi) - T_2\nabla\varphi) + (T_2\eta, \nabla\varphi - R_h\nabla\varphi) \leq Ch\|\varphi\|_2\|\eta\|,\\ I_2 &= (T_1\xi, \nabla\varphi - R_h\nabla\varphi) - (\phi, R_h\nabla\varphi - \nabla\varphi) \leq Ch\left(\|\varphi\|_2\|\xi\| + \|\varphi\|_2\|\phi\|\right),\\ I_3 &= (\phi, \nabla\varphi) = (\nabla\phi, \varphi - P_h\varphi) \leq Ch^2\|\nabla\varphi\|\|\varphi\|_2,\\ I_4 &= (l, R_h\nabla\varphi - \nabla\varphi) + (l, \nabla\varphi) \leq C\|\varphi\|_2\|l\|.\end{aligned}$$

By substituting these inequalities into (25), we complete the proof. □

Lemma 2. *Let $\xi \in W_h, \eta \in \Lambda_h, \phi \in V_h$, then there exists a constant $C > 0$ such that*

$$\|\eta\| \leq C(\|\xi\| + \|l\|).$$

Proof. Choose $v = \phi,\ \xi = div\ \phi,\ \mu = \phi$ in (21)–(23), we obtain $(T_1\xi, \phi) + (T_2\eta, \eta) = (l, \eta)$. Hence, we know $\|\eta\| \leq C(\|\xi\| + \|l\|)$. □

Lemma 3. *Let $\xi \in W_h,\ \eta \in \Lambda_h,\ \phi \in V_h$, then there exists a constant $C > 0$ such that*

$$\|\phi\| \leq C(\|\xi\| + \|\eta\| + \|l\|).$$

Proof. Let $\eta = \phi$ in (21), we get $(T_1\xi, \phi) + (T_2\eta, \phi) + (\phi, \phi) = (l, \eta)$. So we obtain $\|\phi\| \leq C(\|\xi\| + \|\eta\| + \|l\|)$. □

Theorem 1. *There exists one and only one solution of the system (21)–(23).*

Proof. Existence follows from uniqueness since the system is linear. Assume $l = 0$, then implies $\|\xi\| \leq Ch\|\xi\|$. For sufficiently small h, we have $\|\xi\| = 0$, which implies $\xi = 0$, along with η and ϕ. This completes the proof. □

4 Existence and Uniqueness

Theorem 2. *For sufficiently small h, $\varPhi$ defined in (18)–(20) has a fixed point.*

To prove Theorem 2, we give the following Lemma 4.

Lemma 4. *Let $\omega \in L_2(\varOmega)$, $l \in L_2(\varOmega)^2$, if $\xi \in W_h$ satisfies the relations (21)–(23), then there exists a constant $C > 0$, independent of h, such that for h sufficiently small and $2 < \theta < +\infty$,*

$$\|\xi\|_{0,\theta} \leq C\left\{h^{\frac{2}{\theta}}(\|\eta\| + \|\phi\|) + h^{1+\frac{2}{\theta}}\|div\ \phi\| + \|l\|\right\}.$$

Proof. The proof is analogous to our Lemma 1 and Lemma 2.1 of [13]. □

Now let $\overline{V}_h = V_h$ be endowed with the norm $\|v\|_{\overline{\mathbf{V}}_h} = \|v\|_{0,4+\varepsilon} + \|div\ v\|$, and $\overline{W}_h = W_h$ with the norm $\|\omega\|_{\overline{W}_h} = \|\omega\|_{0,4+\varepsilon}$. It follows from the Brouwer's fixed point theorem that Theorem 2 is true if we can show the following result.

Theorem 3. *For $\delta > 0$ sufficiently small (dependent on h), $\varPhi$ maps the ball of radius of δ of $W_h \times \varLambda_h$, centered at $(P_h u, R_h\lambda)$, into itself.*

Proof. Let $\|P_h u - \overline{xx}\| \leq \delta$, $\|R_h\lambda - \overline{yy}\| \leq \delta$, $\|\pi_h\sigma - \sigma_h\| \leq \delta$, and $s = 1 + \varepsilon + \frac{\varepsilon}{4+\varepsilon}$, $\theta = 4+\varepsilon$. Then $s - \dfrac{2}{\theta} = \dfrac{1}{2} + \varepsilon_0$, when $0 < \varepsilon_0 = \varepsilon + \dfrac{3\varepsilon}{8+2\varepsilon} \ll 1$. The Sobolev embedding theorem implies that $H^{\frac{3}{2}+\varepsilon_0}(\varOmega) \subset W^{s,\theta}(\varOmega)$, $\quad \|\chi\|_{s,\theta} \leq C\|\chi\|_{\frac{3}{2}+\varepsilon_0}$. Applying Lemma 4 with $\xi = P_h u - \overline{x}$, $\quad \eta = R_h\lambda - \overline{y}$, $\quad \phi = \pi_h\sigma - \sigma_h$ and

$$l = T_1(P_h u - u) + T_2(R_h\lambda - \lambda) + Q(u - \overline{xx}, \lambda - \overline{yy}) + (\pi_h\sigma - \sigma),$$

then we know

$$\begin{aligned}
\|P_h u - \overline{x}\|_{0,\theta} \leq{}& C(h^{\frac{2}{\theta}}\|\pi_h\sigma - \sigma_h\| + h^{1+\frac{2}{\theta}}\|div\ (\pi_h\sigma - \sigma_h)\| \\
&+ h^{\frac{2}{\theta}}\|R_h\lambda - \overline{y}\| + h^s\|\lambda\|_s + h^s\|u\|_s + \|u - \overline{xx}\|_{0,4}^2 \\
&+ \|u - \overline{xx}\|_{0,4}\|\lambda - \overline{yy}\|_{0,4} + \|\lambda - \overline{yy}\|_{0,4}^2) \\
\leq{}& C(h^{\frac{2}{\theta}}\|\pi_h\sigma - \sigma_h\| + h^{\frac{2}{\theta}}\|R_h\lambda - \overline{y}\| + h^s\|\lambda\|_s + h^s\|u\|_s \\
&+ \|u - P_h u\|_{0,\theta}^2 + \|P_h u - \overline{xx}\|_{0,\theta}^2 \\
&+ \|\lambda - R_h\lambda\|_{0,\theta}^2 + \|R_h\lambda - \overline{yy}\|_{0,4}^2) \\
\leq{}& C(h^{\frac{2}{\theta}}\|\pi_h\sigma - \sigma_h\| + h^{\frac{2}{\theta}}\|R_h\lambda - \overline{y}\| \\
&+ (h^s + \delta^2)(1 + \|u\|_{s,\theta} + \|\lambda\|_{s,\theta})^2)
\end{aligned} \tag{26}$$

Next, using Lemma 2 and Sobolev embedding theorem, we get

$$\|\pi_h\sigma - \sigma_h\| \leq C\left\{\|P_h u - \overline{x}\| + (h^s + \delta^2)(1 + \|u\|_{\frac{3}{2}+\varepsilon_0} + \|\lambda\|_{\frac{3}{2}+\varepsilon_0})^2\right\}, \tag{27}$$

$$\|R_h\lambda - \overline{y}\| \leq C\left\{\|P_h u - \overline{x}\| + (h^s + \delta^2)(1 + \|u\|_{\frac{3}{2}+\varepsilon_0} + \|\lambda\|_{\frac{3}{2}+\varepsilon_0})^2\right\}. \tag{28}$$

Hence (26)–(28) imply that for sufficiently small h, we know $\|P_h u-\overline{x}\|_{0,\theta} \le K_1(h^2+\delta^2)$. Note that the following inverse estimates conclusion [9]: for $0 \le \nu \le \theta$,

$$\|\pi_h\sigma-\sigma_h\|_{0,\theta} \le Ch^{\frac{2}{\theta}-\frac{2}{\nu}}\|\pi_h\sigma-\sigma_h\|_{0,\nu}, \tag{29}$$

$$\|R_h\lambda-\overline{y}\|_{0,\theta} \le Ch^{\frac{2}{\theta}-\frac{2}{\nu}}\|R_h\lambda-\overline{y}\|_{0,\nu}, \tag{30}$$

$$\|P_h u-\overline{x}\|_{0,\theta} \le Ch^{\frac{2}{\theta}-\frac{2}{\nu}}\|P_h u-\overline{x}\|_{0,\nu}, \tag{31}$$

combining (27)–(31), we see

$$\|\pi_h\sigma-\sigma_h\|_{0,\theta} \le Ch^{\frac{2}{4+\varepsilon}-1}(h^s+\delta^2) = C(h^{s-\frac{2+\varepsilon}{4+\varepsilon}}+h^{-\frac{2+\varepsilon}{4+\varepsilon}}\delta^2). \tag{32}$$

Then the choice $\delta = 2K_3 h^{\frac{1}{2}+\varepsilon+\frac{\varepsilon}{8+2\varepsilon}}$ leads to the bound $\|\pi_h\sigma-\sigma_h\| \le \delta$. Similarly, (30)–(31) imply $\|P_h u-\overline{x}\| \le \delta$ *and* $\|R_h\lambda-\overline{y}\| \le \delta$, which complete the proof. □

Theorem 4. *For sufficiently small h, there is a unique solution of (7)–(9) near the solution $\{u,\lambda,\sigma\}$ of (4)–(6).*

Proof. Let $(u_h^{(i)},\lambda_h^{(i)},\sigma_h^{(i)}) \in W_h \times \Lambda_h \times V_h,\ i=1,2,$ be the solution of (7)–(9), and

$$\overline{u}=u_h^{(1)}-u_h^{(2)},\quad \overline{\lambda}=\lambda_h^{(1)}-\lambda_h^{(2)},\quad \overline{\sigma}=\sigma_h^{(1)}-\sigma_h^{(2)},$$
$$\xi^{(i)}=u-u_h^{(i)},\quad \eta^{(i)}=\lambda-\lambda_h^{(i)},\quad \phi^{(i)}=\sigma-\sigma_h^{(i)},\quad i=1,2.$$

Note that (13)–(16), we obtain

$$\begin{aligned}&(T_1\overline{u},\mu)+(T_2\overline{\lambda},\mu)+(\overline{\sigma},\mu)\\ &\qquad\qquad\qquad= (Q(\xi^{(2)},\eta^{(2)})-Q(\xi^{(1)},\eta^{(1)}),\mu),\quad \mu\in\Lambda_h,\\ &(\overline{\lambda},v)+(\overline{u},div\ v)=0,\ v\in V_h.\ \ (div\ \overline{\sigma},\omega)=0,\ \omega\in W_h.\end{aligned}$$

Then from Lemma 2, we know

$$\|\overline{\lambda}\|,\|\overline{\sigma}\| \le C\left\{\|\overline{u}\|+\|Q(\xi^{(2)},\eta^{(2)})-Q(\xi^{(1)},\eta^{(1)})\|\right\}. \tag{33}$$

By Lemma 1, we have

$$\|\overline{u}\| \le C\left\{h(\|\overline{\lambda}\|+\|\overline{\sigma}\|)+\|Q(\xi^{(2)},\eta^{(2)})-Q(\xi^{(1)},\eta^{(1)})\|\right\}. \tag{34}$$

Note that the following inequalities holds [16]:

$$\|Q(\xi^{(2)},\eta^{(2)})-Q(\xi^{(1)},\eta^{(1)})\| \le Ch^{\frac{1}{2}\varepsilon_0}\left(\|\overline{u}\|+\|\overline{\lambda}\|\right). \tag{35}$$

Combining (33)–(35), we see that for sufficiently small h, $\|\overline{\lambda}\| \le C\|\overline{u}\| \le Ch^{\frac{1}{2}\varepsilon_0}\|\lambda\|$, which forces $\overline{u}=0$ and $\overline{\lambda}=0$, so we have $\overline{\sigma}=0$ and complete the proof. □

5 L^2-Error Estimates

Theorem 5. *Let $k > 0$, assume that the solution of (4)–(6) is sufficiently smooth, then for sufficiently small h there is a constant $C > 0$ such that*

$$\begin{aligned}
\|u-u_h\| &\le Ch^\alpha(\|u\|_\alpha+\|\lambda\|_\alpha+\|\sigma\|_\alpha),\quad 1\le\alpha\le k+1,\\
\|\lambda-\lambda_h\| &\le Ch^\alpha(\|u\|_\alpha+\|\lambda\|_\alpha+\|\sigma\|_\alpha),\quad 1\le\alpha\le k+1,\\
\|\sigma-\sigma_h\| &\le Ch^\alpha(\|u\|_\alpha+\|\lambda\|_\alpha+\|\sigma\|_\alpha),\quad 1\le\alpha\le k+1,\\
\|div(\sigma-\sigma_h)\| &\le Ch^\alpha\left(\|u\|_{\alpha+1}+\|\lambda\|_{\alpha+1}+\|\sigma\|_{\alpha+1}\right),\quad 0\le\alpha\le k+1.
\end{aligned}$$

Proof. Let $\xi = P_h u - u_h,\ \eta = R_h\lambda - \lambda_h, \phi = \pi_h\sigma - \sigma_h$. From Lemma 3, we see

$$\begin{aligned}
\|\phi\| &\le C(\|\xi\|+\|u-P_hu\|+\|\lambda-R_h\lambda\|\\
&\quad+\|\sigma-\pi_h\sigma\|+\|Q(u-u_h,\lambda-\lambda_h)\|)\\
&\le C(\|\xi\|+\|u-P_hu\|_0+\|\lambda-R_h\lambda\|\\
&\quad+\|\sigma-\pi_h\sigma\|+\|u-u_h\|_{0,4}^2+\|\lambda-\lambda_h\|_{0,4}^2)\\
&\le C(\|\xi\|+\|u-P_hu\|+\|u-P_hu\|_{0,4}^2\\
&\quad+\|\lambda-R_h\lambda\|+\|\lambda-R_h\lambda\|_{0,4}^2+\|\sigma-\pi_h\sigma\|+\|\eta\|_{0,4}^2+\|\xi\|_{0,4}^2)\\
&\le C(\|\xi\|+h^{-1}\|\xi\|^2+h^{-1}\|\eta\|^2\\
&\quad+\|u-P_hu\|+\|u-P_hu\|_{0,4}^2+\|\lambda-R_h\lambda\|\\
&\quad+\|\lambda-R_h\lambda\|_{0,4}^2+\|\sigma-\pi_h\sigma\|).
\end{aligned}\tag{36}$$

Using the properties of three projection operators, we get

$$\begin{aligned}
\|\phi\| &\le C(\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\eta\|\\
&\quad+h^{2\gamma-1}\|u\|_{\gamma-\frac{1}{2},4}^2++h^r\|u\|_r+h^{2\beta-1}\|u\|_{\beta-\frac{1}{2},4}^2+h^\alpha\|\lambda\|_\alpha+h^\alpha\|\sigma\|_\alpha)\\
&\qquad(\frac{1}{2}\le\gamma\le k+\frac{3}{2},0\le r\le k+1,\frac{3}{4}\le\beta\le k+\frac{3}{2},\frac{1}{2}\le\alpha\le k+1)\\
&\le C(\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\eta\|\\
&\quad+h^{2\gamma-1}\|u\|_{\gamma-\frac{1}{2},4}^2+h^r\|u\|_r+h^{2\beta-1}\|u\|_{\beta-\frac{1}{2},4}^2+h^\alpha\|\lambda\|_\alpha+h^\alpha\|\sigma\|_\alpha)\\
&\le C(\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\eta\|\\
&\quad+h^{2\gamma-1}\|u\|_\gamma^2+h^r\|u\|_r+h^{2\beta-1}\|u\|_\beta^2+h^\alpha\|\lambda\|_\alpha+h^\alpha\|\sigma\|_\alpha)\\
&\le C\{\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\xi\|+Kh^{\frac{1}{2}\varepsilon_0}\|\eta\|\\
&\quad+h^r\|u\|_r(\|u\|_{\frac{r}{2}+1}+1)+h^\alpha(\|\lambda\|_\alpha+\|\sigma\|_\alpha)(\|\lambda\|_{\frac{r}{2}+1}+1)\}.
\end{aligned}\tag{37}$$

From Lemma 1, we know

$$\|\phi\|\le C(\|u\|_{\frac{r}{2}+1}+\|\lambda\|_{\frac{r}{2}+1}+1)\left(h^r\|u\|_r+h^\alpha(\|\lambda\|_\alpha+\|\sigma\|_\alpha)\right).\tag{38}$$

So we have $\|\phi\| \le Ch^\alpha(\|u\|_r + \|\lambda\|_\alpha + \|\sigma\|_\alpha),\ 1 \le \alpha \le k + 1$. Applying Lemmas 1 and 2, we get

$$\|\xi\| \leq Ch^{\alpha}(\|u\|_r + \|\lambda\|_{\alpha} + \|\sigma\|_{\alpha}), \quad 1 \leq \alpha \leq k+1, \tag{39}$$

$$\|\eta\| \leq Ch^{\alpha}(\|u\|_r + \|\lambda\|_{\alpha} + \|\sigma\|_{\alpha}), \quad 1 \leq \alpha \leq k+1, \tag{40}$$

which complete the proof. □

6 Numerical Examples

We carry out numerical examples using the RT^0 rectangular element.

Example 1. Let $s = \nabla u, a = s/\sqrt{1+|s|^2}$ and $\Omega = [0,\pi] \times [0,\pi]$. We consider the equation of prescribed mean curvature

$$\begin{cases} -\nabla \cdot \left(\dfrac{1}{\sqrt{1+|\nabla u|^2}} \nabla u \right) = f(x), & x \in \Omega, \\ u = 0, & x \in \partial\Omega. \end{cases} \tag{41}$$

The analytical solution is chosen to be $u = \sin x \sin y$. We use the maximum norm in error control of the Picard's iterative algorithm, define

$$eps = max\left\{ ({u_h}^n, \lambda_h^n, \sigma_h^n)^T - ({u_h}^{n+1}, \lambda_h^{n+1}, \sigma_h^{n+1})^T \right\}.$$

Let $eps = 10^{-4}$, the error results are listed in Table 2. It is easy to see the convergence rate is almost first order (Fig. 1).

Table 2. L^2-error of Example 1

Partition	$\|\|u - u_h\|\|$	*Rate*	$\|\|\lambda - \lambda_h\|\|$	*Rate*	$\|\|\sigma - \sigma_h\|\|$	*Rate*
2 × 2	0.9712	—	1.1084	—	0.8350	—
4 × 4	0.5082	0.93	0.5341	1.05	0.4254	0.97
8 × 8	0.2605	0.96	0.2611	1.03	0.2168	0.97
16 × 16	0.1318	0.98	0.1291	1.02	0.1093	0.99
32 × 32	0.0658	1.00	0.0639	1.01	0.0549	0.99

Example 2. Let $s = \nabla u, a = s/(1+|s|^2)$ and $\Omega = [0,\pi] \times [0,\pi]$. We consider the following nonlinear second-order Dirichlet problem

$$\begin{cases} -\nabla \cdot \left(\dfrac{1}{1+|\nabla u|^2} \nabla u \right) = f(x), & x \in \Omega, \\ u = 0, & x \in \partial\Omega. \end{cases} \tag{42}$$

The analytical solution is chosen to be $u = \sin x \sin y$. The error results are listed in Table 3 ($eps = 10^{-4}$). It is easy to see the convergence rate is almost first order (Fig. 2).

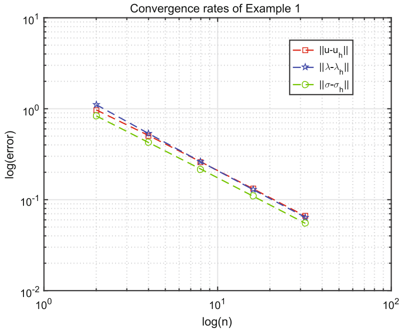


Fig. 1. Convergence rates of Example 1

Table 3. Error of Example 2

Partition	$\|\|u - u_h\|\|$	*Rate*	$\|\|\lambda - \lambda_h\|\|$	*Rate*	$\|\|\sigma - \sigma_h\|\|$	*Rate*
3 × 3	0.6576	—	0.6969	—	0.4246	—
6 × 6	0.3347	0.97	0.3414	1.03	0.2276	0.90
12 × 12	0.1678	1.00	0.1687	1.02	0.1127	1.01
24 × 24	0.0839	1.00	0.0840	1.01	0.0562	1.00
48 × 48	0.0420	1.00	0.0420	1.00	0.0281	1.00

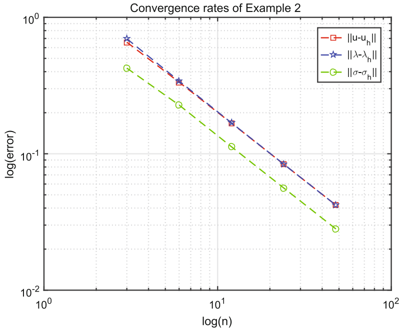


Fig. 2. Convergence rates of Example 2

7 Conclusion

We have shown the efficiency of expanded mixed element method for the nonlinear elliptic problem both theoretically and numerically. We get optimal-order error estimates for nonlinear problems. The proof of existence and uniqueness for the nonlinear discrete problem is given. In the future, we will research the related application in real life and nonlinear parabolic problem.

Acknowledement. The work of first author was supported by the Science and Technology Project for the Universities of Shandong Province (No. J14LI53), the Doctoral Foundation of Shandong Jianzhu University (No. XNBS1441). The work of second author was supported by NNSF of China (Nos. 11471195 and 11101247), and China Postdoctoral Science Foundation (Nos. 2015T80703 and 2013M540541). The work of third author was supported by the Fundamental Research Funds for the Central Universities (Nos. 15CX05065A, 15CX05003A).

References

1. Adams, R.A.: Sobolev Spaces. Academic Press, New York (1975)
2. Arbogast, T., Dawson, C.T., Keenan, P.T., Wheeler, M.F., Yotov, I.: Enhanced cell-centered finite differences for elliptic equations on general geometry. SIAM J. Sci. Comput. **19**, 402–425 (1998)
3. Arbogast, T., Wheeler, M.F., Yotov, I.: Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite difference. SIAM J. Numer. Anal. **34**, 828–852 (1997)
4. Brezzi, F., Douglas Jr., J., Duran, R., Fortin, M.: Mixed finite elements for second order elliptic problems in three variables. Numer. Math. **51**, 237–250 (1987)
5. Brezzi, F., Douglas Jr., J., Fortin, M., Marini, L.: Efficient rectangular mixed finite elements in two and three space variables. RAIRO Math. Model. Anal. Numer. **21**, 581–604 (1987)
6. Chen, Z.: BDM Mixed Methods for a Nonlinear Elliptic Problem. IMA Preprint Series 1079, Institute for Mathematics and Its Applications, Minneapolis, MN, December 1992
7. Chen, Z.: Expanded mixed finite element methods for linear second-order elliptic problems. RAIRO Math. Model. Anal. Numer. **32**, 479–499 (1998)
8. Chen, Z.: Expanded mixed finite element methods for quasilinear second-order elliptic problems. RAIRO Math. Model. Anal. Numer. **32**, 500–520 (1998)
9. Ciarlet, P.G.: The Finite Element Method for Elliptic Equations. North-Holland, Amsterdam (1978)
10. Douglas Jr., J., Wang, J.: A new family of mixed finite element spaces over rectangles. Mat. Aplic. Comput. **12**, 183–197 (1993)
11. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, 2nd edn. Springer, Berlin (1983)
12. Kim, D., Park, E.J.: A priori and a posteriori analysis of mixed finite element methods for nonlinear elliptic equations. SIAM J. Numer. Anal. **48**, 1186–1207 (2010)
13. Milner, F.A.: Mixed finite element methods for quasilinear second-order elliptic problems. Math. Comp. **44**, 303–320 (1985)
14. Milner, F.A., Park, E.J.: A mixed finite element method for a strongly nonlinear second-order elliptic problem. Math. Comp. **64**, 973–988 (1995)
15. Nedelec, J.C.: A new family of mixed finite elements in R^3. Numer. Math. **50**, 57–81 (1986)
16. Park, E.J.: Mixed finite element methods for nonlinear second-order elliptic problems. SIAM J. Numer. Anal. **32**, 865–885 (1995)
17. Raviart, P.A., Thomas, J.M.: A mixed finite element method for second order elliptic problems. In: Galligani, I., Magenes, E. (eds.) Mathematical Aspects of Finite Element Methods. Lecture Notes in Math, vol. 606, pp. 292–315. Springer, Berlin (1977)

18. Roberts, J.E.: Global estimates for mixed methods for second order elliptic equations. Math. Comp. **44**, 39–52 (1985)
19. Pllana, S., Benkner, S., Xhafa, F., Barolli, L.: A novel approach for hybrid performance modelling and prediction of large-scale computing systems. Int. J. Grid Util. Comput. **9**, 316–327 (2009)

Encrypted Image-Based Reversible Data Hiding with Public Key Cryptography from Interpolation-Error Expansion

Fuqiang Di, Junyi Duan, Minqing Zhang(✉), Yingnan Zhang, and Jia Liu

Engineering University of the People's Armed Police, Xi'an, China
18710752607@163.com

Abstract. This paper proposes an improved version of Shiu's encrypted image-based reversible data hiding with public key cryptography (EIRDH-P). The original work vacates embedding room by difference expansion technique and embeds one bit into each pair of adjacent encrypted pixels. The data extraction and image recovery can be achieved by comparing all pairs of decrypted pixels. Shius' work did not fully exploit the correlation inherent in the neighborhood of a pixel and required side information to record the location map. These two issues could reduce the amount of differences and in turn lessen the potential embedding capacity. This letter adopts a better scheme for vacating room before public key encryption using prediction-error expansion method, in which the pixel predictor is utilized by interpolation technique. The experimental results reveal that the proposed method offers better performance over Shiu's work and existing EIRDH-P schemes. For example, when the peak signal-to-noise ratio of the decrypted Lena image method is 35, the payload of proposed method is 0.74 bpp, which is significantly higher than 0.5 bpp of Shius's work.

Keywords: Reversible data hiding · Interpolation-error expansion · Encrypted image · Public key cryptography

1 Introduction

Reversible data hiding (RDH) [1–3] aims to embed some additional information into a carrier image, while the original image can be recovered by one hundred percent after data extraction. In many scenarios, since cryptography is used to convert normal image data to cipher form for secure communication, encrypted image-based reversible data hiding (EIRDH) has attracted much attention in recent years and has a lot of important applications in medical, military and other fields [4–6]. For example, medical images of patients which have been uploaded to the hospital servers or the cloud are encrypted so as to protect privacy of patients. On the one hand, managers need to embed relevant information such as the owner information and recording time into the corresponding cipher text; On the other hand, the original medical images must be recovered without error.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_14

Some classic reversible data hiding algorithms including difference expansion [7], histogram shifting [8] and redundancy compression [9] are not so suitable for encrypted covers as unencrypted covers. Zhang [10] proposed the first EIRDH algorithm with flipping pixel values. He embedded additional data in the image encrypted by stream cipher, and recovered the original content using the correlation between pixels. An improved version of Zhang's method was proposed by Hong et al. [11], but the algorithm does not work when the block size is small. Lots of EIRDH algorithms have been present [12–17] to improve embedding payload and image quality.

However, symmetric cryptosystem based EIRDH algorithms have the drawbacks such as difficulty of key management and unsuitable for multi-party computation problems, while encrypted image-based reversible data hiding with public key cryptography (EIRDH-P) is a natural issue. The first EIRDH-P algorithms is proposed by Chen et al. [18], which encrypts image using the public key and decrypts embedded image by the secret key of receiver. Each pixel is divided into an even integer and a bit, and both of them are encrypted by homomorphic encryption. In the embedding phase, the second parts of two adjacent pixels are modified to embed a bit. Due to the shared key, Chen's scheme no longer depends on a secure channel among the image provider, the data-hider and the receiver, but it has the inherent overflow since the summation of two adjacent pixel values may be overflow. Some improved EIRDH-P algorithms [19, 20] based on additive homomorphic encryption are proposed, but these schemes provides low embedding capacity, and the directly decrypted images is distorted significantly.

To overcome the weakness of the above schemes, Shiu et al. [21] constructs an efficient EIRDH-P scheme from difference expansion (DE). In this scheme, a preprocessing is needed so as to vacate room for data embedding procedure. Then, by side information and pixel difference expansion, the additional bits are hidden. Concerning the additive homomorphic encryption, Shiu et al. embed additional data in decrypted domain by vacating room for hiding data before encryption. However, the scheme does not fully exploits the correlation inherent in the neighborhood of a pixel and the side information required to record the location map can considerably lower embedding capacity.

In this paper, a new EIRDH-P scheme is introduced, in which one quarter of the total pixels are used to predict other pixels based on interpolation technique and the interpolation-error expansion is adopted to embed additional data. Then, the embedding data can be extracted perfectly and the original images can be losslessly recovered. In general, the proposed scheme obtains excellent performance compared with the existing algorithms.

The rest of this paper is organized as follows. In Sect. 2, some preliminaries are introduced. The proposed EIRDH-P scheme is shown in Sect. 3. In Sect. 4, the experimental results are provided. Finally, conclusions of our work are given in Sect. 5.

2 Preliminaries

2.1 Prediction Error Expansion

Prediction error expansion (PEE) [22–24] is a new approach firstly proposed by Thodi et al. [22] to improve the difference expansion (DE). Here we review this method. For a pixel $I(i,j)$, let $I^*(i,j)$ be the predicted pixel value derived from a prediction algorithm, then the prediction error is defined as follows:

$$e = I(i,j) - I^*(i,j) \tag{1}$$

If the additional bit to be embedded is $w \in \{0, 1\}$, the expansion and embedding process is described by

$$e^* = 2e + w \tag{2}$$

where e^* is the new prediction error. Then, the new pixel value after embedding is

$$I_e = I^*(i,j) + e^* \tag{3}$$

It is easy to show that

$$I_e = 2I(i,j) - I^*(i,j) + w \tag{4}$$

After receiving I_e, the receiver compute the predicted value $I^*(i,j)$ using the same prediction algorithm, and compute

$$I_e^* = I_e - I^*(i,j) = 2I(i,j) - 2I^*(i,j) + w \tag{5}$$

Since $I(i,j)$ and $I^*(i,j)$ are both integers, I_e^* and w have the same parity. The extraction of data can be cast as

$$w = \begin{cases} 0, & \text{if } I_e^* \bmod 2 = 0 \\ 1, & \text{if } I_e^* \bmod 2 = 1 \end{cases} \tag{6}$$

Then the image recovery process can be described by

$$I(i,j) = \frac{I_e^* - w}{2} + I^*(i,j) \tag{7}$$

2.2 Paillier Encryption

Homomorphic encryption [25, 26] is a very useful tool that allows computations to be carried out on ciphertext. However, the decrypted results matches the results of operations performed on the plaintext. Paillier encryption [27] is a classical homomorphic encryption with additive homomorphic property. The algorithm can be described as follows.

Select two large primes a and b, and computes

$$p = a \cdot b \tag{8}$$

$$\lambda = lcm(a-1, b-1) \tag{9}$$

where $lcm(x, y)$ means the least common multiple of x and y. The private key is λ, and the public key is composed of p and a randomly selected integer g. If the plaintext is m, then it can be encrypted by

$$c = E[m, r] = g^m \cdot r^p \bmod p^2 \tag{10}$$

where r represents a randomly selected small integer. The decryption process can be described as

$$D(c) = \frac{L(c^\lambda \bmod p^2)}{L(g^\lambda \bmod p^2)} \bmod p \tag{11}$$

where $L(\bullet)$ is defined as

$$L(u) = \frac{u-1}{p} \tag{12}$$

The additive homomorphic property of Paillier encryption can be shown that

$$D[E[m_1, r_1] \bullet E[m_2, r_2] \bmod p^2] = (m_1 + m_2) \bmod p^2 \tag{13}$$

3 The Proposed Algorithm

In this section, the details of the proposed reversible data hiding algorithm in encrypted images using interpolation-error expansion and homomorphic encryption are illustrated, which is made up of image provider, data-hide and receiver. First of all, a preprocessing is employed to vacate room for data embedding procedure and the image is encrypted with public key based on homomorphic cryptosystem. Then, the data-hider embeds some additional data into the carrier image. The receiver can perfectly extract embedding data and obtain the recover image at last. Figure 1 shows the sketch of the proposed EIRDH-P.

3.1 Preprocessing

We assume that the original image I is a 8 bit grayscale image of size $N \times M$, and all pixel values belong to the range $[0, 255]$. Represent each pixel value with $X(i,j)$, $1 \le i \le N$, $1 \le j \le M$. In our work, an interpolation technique in [28] is adopted for pixel prediction. We classify all pixels in the original image into two sets: sample

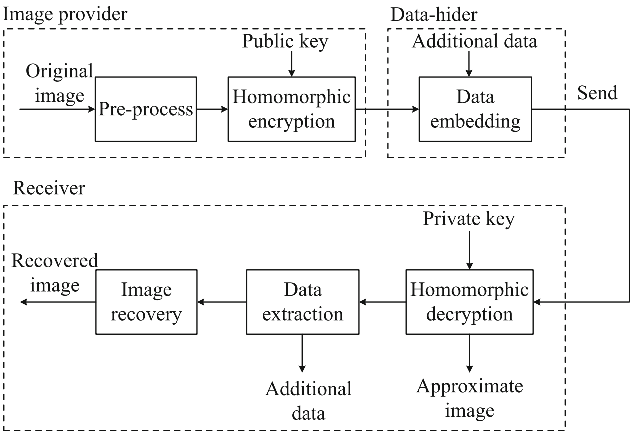


Fig. 1. Sketch of the proposed EIRDH-P scheme

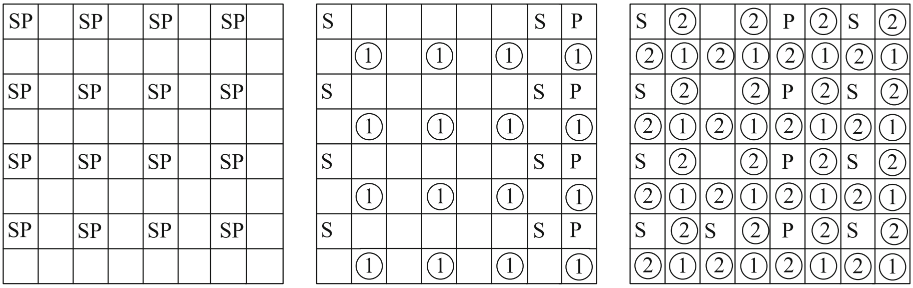
pixels (SP) and non-sample pixels (NSP). Then, the pixels in the set of SP consists of $X(2n-1, 2m-1)$ with $n = 1, 2, \ldots, N/2$, $m = 1, 2, \ldots, M/2$, as depicted in Fig. 2(a). The sample pixels are used to predict non-sample pixels. The prediction process consists two rounds. In the first round, the pixels $X(2n, 2m)$ marked as '①' in Fig. 2(b) can be estimated by the four nearest sample pixels $X(2n-1, 2m-1)$, $X(2n-1, 2m+1)$, $X(2n+1, 2m-1)$, $X(2n+1, 2m+1)$. We compute two prediction value $X_{45}(2n, 2m)$ and $X_{135}(2n, 2m)$ along two orthogonal directions: 45° diagonal and 135° diagonal by

$$X_{45}(2n, 2m) = (X(2n-1, 2m+1) + X(2n+1, 2m-1))/2 \tag{14}$$

$$X_{135}(2n, 2m) = (X(2n-1, 2m-1) + X(2n+1, 2m+1))/2 \tag{15}$$

Select an optimal pair of weights w_{45} and w_{135} to give a good estimate value $X^*(2n, 2m)$ with

$$X^*(2n, 2m) = w_{45} \cdot X_{45}(2n, 2m) + w_{135} \cdot X_{135}(2n, 2m) \tag{16}$$



SP		SP		SP		SP	
SP		SP		SP		SP	
SP		SP		SP		SP	
SP		SP		SP		SP	

(a) Sample pixels

S						S	P
	①		①		①		①
S						S	P
	①		①		①		①
S						S	P
	①		①		①		①
S						S	P
	①		①		①		①

(b) The first round of prediction

S	②		②	P	②	S	②
②	①	②	①	②	①	②	①
S	②		②	P	②	S	②
②	①	②	①	②	①	②	①
S	②		②	P	②	S	②
②	①	②	①	②	①	②	①
S	②	S	②	P	②	S	②
②	①	②	①	②	①	②	①

(c) The second round of prediction

Fig. 2. Illustration of image interpolation

According to [28],

$$w_{45} = \frac{\sigma_{135}}{\sigma_{135} + \sigma_{45}},\ w_{135} = 1 - w_{45} \tag{17}$$

where

$$\begin{cases} \sigma_{45} = \frac{1}{3}\sum\limits_{k=1}^{3} \left(S_{45}(k) - u\right)^2 \\ \sigma_{135} = \frac{1}{3}\sum\limits_{k=1}^{3} \left(S_{135}(k) - u\right)^2 \end{cases} \tag{18}$$

and

$$\begin{cases} S_{45} = \{\mathrm{X}(2n-1, 2m+1),\ X_{45}(2n, 2m),\ X(2n+1, 2m-1)\} \\ u = \frac{1}{4}(\mathrm{X}(2n-1, 2m+1) + X(2n+1, 2m-1) + \mathrm{X}(2n-1, 2m-1) + X(2n+1, 2m+1)) \\ S_{135} = \{\mathrm{X}(2n-1, 2m-1), X_{135}(2n, 2m), X(2n+1, 2m+1)\} \end{cases} \tag{19}$$

In the second round, the non-sample pixels $X(2n-1, 2m)$ and $X(2n, 2m-1)$ marked as '②' in Fig. 2(c) can be estimated by the four nearest pixels along two orthogonal directions: 0° diagonal and 90° diagonal by the same method. After two prediction rounds, the predicted values of all the non-sample pixels can be obtained. Assume the original pixels value and the predicted pixels value are respectively $X(i,j)$ and $X^*(i,j)$, then the interpolation-error is

$$e(i,j) = X(i,j) - X^*(i,j) \tag{20}$$

Since overflow or under flow will happen when the interpolation-error is high, we set a parameter θ to overcome this problem. Then, the preprocess of non-sample pixels can be described as

$$X_p^*(i,j) = \begin{cases} 2X(i,j) - X^*(i,j), & |e(i,j)| \le \theta \\ X(i,j), & |e(i,j)| > \theta \end{cases} \tag{21}$$

where $X_p^*(i,j)$ is the new non-sample pixel after preprocessing. However, all the sample pixels remain unchanged after preprocessing.

3.2 Encryption and Embedding

Since the operation of data embedding based on interpolation-error mainly includes the addition, our algorithm adopts Paillier encryption. Let $X_p^*(i,j)$ be pixel after preprocessing, and the encrypted pixel is calculated by

$$c(i,j) = E[X_p^*(i,j), r(i,j)] = g^{X_p^*(i,j)} \cdot (r(i,j))^p \bmod p^2 \tag{22}$$

where $r(i,j)$ is a randomly selected small integer, and $c(i,j)$ is the encrypted pixel. To embed data by $c(i,j) + m(i,j)$, the corresponding operation in encrypted domain is

$$c_e(i,j) = \begin{cases} c(i,j) \cdot g^{m(i,j)} \cdot (r_e(i,j))^p \bmod p^2, & (i,j) \in SP \text{ and } |e(i,j)| \le \theta \\ c(i,j), & (i,j) \in NSP \end{cases} \tag{23}$$

where $m(i,j) \in \{0,1\}$ is the additional message, $r_e(i,j)$ is a randomly selected small integer, and $c_e(i,j)$ is the pixel value after data embedding.

3.3 Data Extraction and Image Recovery

After receiving the encrypted image, the receiver need decrypt the image using The private key λ firstly. Assume the pixels before and after decryption is $c_e(i,j)$ and $m^*(i,j)$ respectively, then the decryption process is

$$m^*(i,j) = \frac{L([c_e(i,j)]^\lambda \bmod p^2)}{L(g^\lambda \bmod p^2)} \bmod p \tag{24}$$

where $L(\bullet)$ is defined as

$$L(u) = \frac{u-1}{p} \tag{25}$$

As the embedding distortion is very little, the image directly after decrypting can be used as an approximate image in some special scenarios. According to Eqs. (21)–(24), the relationship between $m^*(i,j)$ and $X(i,j)$ is

$$m^*(i,j) = \begin{cases} 2X(i,j) - X_p(i,j) + m(i,j) & (i,j) \in NSP \\ X(i,j) & (i,j) \in SP \end{cases} \tag{26}$$

The receiver can obtain the same predicted non-sample pixel values using the same method, since the sample pixels are not embedded and remain unchanged after decryption. For the non-sample pixel $m^*(i,j)$, $(i,j) \in NSP$, the new interpolation-error $m_t^*(i,j)$ can be compute as follows

$$m_t^*(i,j) = m^*(i,j) - X_p(i,j) = 2X(i,j) - 2X_p(i,j) + m(i,j) \tag{27}$$

Then the data extraction process can be described as

$$m(i,j) = \begin{cases} 0, & m_t^*(i,j) \bmod 2 = 0 \\ 1, & m_t^*(i,j) \bmod 2 = 1 \end{cases} \tag{28}$$

and the original value of the non-sample pixel can be recovered by

$$X(i,j) = \begin{cases} \frac{m_t^*(i,j) - m(i,j)}{2} + X_p(i,j), & (i,j) \in NSP \\ m_t^*(i,j), & (i,j) \in SP \end{cases} \tag{29}$$

4 Experimental Results

The proposed method is verified in this section with the experimental environment of MATLAB R2012b under windows 7. The test is conducted on a 2.6 GHz, Intel(R) Core(TM) i7-6700 HQ system with 8 GB RAM running. We measure the embedding capacity and image quality respectively by *Payload* (bpp) and *PSNR* (Peak signal-to-noise ratio, dB). The *Payload* is the proportion of the total number of embedded bits to the total number of pixels in original image. Moreover, we consider *PSNR* and *MSE* as

(a) (b) (c) (d) (e) (f)

Fig. 3. Comparisons of payload and PSNR among original Lena, directly decrypted Lena with different values of θ, and recovered Lena. (a) Original image. (b) $\theta = 0$, 0.09 bpp, 61.45 dB. (c) $\theta = 4$, 0.56 bpp, 43.50 dB. (d) $\theta = 7$, 0.66 bpp, 40.13 dB. (e) $\theta = 31$, 0.75 bpp, 34.22 dB. (f) Recovered image, PSNR = +∞

Table 1. Experimental results with different values of θ

Value of θ		0	1	2	3	4	5	7	10	20	30	40	50
Lena	*payload*	0.093	0.261	0.394	0.491	0.558	0.603	0.657	0.695	0.736	0.746	0.749	0.750
	PSNR	61.44	53.39	48.62	45.56	43.49	42.03	40.13	38.46	35.61	34.32	33.67	33.39
Peppers	*payload*	0.067	0.195	0.312	0.410	0.489	0.55	0.631	0.689	0.734	0.743	0.746	0.748
	PSNR	62.92	54.6	49.38	45.87	43.4	41.56	39.12	37.14	34.82	33.86	33.26	32.79
Plane	*payload*	0.134	0.338	0.459	0.533	0.578	0.609	0.648	0.680	0.725	0.739	0.744	0.745
	PSNR	59.87	52.33	48.39	45.95	44.33	43.12	41.28	39.45	35.83	34.02	33.13	32.66
Lake	*payload*	0.060	0.170	0.260	0.333	0.394	0.445	0.525	0.604	0.704	0.732	0.743	0.747
	PSNR	63.37	55.25	50.35	47.00	44.51	42.57	39.7	36.9	32.75	30.99	30.03	29.52
Baboon	*payload*	0.028	0.082	0.135	0.185	0.230	0.271	0.341	0.422	0.573	0.649	0.693	0.719
	PSNR	66.70	58.30	52.92	49.09	46.22	43.99	40.66	37.31	31.43	28.37	26.37	25.07
Boat	*payload*	0.089	0.248	0.368	0.448	0.501	0.538	0.590	0.641	0.716	0.738	0.745	0.748
	PSNR	61.68	53.62	48.99	46.13	44.21	42.79	40.64	38.26	34.01	32.19	31.28	30.78

$$PSNR = 10\log_{10}(\frac{255^2}{MSE}) \tag{30}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - p_i^*)^2 \tag{31}$$

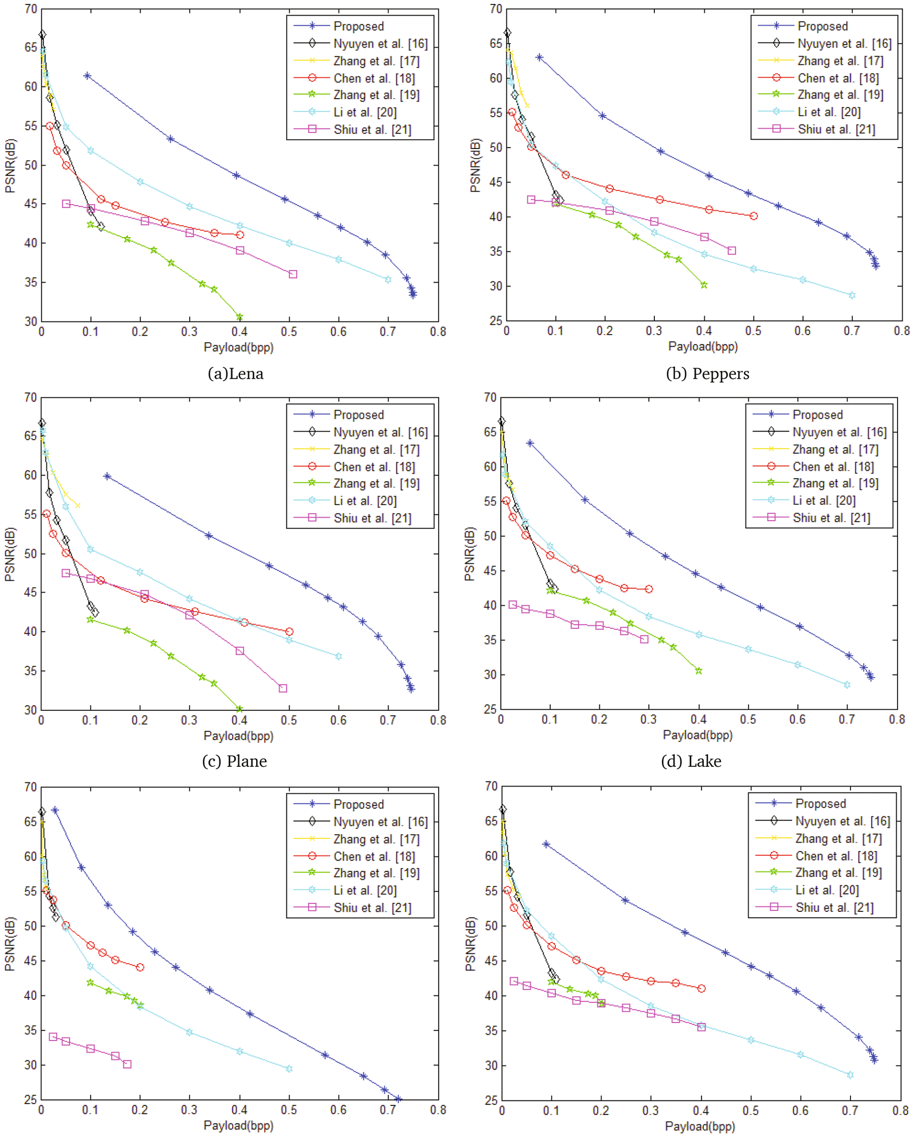
where n is the total number of pixels in original image, p_i and p_i^* are respectively the original pixel value and embedded pixel value. We firstly choose the 8-bit grayscale

(a)Lena (b) Peppers (c) Plane

(d) Lake (e) Baboon (f) Boat

Fig. 4. Six 512 × 512 grayscale test images

image *Lena* of size 512 × 512 to verify the feasibility of the proposed algorithm. Figure 3 shows the experimental comparisons among original Lena, directly decrypted Lena, and recovered Lena. Here, we set respectively $\theta = 0$, $\theta = 4$, $\theta = 7$, $\theta = 31$ and get different *Payload* and *PSNR*. The notation "+∞" shows that the original image can be reconstructed perfectly without any distortion. Moreover, the receiver can coordinate the relationship between embedding capacity and image quality by parameter θ.



(a)Lena

(b) Peppers

(c) Plane

(d) Lake

(e) Baboon

(f) Boat

Fig. 5. Comparisons of the performances of different schemes on six images

Table 1 lists the *Payload* and *PSNR* of the directly decrypted images when different values of θ are used for six standard gray text images shown in Fig. 4. The parameter θ is used to determine the *Payload* and *PSNR* of the directly decrypted images. As can be seen in Table 1, the higher value of θ, the higher *Payload* and the lower *PSNR* we will get.

Further, we compare the experimental results of the proposed algorithm and six existing schemes in [16–21]. The comparison results are shown in Fig. 5. The parameter θ is used to determine the *Payload* and *PSNR* of the directly decrypted images. By observing the results, the proposed algorithm is better than the existing schemes with respect to the embedding capacity and image quality.

5 Conclusions

This work proposes a EIRDH-P algorithm with interpolation-error expansion and homomorphic encryption. On the one hand, the existing scheme introduces obvious distortion when the embedding date is high while the proposed method improves the embedding capacity and image quality of the directly decrypted image. On the other hand, the proposed method overcomes the overflow or underflow problem and does not need side information. Meanwhile, the additional data can be only extracted after image decryption, which is not flexible enough. In the future, the research on homomorphic encrypted algorithm will be carried on to study separable EIRDH-P algorithm with interpolation-error expansion.

References

1. Shi, Y., Li, X., Zhang, X., et al.: Reversible data hiding: advances in the past two decades. IEEE Access (2016). doi:10.1109/ACCESS.2016.2573308
2. Wang, J., Ni, J., Zhang, X., et al.: Rate and distortion optimization for reversible data hiding using multiple histogram shifting. IEEE Trans. Cybern. (2016). doi:10.1109/TCYB.2015.2514110
3. Ma, B., Shi, Y.: A reversible data hiding scheme based on code division multiplexing. IEEE Trans. Inf. Secur. Forensics **11**(9), 1914–1927 (2016)
4. Qian, Z., Zhang, X.: Reversible data hiding in encrypted images with distributed source encoding. IEEE Trans. Circuits Syst. Video Technol. **26**(4), 636–646 (2016)
5. Zhang, W., Wang, H., Hou, D., et al.: Reversible data hiding in encrypted images by reversible image transformation. IEEE Trans. Multimedia. doi:10.1109/TMM.2016.2569497
6. Wu, H., Shi, Y., Wang, H., et al.: Separable reversible data hiding for encrypted palette images with color partitioning and flipping verification. IEEE Trans. Circuits Syst. Video Technol. (2016). doi:10.1109/TCSVT.2016.2556585
7. Tian, J.: Reversible data embedding using a difference expansion. IEEE Trans. Circuits Syst. Video Technol. **13**(8), 890–896 (2003)
8. Dragoi, L., Coltuc, D.: Local-prediction-based difference expansion reversible watermaking. IEEE Trans. Image Process. **23**(4), 1779–1790 (2014)
9. Jarali, A., Rao, J.: Unique LSB compression data hiding method. Int. J. Emerg. Sci. Eng. **2** (3), 17–21 (2013)

10. Zhang, X.: Reversible data hiding in encrypted image. IEEE Signal Process. Lett. **18**(4), 255–258 (2011)
11. Hong, W., Chen, T., Wu, H.: An improved Reversible data hiding in encrypted images using side match. IEEE Signal Process. Lett. **19**(4), 199–202 (2012)
12. Ma, K., Zhang, W., Zhao, X., et al.: Reversible data hiding in encrypted images by reserving room before encryption. IEEE Trans. Inf. Secur. Forensics **8**(3), 553–562 (2013)
13. Zhou, J., Sun, W., Dong, L., et al.: Secure reversible image data hiding over encrypted domain via key modulation. IEEE Trans. Circuits Syst. Video Technol. **26**(3), 441–452 (2016)
14. Cao, X., Du, L., Wei, X., et al.: High capacity reversible data hiding in encrypted images by patch-level sparse representation. IEEE Trans. Cybern. **46**(5), 1132–1143 (2016)
15. Xu, D., Wang, R.: Separable and error-free reversible data hiding in encrypted imaged. Signal Process. **123**, 9–21 (2016)
16. Nyuyen, T., Chang, C., Chang, W.: High capacity reversible data hiding scheme for encrypted images. Signal Process. Image Commun. **44**, 52–64 (2016)
17. Zhang, W., Ma, K., Yu, N.: Reversibility improved data hiding in encrypted images. Signal Process. **94**(1), 118–127 (2014)
18. Chen, Y., Shiu, C., Horng, G.: Encrypted signal-based reversible data hiding with public key cryptosystem. J. Vis. Commun. Image Represent. **25**, 1164–1170 (2014)
19. Zhang, X., Long, J., Wang, Z., Cheng, H.: Lossless and reversible data hiding in encrypted images with public key cryptography. IEEE Trans. Circuits Syst. Video Technol. (2015). doi:10.1109/TCSVT.2015.2433194
20. Li, M., Xiao, D., Zhang, Y., Nan, H.: Reversible data hiding in encrypted images using cross division and additive homomorphism. Signal Process. Image Commun. **39**, 234–248 (2015)
21. Shiu, C., Chen, Y., Hong, W.: Encrypted image-based reversible data hiding with public key cryptosystem from difference expansion. Signal Process. Image Commun. **39**, 226–233 (2015)
22. Thodi, D.M., Rodriguez, J.: Expansion embedding techniques for reversible watermarking. IEEE Trans. Image Process. **16**(3), 721–730 (2007)
23. Dragoi, I., Coltuc, D.: On local prediction based reversible watermarking. IEEE Trans. Image Process. **24**(4), 1244–1246 (2015)
24. Xiang, S., Wang, Y.: Non-inter expansion embedding techniques for reversible image watermarking. EURASIP J. Adv. Signal Process. **2015**, 56–68 (2015)
25. Aguilar, C., Fau, S., Fontaine, C., Gogniat, G.: Recent advances in homomorphic encryption: a possible future for signal processing in the encrypted domain. IEEE Signal Process. Mag. **30**(2), 108–117 (2013)
26. Wang, W., Hu, Y., Chen, L., Huang, X., Sunar, B.: Exploring the feasibility of fully homomorphic encryption. IEEE Trans. Comput. **64**(3), 698–706 (2015)
27. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques Prague, Czech Republic, pp. 223–238 (1999)
28. Luo, L., Chen, Z., Chen, M., Zeng, X., Xiong, Z.: Reversible image watermarking using interpolation technique. IEEE Trans. Inf. Secur. Forensics **5**(1), 187–193 (2010)

Reversible Image Data Hiding with Homomorphic Encryption and Contrast Enhancement

Fuqiang Di, Junyi Duan, Minqing Zhang(✉), Yingnan Zhang,
and Jia Liu

Engineering University of the People's Armed Police, Xi'an, China
18710752607@163.com

Abstract. This paper proposes a novel reversible data hiding algorithm with image contrast enhancement based on homomorphic public key cryptosystem. The additional data is embedded based on histogram shifting after preprocessing procedure. Then the image is encrypted using public key and side information is embedded. On the receiver side, the image with contrast enhancement is obtained directly after image decryption using private key. Due to the correlation between adjacent pixels, date extraction and image recovery can be implemented. To our best knowledge, it is the first reversible data hiding in encrypted image algorithm with image contrast enhancement. Experimental results have demonstrated the feasibility and effectiveness of the proposed method.

Keywords: Reversible data hiding · Image encryption · Homomorphic encryption · Contrast enhancement

1 Introduction

Reversible data hiding (RDH) [1–3] aims to embed some additional information into a carrier image, while the original image can be recovered by one hundred percent after data extraction. In many scenarios, cryptography is used to convert normal image data to cipher form for secure communication, reversible data hiding in the encrypted images (RDHEI) has attracted much attention in recent years and has a lot of important applications in medical, military and other fields [4–6]. For example, medical images of patients which have been uploaded to the hospital servers or the cloud are encrypted so as to protect privacy of patients. On the one hand, managers need to embed relevant information such as the owner information and recording time into the corresponding cipher text; On the other hand, the original medical images must be recovered without error.

Some classic reversible data hiding algorithms including difference expansion [7], histogram shifting [8] and redundancy compression [9] are not so suitable for encrypted covers as unencrypted covers. Zhang et al. [10] proposed the first RDHEI algorithm

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_15

with flipping pixel values. They embedded additional data in the image encrypted by stream cipher, and recovered the original content using the correlation between pixels. In [11, 12], improved versions of [10] were proposed, but images in both algorithms should be decrypted firstly before data extraction. Works [13] presented a novel scheme that based on side information and source coding to implement separable reversible data hiding in encrypted domain.

However, symmetric cryptosystem based RDHEI algorithms have the drawbacks such as difficulty of key management and are not suitable for multi-party computation problems, while homomorphic public key cryptosystem based RDHEI algorithms have the better security and are more suitable for multi-party computation scenarios. In [14], three data hiding schemes for ciphertext images encrypted by public key cryptosystems with homomorphic and probabilistic property were proposed, but could not meet the application requirements of image enhancement. Wu et al. [15] proposed the first algorithm achieving image contrast enhancement by RDH, which was easy to cause image distortion problems.

The proposed algorithm aims to perform reversible data hiding in homomorphic encrypted image and contrast enhancement at the same time. The idea is inspired by [14, 15], and avoids the image distortion problems in [15] using pre-processing operations. The rest of this paper is organized as follows. Section 2 presents the details of image encryption and data embedding. The details of data extraction and image recovery are described in Sect. 3. Section 4 presents the experimental results and Sect. 5 concludes this paper.

2 Image Encryption and Data Embedding

In this section, the details of image encryption and data embedding are illustrated, which is made up of image provider and data-hide. First of all, a preprocessing is employed to avoid overflow or underflow in data embedding procedure. Then, the data-hider embeds some additional data into the carrier image using histogram shifting technology, and encrypts the image using the public key based on homomorphic cryptosystem. To perfectly extract embedding data and recover image, side information is embedded at last. The sketch of proposed method in sending end is given in Fig. 1.

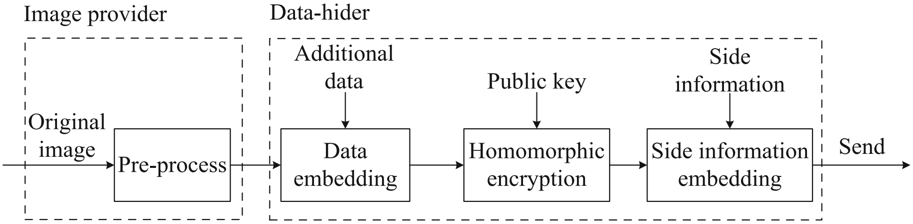


Fig. 1. Sketch of data embedding and image encryption

2.1 Preprocessing

We assume that the original image I is a 8-bit grayscale image of size $m \times n$, and all pixel values belong to the range $[0, 255]$. Represent each pixel value with $m(i, j)$, and $i,j \in [m, n]$. We use b_k to denote additional data encoded as a stream of bits where $k = \{1, 2, 3, \ldots\}$. An iterative parameter λ selected by image provider will be used in this section to ensure all pixel value fall into $[\delta, 255{-}\delta]$. The purpose of this processing operation is to avoid overflow or underflow in data embedding procedure. To achieve image contrast enhancement, the histogram specification operation is implemented using a vector $Q = \{q(0), q(1), \ldots, q(255)\}$ and the vector is calculated by Eq. (1):

$$q(x) = \begin{cases} 0, & x < \lambda \\ (m \times \mathrm{n})/(256 - 2\lambda), & \lambda \leq x \leq 255 - \lambda \\ 0, & x > 255 - \lambda \end{cases} \tag{1}$$

where $1 \leq \lambda \leq 64$. To successfully recover original image, a location map is generated by assigning 1 to the modified pixel, and 0 to the others. The location map and the original pixel values can be compressed and embedded as a part of side information into encrypted image in Sect. 2.4.

2.2 Embedding in Plain Domain

First of all, pixels values of the image after preprocessing are counted. We assume that h_j represents the number of pixels with value j where $\lambda \leq j \leq 255 - \lambda$. Suppose the first and second largest values are chosen and the corresponding values are denoted by s_1 and s_2. Then, the first round of embedding procedure can be described as:

$$m(i,j) = \begin{cases} m(i,j) - 1, & m(i,j) < s_1 \\ s_1 - b_k, & m(i,j) = s_1 \\ m(i,j), & s_1 < m(i,j) < s_2 \\ s_1 + b_k, & m(i,j) = s_2 \\ m(i,j) + 1, & m(i,j) > s_2 \end{cases} \tag{2}$$

The reversible data hiding with large embedding capacity can be achieved after repeating Eq. (2) for λ rounds. The embedding parameters s_1 and s_2 in every rounds are saved as a part of side information to embed into encrypted image in Sect. 2.4. To avoid overflow or underflow problems in data embedding procedure in encrypted image, operation by Eq. (1) is performed again before image encryption.

2.3 Paillier-Based Image Encryption

With homomorphic public key cryptosystem, cipher text can directly perform arithmetic calculations without decryption. In this section, Paillier algorithm [16] is chosen to image encryption. Firstly, select large prime number p and q randomly, and

$$N = pq \tag{3}$$

$$k = lcm(p-1, q-1) \tag{4}$$

where $lcm(\bullet)$ represents least common multiple function. Assume that pixel values before and after encryption respectively are denoted as $m(i,j)$ and $c(i,j)$. The sending end selects a large integer $g(i,j) \in Z_{N^2}^*$ and a small integer randomly selected $r(i,j) \in Z_{N^2}^*$ where $Z_{N^2}^*$ represent set of integers less than N^2 and N^2 prime. Then, the encryption procedure can be described as:

$$c(i,j) = g^{m(i,j)} \cdot (r(i,j))^N \bmod N^2 \tag{5}$$

where g and N are the public keys. According to Paillier algorithm, the semantic security is achieved.

2.4 Embedding in Encrypted Domain

In some scenarios, database administrator wants to embed the label data into the encrypted image, and the data-hider needs to embed some side information. Embedding method in encrypted domain will be given in this section. We divide all pixels into two sets: Set A including $c(i,j)$ with odd value of $(i+j)$, and the rest pixels belong to Set B. To make use of pixel correlation, pixels in Set A are embedded in one-to-one manner while pixels belonging to Set B are remained unchanged. Assume that pixel values before and after embedding respectively are denoted as $c(i,j)$ and $c^*(i,j)$. The data embedding in Set A can be conducted by

$$c^*(i,j) = \begin{cases} c(i,j) \cdot g^{N-\lambda} \cdot (r^*(i,j))^N \bmod N^2, & B_k = 0 \\ c(i,j) \cdot g^{\lambda} \cdot (r^*(i,j))^N \bmod N^2, & B_k = 1 \end{cases} \tag{6}$$

where B_k is the bitstream data to be embedded and $r^*(i,j)$ is a small integer randomly selected. The definitions of g, N and λ are given in Sect. 2.3.

3 Data Extraction and Image Recovery

In this section, the details of data extraction and image recovery are illustrated. First of all, the receiver decrypts the received image using private key and an image with contrast enhancement is obtained. Then, the additional data can be perfectly extracted using pixels estimate technology and the original image can be recovered if it is needed. The sketch of data extraction and image recovery in receiver end is given in Fig. 2.

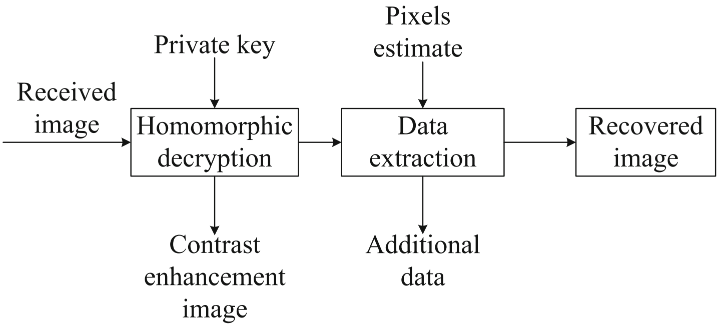


Fig. 2. Sketch of data extraction and image recovery

3.1 Image Decryption

On the receiver end, with the private key of receiver, the image containing additional data with contrast enhancement can be obtained. Assume that pixel values before and after image decryption respectively are denoted as $c^*(i,j)$ and $m^*(i,j)$. According Paillier algorithm, the image decryption procedure can be conducted by

$$m^*(i,j) = \frac{L([c^*(i,j)]^{k_s} \bmod N^2)}{L(g^{k_s} \bmod N^2)} \bmod N \tag{7}$$

where the definition of $L(\bullet)$ is

$$L(u) = \frac{u-1}{N} \tag{8}$$

On the one hand, the distortion between the image after decryption and the original image is very small due to the small value of λ. Thus, the image can be used in fields with not too high image quality requirement. On the other hand, the image with contrast enhancement is more suitable for areas sensitive to image contrast, such as medical diagnosis.

3.2 Data Extraction

We denote the encryption and the decryption procedure as $E[\bullet]$ and $D[\bullet]$ respectively. According to homomorphic property of Paillier algorithm, there is

$$D[E[m(i_1,j_1), r(i_1,j_1)] \cdot E[m(i_2,j_2), r(i_2,j_2)] \bmod N^2] = (m_1 + m_2) \bmod N \tag{9}$$

Assume that pixel values after decryption and before encryption respectively are denoted as $m^*(i,j)$ and $m(i,j)$. B_k represents the bitstream data embedded in Sect. 2.4. For the pixels in Set A, there is

$$m^{*}(i,j) = \begin{cases} m(i,j) + \lambda, & B_k = 1 \\ m(i,j) - \lambda, & B_k = 0 \end{cases} \tag{10}$$

With the neighbor pixels, the pixel value in Set A can be estimated by

$$\overline{m(i,j)} = \frac{m(i-1,j) + m(i+1,j) + m(i,j-1) + m(i,j+1)}{4} \tag{11}$$

Owing to the correlation between adjacent pixels, the bitstream data B_k embedded in Sect. 2.4 can be calculated by

$$B_k = \begin{cases} 1, & m^{*}(i,j) > \overline{m(i,j)} \\ 0, & m^{*}(i,j) \le \overline{m(i,j)} \end{cases} \tag{12}$$

The pixel value after embedding in Sect. 2.2 can be obtained via

$$m(i,j) = \begin{cases} m^{*}(i,j) - \lambda, & m^{*}(i,j) > \overline{m(i,j)} \\ m^{*}(i,j) + \lambda, & m^{*}(i,j) \le \overline{m(i,j)} \end{cases} \tag{13}$$

Then, the additional data embedded in Sect. 2.2 can be extracted using side information. The extraction procedure in each round can be described as

$$b_k = \begin{cases} 1, & m(i,j) = s_1 - 1 \\ 0, & m(i,j) = s_1 \\ 0, & m(i,j) = s_2 \\ 1, & m(i,j) = s_2 + 1 \end{cases} \tag{14}$$

3.3 Image Recovery

In this section, the image recovery method is described in case that the receiver wants to get the original image without distortion. Firstly, the pixels values before preprocessing can be obtained using parameter λ, s_1 and s_2 in each round. The recovery operation in each round is

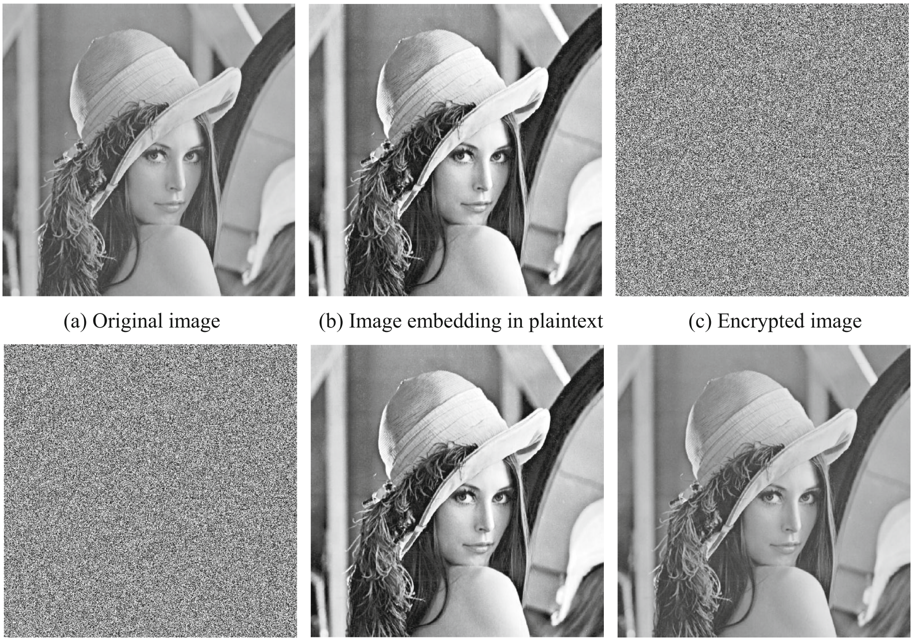
$$m(i,j) = \begin{cases} m(i,j) + 1 \;, & m(i,j) < s_1 - 1 \\ s_1 \;, & m(i,j) = s_1 \text{ 或 } m(i,j) = s_1 - 1 \\ s_2 \;, & m(i,j) = s_2 \text{ 或 } m(i,j) = s_2 + 1 \\ m(i,j) - 1 \;, & m(i,j) > s_2 + 1 \end{cases} \tag{15}$$

Then, the original image can be recovered without distortion using the side information of the location map and the original pixel values in Sect. 2.1.

4 Experimental Results and Analysis

4.1 Feasibility

The proposed method is verified in this section with the experimental environment of MATLAB R2012b under windows 7. We firstly choose image *Lena* which is a 8-bit grayscale image of size 512 × 512 to verify the feasibility of the proposed algorithm. The selected iterative parameter is $\lambda = 40$, and the parameters in the encryption procedure are respectively $p = 57$, $q = 59$, $N = pq = 3363$. Figure 3 shows a group of experimental results with image *Lena*. Figure 3(a) and (b) show the image before and after additional data embedding. After Paillier algorithm encrypted, the encrypted image is shown as Fig. 3(c). Figure 3(d) shows the image containing embedded the side information. On the receiver side, image Fig. 3(e) is obtained after direct image decryption. Although the PSNR (Peak Signal to Noise Ratio) value of the decrypted image is only 22.27, it has better visual quality and image contrast which is crucial in many scenarios. Figure 3(f) shows the recovery image after data extraction.



(a) Original image (b) Image embedding in plaintext (c) Encrypted image

(d) Image embedding in ciphertext (e) Decrypted image (f) Recovered image

Fig. 3. Experimental results of image *Lena*

4.2 Visual Quality

Furthermore, Fig. 4 compares the visual quality of our proposed algorithm and existing algorithm. Three standard test images, *Man, Crowd, and Couple* are listed in Fig. 4. When the value of L in [15] or the λ in the proposed method is low, both algorithms have high visual quality and the contrast enhancement effects are similar. However, the over-enhanced contrast introduces obvious distortion in [15] when the amount of additional data is large and partial experimental results are shown as Fig. 5. Figure 6 imply that the proposed method can effectively avoid over-enhanced contrast problems and has the more satisfying visual quality due to the preprocessing procedure especially when the payload is high.

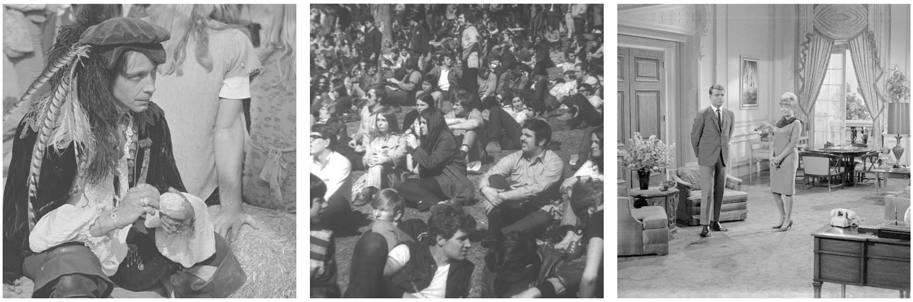


Fig. 4. Original images

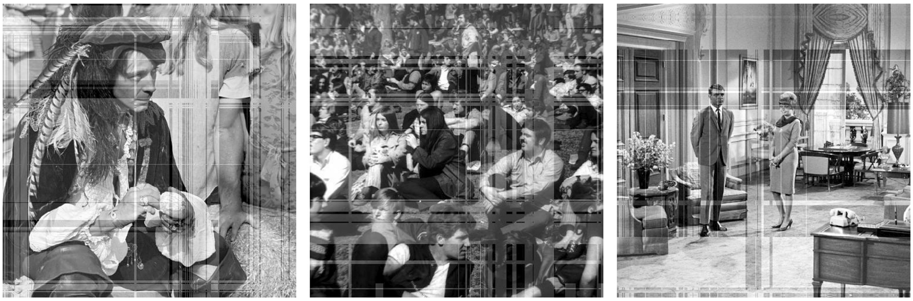


Fig. 5. In [15], $L = 40$

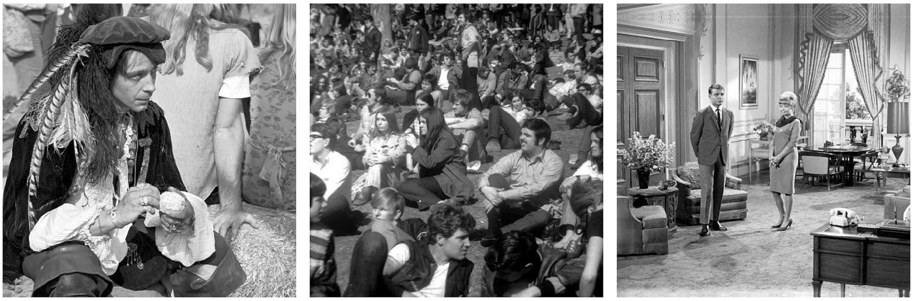


Fig. 6. In the proposed method, $\lambda = 40$

4.3 Embedding Capacity

Table 1 shows the comparisons of max embedding capacity between existing methods and the proposed method. The max embedding capacity in this section means that the embedding rate which can ensure the distortion of image after embedding is in a certain range. It is clear that the proposed algorithm can embed more additional data while having better visual quality because it can overcome the distortion defect with high payload.

Table 1. Comparisons of max embedding capacity of partial images using different methods

Image name	*Lena*	*Man*	*Crowd*	*Couple*	*Hill*
Payload (/bpp) in [15]	0.4176	0.6015	0.4357	0.3641	0.5062
Payload (/bpp) in [16]	0.4083	0.3497	0.3282	0.3862	0.3705
Payload (/bpp) in Proposed	0.8706	1.0881	1.9107	1.4175	1.2894

5 Conclusions

This work proposes a novel reversible image data hiding algorithm with homomorphic encryption and image contrast enhancement. On the one hand, the existing reversible data hiding algorithm with contrast enhancement introduces obvious distortion when the embedding date is high while the proposed method solves this problems using preprocessing procedure. On the other hand, the proposed method improves the max payload with satisfactory visual quality and has proper image contrast enhancement degree when the embedding rate is very high. The experimental results imply that the proposed method can improve both visual quality and max payload. Meanwhile, the additional data can be only extracted after image decryption, which is not flexible enough. In the future, the research on homomorphic encrypted algorithm will be carried on to study separable reversible data hiding methods in encrypted images with contrast enhancement.

References

1. Asifullah, K., Ayesha, S., Summuyya, M.: A recent survey of reversible watermarking techniques. Inf. Sci. **279**, 251–272 (2014)
2. Zhang, W., Hu, X., Yu, N.: Optimal transition probability of reversible data hiding for general distortion metrics and its applications. IEEE Trans. Image Process. **24**(1), 294–304 (2015)
3. Ou, B., Li, X., Zhao, Y., Ni, R., Shi, Y.Q.: Pairwise prediction-error expansion for efficient reversible data hiding. IEEE Trans. Image Process. **22**(12), 5010–5021 (2013)
4. Qian, Z., Zhang, X.: Reversible data hiding in encrypted images with distributed source encoding. IEEE Trans. Circ. Syst. Video Technol. **26**(4), 636–646 (2016)
5. Qian, Z., Zhang, X., Wang, S.: Reversible data hiding in encrypted JPEG Bitstream. IEEE Trans. Multimedia **16**(5), 1486–1491 (2014)

6. Zhang, W., Ma, K., Yu, N.: Reversibility improved data hiding in encrypted images. Sig. Process. **94**, 118–127 (2014)
7. Tian, J.: Reversible data embedding using a difference expansion. IEEE Trans. Circ. Syst. Video Technol. **13**(8), 890–896 (2003)
8. Dragoi, L., Coltuc, D.: Local-prediction-based difference expansion reversible watermaking. IEEE Trans. Image Process. **23**(4), 1779–1790 (2014)
9. Jarali, A., Rao, J.: Unique LSB compression data hiding method. Int. J. Emerg. Sci. Eng. **2** (3), 17–21 (2013)
10. Zhang, X.: Reversible data hiding in encrypted image. IEEE Sig. Process. Lett. **18**(4), 255–258 (2011)
11. Hong, W., Chen, T., Kao, Y.: Reversible data embedment for encrypted cartoon images using unbalanced bit flipping. In: Advances on Swarm Intelligence. LNCS, vol. 7929, pp. 208–214 (2013)
12. Hong, W., Chen, T., Wu, H.: An improved reversible data hiding in encrypted cartoon images using side match. IEEE Signal Process. Lett. **19**(4), 199–202 (2012)
13. Zhang, X.: Separable reversible data hiding in encrypted images. IEEE Trans. Inf. Forensics Secur. **7**(2), 826–832 (2012)
14. Zhang, X., Long, J., Wang, Z., Cheng, H.: Lossless and reversible data hiding in encrypted images with public key cryptography. IEEE Trans. Circ. Syst. Video Technol. **26**, 1622–1631 (2016). doi:10.1109/TCSVT.2015.2433194
15. Wu, H.T., Dugelay, J., Shi, Y.Q.: Reversible image data hiding with contrast enhancement. IEEE Signal Process. Lett. **22**(1), 81–85 (2015)
16. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Proceeding of the Advances Cryptology. LNCS, pp. 223–238 (1999)

A Deep Network with Composite Residual Structure for Handwritten Character Recognition

Zheheng Rao[1,2], Chunyan Zeng[1,2], Nan Zhao[1,2], Min Liu[1,2],
Minghu Wu[1,2](✉), and Zhifeng Wang[3]

[1] Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy
and Operation Control of Energy Storage System,
Hubei University of Technology, Wuhan, China
zh_rao123@163.com, swallow_chunyan@163.com,
[2] Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar
Energy, Hubei University of Technology, Wuhan 430068, China
wuxx1005@mail.hbut.edu.cn
[3] School of Educational Information Technology,
Central China Normal University, Wuhan 430079, China
zfwang@mail.ccnu.edu.cn

Abstract. This paper presents a new deep network (non – very deep network) with composite residual for handwritten character recognition. The main network design is as follows: (1) Introduces an unsupervised FCM clustering algorithm to preprocess the experimental data. (2) By exploiting a composite residual structure the multilevel shortcut connection is proposed which is more suitable for the learning of residual. (3) In order to solve the problem of overfitting and time-consuming for training the network parameters, a dropout layer is added after the completion of all convolution operations of each extended nonlinear residual kernel. Comparing with general deep network structures of same deep on handwritten character MNIST database, the proposed algorithm shows better recognition accuracy and higher recognition efficiency.

1 Introduction

With the development of technology, intelligent recognition bring us a lot of challenges [1]. Handwritten character recognition methods are mainly divided into two categories: handwritten character recognition method based on traditional feature extraction and pattern classification [2], handwritten character recognition method based on deep learning [3].

In this paper, we propose an algorithm for handwritten character recognition based on composite residual structure [4, 5]. In the remaining part of this paper, Firstly, we introduces the handwriting recognition framework based on deep learning, then illustrate the design of the structure, and propose the handwritten character recognition algorithm based on composite residual structure. The algorithm solves the classification problem of handwritten numeral recognition. Then, the experimental scheme is

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_16

designed, and compares our method and convolutional neural network. Finally, some useful conclusions are obtained, and the further research work is prospected [6, 7].

2 Handwriting Recognition Framework Based on Deep Learning

The frame of handwritten character recognition based on deep learning [8] is shown in Fig. 1.

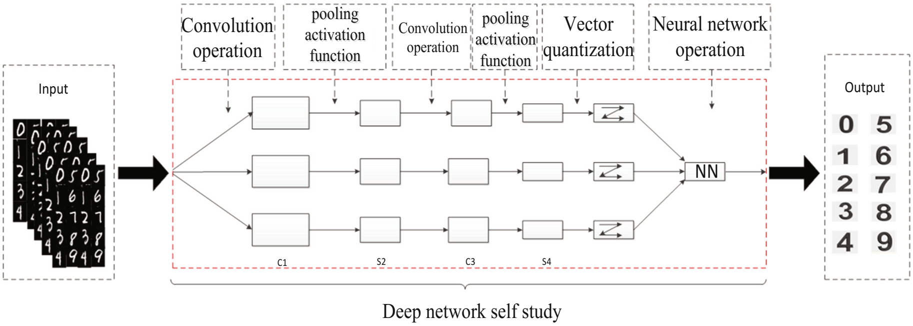


Fig. 1. Handwritten character recognition method based on deep learning

3 Handwritten Character Recognition Algorithm Based on Composite Residual Structure

At present [9, 10], the deep learning method has some problems need to be solved: First problem: Deep learning model network parameter training is time-consuming, and the latest research shows that: very deep network is not the greatest impact on the overall performance of the factors, but will affect the other components of the network; The second: The rapid development of industrial and academic circles, recognition of handwritten character recognition accuracy and recognition efficiency of the increasingly high demand. Therefore, it is possible to construct a structure to make it more excellent performance in a certain depth (non deep) network. In order to solve the above problems [11], this paper presents a framework of the handwritten character recognition algorithm based on composite residual structure, and its structure is shown in Fig. 2 below. The framework is characterized by:

(1) FCM clustering algorithm is introducing, and the clustering results are optimized.
(2) The feature extraction and morphological classification of handwritten characters are carried out by using the characteristics of sparse connection and weight sharing of convolutional neural network.
(3) The composite residual kernel is constructed. The multilevel short connection is introduced. Then the Dropout layer is added after the optimization parameter.

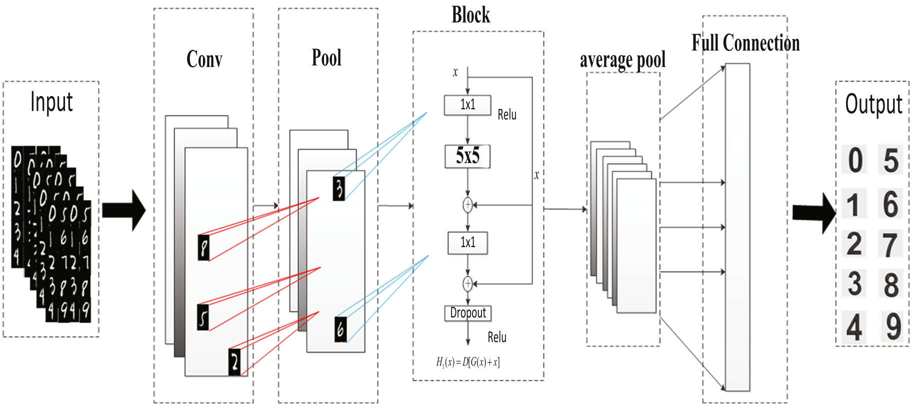


Fig. 2. Frame of handwritten character recognition based on composite residual structure

Specifically, the framework of the proposed algorithm can be divided into 3 parts: data preprocessing [12], neural network architecture for expanding nonlinear kernel, and composite residual structure kernel structure.

3.1 Data Preprocessing

In this paper, based on composite residual structure kernel structure, we introduce an unsupervised clustering algorithm to preprocess the experimental data–FCM clustering algorithm. FCM algorithm is an algorithm to determine belongs to a certain center of clustering algorithm by membership degree of each data point, which is an improvement of the traditional hard clustering algorithm.

3.2 Convolution Neural Network with Composite Residual Structure Kernel Structure

Based on composite residual structure kernel convolution neural network architecture shows in Fig. 3. We introduce this structure in the algorithm: the performance will be

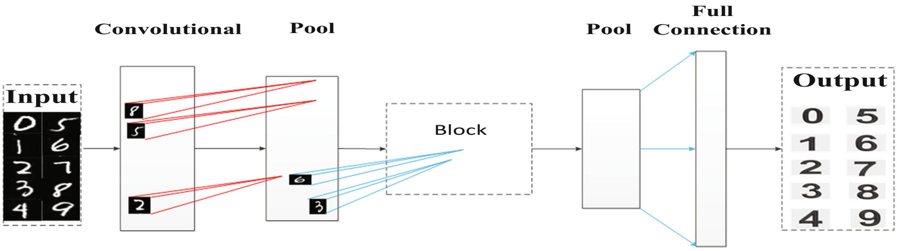


Fig. 3. Composite residual structure for handwritten character recognition convolution neural network architecture

more obvious and the image can be directly used as the input of the network, so as to avoid the complex image feature extraction and data reconstruction in the traditional recognition algorithm.

3.3 Composite Residual Structure Kernel Structure

This paper present a framework of the handwritten character recognition algorithm based on composite residual structure kernel structure, and the main network design of composite residual structure kernel lies in:

(1) Proposing the compound residual structure, and adding the multilevel short connection;
(2) Adding dropout layer after optimizing parameters.
(3) Replace the new residual block with a value of 1 × 1 + n * n + 1 × 1.

Based on the proposed scheme, this paper proposes a composite residual structure kernel. The principle structure is shown in Fig. 4 below.

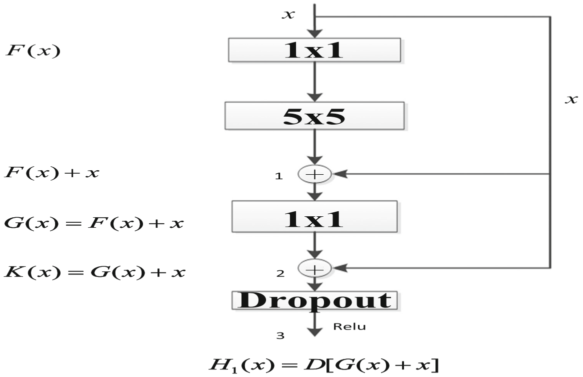


Fig. 4. Composite residual structure kernel structure framework

A description of the formula used in Fig. 4, with x as the input of the first layer of the residual kernel.

Node 1:

$$G(x) = F(x) + x \tag{1}$$

$F(x)$ is the normal deep structure network, x is the first level shortcut connection;

Node 2:

$$K(x) = G(x) + x \tag{2}$$

$G(x)$ is the normal deep structure network, x is the second level shortcut connection;

Node 3:

$$H_1(x) = D[G(x) + x] \tag{3}$$

$D(x)$ is the Dropout layer, x of the $D(x)$ can be adjusted according to the settings of different parameters. $G(x) + x$ relative to node 3, for normal deep structure.

4 Experimental Comparison

This experiment using the Ubuntu16.04 system, TensorFlow platform, using the MNIST standard library (training database of handwritten characters: 60000 pictures, testing database of handwritten characters: 10000 pictures) to different network structures in the experimental verification the same effect of network layers, the same data set.

(1) Convolution neural network model identification accuracy: 0.9465.

(2) Model based on composite residual structure, as shown in Fig. 2, the model identification accuracy: 0.988281. And the training process records of the model, as shown in Figs. 5 and 6. The horizontal axis represents the number of cluster training and with the increase in the number of labeled training clusters, training accuracy gradually increased and stable.

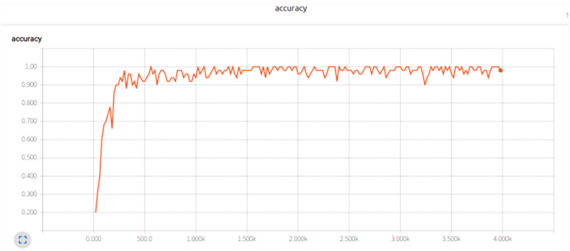


Fig. 5. Accuracy of the training process based on composite residual structure model

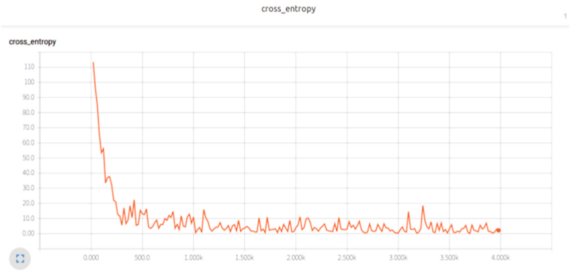


Fig. 6. Cross_entropy of the training process based on composite residual structure model

5 Conclusion

In this paper, we presents a new deep network structure based on composite residual structure network handwritten character recognition algorithm network. Its main idea lies in: Proposed a new network structure which is tested under the same conditions, compared with the method of handwritten character classification, has better character recognition accuracy and higher recognition efficiency. First of all, we introduce an FCM unsupervised clustering algorithm to preprocess the experimental data. Then, on the basis of the basic framework, this paper puts forward two kinds of structural design:

(1) The composite residual structure is introduced and the multilevel shortcut connection is proposed,
(2) After the completion of all the convolution operations of each composite residual structure kernel, a dropout layer is added.

Acknowledgments. This research was supported by National Natural Science Foundation of China (No. 61471162, No. 61501178, No. 61501199, No. 61601177); Program of International science and technology cooperation (2015DFA10940); Science and technology support program (R & D) project of Hubei Province (2015BAA115); PhD Research Startup Foundation of Hubei University of Technology (No. BSQD13037, No. BSQD14028); Open Foundation of Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar Energy (HBSKFZD2015005, HBSKFTD2016002).

References

1. Pereira, R., Pereira, E.G.: Future internet: trends and challenges. Int. J. Space-Based Situated Comput. **5**(3), 159–167 (2015)
2. Mahesha, P., Vinod, D.S.: Support vector machine-based stuttering dysfluency classification using GMM supervectors. Int. J. Grid Util. Comput. **6**(3–4), 143–149 (2015)
3. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
4. Al-Jumeily, D., Hussain, A., Fergus, P.: Using adaptive neural networks to provide self-healing autonomic software. Int. J. Space-Based Situated Comput. **5**(3), 129–140 (2015)
5. Zhu, X.D., Li, H., Li, F.H.: Privacy-preserving logistic regression outsourcing in cloud computing. Int. J. Grid Util. Comput. **4**(2–3), 144–150 (2013)
6. Varaprasad, G., Murthy, G.S., Jose, J., et al.: Design and development of efficient algorithm for mobile ad hoc networks using cache. Int. J. Space-Based Situated Comput. **1**(2–3), 183–188 (2011)
7. Wu, K., Kang, J., Chi, K.: Research on fault diagnosis method using improved multi-class classification algorithm and relevance vector machine. Int. J. Inf. Technol. Web Eng. (IJITWE) **10**(3), 1–16 (2015)
8. Wu, Z., Lin, T., Tang, N.: Explore the use of handwriting information and machine learning techniques in evaluating mental workload. Int. J. Technol. Hum. Interact. (IJTHI) **12**(3), 18–32 (2016)
9. Liu, C.L., Yin, F., Wang, D.H., et al.: Online and offline handwritten Chinese character recognition: benchmarking on new databases. Pattern Recogn. **46**(1), 155–162 (2013)

10. Delaye, A., Liu, C.L.: Contextual text/non-text stroke classification in online handwritten notes with conditional random fields. Pattern Recogn. **47**(3), 959–968 (2014)
11. Zhang, X.Y., Bengio, Y., Liu, C.L.: Online and offline handwritten Chinese character recognition: a comprehensive study and new benchmark. Pattern Recogn. **61**, 348–360 (2017)
12. Alamareen, A., Al-Jarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**(3), 1–14 (2016)

An Ensemble Hashing Framework for Fast Image Retrieval

Huanyu Li(✉) and Yunqiang Li

Air Force Engineering University, Xi'an, Shanxi, China
lihuanyu1984@163.com, 1782545475@qq.com

Abstract. Binary hashing has been widely used for efficient similarity search due to its query and storage efficiency. In this paper, we attempt to exploit ensemble approaches to tackle hashing problem. A flexible ensemble hashing framework is proposed to guide the design of hashing methods, which takes into account three important principles namely higher accuracy, larger diversity and the optimal weights for predictors simultaneously. Next, a novel hashing method is designed by the proposed framework. In this work, we first use the weighted matrix to balance the variance of hash bits and then exploit bagging method to inject the diversity among hash tables. Under the same code length, the experimental results show that the proposed method achieves better performance than several other state-of-the-art methods on two image benchmarks CIFAR-10 and LabelMe.

1 Introduction

With the development of artificial intelligence and computer vision, image processing technology has been more important in many fields [1–3]. Due to its low storage cost and fast query speed, hashing has been widely adopted for approximate nearest neighbour (ANN) search in large-scale datasets, especially in image retrieval [4]. Hashing is an approach of transforming the data item to a low-dimensional representation, or equivalently a short code consisting of a sequence of bits. The learning-based hashing methods can be divided into three main streams: unsupervised, semi-supervised and supervised methods.

The first stream is unsupervised methods which only use unlabeled data to learn hash functions. The representative algorithms in this category include Locality Sensitive Hashing (LSH), Spectral Hashing (SH) [5], Iterative quantization (ITQ) [6], Isotropic Hashing (IsoH) [7], Anchor graph hashing (AGH) [8], Harmonious Hashing [9] and Learning Binary Codes with Bagging PCA (BPCAH) [10].

The other two streams are semi-supervised and supervised methods, which have been further developed to improve the quality of hashing by incorporating supervisory information in form of class labels, including Semi-supervised hashing (SSH) [11], Binary Reconstructive Embedding (BRE) [12], minimal loss hashing (MLH) [13], Kernel-Based Supervised Hashing (KSH) [14], Fast Supervised Hashing (FastHash) [15] and Supervised Discrete Hashing (SDH) [16].

Note that the methods which attempt to preserve label similarity, including semi-supervised and supervised hashing, usually outperform the unsupervised methods

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_17

in performance which not exploits the label information. In addition, some unsupervised methods, such as SH and Harmonious Hashing which integrate the local information of data, commonly perform better than those, such as PCAH [17] and ITQ which only consider the global information. Also the data-dependent methods which utilize the training data information usually have better performance than data-independent methods. Therefore, an important criterion guiding the design of hashing methods is that the generated hash bits should take as much information as possible.

Recently, some researchers propose to assign different bit-level weights to different hash bits to calibrate the Hamming distance ranking. [18] proposes a query-adaptive Hamming distance ranking method for image retrieve using the learned bitwise weights for a diverse set of predefined semantic concept classes. [19] studies query-sensitive hash code ranking algorithm (QsRank) for PCA-based hashing algorithms. [20] also presents a weighted Hamming distance ranking algorithm (WhRank) based on the data-adaptive and query-sensitive bitwise weight.

To maximally cover the nearest neighbors, recently, multi-table methods have been further developed, which attempt to build multiple complementary hash tables. As one of the best known methods, LSH constructs multiple hash tables independently, aiming to enlarge the probability that similar data points are mapped to similar hash codes. However, it often suffers from severe redundancy of the hash tables. Complementary Hashing (CH) [21] proposes a sequential learning method to build complementary hash tables, and obtained promising performance with much less tables. BPCAH tries to learn several pieces of diverse short codes with Principal Component Analysis (PCA) [22] and concatenates them into one piece of long codes, where a piece of short code can be seen as a hash table with small code size.

In this paper, we attempt to exploit ensemble approaches, which have been widely used in classification and regression, to tackle hashing problem. A flexible ensemble hashing framework is proposed to guide the design of hashing methods. The framework places three important principles namely higher accuracy, larger diversity and the optimal weights for predictors into a common framework, which can provide insight into their interrelation and would be helpful in designing more effective hashing methods. Next, a novel hashing method dubbed weighted bagging PCA-ITQ is designed by our ensemble hashing framework. Firstly, a weighted matrix is utilized to balance the variance of hash bits in a single hash table; and then bagging technique is exploited to inject diversity among hash tables; Finally, we concatenate the binary codes generated by multiple diverse hash tables. Under the same code length, our proposed method outperforms most state-of-the-art methods in retrieval performance.

2 Related Work

Most existing hashing methods are projections-based methods which this paper focuses on. The PCAH adopts PCA to learn the projection functions. However, since the variance of the data in each PCA direction is different, it is unreasonable to use the same number of bits for different projected dimensions. As noted in [23], one simply way to balance the variance is to apply a random orthogonal transformation to the PCA-projected data. Motivated by [23], ITQ tries to learn an orthogonal rotation matrix to refine the initial projection matrix learned by PCA, so that the quantization error of mapping the data to the vertices of binary hypercube is minimized. [9] uses locality preserving projection (LPP) to replace the PCA projection which integrates more data information. These methods are all data-dependent methods that the projection functions are learned from training data.

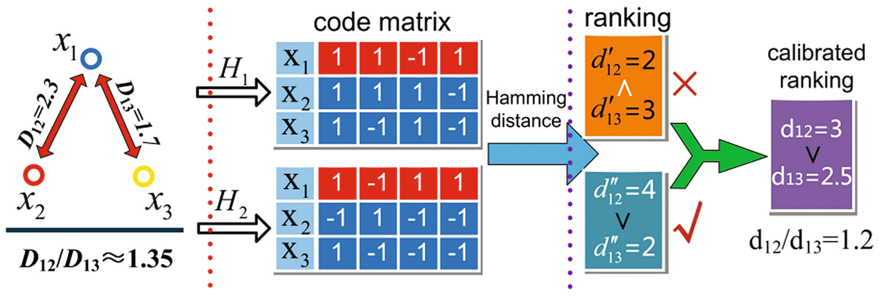
Another class of projection-based hash methods are called data-independent methods, whose projection functions are independent of training data, including LSH, Kernelized locality-sensitive hashing for scalable image search (KLSH) [24], Non-metric locality-sensitive hashing [25] and Locality-sensitive binary codes from shift-invariant kernels (SKLSH) [26]. They use simple random projections for hash functions. The data-dependent methods usually have better performance than data-independent methods since the learned hash functions integrate data information.

Ensemble approaches to classification and regression have attracted a great deal of interest in recent years [27–30]. While ensemble approaches to classification usually make use of non-linear combination methods like majority voting; regression problems are naturally tackled by linearly weighted ensembles. Two popular methods for creating accurate ensembles are bagging and boosting [31]. There are a few hashing methods that use ensemble techniques to improve the retrieve performance. CH proposes to adopt boosting-based approach to build complementary hash tables. BPCAH exploits bagging method to learn several pieces of diverse short codes and concatenate them into one piece of long codes. [32] attempts to train each hash function independently using supervised information but introduce diversity among them using techniques from classier ensembles. However, these methods are usually pretty one-sided, which will neglect some important factors in designing hash approaches. This paper proposes a comprehensive and flexible ensemble hashing framework to guide the design of hash methods.

3 The Proposed Ensemble Hashing Framework

This section will introduce our ensemble hashing framework in detail. Let us first introduce a set of notations. Given a set of n data points $\left\{x_i \in \mathbb{R}^d\right\}_{i=1}^n$, and the matrix form is $\mathbf{X} \in \mathbb{R}^{n \times d}$. Let $\{H_l\}_{l=1}^L$ denote L hash tables and each hash table is $H_l(x) = \{h_i(\bullet)\}_{i=1}^k$ consisting of k hash functions. The output of one hash table is a k-bit binary code matrix $\mathbf{B}_l(x) \in \{1, -1\}^{n \times k}$. A single hash bit can be expressed as $h_i(x) = \operatorname{sgn}[g(x)]$, where $g(x)$ denotes prediction functions and $\operatorname{sgn}(v) = 1$ if $v \geq 0$ and -1 otherwise.

Figure 1 illustrates our observation. Firstly, we randomly sample three data points $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ from CIFAR-10 dataset, and compute their Euclidean distance as $D(\mathbf{x}_1, \mathbf{x}_2) = 2.3 > D(\mathbf{x}_1, \mathbf{x}_3) = 1.7$, which denotes that $\mathbf{x}_3$ is more similar to $\mathbf{x}_1$ than $\mathbf{x}_2$, their similarity proportion is 1:35. Then, ITQ method is utilized to construct two hash tables $H_1(\mathbf{x})$ and $H_2(\mathbf{x})$, and the above three data points can be encoded into two group of 4-bit binary code matrices $\mathbf{B}_1$ and $\mathbf{B}_2$. Let $d_{ij}\left(\mathbf{x}_i, \mathbf{x}_j\right)$ denote the Hamming distance between binary codes $\mathbf{x}_i$ and $\mathbf{x}_j$. From Fig. 1(c), we can find that the ranking obtained by the first hash table is false $d'_{12} < d'_{13}$. Although the second ranking is correct, it is not accurate since the similarity proportion is $d''_{12}/d''_{13} = 2$, which is far away from the



(a) Original space (b) Hamming distance ranking (c) Calibrate ranking

Fig. 1. Our observation: two hash tables can not necessary achieve unique Hamming distance ranking.

similarity proportion in original space. From Fig. 1(b), we find that the binary code matrices $\mathbf{B}_1$ and $\mathbf{B}_2$ are not identical. In other word, the learned hash tables $H_1(\mathbf{x})$ and $H_2(\mathbf{x})$ are diverse.

The observation above indicates that different hash tables may generate non-unique Hamming distance ranking, false or correct, all in all, they are not the best ranking.

4 A Novel Weighted Bagging PCA-ITQ Method

4.1 PCA-ITQ

ITQ tries to learn an orthogonal rotation matrix **R** to refine the initial projection matrix learned by PCA, so that the quantization error of mapping the data to the vertices of binary hypercube is minimized. Its objective function can be written as:

$$\begin{aligned}&\min Q(\mathbf{B},\mathbf{R}) = \|\mathbf{B}-\mathbf{VR}\|_F^2\\&\text{s.t. } \mathbf{R}^{\mathbf{T}}\mathbf{R}=\mathbf{I}\end{aligned} \tag{1}$$

Where $\mathbf{V}=\mathbf{XW}^*$ denotes the PCA-projected data, **B** is binary codes matrix, $\|\bullet\|_F^2$ denotes the Frobenius norm, and **I** is a d-by-d identity matrix.

Weight: In practice, ITQ expects to balance the variances of different PCA-projected dimensions using orthogonal transformation. However, it cannot necessarily guarantee equal variances. We find that the variances are more uniform with the increase of iterations, but cannot be equal. Therefore, it is unreasonable to assign the same weight $\omega_l = 1/k$ to different hash bits in ITQ.

Diversity: We randomly select m data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$ from the original data **X** X as the test set to evaluate the Ambiguity term. For the element (x_a, y_a), its similarity measured by the i-th hash bit is denoted as $S_i^{(a)}$, and the convex combination is $S_{en}^{(a)} = \sum_{i=1}^{k} S_i^{(a)} \Big/ k$. The average Ambiguity term over the test data can be written as:

$$\gamma = \frac{1}{km} \sum_{a=1}^{m} \sum_{i=1}^{k} \left(S_i^{(a)} - S_{en}^{(a)} \right)^2 \tag{2}$$

Figure 2 illustrates the Ambiguity term of top k hash bits (ensembles). Here, we set the size of test set as m = 5000. We find that the Ambiguity term curve increases relatively fast for the first few ensembles, then it tends to be smooth and up to the boundary as the ensembles increases.

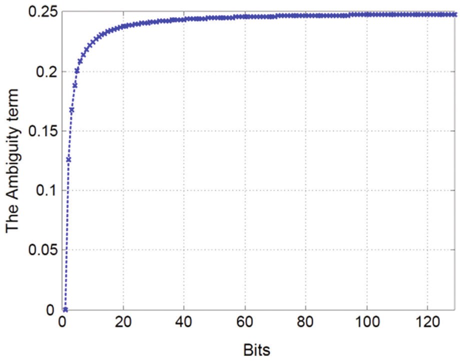


Fig. 2. The evaluation of the Ambiguity term on the CIFAR-10 dataset.

4.2 A Simple Weighted Method

Let $\{\lambda_t\}_{t=1}^{k}$ denote the top k eigenvalues after applying PCA on training data. It is known that the projection dimension with larger eigenvalue λ should contain more information. Define $\{z_u\}_{i=1}^{n} \in \mathbb{R}^k$ as the PCA-projected data, its mean is $\bar{z} = \frac{1}{n} \sum z_i = 0$ and its variance is $\{v_t\}_{t=1}^{k}$ with $v_t = \mathrm{var}(z^t)$. According to the definition of PCA which projects the data along the directions of maximal variances, we have $v_t \Big/ \sum_{t=1}^{k} v_t = \lambda_t \Big/ \sum_{t=1}^{k} \lambda_t$. It denotes that the dimension with larger variance should contain more information. Therefore, more weight should be assigned to the hashing bit with larger variance and we propose to follow the proportion principle that the weight of a hashing bit should be proportional to its variance which is denoted as:

$$\omega_i = v_i \Big/ \sum_{t=1}^{k} v_t \tag{3}$$

4.3 Diverse Hash Tables Learning

From Fig. 2, we find that when enough bits are assigned, the Ambiguity term tends to be invariable in ITQ. Here, we attempt to utilize bagging method to inject diversity among hash tables. Bagging is based on random sampling in all the training data with replacement. It trains a number of base predictors each from a different bootstrap sample by calling a base learning algorithm. Next we will build our model formulation and optimize it in a single hash table.

Model formulation: Firstly, we randomly sample a bootstrap sample $\mathbf{X}^{(l)} = \left\{x_1^{(l)}, x_2^{(l)}, \ldots, x_p^{(l)}\right\}^{\mathrm{T}}$ from the training set $\mathbf{X}$. Combining the weighted method with this bootstrap training set, the optimization function can be written as:

$$\begin{aligned} &\max_{\mathbf{G}^{(l)},\mathbf{W}^{(l)}} \frac{1}{p} tr\left(\mathbf{G}^{(l)}\mathbf{W}^{(l)\mathrm{T}}\mathbf{X}^{(l)\mathrm{T}}\mathbf{X}^{(l)}\mathbf{W}^{(l)}\right) \\ &\quad \text{s.t. } \mathbf{W}^{(l)\mathrm{T}}\mathbf{W}^{(l)}{=}\mathbf{I}_k, \ \sum_i \omega_i = 1 \\ &\quad \omega_i = v_i \Big/ \sum\nolimits_{t=1}^{k} v_t \end{aligned} \tag{4}$$

The constraint $\mathbf{W}^{(l)\mathrm{T}}\mathbf{W}^{(l)}{=}\mathbf{I}_k$ requires the hashing hyperplanes to be orthogonal to each other. $\mathbf{G}^{(l)}$ is a $k \times k$ weighted matrix with diagonal elements to be the weight vector ω_i.

Optimization: To solve the problem above, an alternating way is put forward: updating one variable with others fixed. The relaxed problem can be solved by doing the following two steps iteratively.

(1) Fix $\mathbf{G}$ and update $\mathbf{W}$.

Firstly, we initialize the weight matrix as: $\mathbf{G}^{(l)} = k^{-1}\mathbf{E}$, where $\mathbf{E}$ is an identity matrix. The objective function above can be equivalently shown as:

$$\begin{aligned} &\max_{\mathbf{W}^{(l)} \in \mathbb{R}^{d \times k}} \quad \frac{1}{p} tr\left(\mathbf{W}^{(l)\mathrm{T}}\mathbf{X}^{(l)\mathrm{T}}\mathbf{X}^{(l)}\mathbf{W}^{(l)}\right) \\ &\text{s.t.} \quad \mathbf{W}^{(l)\mathrm{T}}\mathbf{W}^{(l)} = \mathbf{I}_k \end{aligned} \tag{5}$$

This objective function is exactly the same as that of PCA. For a code of k bits, the initial projection matrix $\tilde{\mathbf{W}}^{(l)}$ can be obtained by taking the top k eigenvectors of the data covariance matrix $\mathbf{X}^{(l)\mathrm{T}}\mathbf{X}^{(l)}$.

Inspired by ITQ, after obtaining $\tilde{\mathbf{W}}^{(l)}$, we propose to find an orthogonal transformation to minimum the quantization error measured as $\|\mathbf{B} - \mathbf{V}\mathbf{R}\|_F^2$ with $\mathbf{V} = \mathbf{X}\tilde{\mathbf{W}}^{(l)}$.

The above optimization problem can be solved with respect to **B** and **R** alternatively, these two alternating steps are as follows: The first step is to fix **R**, which begins with a random orthogonal matrix, and update **B** as $\mathbf{B} = \mathrm{sgn}(\mathbf{VR})$. In the second step, for a fixed **B**, we compute the SVD of the matrix $\mathbf{B}^{\mathrm{T}}\mathbf{V}$ as $\mathbf{B}^{\mathrm{T}}\mathbf{V} = \mathbf{S}\boldsymbol{\Omega}\bar{\mathbf{S}}^{\mathrm{T}}$, and then let $\mathbf{R} = \mathbf{S}\bar{\mathbf{S}}^{\mathrm{T}}$. Performing the above two steps alternatively to obtain the optimal hashing codes and the orthogonal transform matrix. The final projection matrix is denoted as $\mathbf{W}^{(l)} = \tilde{\mathbf{W}}^{(l)}\mathbf{R}$. In this paper, we use 50 iterations for all experiments.

(2) Fix **W** and update **G**.

To balance the importance of each hash bit, we attempt to assign different weights to different hashing bits. The variance can be gained directly from the learning hashing function that $v_i = \mathrm{var}(z^i)$, and the weight is computed as $\omega_i = v_i \Big/ \sum_{t=1}^{k} v_t$. Therefore we obtain the weight matrix $\mathbf{G}^{(l)}$.

Multiple diverse hash tables: Repeating the process for L times, we can obtained L diverse hash tables $\mathbf{W} = \left\{\mathbf{W}^{(l)}\right\}_{l=1}^{L}$ and its corresponding weight matrix. Since the hash tables are learned by the same base learning algorithm, the weights to hash tables are uniform that $\omega = 1/L$. Given a new entity $x \in \mathbb{R}^d$, the binary codes matrix generated by one hash tables can be written as $\mathbf{B}_l = \mathrm{sgn}\left(x\mathbf{W}^{(l)}\right)$. We concatenate the L diverse binary codes matrices to form the final codes matrix as: $\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_L] \in \mathbb{R}^{Lk \times 1}$. Finally, our WBPCA-ITQ algorithm is summarized in Algorithm 1.

Algorithm 1 . The WBPCA-ITQ algorithm

```
Input: The original data: $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, $\boldsymbol{x}_i \in \mathbb{R}^d$; number of hash tables $L$;
code length $k$; the size of bootstrap sample $p$; iterations: $iter$.
Generate $L$ bootstrap sample $\left\{\boldsymbol{X}^{(l)} \in \mathbb{R}^{p \times d}\right\}_{l=1}^{L}$;
For $l = 1, 2, \ldots, L$
  Compute initial projection matrix $\tilde{\boldsymbol{W}}^{(l)} \leftarrow PCA\left(\boldsymbol{X}^{(l)}\right)$;
  Initialize the random orthogonal matrix $\boldsymbol{R}$; the weight
  matrix $\boldsymbol{G}^{(l)} = k^{-1}\boldsymbol{E}$;
  Fix $\boldsymbol{G}^{(l)}$ and update $\boldsymbol{W}^{(l)}$.
    for $i = 1{:}iter$ do
        Step-1: Fix $\boldsymbol{R}$, update $\boldsymbol{B} = sgn(\boldsymbol{X}\tilde{\boldsymbol{W}}^{(l)})$;
        Step-2: Fix $\boldsymbol{B}$, update $\boldsymbol{R}$.
                    $SVD(\boldsymbol{B}^T\boldsymbol{V}) = \boldsymbol{S}\boldsymbol{\Omega}\bar{\boldsymbol{S}}^T$
                    $\boldsymbol{R} = \bar{\boldsymbol{S}}\boldsymbol{S}^T$
    end for
  The final projection matrix $\boldsymbol{W}^{(l)} = \tilde{\boldsymbol{W}}^{(l)}\boldsymbol{R}$;
  Fix $\boldsymbol{W}^{(l)}$ and update $\boldsymbol{G}^{(l)}$ using Eqn.(13);
End For
Output: The projection matrix: $\boldsymbol{W} = \left\{\boldsymbol{W}^{(l)}\right\}_{l=1}^{L}$; the weight matrix: $\boldsymbol{G} = \left\{\boldsymbol{G}^{(l)}\right\}_{l=1}^{L}$.
```

5 Experiments

We evaluate our methods on two real-world image data sets, CIFAR-10 [33] and LabelMe [34]. The first dataset is CIFAR-10 which consists of 60K images categorized into 10 classes namely air- plane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The size of each image is 32 × 32 pixels, we represent them with 320-dimensional gray-scale GIST descriptors [35]. The second dataset is 22K LabelMe used in [13, 34] that consists of 22,019 images. The images are scaled to 32 × 32 pixels, and then represented by 512-dimensional GIST descriptors. For each dataset, we randomly select 1,000 data points as queries and use the rest as gallery database and training set.

To perform fair evaluation, we adopt the Hamming Ranking search commonly used in the literature. All points in the database are ranked according to their Hamming distance to the query and the top K samples will be returned. The groundtruth of each query instance is defined as its 50 nearest neighbours based on Euclidean neighbours [36]. The retrieval performance is measured with three widely used metrics: mean average precision (MAP), precision of the top K returned examples and precision-recall curves.

We compare the proposed WBPCA-ITQ with several state-of-the-art hashing algorithms including LSH, SH, PCAH, ITQ, BPCAH. We also compare it with our WPCAH and WPCA-ITQ which just use the simple weighted scheme on PCAH and ITQ respectively.

There are two parameters to be set, the size of each bootstrap training set p and the size of each individual (ensemble) k. The value of code length k has great effect on the ensemble performance. Here, we set $p = 20\% \times n$ and $k = 16$ for all the comparisons. The ensemble code length is denoted as $L \times k$. All the comparison methods are under the same code length with our proposed method.

Figure 3 evaluates the retrieval precision for top 500 returned images with different number of bits on the datasets CIFAR-10 and LabelMe. As for our proposed WBPCA-ITQ method, it consistently performs better than the competitors BPCAH and ITQ. For the ITQ method, we find that its performance increases rapidly for the

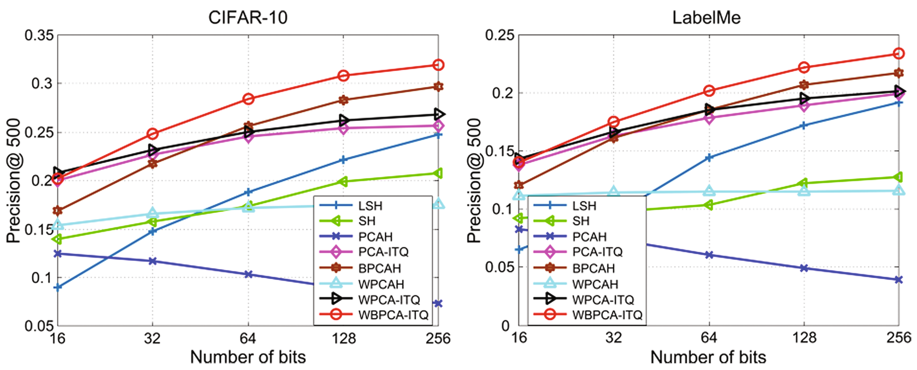


Fig. 3. Precision of top K returned images with different number of bits on the two datasets.

smallest code sizes, but then begin to level off as the code size increases. The cause is that when enough bits are assigned, the Ambiguity term tends to be invariable, so does the final ensemble result. Both BPCAH and our proposed method attempt to utilize bagging method to inject diversity among hash tables, where BPCAH can be seen as a specific example of our ensemble hashing framework whose base learner is PCA-RR rather than PCAH. Because our proposed method is based on the more effective base learner WPCA-ITQ, its ensemble performance can be more effective. Specially, we also find that increasing number of bits leads to poorer precision performance for PCAH. However, by intuition, more hash codes should catch more information and should give better retrieval performance. The phenomenon can be visually illustrated that the increasing weighted average error of hash bits will increase the final ensemble error, so decrease the ensemble performance.

Figure 4 illustrates the precision and precision-recall curves respectively using 64 and 128 bits on CIFAR-10 and LabelMe data sets. The curves confirm the trends seen in Fig. 3. Our proposed method WBPCA-ITQ consistently achieves the superior performance to other methods. We observe that WBPCA-ITQ achieves the highest MAP scores with different code lengths on all the datasets.

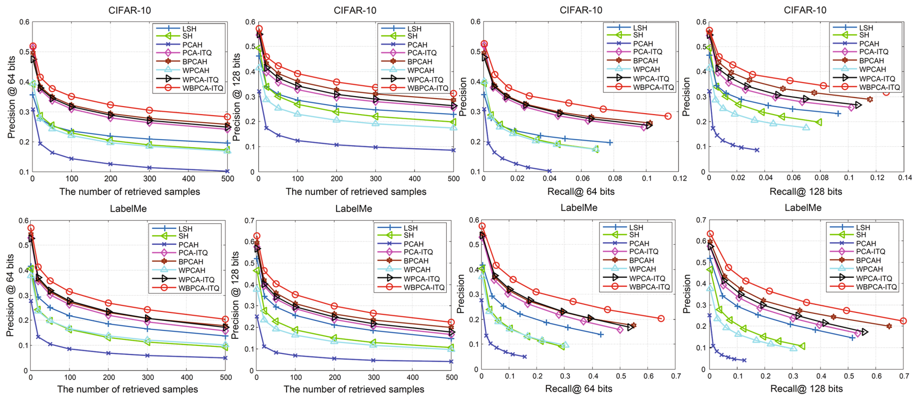


Fig. 4. The precision and precision-recall curves respectively using 64, 128 and 256 bits on CIFAR-10 and LabelMe datasets.

6 Conclusion

We have shown that the ensemble hashing framework, which places three important principles namely higher accuracy, larger diversity and the optimal weights coefficient for predictors into a common framework, is effective to guide the design of hashing methods. There are several possible area to explore in the future research. Here we will describe briefly about our future ideas as follows: (1) The local information of data can be exploited to design more effective ensemble hashing methods in the future work. (2) As noted in Deep Supervised Hashing (DSH) [37], the authors investigated network ensembles with different random initializations for retrieval problem. Under the same code length, the ensemble codes further improve the retrieval performance. Although this discovery is very interesting and useful, the authors failed to offer further theoretical analysis to it. In the future, we would like to use the robust CNN structure with our ensemble hashing framework to further boost retrieval performance.

Acknowledgments. This research was supported by the national natural science fund (No. 61472442, 61502522, and 61502523).

References

1. Alamareen, A., Al-Jarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**(3), 1–14 (2016)
2. Kong, W., Lei, Y., Zhao, R.: Fusion technique for multi-focus images based on NSCT–ISCM. Optik Int. J. Light Electron Opt. **126**(21), 3185–3192 (2015)
3. Kong, W., Lei, Y., Ren, M.: Fusion technique for infrared and visible images based on improved quantum theory model. Neurocomputing **35**(2), 1637–1640 (2016)
4. Li, W., Zhou, Z.: Learning to hash for big data: current status and future trends. Chin. J. **60** (5–6), 485 (2015)
5. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: NIPS, pp. 490–512 (2008)
6. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2916–2929 (2013)
7. Kong, W., Li, W.: Isotropic hashing. In: Advances in Neural Information Processing Systems (2012)
8. Liu, W., Wang, J., Kumar, S., Chang, S.F.: Hashing with graphs. In: ICML (2011)
9. Xu, B., Bu, J., Lin, Y., Chen, C., He, X., Cai, D.: Harmonious hashing. In: Proceedings of International Joint Conference on Artificial Intelligence (2013)
10. Leng, C., Cheng, J., Yuan, T., Bai, X., Lu, H.: Learning binary codes with bagging PCA. In: ECML (2014)
11. Wang, J., Kumar, S., Chang, S.-F.: Semi-supervised hashing for large-scale search. TPAMI **34**, 2393–2406 (2012)
12. Kulis, B., Darrell, T.: Learning to hash with binary reconstructive embeddings. In: NIPS, pp. 1042–1050 (2009)
13. Norouzi, M., Fleet, D.J.: Minimal loss hashing for compact binary codes. In: Proceedings of the 28th International Conference on Machine Learning (ICML) (2011)
14. Liu, W., Wang, J., Ji, R., Jiang, Y., Chang, S.: Supervised hashing with kernels. In: Conference on Computer Vision and Pattern Recognition, Providence, pp. 2074–2081 (2012)
15. Lin, G., Shen, C., Shi, Q., van den Hengel, A., David, S.: Fast supervised hashing with decision trees for high-dimensional data. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, Providence (2014)
16. Shen, F., Shen, C., Liu, W., Shen, H.T.: Supervised discrete hashing. In: CVPR, pp. 37–45 (2015)
17. Wang, J., Kumar, S., Chang, S.F.: Semi-supervised hashing for large-scale search, pp. 2393–2406 (2012)

18. Jiang, Y., Wang, J., Chang, S.F.: Lost in binarization: query-adaptive ranking for similar image search with compact codes. In: ICMR (2011)
19. Zhang, X., Zhang, X., Shum, H.Y.: Qsrank: Querysensitive hash code ranking for efficient ε-neighbor search. In: CVPR (2012)
20. Zhang, X., Zhang, Y., Tang, J., Lu, K., Tian, Q.: Binary code ranking with weighted hamming distance. In: CVPR (2013)
21. Xu, H., Wang, J., Li, Z., Zeng, G., Li, S., Yu, N.: Complementary hashing for approximate nearest neighbor search s. In: ICCV, pp. 1631–1638 (2011)
22. Jollie, I.T.: Principal Component Analysis. Springer, New York (1989)
23. Jégou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: CVPR (2010)
24. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing. IEEE Trans. Pattern Anal. Mach. Intell. **34**, 1092–1104 (2011)
25. Mu, Y., Yan, S.: Non-metric locality-sensitive hashing. In: AAAI (2010)
26. Raginsky, M., Lazebnik, S.: Locality-sensitive binary codes from shift-invariant kernels. In: NIPS, pp. 1509–1517 (2009)
27. Breiman, L.: Bagging predictors. Mach. Learn. **24**, 123–140 (1996)
28. Dietterich, T.: Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems, pp. 1–15 (2000)
29. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: NIPS, pp. 231–238 (1995)
30. Ueda, N., Nakano, R.: Generalization error of ensemble estimators. In: Proceedings of International Conference on Neural Networks, pp. 90–95 (1996)
31. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Machine Learning, pp. 148–156 (1996)
32. Miguel, A.C.P., Ramin, R.: An ensemble diversity approach to supervised binary hashing (2016). arXiv:1602.01557 [cs.LG]
33. Krizhevsky, A.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto, 8 April 2009
34. Torralba, A., Fergus, R., Weiss, Y.: Small codes and large image databases for recognition. In: Proceedings of Computer Vision and Pattern Recognition (2008)
35. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. **42**, 145–175 (2001)
36. Yu, F.X., Kumar, S., Gong, Y., Chang, S.: Circulant binary embedding. In: ICML, pp. 946–954 (2014)
37. Haomiao, L., Ruiping, W., Shiguang, S., Xilin, C.: Deep supervised hashing for fast image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

A Novel Construction and Design of Network Learning Platform in Cloud Computing Environment

Huanyu Li[1(✉)], Jing Wang[2], and Hua Shen[1]

[1] Air Force Engineering University, Xi'an, Shanxi, China
{lihuanyu1984, king448705776}@163.com
[2] Xi'an Aerospace Middle School, Xi'an, Shanxi, China
Wangjing63363@163.com

Abstract. With the rapid development of computer technology, educational information technology plays a more and more important role in modern education. The development of cloud computing technology has brought great influence and change to the construction of network learning platform. Aiming at the problem of design and implementation under the environment of network learning platform for cloud computing, through in-depth analysis of the current network learning under cloud computing environment, learning theory, learning object and system function, system construction is given a practical significance of the cloud computing network learning platform and system design method, provides the theoretical basis and reference for promoting the application of cloud computing technology in modern education.

1 Introduction

With the rapid development of information technology, cloud computing has been widely used in the field of education for its advantage and characteristics. However, the application of cloud computing in higher education is still in its early stage in China [1–3]. As the organization of knowledge and the resource center of education information, colleges and universities need to provide information services through modern information technology today [4, 5]. At the same time, as a novel web application mode which is of high reliability, high cost-effective and high scalability, the cloud computing technology exactly meet the demands of effective utilization on higher education information. So it is an inevitable choice to apply the cloud computing technology in higher education field.

At present, some colleges and universities use the cloud computing technology to integrate the massive educational information resources, and develop education platforms of various functions, such as the resource sharing platform designed by Hongmin Gong [6], the education service platform designed by Tiebin Tang [7] and the digital resource management platform designed by Junke Song [8], and so on [9–13]. These educational platforms have improved the education level of colleges and universities, enrich the education form in colleges and universities, and expand the pattern of personality study. Therefore, addressing at the problem of multimedia teaching in

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_18

colleges and universities, by analyzing the advantages of the cloud computing technology deeply, a novel construction and design of multimedia learning platform of colleges and universities has been proposed in this paper, which is based on the cloud computing technology. The theoretical foundation, object analysis and design principle are discussed in this paper to illuminate the rationality and validity of the construction and design. And the functions and applies of the multimedia learning platform are analyzed and demonstrated. This work will provide several theoretical references for the application of the cloud computing technology in higher education in the further.

2 Theoretical Foundation

2.1 Learning Theory

The network learning platform based on cloud computing is a platform that based on multimedia resources in the network environment. The traditional learning theory and the new learning theory under the network environment have a very important practical significance for the design and construction of the cloud based network learning platform [14].

The purpose of using network learning platform is to improve the quality of education. To cultivate college students' learning interest and learning ability, also to solve the problems and problems in classroom learning, to cultivate their creativity and self-learning ability. Therefore, to guide the network learning platform for cloud computing based on the theory of learning should be diversified, mainly including behaviorism learning theory, constructivist learning theory, humanistic learning theory, educational communication theory, learning theory and learning theory of micro Unicom etc.

2.2 Object Analysis

Object analysis, also called learner analysis, is a key part of instructional design. The purpose of this study is to understand the learner's learning readiness and learning style. In order to determine the students' learning task, importance and difficulty of teaching, from teaching a starting point, clear target system, the design of teaching activities, teaching contents, teaching strategies of the organization and arrangement of the selection and use of teaching media, teaching evaluation and the use of the design to provide the basis. It is the concrete embodiment of the teaching idea of "student is the foundation".

For network learning this form of learning, object analysis in addition to the need to analyze on the learners, learning environment should also be analyzed, because only by understanding and considering the particularity of the learning environment, learners can better analyze play its role in guiding teaching. Through the analysis of the object, it can be seen that the network learners are more diverse, individual, independent and interactive than the traditional learners. At the same time, it is found that autonomous learning is a process of active and independent self-knowledge construction [13]. The characteristics of online learning requires learners to have a certain amount of knowledge, positive learning motivation and correct learning strategies, and the need for efficient hardware facilities and a wealth of learning resources. Therefore, it is very important to carry out a comprehensive analysis of the design of the entire network learning platform.

2.3 Design Principles

The network learning platform based on cloud computing is an open learning resource integration environment, which can integrate various resources. The combination of teaching and learning by means of information technology [15]. Therefore, it should be designed in accordance with the following principles:

(1) It should be an interactive, heuristic and autonomous online learning system, so as to provide effective support for the cultivation innovative ability of students.
(2) It should be independent, flexible convenient, and with a powerful background management system. Then all teachers can easily modify their teaching plans and manage their teaching resources.
(3) It should establish a flexible dynamic management for all kinds of multimedia source materials, such as text, film and television, animation, sound, pictures and other multimedia materials. And it has to provide a flexible retrieval mode.
(4) It can realize scientific, reasonable and effective online instruction that online or asynchronous based on expert knowledge and teacher experience.
(5) It should be of fully functional, and easy to operate.

The function of the platform should be able to meet the requirements of students' autonomous learning, such as course learning, data downloading, online testing, collaborative communication and so on. The use of the platform should be as simple as possible, to take into account the level of computer literacy of junior high school students.

3 Demand Analysis

3.1 Network Learning Environment in Cloud Computing

The network learning environment in cloud computing environment contains four important components: client, network learning system, server cluster in the cloud, database in the cloud. The relationship and data flow between them can be described as Fig. 1.

3.2 Application Analysis

Application analysis is an important stage of software development, and it is an effective guarantee for the whole design process of the software. So it is necessary to pay more attention to application analysis at first. Here, the application analysis will achieve a series of related tasks in software development, such as the specific

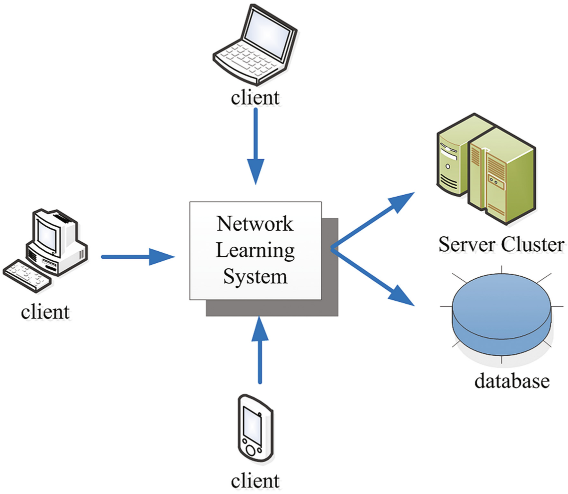


Fig. 1. The relationship of components in network learning environment

stipulation of technology requirements, the explicit functions of the each modules, so as to reduce the workload in design, and improve the efficiency in development and test, reduce the workload caused by rework.

In network learning platform in cloud computing environment, applications can be divided into two categories, functional and non-functional. Functional applications are primarily user oriented, they are directly related to the user's actual application. The demarcation and fundamental function of each module contained in the platform should be clarified. They are the most concerned contents by the user. Non-functional applications is developer oriented, they are designed to analysis the technical requirements, which are used to guarantee the normal operation and the post-maintenance of the system.

As in common system in cloud computing environment, application layer and resource layer are the most important constituent part in our proposed network learning platform. Application are realized in application layer, this layer provide user interface to all kinds of user, and do data read-write with resource layer. Overall, the most important functional application contain three categories. The class about knowledge learning that contains instructional design, teaching management and knowledge hierarchy. The class about learning place that contains distance training, virtual classroom, and network course. The class about entrance that contains several network college and universities. The applications can be described in a block diagram as shown in Fig. 2.

3.3 Function Analysis

Through a concrete analysis of application of network learning platform system based on cloud computing, the system function is designed for the user management module, public information module, online learning support module, tool module and data processing and analysis module of interactive learning module, as shown in Fig. 3.

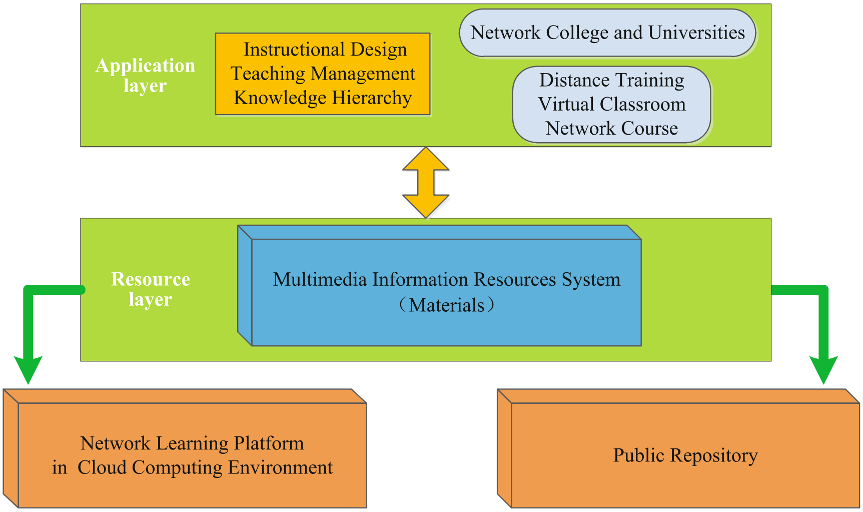


Fig. 2. Application block diagram of proposed network learning platform

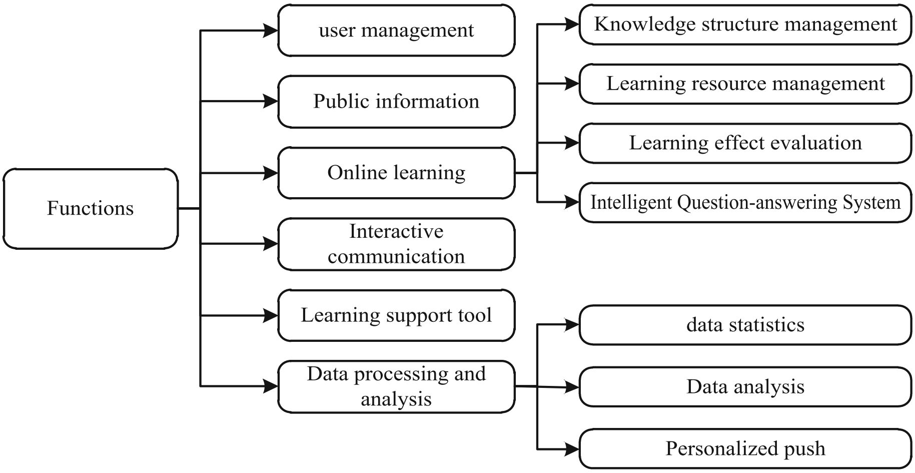

Functions
user management
Public information
Online learning
Interactive communication
Learning support tool
Data processing and analysis
Knowledge structure management
Learning resource management
Learning effect evaluation
Intelligent Question-answering System
data statistics
Data analysis
Personalized push

Fig. 3. Function block diagram of proposed network learning platform

The user management module diagram to ensure the legality of access and operation for the learning of the users of the system; public information module has issued the basic information of the website maintenance and information announcement function; online learning module is a network learning platform for cloud computing core module based on the realization of it is responsible for the teaching function. The interactive communication module is a module of communication between students and students, between students and teachers and between teachers and teachers. Learning support tool mainly consists of teachers' teaching experience according to their own design, and in accordance with the provisions of the standard complete summary information files and documentation in the form of attachment, then uploaded to the platform through the network system. Data processing and analysis module is to achieve intelligent learning process management, intelligent learning content recommendation, the learning platform for the intelligent check missing trap and other aspects of the foundation to realize.

4 System Architecture

4.1 The Architecture of Our Proposed Platform

Network learning platform for cloud computing using the hierarchical architecture based design, according to the application as the center, the construction of ideas to function as the goal, building architecture, the hierarchical idea extends from a single business application architecture to the system, according to different components of the system physical and logical characteristics in the system within the scope of the level of stratification.

The layered structure of this system, can give full play to the function of the cloud computing platform for large-scale distributed system resource gathering, management and scheduling, can be extended to provide high performance communication, distributed storage and computing ability, and integration of the concept of SOA, to provide a unified support for the data in the system range, life cycle management, support service interaction management, reliability and availability management, realize the loose coupling architecture within the scope of system.

Flexible, transparent, building blocks, dynamic, universal and multi lease, is the six core technology ideas of cloud computing platform. According to the characteristics of cloud computing technology and the practical needs of the construction of network learning platform, this paper designs the network learning platform system as shown in Fig. 4. The learning platform system is composed of learning resources, storage space, computing resources, application system and operating interface. Among them, the user terminal, including user and client is the cloud computing system of the consumer, the application layer, platform layer, data management layer and infrastructure layer is the cloud computing system's supporters and service providers.

4.2 The Modular Construction of Our System in the Cloud

The network learning platform design uses a top-down idea of modular design, in order to make system toward distributed, miniaturization, direction, and enhance the system scalability and operation stability.

In our proposed network learning platform, it is composed by six modules, as shown in Fig. 5.

Operation Maintenance Module. This module is responsible for the entire system configuration and control, which including rights management, monitoring alarm, applications management, faults management and so on.

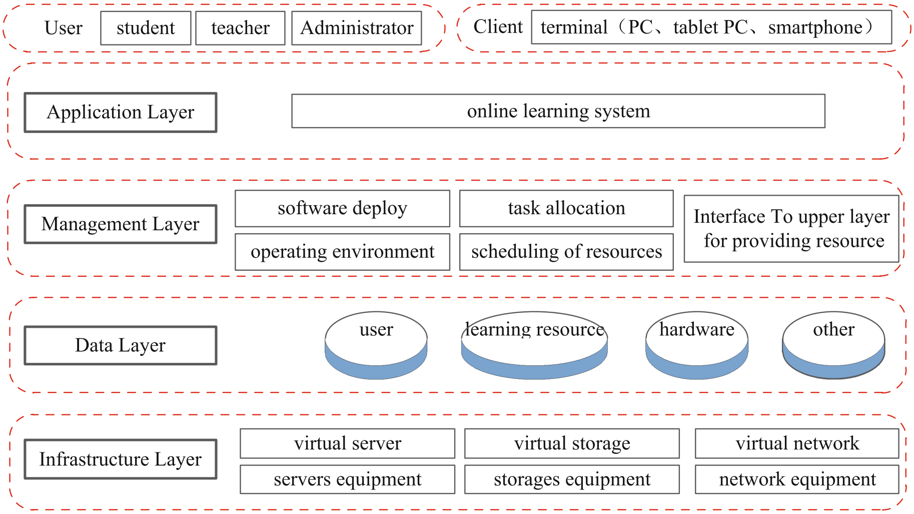


Fig. 4. The system design of our network learning platform based on cloud computing

Data Acquisition Module. This module is responsible for collecting data from external interface. Configuration management, data filtering, data preprocessing, and other tasks are implemented in this module.

Data Storage Module. This module is most important part of the whole system. Its kernel is integrated data management cell. This module contains relational database cluster, distributed real-time data, distributed file systems. This module is responsible for distributing the data synchronization of the whole system, database management, access control, redundant strategy implementation, and so on.

Data Services Module. There is a web service APT in this module, to exchange important date with service delivery module. This module contains a service management system, which is responsible for service registration, alteration, design, review, distribution and cancellation. This module also provides data access services and business logic services.

Service Delivery Module. This module achieves load balancing. And it also provides services for the management of the former servers and functional business delivery management of all data.

Data Analysis Module. This module use a distributed computing model to do data transformation, data aggregation, data correlation, and data mining.

All the six modules are mounted on a uniform distributed data bus. Using the distributed data bus to exchange data and order with each other.

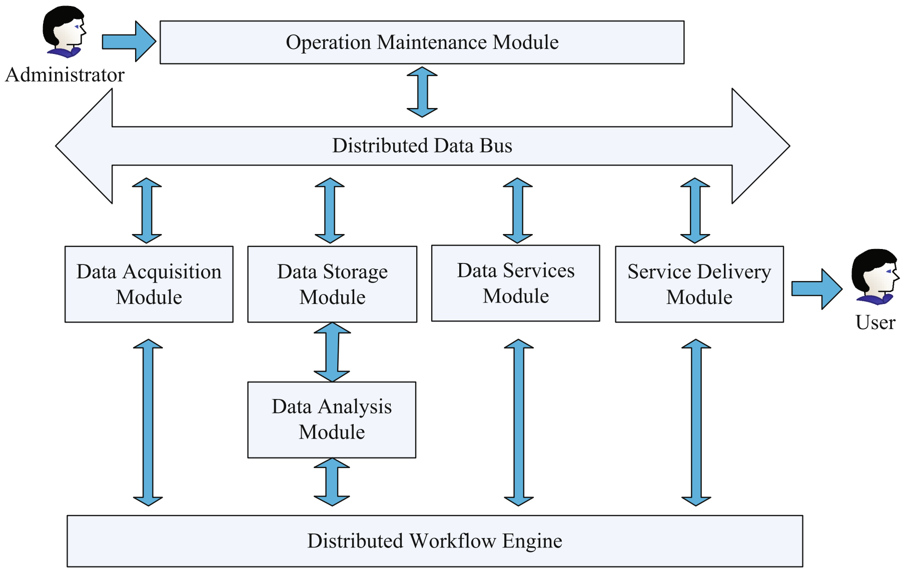


Fig. 5. The modular construction of our system

5 Conclusion

The development of cloud computing technology is gradually changing the way of higher education today, all kinds of education resources platform construction based on cloud computing is an important part of the application of cloud computing in higher education. This paper focuses on the advantages of cloud computing and the demand for higher education, research and analysis learning theory and application requirements of network learning platform based on the cloud computing and detailed design functional requirements of each module, and on this basis, construct the multi-media learning platform architecture and system, which can provide theoretical and technical reference for the integration, sharing and maximum utilization of information resources in colleges and universities.

Acknowledgments. This research was supported by the national education information technology research subject (No. 144442192), the national natural science fund (No. 61472442, 61502522, and 61502523).

References

1. Qian, L., Luo, Z., Du, Y.: Cloud Computing: An Overview. LNCS, vol. 5931, pp. 626–631 (2009)
2. Wang, Y., Chen, I.-R., Wang, D.-C.: A survey of mobile cloud computing applications: perspectives and challenges. Wireless Pers. Commun. **80**(4), 1607–1623 (2015)

3. Ma, K., Zhang, L.: Bookmarklet-triggered unified literature sharing services in the cloud. Int. J. Grid Util. Comput. (IJGUC) **5**(4), 217–226 (2014)
4. Serhani, M.A., Atif, Y., Benharref, A.: Towards an adaptive QoS-driven monitoring of cloud SaaS. Int. J. Grid Util. Comput. (IJGUC) **5**(4), 263–277 (2014)
5. Jin, M.: Review on the application of cloud computing in China. Commun. World **06**, 325 (2015)
6. Gong, H.: Research on High Quality Resource Sharing Platform Based on Cloud Computing Environment. Shanxi Normal University (2013)
7. Tang, T., Hao, Q.: Research and design of cloud computing education service platform. J. Changsha Soc. Work Coll. **20**(4), 160–161 (2013)
8. Song, J.: Research and implementation of digital resource management platform in educational cloud. East China Normal University (2015)
9. Yao, Z., Xiong, J., Ma, J., et al.: Access control requirements for structured document in cloud computing. Int. J. Grid Util. Comput. (IJGUC) **4**(2/3), 95–102 (2013)
10. Alamareen, A., Al-Jarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**(3), 1–14 (2016)
11. Khan, N., Al-Yasiri, A.: Cloud security threats and techniques to strengthen cloud computing adoption framework. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**(3), 50–64 (2016)
12. Mezghani, K., Ayadi, F.: Factors explaining IS managers attitudes toward cloud computing adoption. Int. J. Technol. Hum. Interact. (IJTHI) **12**(1), 1–20 (2016)
13. Yuriyama, M., Kushida, T.: Integrated cloud computing environment with IT resources and sensor devices. Int. J. Space-Based Situated Comput. (IJSSC), **1**(2/3), 163–173 (2011). doi:10.1504/IJSSC.2011.040342
14. Raekow, Y., Simmendinger, C., Jenz, D., et al.: On-demand software licence provisioning in grid and cloud computing. Int. J. Grid Util. Comput. (IJGUC) **4**(1), 10–20 (2013)
15. Achuthan, S., Chang, M., Shah, A.: SPIRIT-ML: a machine learning platform for deriving knowledge from biomedical datasets. LNCS, vol. 9162, pp. 240–250 (2015)

Automatic Kurdish Text Classification Using KDC 4007 Dataset

Tarik A. Rashid[1(✉)], Arazo M. Mustafa[2], and Ari M. Saeed[3]

[1] Department of Computer Science and Engineering,
University of Kurdistan Hewler, Erbil, Kurdistan, Iraq
tarik.ahmed@ukh.edu.krd
[2] School of Computer Science, College of Science,
University of Sulaimania, Sulaymaniyah, Kurdistan, Iraq
arazo.2007@yahoo.com
[3] Department of Computer Science, College of Science,
University of Halabja, Halabja, Kurdistan, Iraq
arimohsaeed@gmail.com

Abstract. Due to the large volume of text documents uploaded on the Internet daily. The quantity of Kurdish documents which can be obtained via the web increases drastically with each passing day. Considering news appearances, specifically, documents identified with categories, for example, health, politics, and sport appear to be in the wrong category or archives might be positioned in a nonspecific category called others. This paper is concerned with text classification of Kurdish text documents to placing articles or an email into its right class per their contents. Even though there are considerable numbers of studies directed on text classification in other languages, and the quantity of studies conducted in Kurdish is extremely restricted because of the absence of openness, and convenience of datasets. In this paper, a new dataset named KDC-4007 that can be widely used in the studies of text classification about Kurdish news and articles is created. KDC-4007 dataset its file formats are compatible with well-known text mining tools. Comparisons of three best-known algorithms (such as Support Vector Machine (SVM), Naïve Bays (NB) and Decision Tree (DT) classifiers) for text classification and TF × IDF feature weighting method are evaluated on KDC-4007. The paper also studies the effects of utilizing Kurdish stemmer on the effectiveness of these classifiers. The experimental results indicate that the good accuracy value 91.03% is provided by the SVM classifier, especially when the stemming and TF × IDF feature weighting are involved in the preprocessing phase. KDC-4007 datasets are available publicly and the outcome of this study can be further used in future as a baseline for evaluations with other classifiers by other researchers.

1 Introduction

In recent years, there is an enormous amount of machine readable data stockpiled in files and databases in a form of text documents. Text classification is one of the most common and convenient techniques for information exchange, in the meanwhile, much of the world's data can be found in text forms such as newspaper articles, emails,

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_19

literature, web pages, etc. The rapid growth of the text databases is due to the increased amounts of information available in electronic forms such as e-mails, the World Wide Web, electronic publications, and digital libraries. Text mining can be defined as the process of discovering meaningful and interesting linguistic patterns from a large collection of textual data, and it is relevant to both information retrieval and knowledge discovery in databases [1–3].

In general, data mining is an automatic process of finding useful and informative patterns among large amounts of data or detecting new information in terms of patterns or rules from that enormous amount of data. Data mining usually deals with structured data, but information stored in text files is usually unstructured and difficult to deal with, and to deal with such data, a pre-processing is required to convert textual data into an appropriate format for automatic text processing. The purpose of text mining is to process unstructured textual, extract non-trivial patterns or meaningful pattern from the text, make the information included in the text accessible to the different data mining algorithms and reduce the effort required from users to obtain useful information from large computerized text data sources [1, 3]. Text mining generally multidisciplinary domain, thus, research works in texts involve dealing with problems such as text representation, text analysis, text summarization, information retrieval, information extraction, text classification and document clustering. In all these problems, data mining and statistical techniques are used to process textual data [1, 2].

One of the most widely utilized procedures in the text mining studies is the procedure of text classification which is addressed in general as is the task of learning under the supervisor. Text classification is a process of automatically classifying unstructured documents into one or more pre-defined categories such as science, art or sport… etc., based on linguistic features and content. The procedure can be depicted as a natural language problem and the aim of this paper is to reduce the need of manually organizing the huge amount of text documents. In the field of text classification problem, very few research works have been studied for the Kurdish Sorani language. Therefore, this field is at initial stages. It should be noted that due to the progress of the World Wide Web, and the increased number of non-English users, many research efforts for applying pre-processing approaches for other languages have been documented in literature. One of the most criteria in this framework is applying the text classification problem on Kurdish Sorani text documents.

2 Literature Survey

Since Kurdish Sorani script is considered as the closest to the Arabic language, and technically both have a written system that is from right to left. Thus, in this literature study, the research works in the text classification field have been sorted starting with the Kurdish language, shadowed by Arabic language, and then followed by English language.

Mohammed et al. in 2012 used the N-gram Frequency Statistics for classifying Kurdish text. An algorithm called Dice’s measure of similarity was employed to classify the documents. A corpus of Kurdish text documents was build using Kurdish Sorani news which consisted of 4094 text files divided into 4 categories: art, economy, politics, and sport. Each category was divided equally per their sizes (50% as a training set and 50% as a testing set). The Recall, Precision and F1-measure were used to compare the performance. The results showed that N-gram level 5 outperformed the other N-gram levels [4].

In [5], Al-Kabi M et al., in 2011, conducted comparison between three classifiers Naïve Bayes classifier, Decision Tree using C4.5 Algorithm and Support Vector Machine to classify Arabic texts. An in-house collected Arabic dataset from different trusted websites is used to estimate the performance of those classifiers. The dataset consisted of 1100 text documents and divided into nine categories: Agriculture, Art, Economics, Health, Medicine, Law, Politics, Religion, Science, and Sports. Additionally, pre-processing (which included word stemming and stop words removing) was conducted. The experiments showed that three classifiers achieved the highest accuracy in cases that did not include stemming. While the accuracy was decreased when using stemming. This means that the stemming had impacted negatively on the performance of the classification accuracy of the three classifiers.

In [6], Mohammad AH et al., in 2016 studied the performance of three well-known machine learning algorithms Support vector machine, Naïve Bayes and Neural Network (NN) on classifying Arabic texts. The datasets consisted of 1400 Arabic documents divided into eight categories collected from three Arabic news articles namely: Aljazeera news, Saudi Press Agency (SPA), Alhayat. In terms of performance, three evaluation measures were used (recall, precision and F1-measure). The results indicated that SVM algorithm outperformed NB and NN and F1-measure for three classifiers were 0.778, 0.754, and 0.717 respectively.

In [7], Mohsen AM et al., in 2016 conducted study to compare the performance of different well known machine classifiers to classifying emotion documents. The ISEAR dataset was applied. It consisted of 7,666 documents belonging to five categories namely: Anger, Disgust, Fear, Joy and Guilt. Tokenization, stop word removal, stemming and lemmatization as preprocessing tasks and TF-IDF as term weighting. Also, two lexicons were used which are NRC emotion lexicons (National Research Council of Canada) and SentiWordNet sentiment lexicons. Based on the obtained results, the authors concluded that Logistic Model Tree (LMT) is the most appropriate classifier in comparison with the other algorithms for English emotion documents classification.

3 Text Mining Functionalities

In this paper, three types of classification techniques are used as described in the following subsections.

3.1 Naive Bayes Classifier

Naive Bayes classifier is a probabilistic based approach which is based on Baye's theorem. It is a simple and efficient to implement [8, 9]. Naive Bayes classifier underlies on the assumption that the features (words) in the dataset are conditionally independent which is computing the probability of each by figuring the frequency of features (words) and the relevance between them in the dataset [8]. Despite that the features independence assumption is unrealistic, Naive Bayes has been discovered extremely effective for many functional applications, for example, text classification and medical diagnosis, even when the dimensionality of the input is high [9]. Advantages of NB would include simplicity, efficiency, robustness and interpretability, while the main disadvantage of NB is that it does not work properly with data having noises. Thus, remove all the noises before applying NB classifier is the need [10].

3.2 Decision Tree Classifier

Decision tree algorithm is widely used in machine learning and data mining. It is also simple and can be easily understandable and converted into a set of humanly readable if-then rules [11]. The decision tree mechanism is used to test some feature values of unseen instances at each node for classifying or finding the class of a given unseen instance where the test starts at the root node and goes down to a leaf node. Information gain is a suitable measure for choosing the best feature where the feature with highest information gain is chosen to be the root node [12]. Advantages of the decision tree can include straightforwardness, interpretability and capacity to handle feature interactions. In addition, the decision tree is nonparametric, which makes issues like exceptions and whether the dataset is linearly divisible [8]. Disadvantages of the decision tree can include the lack of support for online learning and suffer from the issue of over fitting, which can be handled using different strategies like random forests (or boosted trees) or perhaps the problem of over fitting could be avoided by pruning the tree [8].

3.3 Support Vector Machine Classifier

SVM is a supervised machine learning algorithm and was proposed for text classification by [13]. Researchers have used SVM widely in a text categorization task, such as in [8]. In N-dimensional space, input points are mapped into a higher dimensional space and then a maximal separating hyperplane is found. SVM technique classification depends on the Structural Risk Minimization principal [14]. The linear Kernel function is used in this paper scope as there is a very large number of features in the document classification problem. Thus, SVM is suitable for text categorization problems due to their ability to learn [8]. Advantages of SVM can involve high accuracy, and has great theoretic guarantees about overfitting [8]. Similarly, they work well regardless of whether the data is linearly separable or not. Disadvantages of SVM can involve complexity, poor interpretability and high memory requirements.

4 Methods and Materials

Before plunging into the details of the used methods and material in this paper, it is worth mentioning an overview about Kurdish language. The Kurdish language belongs to the Indo-European family of languages. This language is spoken in the geographical area spanning the intersections of Iran, Iraq, Turkey, and Syria [15]. However, Kurds have lived in other countries such as Armenia, Lebanon, Egypt, and some other countries since several hundred years ago, [16]. The Kurdish language is generally divided into two widely spoken and most dominant dialects, namely; Sorani and Kurmanji. Kurdish is written using four different scripts, which are modified Persian/Arabic, Latin, Yekgirtû (unified), and Cyrillic [16]. The Persian/Arabic script is mainly used in Sorani dialect for writing. Kurdish Sorani text documents have used in this research. The Sorani text is more complex with its reading from right-to-left and its concatenated writing style. The Kurdish Sorani character set is consisted of 33 letters which are shown in Fig. 1.

ا ە ب پ ت ج چ ح خ د ر ڕ ز ژ س ش ع غ ف ڤ ق ک گ ل ڵ م ن و ۆ وو ھ ی ێ

Fig. 1. The Kurdish Sorani alphabets.

In the next subsections, Kurdish Sorani pre-processing steps, data representation and term weighting are explained.

4.1 Kurdish Sorani Pre-processing Steps

Dataset pre-processing is an important stage in text mining. A huge number of features or keywords in the documents can lead to a poor performance in terms of both accuracy and time. For the problem of text classification, a document, which typically has high dimensionality of feature space and most of the features (i.e., terms) are irrelevant to the classification task or non-informative terms. The main objective of the pre-processing steps is to prepare text documents which are represented by many features for the next step in text classification. The proposed model of the pre-processing steps for Kurdish Sorani text documents was introduced by authors in [17]. The most common steps for the Kurdish pre-processing steps are tokenization, normalization, stop-word filtering, and Kurdish stemming. The proposed Kurdish stemming-step module of the pre-processing stage is a step-based approach to stages via which a word goes through before arriving at the extracted root of the word. This stemmer determines the words that have several affixes (e.g. a word that has 'prefix' + 'root' + 'suffix1' + 'suffix2' + ••• + 'suffixN'). This approach is not only utilized for stripping affixes from nouns and verbs as it is used in other languages, but it is also used to strip affixes from the stop words. A list that contains nearly 240 stop words (words that are widely used in Kurdish Sorani) has been used [17].

4.2 Data Representation and Term Weighting

The text representation model is a process of transforming a document from a series of characters into sequences of words so that to be appropriate for learning the algorithm and the classification task. The representation of text document can be coded as a form of a matrix, where columns indicate words that distinguish the objects (text documents) stored in each row, where each apparent word is a feature and the number of times the word occurs in the text document is its value [18]. The most common technique for text representation in the text classification task is the bag of words (BOW) [18]. This method of document representation is also known as Vector Space Model (VSM); this is the most common and easy way of text representation [19]. Each term that appears in documents must be represented to a machine learning classifier as real-number vectors of weights [19]. Thus, per a text classification research, the weighting of the term can be divided into three major approaches [20]:

4.2.1 Boolean or Binary Weighting

This is the simplest way of encoding for the term weighting. If the corresponding word (term) is used in the document d_j at least once, then it is set to 1 otherwise to 0 [19, 20].

4.2.2 Term Frequency (TF)

In term frequency weighting scheme, an integer value indicates the number of times that the term appears in a specific document d_j [19].

4.2.3 Term Frequency Inverse Document Frequency (TF × IDF)

TF-IDF can be considered as the most accurate application for text categorization domains with more than two categories. The TF-IDF weights are typically preferred over the other two options [20]. It is a straightforward and efficient method for weighting the terms in text documents categorization purposes [18]. In this work, the TF-IDF weighting function is used which is based on the distribution of the terms within the document, and within the collection, where the higher value indicates that the word occurs in the document, and does not occur in many other documents, and in inverse amount to the number of documents in the collection for which, the word occurs at least one time [20]. This is can be calculated as follows:

$$\mathrm{TF.IDF}\left(\mathrm{t_i}, \mathrm{d_j}\right) = \mathrm{TF}\left(\mathrm{t_i}, \mathrm{d_j}\right) * \log(\mathrm{N}/\mathrm{DF}(\mathrm{t_i})) \tag{1}$$

where *TF* is the frequency of the term in document d_j and *DF* (t_i) is the number of documents that contain term t_i, after stopping word removal and word stemming, and N is the total number of documents.

5 Dataset, Experimentations, and Evaluation

Details of data set, experimental studies, different test options and various evaluation metrics are explained in the following subsections.

5.1 Dataset

Since datasets in general are not accessible for tests and text classification studies. Thus, this new dataset called KDC-4007 is created. The most important feature of this dataset is its simplicity and it is well-documented. The data set can be accessed and widely used in various studies of text classification regarding Kurdish Sorani news and articles. The documents consisted of eight categories, which are Sports, Religions, Arts, Economics, Educations, Socials, Styles, and Health, each of which is consisted of 500 text documents, where the total size of the corpus is 4,007 text files. The dataset and documents can become freely accessible to have original outcomes via future experimental assessments on KDC-4007 dataset (KDC-4007 dataset can be accessed through: https://github.com/arazom/KDC-4007-Dataset/blob/master/Kurdish-Info.txt). Table 1, gives a full detail about the KDC-4007 dataset version. In datasets, the ST-Ds is just stop words elimination is performed by using Kurdish preprocessing-step approach. In the Pre-Ds dataset, Kurdish preprocessing-step approach is used. In the Pre + TW-Ds dataset, *TF* × *IDF* term weighting on Pre-Ds dataset is performed. In the Orig-Ds datasets, no process is used which is original dataset.

Table 1. KDC-4007 Dataset Experimentation.

Dataset Name	K-Preprocessing-Step Module	Stop-word Filtering	TF-IDF Weighting	No. of DOC's	# of Features
Orig-Ds	No	No	No	4,007	24,817
ST-Ds	No	Yes	No	4,007	20,150
Pre-Ds	Yes	Yes	No	4,007	13,128
Pre + TW-Ds	Yes	Yes	Yes	4,007	13,128

5.2 Experimentations

Generally, the point of performing text classification is to classify uncategorized documents into predefined categories. However, when we look from machine learning point of view, the objective of text classification is to learn classifiers from labeled documents and satisfy categories on unlabeled documents. In literature, there is an affluent set of machine learning classifiers for text classification. The determination of the best performing classifier relies on various parameters, for example, dimensionality of the feature space, number of training examples, over-fitting, feature independence, straightforwardness and system's requirements. Taking into consideration the high dimensionality and over-fitting aspects, three well-known classifiers (C4.5, NB and SVM) are chosen among all classifiers in our experimentation.

Each classifier is tested with the 10-fold cross validation technique, which is a very common strategy for the estimate of classifier performance, the data is divided into 10 folds; nine folds of the data are used for the training, and one-fold of the data is used for testing.

5.3 Evaluation

In the field of machine learning, there are diverse evaluation criteria that can be used to appraise classifiers. In this study, the four popular evaluations; accuracy (ACC), precision, recall and F1-measure are utilized. Their mathematical equations are illustrated below:

$$\text{Accuracy} = \text{TN} + \text{TP}/\text{TP} + \text{FP} + \text{TN} + \text{FN} \tag{2}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{3}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{4}$$

$$\text{F} - \text{Measure} = 2 * (\text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \tag{5}$$

Accuracy it is the most widely used on a large scale to assess the standard of performance, which is the proportion of the total number of class files that are properly classified. In addition, the time to build the model is involved in the comparatives analysis. The classifiers compare the effectiveness of the proposed approach to measure how accurate the classification was by counting the number of correctly classified instances and the number of incorrectly classified instances. It is worth noticing that the same datasets are applied on all classifiers.

6 Results and Discussion

In this section, three different classifiers are used to study the effect of each of the preprocessing tasks. The three classifiers are used with four different representations of the same datasets. After conducting comparisons on the datasets, some insightful thoughts and conclusions can be discussed. The objective of this set of experiments was to compare the performance of the considered classifiers for each of the four different tests of the dataset.

Tables 2, 3 and 4 show the accuracy for the four different representations with three classifiers, the number of correctly classified instances (CCI), the number of incorrectly classified instances (ICI), and time spent to build mode (TB). Per the proposed technique, using normalization, stop-word removal, and Kurdish Stemming-step module produced a positive impact on classification accuracy in general. As shown in Table 2, the Kurdish preprocessing-step module provided a dominant impact and generated a significant improvement (in terms of classification accuracy) with the SVM classifier. This can be seen from the experiences of the Pre-Ds and the Pre + TW-Ds datasets respectively. On the other hand, stop word removal provided a slight improvement with the SVM classifier which can be seen from the ST-Ds dataset. However, stemming helped in gathering the words that contained similar importance, a smaller number of features with further discrimination were achieved. For any classification system, the model building time is a critical factor. As expected, the learning (model building) times for the four tests were generally low compared with the NB and DT (C4.5). Utilizing Kurdish Stemming-step module reduced the building times for the classifier compared with the Orig-Ds dataset. In addition, the average precision and recall of the eight categories for the Pre-Ds were satisfactory compared to the Orig-Ds dataset in which stemming processing was used which reduced the size of feature that effected the final performance of Kurdish text classification.

Table 2. Accuracies for the SVM Classifier on KDC-4007 Dataset

Trails	ACC%	CCI	ICI	Precision	Recall	F1-Measure	TB (Sec)
Orig-Ds	87.17	3493	514	0.87	0.87	0.87	4:27
ST-Ds	87.62	3511	496	0.88	0.87	0.87	4:20
Pre-Ds	91.44	3664	343	0.92	0.91	0.91	3:33
Pre + TW-Ds	91.48	3666	341	0.92	0.91	0.91	4:23

On the other hand, it was noticed that the precision, recall and F-measure for the ST-Ds dataset were slightly effected. The Pre + TW-Ds results using the SVM with the TF-IDF term weighting yielded better than using the DT (C4.5) and the NB with the TF-IDF term weighting.

From Table 3, on the DT classifier, it can be concluded that the Pre-Ds had the best performance in general. On the other hand, the Pre + TW-Ds included feature-weighting $TF \times IDF$ and produced accuracy that was almost the same as the Pre-Ds. The results in the Orig-Ds (the original dataset is used) were very small, whereas, the performance for same dataset and the same classifier used with Pre-Ds dataset increased significantly compared to the Orig-Ds dataset. The reason for this is that the test in the preprocessing step contained Kurdish Stemming-step module technique; whereas, the performance for the ST-Ds increased marginally which contained stop word removal in the preprocessing stage. Another measure which obtained from the experiments was the amount of time taken for building the models. As shown in the Table 3, the DT required a huge amount of time to build the needed model for four different datasets in general. While the time for building the models in tests contained the preprocessing stage decreased very significantly compared to the original dataset. The weighted averages for the precision, recall and F1- measure in the Orig-Ds dataset are very small. Though the F1- measure for the same dataset and the same classifier used in the Pre-Ds and the Pre + TW-Ds increased significantly compared with the Orig-Ds dataset. The reason for this is that the two tests in preprocessing step contained stemming, thus it can be inferred that the Kurdish Stemming-step module improved the Precision and Recall for the classifier.

As indicated by the data introduced in Table 4 for the NB classifier, the highest accuracy (86.42%) achieved when the pre-processing steps were used with the Pre-Ds dataset. After performing feature weighting $TF \times IDF$, the NB classifier obtained the worst accuracy results with the Pre + TW-Ds compared to the values obtained from the Pre-Ds dataset, where they were unexpected. As expected, the building times for classifier like the NB, required a small amount of time to complete the model compared to the DT. Consequently, it can be noticed from Table 4, that the results gave the highest average Precision, Recall and F1- measure when the pre-processing steps were used.

Table 3. Accuracies for the DT Classifier on KDC-4007 Dataset

Trails	ACC%	CCI	ICI	Precision	Recall	F1-Measure	TB (Sec)
Orig-Ds	64.88	2600	1407	0.65	0.64	0.65	231:33
ST-Ds	64.26	2575	1432	0.68	0.64	0.64	228:49
Pre-Ds	80.58	3229	778	0.81	0.80	0.80	150:29
Pre + TW-Ds	80.53	3227	780	0.81	0.80	0.80	164:29

Table 4. Accuracies for the NB Classifier on KDC-4007 Dataset

Trails	ACC%	CCI	ICI	Precision	Recall	F1-Measure	TB (Sec)
Orig-Ds	76.89	3081	926	0.77	0.76	0.77	9:36
ST-Ds	79.13	3129	881	0.78	0.78	0.78	14:5
Pre-Ds	86.42	3464	544	0.86	0.86	0.86	10:36
Pre + TW-Ds	82.48	3305	702	0.82	0.82	0.82	10:50

Also, it was noticed that when the feature-weighting *TF* × *IDF* was used, the F1-measure was decreased for the same dataset and the same classifier.

The dataset was experimented using 10-fold cross validation method. As illustrated in Fig. 2, the best result obtained was through the SVM classifier. From the experimental results, as in Fig. 2, it is obvious that the Kurdish preprocessing-step module technique significantly influenced the performance of the DT classifier on the four datasets. Thus, the range of accuracy in the DT was higher than SVM classifier.

In other words, the dimension of the dataset less influences the range of accuracy in the SVM classifier than the DT and NB classifiers. This is because it works better in a high dimensional environment.

Figure 3, shows the performance comparison of feature weighting *TF* × *IDF* methods in terms of accuracy on datasets. The accuracy performance values of the two classifiers excluding the SVM on datasets were insignificantly decreased after applying feature weighting *TF* × *IDF* method. The accuracy values in the Pre-Ds for the NB

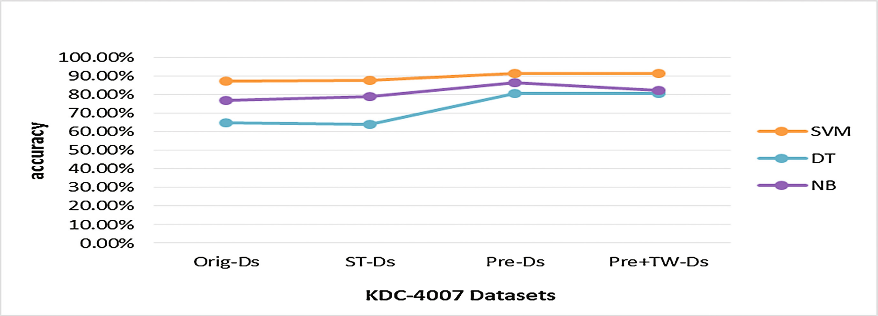


Fig. 2. SVM, NB and DT Results on the four versions of the dataset using Fold = 10.

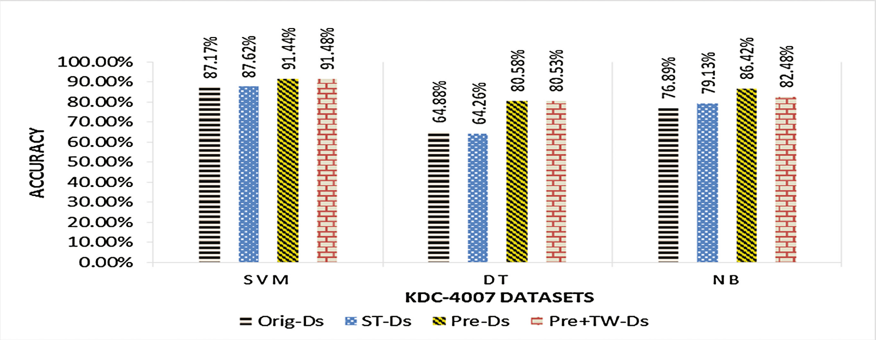


Fig. 3. Feature Weighting TF × IDF on Datasets using the SVM, NB and DT classifiers.

and DT classifiers were 86.42% and 80.58% respectively; however, they became 82.48%, and 80.53% after performing feature weighting TF × IDF method in the Pre + TW-Ds. The only classifier with insignificantly increased performance was the SVM classifier. For example, the accuracy value on the Pre-Ds datasets was increased from 91.44% to 91.48%.

7 Conclusion

In this research, the experiments indicated that the SVM outperformed both NB, and C4.5 classifiers in all tests. Applying normalization and Kurdish Stemming-step module on the original datasets was affected the performance of the three used classifiers, thus, these classifiers provided better classification accuracy compared to the original data. The performance of the classifiers SVM, NB and C4.5 was increased marginally when the stop word filtering approach was used in the preprocessing stage. Term weighting, such as *TF* × *IDF* method was performed after pre-processing steps to determine the impacts of feature weighting methods on Kurdish text classification. The experimental results indicated that; SVM increased the classification accuracy value by 0.25%, but, the classification accuracy was decreased by 5.1 using NB classifier. Besides, the classification accuracy was not affected when DT (C4.5) was used.

References

1. Hotho, A., Nurnberger, A., Paaß, G.: A brief survey of text mining. LDV Forum-GLDV J. Comput. Linguist. Lang. Technol. **20**, 19–62 (2005)
2. Tan, A.: Text mining: the state of the art and the challenges concept-based. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, pp. 65–70 (1999)

3. Chen, K.C.: Text Mining e-complaints data from e-auction store. J. Bus. Econ. Res. **7**(5), 15–24 (2009)
4. Mohammed, F.S., Zakaria, L., Omar, N., Albared, M.Y.: Automatic kurdish sorani text categorization using N-gram based model. In: 2012 International Conference on Computer & Information Science (ICCIS), 12 Jun 2012, vol. 1, pp. 392–395. IEEE (2012)
5. Wahbeh, A., Al-Kabi, M., Al-Radaideh, Q., Al-Shawakfa, E., Alsmadi, I.: The effect of stemming on arabic text classification: an empirical study. Int. J. Inf. Retrieval Res. **1**(3), 54–70 (2011)
6. Mohammad, A.H., Alwada'n, T., Al-Momani, O.: Arabic text categorization using support vector machine, Naïve Bayes and neural network. GSTF J. Comput. (JoC) **5**(1), 108–115 (2016)
7. Mohsen, A.M., Hassan, H.A., Idrees, A.M.: Documents emotions classification model based on tf-idf weighting measure. World Acad. Sci. Eng. Technol. Int. J. Comput. Electric. Automat. Control Inf. Eng. **3**(1), 1795 (2016)
8. Hmeidi, I., Al-Ayyoub, M., Abdulla, N.A., Almodawar, A.A., Abooraig, R., Mahyoub, N. A.: Automatic Arabic text categorization: a comprehensive comparative study. J. Inf. Sci. **41** (1), 114–124 (2015)
9. Rish, I.: An empirical study of the naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 4 August 2001, vol. 3, no. 22, pp. 41–46. IBM, New York (2001)
10. Sharma, R., Gulati, N.: Improving the accuracy and reducing the redundancy in data mining. Int. J. Eng. Sci., 45–75 (2016)
11. Last, M., Markov, A., Kandel, A.: Multi-lingual detection of web terrorist content. In: Chen, H. (ed.) WISI. LNCS, pp. 16–30. Springer (2006)
12. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques, vol. 31, pp. 249–268 (2007)
13. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
14. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
15. Esmaili, K.S., Eliassi, D., Salavati, S., Aliabadi, P., Mohammadi, A., Yosefi, S., Hakimi, S.: Building a test collection for Sorani Kurdish. In: Proceedings of the 10th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2013), Ifrane, Morocco, 27–30 May 2013. IEEE, New York (2013)
16. Hassani, H., Medjedovic, D.: Automatic kurdish dialects identification. Comput. Sci. Inf. Technol., 61 (2016)
17. Mustafa, A.M., Rashid, T.A.: Kurdish stemmer pre-processing steps for improving information retrieval. J. Inf. Sci., 1–14 (2017). doi: 10.1177/0165551510000000, sagepub.co.uk/journalsPermissions.nav, jis.sagepub.com
18. Szymański, J.: Comparative analysis of text representation methods using classification. Cybern. Syst. **45**(2), 180–199 (2014)
19. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
20. Patra, A., Singh, D.: A survey report on text classification with different term weighing methods and comparison between classification algorithms. Int. J. Comput. Appl. **75**(7) (2013)

Outsourcing the Decryption of Ciphertexts for Predicate Encryption via Pallier Paradigm

Bai Ping[1], Zhang Wei[2], Li Zhenlin[1], and Xu An Wang[2(✉)]

[1] Department of Electronic Technique,
Engineering University of PAP, Xi'an Shaanxi, China
bp15771937011@163.com, lizhenlin19921109@163.com
[2] Key Laboratory of Information Security,
Engineering College of PAP, Xi'an Shaanxi, China
zhaangweei@yeah.net, wangxazjd@163.com

Abstract. With the proliferation of computation and storage outsourcing, access control has become one of the vital problems in cloud computing. Supporting more complex and flexible function, predicate encryption (PE) is gaining more and more attention in access control of decryption outsourcing. Based on the somewhat homomorphic encryption Paillier scheme and combined with Green's scheme which supports outsourcing the decryption ciphertexts, we constructed a Paillier type outsourcing the decryption ciphertexts via predicate encryption. In this scheme, decryption can be outsourced to the cloud, and which can greatly reduce the storage and computation costs of user end. Moreover, the scheme supports arbitrary homomorphic addition and one homomorphic multiplication on ciphertexts. IND-AH-CPA security of the scheme is proved under subgroup decision assumption.

1 Introduction

With the proliferation of cloud computing [1], in order to reduce cost, more and more users tend to outsource the complex computing tasks to the cloud. However, security issues associated with cloud computing have become increasingly prominent. When enjoying the convenience of outsourcing, the problem of access control is becoming more and more serious. Li et al. [2], Chen et al. [3], Wei et al. [4] studies this problem from different viewpoint and use different methods. However, research works that combining the homomorphism and predicate encryption scheme are rare. There are still a lot of works to do in this area. If one user save his sensitive data to the cloud server without encryption, then the cloud may access illegally and even distort the data. In order to prevent malicious leakage and illegal access to the sensitive data, users must outsource their data in the encrypted form. And access control for the sensitive data is also an important problem. The traditional encryption and decryption model of cloud computing cannot achieve fine-grained access control. In 1984, Shamir [5] proposed Identity-Based Encryption (IBE), in which the user's public key was generated by a unique identifier that was related to his/her identity. When the user visit server, they did not need to query the user's public key certificate any more. Predicate Encryption (PE), which was proposed by Katz et al. [6], enables more sophisticated and flexible functionality.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_20

In a Predicate Encryption system, a secret key $sk_{\vec{v}}$ corresponds to a predicate $\vec{v}$ and a ciphertext is associated with an attribute vector $\vec{x}$. The predicate vector $\vec{v} \in Z_p^n \backslash \{\vec{0}\}$ and the attribute vector $\vec{x} \in Z_p^n \backslash \{\vec{0}\}$, where if $\vec{x} \cdot \vec{v} = 0 \bmod N$ then $f_{\vec{v}}(\vec{x}) = 1$, else $f_{\vec{v}}(\vec{x}) = 0$. The user will get the correct decryption if and only if his/her attribute meets the access policy. Due to the fine-grained access control on the ciphertexts, PE has gained much attention and some works have been appeared, such as Identity-Based Encryption (IBE) [5, 7], Attribute-Based Encryption (ABE) [8, 9], Hidden-Vector Encryption (HVE) [10]. Clear et al. [11] proposed a predicate encryption scheme to achieve homomorphic multiplication on ciphertexts on Z_2.

In this paper, basing on Paillier's homomorphic encryption scheme [12], and adopting the classic method [13], we build a Paillier type encryption scheme that can outsource the decryption of encrypted ciphertexts via predicate. The main advantage of our scheme is the fine-grained access control property due to the idea of predicate encryption. In our scheme, partial decryption of ciphertexts is outsourced to the cloud, which greatly reduce the computing burden of users. The access control policy is embedded into the ciphertexts, and only the users whose attributes satisfy the access policy can decrypt the ciphertexts. Meanwhile, our scheme can operate on ciphertexts for arbitrary additions and one multiplication. Homomorphic encryption allows the server to operate in case where it do not know the original plaintexts. It is more convenient to operate the data for users. Homomorphic encryption has gradually become a research hotpot and many homomorphic encryption scheme have emerged [11, 14, 15].

In Sect. 2, we give the preliminary knowledge of this paper. We present our construction of outsourcing and analyze the homomorphic properties of the scheme in Sect. 3. In Sects. 4 and 5, its security and performance analysis is described respectively. In the next chapter, we make a conclusion.

2 Preliminares

2.1 Bilinear Map

Let G and G_T be two multiplicative cyclic groups of prime order p. Let g be a generator of G. Define: $e : G \times G \to G_T$ be a bilinear map with the following properties:

1. Bilinearity: for all $\mathrm{r}, \mathrm{s} \in G$ and $a, b \in Z_p$, then $e\left(r^a, s^b\right) = e(r, s)^{ab}$.
2. Non-degeneracy: for arbitrary $\mathrm{r}, \mathrm{s} \in G$, $e(\mathrm{r}, \mathrm{s}) \neq 1$.
3. Computable: for all $\mathrm{r}, \mathrm{s} \in G_T$, $e(\mathrm{r}, \mathrm{s})$ can be computed in polynomial time.

2.2 Access Structures

Definition 1 (Access Structure [16]). Let $P = \{P_1, P_2, \cdots, P_n\}$ be a secret-sharing set of participants. The access structure AS is a non-empty subset of $P = \{P_1, P_2, \cdots, P_n\}$, namely, $AS \subseteq 2^P \backslash \{\varnothing\}$. The monotone property of AS is defined as: If $A \in AS$ and $A \in B$, then $B \in AS$. At the same time, sets in AS is called the authorized sets, otherwise the unauthorized sets.

2.3 Linear Secret Sharing Schemes

Definition 2 (Linear Secret-Sharing Schemes (LSSS) [16]). A secret-sharing scheme Π over a set of participants P is called linear (over Z_p) if the following holds.

1. The shares of the participants form a vector over Z_p.
2. There exists a $l \times n$ matrix $\boldsymbol{M}$ that is called the share-generating matrix for Π. We define a function ρ that maps every row of the share-generating matrix to a related participant, i.e., for $i = 1, 2, \cdots, l$, the value $\rho(i)$ is the participant which is associated with the r_{th} row. And we build a column vector $\boldsymbol{v} = (s, y_2, \cdots y_n)$, in which $(y_2, \cdots y_n) \in Z_p^n$ are chosen randomly, and $s \in Z_p$ is just the secret to be shared, then $\boldsymbol{M} \cdot \boldsymbol{v}$ is the vector of l shares of the secret s according to Π. The share $(\boldsymbol{M} \cdot \boldsymbol{v})_i$ belongs to participant $\rho(i)$.

Definition 3 (Linear Reconstruction [16]). Each linear secret sharing-scheme has the linear reconstruction property: Let Π being an LSSS for the access structure AS. Let $S \in AS$ be an authorized set, and $I = \{i : \rho(i) \in S\}, I \subseteq \{1, 2, \cdots, l\}$. Then, if $\lambda_i = \boldsymbol{M}_i \cdot \boldsymbol{v}$ are valid shares of any secret s according to Π, there must exist constants $\{w_i \in Z_p\}_{i \in I}$ such that $\sum_{i \in I} w_i \boldsymbol{M}_i \boldsymbol{v} = s$, otherwise $\{w_i \notin Z_p\}_{i \in I}$.

2.4 Paillier Scheme

The Paillier encryption scheme [12], named after and constructed by Pascal Paillier in 1999, is a probabilistic public-key algorithm. A notable feature of the Paillier scheme is its homomorphic property. Paillier scheme supports arbitrary homomorphic additions and one homomorphic multiplication The scheme is described as follows:

Key Generation: Given a security parameter γ, Choose two large prime number p and q randomly and independently of each other. If both primes are of equal length then $\gcd(pq, (p-1)(q-1)) = 1$. Let $n = pq, g = n + 1, \eta = \varphi(n), \phi = \varphi(n)^{-1}$ $(\mathrm{mod} n)$. Select a random integer g where $g \in Z_{n^2}^*$. To ensure n dividing the order of g, we combine binomial theorem $(1+n)^x = 1 + nx(\mathrm{mod} n^2)$ to check the existence of the following modular multiplicative inverse: $\phi = (L(g^\eta(\mathrm{mod} n^2)))^{-1} \mathrm{mod} n$, where function L is defined as $L(u) = \frac{u-1}{n}$. Finally, the public (encryption) key is $pk = (n, g)$ and the private (decryption) key is $sk = (\eta, \phi)$.

Encryption: Let m be a message to be encrypted, where $m \in Z_n$. Select random R where $R \in Z_n^*$. Compute ciphertext as $c = g^m \cdot R^n(\mathrm{mod} n^2)$.

Decryption: Let c be the ciphertext to be decrypted. we use private key to compute the plaintext message m as

$$\begin{aligned} L\left(c^{\eta}\left(\bmod n^{2}\right)\right)\cdot\phi &= L\left(\left(g^{m}R^{n}\right)^{\eta}\left(\bmod n^{2}\right)\right)\cdot\eta^{-1} \\ &= L\left((1+n)^{m\eta}\cdot R^{n\eta}\left(\bmod n^{2}\right)\right)\cdot\eta^{-1} \\ &= m\cdot\eta\cdot\eta^{-1} = m(\bmod n) \end{aligned}$$

Homomorphic Properties: A notable feature of the Paillier scheme is its homomorphic properties. Given two ciphertexts $E(m_1, pk) = g^{m_1}R_1^n(\bmod n^2)$ and $E(m_2, pk) = g^{m_2}R_2^n(\bmod n^2)$, where R_1 and R_2 are randomly chosen from Z_n^*.

Homomorphic Addition of Plaintext:

$$\begin{aligned} D\left(E(m_1,pk)\cdot E(m_2,pk)\left(\bmod n^2\right)\right) &= D\left(g^{m_1}R_1^n\right)\left(g^{m_2}R_2^n\right)\left(\bmod n^2\right) \\ &= D(E(m_1+m_2,pk)) = m_1+m_2 \end{aligned}$$

Homomorphic Multiplication of Plaintexts:

$$\begin{aligned} D\left(E(m_1,pk)^{m_2}\left(\bmod n^2\right)\right) &= D\left(\left(g^{m_1}R_1^n\right)^{m_2}\left(\bmod n^2\right)\right) \\ &= D(E(m_1m_2,pk)) = m_1m_2 \end{aligned}$$

2.5 Security Model for PE

We adopt the security model of [6]. Which is described as the following:

Setup: The challenger C runs the Setup algorithm Γ and gives the public parameters to the adversary B;

Phase 1: The adversary B is allowed to adaptively issue queries for private keys $sk_{\vec{v}}$ for many predicates vector $\vec{v}$;

Challenge: The adversary B submits two messages m_0 and m_1 with equal length and two attribute vectors $\vec{x}_0, \vec{x}_1$, where $f_{\vec{v}}(\vec{x}_0) \neq 1$ and $f_{\vec{v}}(\vec{x}_1) \neq 1$ for all the key queried in **Phase1:** The challenger C flips a random coin $b \in \{0, 1\}$ and encrypts m_b with $\vec{x}_b$. The challenge ciphertext c^* is passed to the adversary B;

Phase 2: The adversary B may continue to issue adaptively queries like Phase1, except the key query for predicate $f_{\vec{v}}(\vec{x}_0) = 1$ and $f_{\vec{v}}(\vec{x}_1) = 1$;

Guess: The adversary B outputs a guess $b^{'}$. If $b' = b$, then the adversary B is successful.

The advantage B of an IND-AH-CPA adversary in this game is defined as $Adv_{\Gamma,\mathrm{B}}(\lambda) = \left|\Pr[b' = b] - \frac{1}{2}\right|$, where λ is security parameter.

Definition 4: A Predicate Encryption scheme is IND-AH-CPA secure if all polynomial time adversaries B have at most a negligible advantage in the above security game.

Attribute-hiding requires that a ciphertext hide the associated attribute as well as the plaintext, while payload-hiding only requires that a ciphertext hide the plaintext.

In some applications which require the protection of attribute, payload-hiding is unacceptable, and attribute-hiding meets the requirement.

2.6 Assumption

These assumptions are an extension of [17].

We define a group generator $\mathcal{G}$. An algorithm which takes in a security parameter 1^λ and output $(N = p_1p_2p_3, G, G_T, e)$, where p_1, p_2, p_3 are distinct primes, G and G_T are cyclic groups of order N. Let $e : \mathrm{G} \times \mathrm{G} \rightarrow \mathrm{G}_T$ being a map. We say that G is a bilinear group if the group operation in G and the bilinear map $e : \mathrm{G} \times \mathrm{G} \rightarrow \mathrm{G}_T$ are both efficiently computable. Let $G_{p_1}, G_{p_2}, G_{\mathrm{p}_3}$ denote the subgroups of G. Select random $q \in G_{p_1}$. Introducing the additional term $h \in G_{p_1}$ still does not help to add advantage to Λ, since h is independent of the challenge.

We now state the complexity assumptions that we will use to prove security of our systems.

Assumption 1: Given a group generator $\mathcal{G}$, we define the following distribution:

$$\begin{gathered}(N = p_1p_2p_3, G, G_T, e) \leftarrow \mathcal{G}\\ g, h \leftarrow G_{p_1}, X_3 \leftarrow G_{p_3},\\ D = (G, q, h, X_3),\\ T_0 \leftarrow G_{p_1}, T_1 \leftarrow G_{p_1p_2}.\end{gathered}$$

We define the advantage of an algorithm Λ in breaking Assumption 1 to be:

$$Adv1_{\mathcal{G},\Lambda}(\lambda) = |\Pr[\Lambda(D, T_0) = 0] - \Pr[\Lambda(D, T_1) = 0]|$$

Definition 5: We say that $\mathcal{G}$ satisfies Assumption 1 if $Adv1_{\mathcal{G},\Lambda}(\lambda)$ is a negligible function of 1^λ for any polynomial time algorithm Λ.

Assumption 2: Given a group generator $\mathcal{G}$, we define the following distribution:

$$\begin{gathered}(N = p_1p_2p_3, G, G_T, e) \leftarrow \mathcal{G}\\ g, h, X_1 \leftarrow G_{p_1}, X_2, Y_2 \leftarrow G_{p_2}, X_3, Y_3 \leftarrow G_{p_3},\\ D = (G, q, h, X_3, X_1X_2, Y_2Y_3),\\ T_0 \leftarrow G_{p_1p_3}, T_1 \leftarrow G_{p_1p_2p_3}.\end{gathered}$$

The advantage of an algorithm Λ in breaking Assumption 2 is defined as:

$$Adv2_{\mathcal{G},\Lambda}(\lambda) = |\Pr[\Lambda(D, T_0) = 0] - \Pr[\Lambda(D, T_1) = 0]|.$$

Definition 6: We say that $\mathcal{G}$ satisfies Assumption 2 if $Adv2_{\mathcal{G},\Lambda}(\lambda)$ is a negligible function of 1^λ for any polynomial time algorithm Λ.

Assumption 3: Given a group generator $\mathcal{G}$, we define the following distribution:

$$
\begin{gathered}
(N = p_1p_2p_3, G, G_T, e) \leftarrow \mathcal{G}, k \in Z_N, \\
q, h, \leftarrow G_{p_1}, X_2, Y_2, Z_2 \leftarrow G_{p_2}, X_3 \leftarrow G_{p_3}, \\
D = (G, q, X_3, Z_2, g^k X_2, hY_2), \\
T_0 = e(g, h)^k, T_1 \leftarrow G_T.
\end{gathered}
$$

The advantage of an algorithm Λ in breaking assumption 3 is defined as:

$$Adv3_{\mathcal{G},\Lambda}(\lambda) = |\Pr[\Lambda(D, T_0) = 0] - \Pr[\Lambda(D, T_1) = 0]|.$$

Definition 7: We say that $\mathcal{G}$ satisfies Assumption 3 if $Adv3_{\mathcal{G},\Lambda}(\lambda)$ is a negligible function of 1^λ for any polynomial time algorithm Λ.

2.7 Paillier Type Outsourcing the Decryption of PE Ciphertexts Model

Before structuring the scheme, we briefly introduce the decryption of predicate encryption ciphertexts model:

(1) Cloud: the cloud can provide convenience and shortcut for user because it has powerful computing and storage capacity. However, the cloud server cannot be trusted completely. We must not only encrypted the sensitive data, but also set the necessary access control.

(2) User: the user usually tend to outsource complex data resources to the cloud server, on the other hand, they do not hope these data being stolen and modified by the cloud server (Fig. 1).

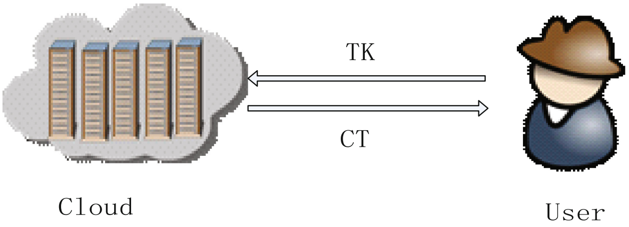


Fig. 1. Diagram of outsourcing scheme

3 Our Construction

Combining Paillier scheme with the idea of outsourcing decryption of predicate-based ciphertexts, we present our construction that can realize access control on the results of cloud outsourcing. Our scheme consists of the following five algorithms:

Setup (π, U): The setup algorithm takes as input a security parameter π and a universe description $U = \{0, 1\}^*$. Then β is a random generator of the subgroup of G of order N. F is a hash function that maps $\{0, 1\}^*$ to G. What's more, it randomly chooses exponents $\alpha, a \in Z_N$ and $h \in G_{p_1}$ Then a, $\{t_i\}_{i=1,\ldots,n} \overset{r}{\leftarrow} Z_N$ and $a \neq t_i$. The algorithm sets $\mathrm{MK} = (g^\alpha, \mathrm{PK})$ as the master secret key. The public parameters of this system is

$$PK = \{g, g_1 = g^a, e(g,g)^\alpha, \{T_i = g^{t_i}\}_{i=1,\ldots,n}, F\}$$

Encrypt $(PK, m, (M, \rho))$: The encryption algorithm takes as input the public parameters PK and a message m. In addition, it takes as input an LSSS access structure (M, ρ). The function ρ associates rows of M to attributes. Let M be an $l \times n$ matrix. The algorithm first chooses a random vector $\boldsymbol{v} = (s, y_2, \cdots, y_n) \in Z_p^n$, and s is the secret to be shared. For $i = 1, 2, \cdots, n$, it computes $\lambda_i = \boldsymbol{M}_i \cdot \boldsymbol{v}$, in which $\boldsymbol{M}_i$ is the vector corresponding to the i_{th} row of $\boldsymbol{M}$. In addition, the algorithm chooses random $\rho, R \in Z_n^*, r_1, \cdots, r_l \in Z_p$. H is a hash function that maps G to $(0,1)$ and $H(\rho) \neq 0$. Finally, it select $\vec{v}_i = \{v_1, \ldots, v_n\}$ to output the ciphertext CT:

$$\begin{aligned}
&c = g^{m/H\left(e(g,g)^{-rx_i}\right)} \cdot R^n(\mathrm{mod} n^2), C' = g^{rs} \\
&\left(C_1 = g^{a\lambda_1} \cdot F(\rho(1))^{-r_1}, D_1 = g^{r_1}\right), \cdots, \left(C_l = g^{a\lambda_l} \cdot F(\rho(l))^{-r_l}, D_l = g^{r_l}\right) \\
&c_1 = \left(g_1^{-1} T_i\right)^{rx_i}
\end{aligned}$$

The final form of ciphertext is $C = (c, c_1)$.

KeyGen (MK, S): This algorithm takes as input MK and an attribute set S to obtain $\mathrm{SK}' = \left(\mathrm{PK}, K' = g^{-x_i/s} g^{\frac{at'}{r}}, L' = g^{t'}, \left\{K'_x = F(x)^{t'}\right\}_{x \in S}\right)$. The predicate vector is $\vec{v}_i = \{v_1, \ldots, v_n\}$. The algorithm chooses a random value $\varepsilon, t \in Z_N^*$. Let $t = t'/\varepsilon$, then publish the transformation key TK as:

$$\mathrm{PK}, K = {K'}^{1/\varepsilon} = g^{-x_i/s\varepsilon} g^{at/r}, L = {L'}^{1/\varepsilon} = g^{(t'/\varepsilon)} = g^t, \{K_x\}_{x \in S} = \left\{{K'_x}^{1/\varepsilon}\right\}_{x \in S}$$

And compute part of the private key as:

$$sk_{\vec{v}} = \{\{d_i = (h^{sv_i} g)^{\frac{1}{a - t_i}}\}_{i=1,\ldots,n}\}$$

Finally the private key is $\mathrm{SK} = \left(\varepsilon, \mathrm{TK}, \mathrm{sk}_{\vec{v}}\right)$.

Transform (TK, CT): The transformation algorithm takes as input a transformation key $\mathrm{TK} = \left(\mathrm{PK}, K, L, \{K_x\}_{x \in S}\right)$ for a set S and a ciphertext $\mathrm{CT} = (C, C', C_1, \cdots, C_l)$ for access structure $(\boldsymbol{M}, \rho)$. If S does not satisfy the access structure, it outputs $\perp$. Suppose that S satisfies the access structure and let $I \subseteq \{1, 2, \cdots, l\}$ be defined as $I = \{i : \rho(i) \in S\}$. Then, let $\left\{\omega_i \in Z_p\right\}_{i \in I}$ be a set of constants such that if $\{\lambda_i\}$ are valid shares of any secret S according to M, then $\sum_{i \in I} \omega_i \lambda_i = s$. The transformation algorithm calculates:

$$
\begin{aligned}
Q &= e(C',K)\Big/\left(e\left(\prod_{i\in I} C_i^{w_i}, L\right)\cdot\prod_{i\in I} e\left(D_i^{w_i}, K_{\rho(i)}\right)\right)\\
&= e(g,g)^{-rx_i/\varepsilon} e(g,g)^{sat}\Big/\left(\prod_{i\in I} e(g,g)^{ta\lambda_i w_i}\right)\\
&= e(g,g)^{-rx_i/\varepsilon}
\end{aligned}
$$

It outputs the partially decrypted ciphertext $\mathrm{CT}' = (C,Q)$.

Decrypt (SK, CT): The decryption algorithm takes as input a private key $\mathrm{SK} = \left(\varepsilon, \mathrm{TK}, \mathrm{sk}_{\vec{v}}\right)$ and a ciphertext CT. If the ciphertext is not partially decrypted, then the algorithm first executes transformation algorithm. If the output is $\perp$, then this algorithm outputs $\perp$ as well. Otherwise, it uses (ε, Q) to obtain $Q^{\varepsilon} = e(g,g)^{-rx_i}$, then uses the partial private key $sk_{\vec{v}}$ and combines Euler's theorem, that is:

$$
\begin{aligned}
&\mathrm{L}\left(c^{\eta}\left(\mathrm{mod} n^2\right)\right)\cdot\phi\ \cdot H(e(c_1,d_i))\\
&\quad = L\left(\left(1+n\right)^{m/H\left(e(g,g)^{-rx_i}\right)}\cdot R^n\right)^{\eta}\left(\mathrm{mod} n^2\right)\cdot\eta^{-1}\cdot H\left(e\left(\left(g_1^{-r}T_i^r\right)^{x_i}, \left(h^{sv_i} g\right)^{\frac{1}{a-t_i}}\right)\right)\\
&\quad = \left(m\eta/H\left(e(g,g)^{-rx_i}\right)\right)\cdot\eta^{-1}\cdot H\left(e(g,g)^{-rx_i}\cdot e(g,h)^{sr\sum x_i v_i}\right)\\
&\quad = \left(m/H\left(e(g,g)^{-rx_i}\right)\right)\cdot H\left(e(g,g)^{-rx_i}\cdot e(g,h)^{sr\sum x_i v_i}\right)
\end{aligned}
$$

We can attain the plaintext m if and only if $\vec{x}\cdot\vec{v} = 0 \bmod N$. The illegal decryptor can not attain the plaintext m if $\vec{x}\cdot\vec{v} \neq 0 \bmod N$, according to the non collision of hash function.

Our outsouring construction is based on the Paillier scheme, so it satisfies the properties of arbitrary additions and one multiplication. Let $E(m_1, PK) = g^{m_1/H\left(e(g,g)^{-rx_i}\right)}\cdot R_1^n(\mathrm{mod} n^2)$ and $E(m_1, PK) = g^{m_2/H\left(e(g,g)^{-rx_i}\right)}\cdot R_1^n(\mathrm{mod} n^2)$.

(1) Additively Homomorphic:

$$
\begin{aligned}
E(m_1, PK)\cdot E(m_2, PK) &= g^{m_1/H\left(e(g,g)^{-rx_i}\right)}\cdot R_1^n\cdot g^{m_2/H\left(e(g,g)^{-rx_i}\right)}\cdot R_2^n\\
&= g^{(m_1+m_2)/H\left(e(g,g)^{-rx_i}\right)}\left(R_1R_2\right)^n\\
&= E(m_1+m_2, PK)
\end{aligned}
$$

The decryptor can obtain the value of $e(g,g)^{-rx_i}$, if and only if the attribute satisfy the ciphertext strategy, then the decryptor can use decryption algorithm to obtain m_1+m_2.

(2) Multiplicatively Homomorphic:

$$\begin{aligned} E(m_1, PK)^{m_2} &= \left(g^{m_1/H(e(g,g)^{-rx_i})} \cdot R_1^n\right)^{m_2} \left(\bmod n^2\right) \\ &= E(m_1 m_2, pk) \end{aligned}$$

In the similar way, the legal users can work out $m_1 m_2$. Since there is no efficient algorithm to make $e : G \times G \to G_T$, so the scheme can operate on ciphertexts for only one multiplication.

4 Security

To prove the security, we will adopt the dual system encryption methodology which was used in [17]. We define two additional structures: semi-functional ciphertexts and semi-functional keys.

Semi-functional Ciphertext: Let g_2 denote a generator of G_{p_2} ,select randomly $c \in Z_N$ and $\{z_i \in Z_N\}_{i=1,\ldots,n}$. A semi-functional ciphertext is formed as follows:

$$\{c_i = (g^{r_i x_i} g_2^{cz_i})^{a-t_i}\}_{i=1,\ldots,n}$$

Semi-functional Key: $d \in Z_N$ and $\{y_i \in Z_N\}_{i=1,\ldots,n}$ are random values. A semi-functional key is formed as follows:

$$\{d_i = (hg^{sv_i} g_2^{dy_i})^{\frac{1}{a-t_i}}\}_{i=1,\ldots,n}$$

A normal key can decrypt both normal and semi-functional ciphertexts, while a normal ciphertexts can be decrypted by both normal and semi-functional keys. When we use a semi-functional key to decrypt a semi-functional ciphertext, there is left with additional term $e(g_2, g_2)^{cd \sum y_i \cdot z_i}$. Notice that if and only if a semi-functional key which is satisfy that $\sum y_i \cdot z_i = 0$ is used to decrypt a semi-functional ciphertext.

Based on the assumptions, we will prove the security of our system using a sequence of games.$Game_{real}$ is the real security game in which both keys and ciphertext are normal. In the second game,$Game_0$, the ciphertext is semi-functional and all keys are normal. In third game,$Game_k$, the first k key queries are semi-functional and the rest are normal. By the final game,$Game_{final}$, all of the key queries are semi-functional key and the ciphertext is a semi-functional encryption message. We will prove these games are indistinguishable in the following lemmas.

Lemma 1: Assume there is a polynomial time adversary A such that $Adv_A^{Game_{real}} - Adv_A^{Game_0} = \varepsilon$. Then we can construct a polynomial time simulator B with advantage ε in breaking Assumption 1.

Proof: B is given a challenge sample of Assumption 1 (G, q, h, X_3, T), which is used as an input of the Setup algorithm.B chooses random value $a \in Z_N, t_i \in Z_N$, $i = 1, \ldots, n$. The public parameters are the same as Setup algorithm. B will simulate $Game_{real}$ and $Game_0$ with A.

As to the key queries $\vec{v}_i$, B can generate normal key by using the KeyGen algorithm, because it knows the MK.

As to the challenge A randomly choose plaintext (m_0, m_1) and attribute $(\vec{x}_0, \vec{x}_1)$, then B will imbeds the Assumption 1 into the challenge ciphertext. In the first place, it flips a random coin b, and then attain the sets of ciphertext:

$$c^* = \{c = g^{m/H\left(e(T,g)^{-\sum x_i^b}\right)} \cdot R^n \left(\text{mod} n^2\right), \{c_i = T^{x_i^b(a-t_i)}\}_{i=1,\ldots,n}\}$$

If $T \in G_{p_1}$, namely $T = g^r$, it is clearly that this is a properly distributed normal ciphertext c^*,else if $T \in G_{p_1p_2}$, namely $T = g^r g_2^c$. This is a properly distributed semi-functional ciphertext. B can use the output of A to gain advantage ε in breaking Assumption 1 after all.

Lemma 2: Assume there is a polynomial time adversary A such that $Adv_A^{Game_{k-1}} - Adv_A^{Game_k} = \varepsilon$. Then we can construct a polynomial time simulator B with advantage ε in breaking Assumption 2.

Proof: B is given a challenge sample of Assumption 2,(q, $h, X_3, T, Y_2Y_3, X_1X_2$), as an input of the algorithm. The public parameters are generated which are just the same in the proof of Lemma 1. B will simulate $Game_{k-1}$ and $Game_k$ with A.

As to the key queries $\vec{v}_i$, for the queries $> k$,B can generate normal key by using the KeyGen algorithm by using its knowledge of MK. For the queries $<k$, B can generate semi-functional key which then be defined as:

$$\{d_i = (h^{sv_i} g(Y_2Y_3)^{dy_i})^{\frac{1}{a-t_i}}\}_{i=1,\ldots,n}$$

where chooses random value $s, d \in Z_N$.

To the k_{th} key, B uses the value of T in the challenge, and choose random value. The key will be set as:

$$\{d_i = (h^{sv_i} T^{v_i})^{\frac{1}{a-t_i}}\}_{i=1,\ldots,n}$$

If $T \in G_{p_1p_3}$, it is clearly that this is a properly distributed normal key, else if $T \in G_{p_1p_2p_3}$, namely, $T = g^s g_2^d g_3^f$. According to the Chinese Remainder Theorem, this is a properly distributed semi-functional key. Now A choose the challenge (m_0, m_1) and attribute $(\vec{x}_0, \vec{x}_1)$, B flips a random coin b, and the sets of ciphertext:

$$c^* = \{c = g^{m/H\left(e(X_1X_2,g)^{-\sum x_i^b}\right)} \cdot R^n \left(\text{mod} n^2\right), \{c_i = (X_1X_2)^{x_i^b(a-t_i)}\}_{i=1,\ldots,n}\}$$

Let $X_1X_2 = g^r g_2^c$, we can know that these values are also actually uncorrelated in the subgroups G_{p_1}, G_{p_2} according to the Chinese Remainder Theorem. This is a properly distributed semi-functional ciphertext. B can use the output of A to gain advantage ε in breaking Assumption 2.

Lemma 3: Assume there is a polynomial time adversary A such that $Adv_A^{Game_q} - Adv_A^{Game_{final}} = \varepsilon$. Then we can constrct a polynomial time simulator with advantage ε in breaking Assumption 3.

Proof: B is given a challenge sample of Assumption 3, $(\mathrm{q}, X_3, Z_2, g^r X_2, hY_2, T)$ as an input of the algorithm. B chooses random $a \in Z_N, t_i \in Z_N, i = 1, \ldots, n$. The public parameters are set as: $PK = \{g, hY_2, g_1 = g^a, \{T_i = g^{t_i}\}_{i=1,\ldots,n}\}$. A will not distinguish h from hY_2, because it is hard to find a valid factor of $N.B$ will simulate $Game_q$ and $Game_{final}$ with A.

As to the key queries $\vec{v}_i$,B choose random $s, y_i^{'} \in Z_N$, and sets the semi-functional key as:

$$\{d_i = (h^{sv_i} Y_2 g Z_2^{y_i'})^{\frac{1}{a-t_i}}\}_{i=1,\ldots,n}\}$$

Let $Y_2 = g_2^f, Z_2 = g_2^d$, and then set $y_i = y_i' + f/d$. Thus, this is a properly distributed semi-functional ciphertext. A randomly choose the challenge (m_0, m_1) and attribute $(\vec{x}_0, \vec{x}_1)$, B will imbeds the Assumption 3 into the challenge ciphertext. B flips a random coin b, and the sets of ciphertext:

$$c^* = \{c = g^m \cdot R^n \cdot T^{-\sum x_i^b}, \{c_i = (g^r X_2)^{x_i^b (a-t_i)}\}_{i=1,\ldots,n}\}$$

if $T = g^{m/H(e(g,h)^r)-m}$, the ciphertext c^* is valid semi-functional, else if $T \in G_T$, the ciphertext c^* will be a semi-functional encryption of a random message and it is a perfect simulation of $Game_{final}$.

Theorem 1: If assumptions 1, 2, and 3 hold, then our system is IND-AH-CPA secure.

Proof: If assumptions 1, 2, and 3 hold, the real security game is indistinguishable from $Game_{final}$, according to the previous lemmas. In $Game_{final}$, the challenge ciphertext will give no information about b. Therefore, A only can attain negligible advantage in breaking our construction. This is clear that our system is IND-AH-CPA secure.

5 Summary

In this article, we bring the thought of outsourcing the decryption of ABE ciphertexts into Paillier scheme, and propose our Paillier type outsourcing the decryption of predicate encryption ciphertexts scheme, which is suitable for the cloud environment. By using the method of attribute-based encryption, we can solve the problem of access control on cloud computing results, and the users' computation overhead in decryption reduces remarkably, because the process of outsourcing improves users' decrypting efficiency.

Further work is to explore the combination of outsourcing the decryption of PE ciphertexts with the full homomorphic encryption, and to construct a more efficient and practical outsourcing scheme for the full homomorphic encryption based on the cloud.

References

1. Mezghani, K., Ayadi, F.: Factors explaining IS managers attitudes toward cloud computing adoption. Int. J. Technol. Hum. Interact. (IJTHI) **12**(1), 1–20 (2016)
2. Li, X., He, Y., Niu, B., Yang, K., Li, H.: An exact and efficient privacy-preserving spatiotemporal matching in mobile social networks. Int. J. Technol. Hum. Interact. (IJTHI) **12**(2), 36–47 (2016)
3. Chen, Q., Wu, L., Li, L., Ma, X., Wang, X.A.: Method for improving data security in register files based on multiple pipeline restart. Int. J. Inf. Technol. Web Eng. (IJITWE) **10** (3), 17–32 (2015)
4. Wei, Z.: A pairing-based homomorphic encryption scheme for multi-user settings. Int. J. Technol. Hum. Interact. (IJTHI) **12**(2), 72–82 (2016)
5. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakley, G.R., Chaum, D. (eds.) CRYPTO 1984. LNCS, vol. 196, pp. 47–53. Springer, Heidelberg (1985). doi:10.1007/3-540-39568-7_5
6. Katz, J., Sahai, A., Waters, B.: Predicate encryption supporting disjunctions, polynomial equations, and inner products. In: Smart, N. (ed.) EUROCRYPT 2008. LNCS, vol. 4965, pp. 146–162. Springer, Heidelberg (2008). doi:10.1007/978-3-540-78967-3_9
7. Waters, B.: Efficient identity-based encryption without random oracles. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 114–127. Springer, Heidelberg (2005). doi:10.1007/11426639_7
8. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 457–473. Springer, Heidelberg (2005). doi:10.1007/11426639_27
9. Goyal, V., Pandey, O., Sahai, A., Waters, B.: Attribute-based encryption for fine-grained access control of encrypted data. In: ACM Conference on Computer and Communications Security, pp. 89–98 (2006)
10. Boneh, D., Waters, B.: Conjunctive, subset, and range queries on encrypted data. In: Vadhan, Salil P. (ed.) TCC 2007. LNCS, vol. 4392, pp. 535–554. Springer, Heidelberg (2007). doi:10.1007/978-3-540-70936-7_29
11. Clear, M., Hughes, A., Tewari, H.: Homomorphic encryption with access policies: characterization and new constructions. In: Youssef, A., Nitaj, A., Hassanien, A.E. (eds.) AFRICACRYPT 2013. LNCS, vol. 7918, pp. 61–87. Springer, Heidelberg (2013). doi:10.1007/978-3-642-38553-7_4
12. Paillier, P.: Public-key cryptosystems based on composite degree residuosity classes. In: Stern, J. (ed.) EUROCRYPT 1999. LNCS, vol. 1592, pp. 223–238. Springer, Heidelberg (1999). doi:10.1007/3-540-48910-X_16
13. Green, M., Hohenberger, S., Waters, B.: Outsourcing the decryption of ABE ciphertexts. In: USENIX Security Symposium, vol. 2011, no. 3 (2011)
14. Gentry, C., Sahai, A., Waters, B.: Homomorphic encryption from learning with errors: conceptually-simpler, asymptotically-faster, attribute-based. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013. LNCS, vol. 8042, pp. 75–92. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40041-4_5

15. Gentry, C., Halevi, S., Smart, N.P.: Fully homomorphic encryption with polylog overhead. In: Pointcheval, D., Johansson, T. (eds.) EUROCRYPT 2012. LNCS, vol. 7237, pp. 465–482. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29011-4_28
16. Beimel, A.: Secure schemes for secret sharing and key distribution. Int. J. Pure Appl. Math. (1996)
17. Lewko, A., Waters, B.: New techniques for dual system encryption and fully secure HIBE with short ciphertexts. In: Micciancio, D. (ed.) TCC 2010. LNCS, vol. 5978, pp. 455–479. Springer, Heidelberg (2010). doi:10.1007/978-3-642-11799-2_27
18. Waters, B.: Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization. In: Catalano, D., Fazio, N., Gennaro, R., Nicolosi, A. (eds.) PKC 2011. LNCS, vol. 6571, pp. 53–70. Springer, Heidelberg (2011). doi:10.1007/978-3-642-19379-8_4

A New Middleware Architecture for RFID Data Management

Weiwei Shen[1,2], Han Wu[3], He Xu[1,2](✉), and Peng Li[1,2]

[1] School of Computer, Nanjing University of Posts and Telecommunications,
Nanjing, China
{1214042932,xuhe,lipeng}@njupt.edu.cn
[2] Jiangsu High Technology Research Key Laboratory
for Wireless Sensor Networks, Nanjing, China
[3] Overseas Education College,
Nanjing University of Posts and Telecommunications, Nanjing, China
1308790676@qq.com

Abstract. With the developments of RFID Technology, the unreliability of RFID devices result in a large number of tag data's redundancy data, and the reliability of RFID source data is gradually harder to get the safeguard. The reliability means the correctness of the RFID source data such as multi and leakage of reading data. This paper designs a new RFID middleware architecture, which introduces the system structure of middleware and shows the details of each significant part of system. The data processing module and data transmission module are the two main parts designated to get reliable RFID data. The data processing module uses L-NCD algorithm to process data redundancy. The data transmission module takes the responsibility of transferring data between RFID middleware and the sever through JMS message transferring mechanism.

1 Introduction

With the developments of the RFID technology, the size of the RFID systems became larger and more complex [1]. Because of the loss of united standard, the communication interfaces between RFID reader and system are mostly decided by the reader manufactures. This situation results in the problem that the RFID system is difficult to efficiently manage the readers [2] who have different protocols when they are large-scaled deployed. During the early RFID applications, readers were directly connected to the applications. The RFID data is dealt by applications to logical data, where this method satisfies the need of simple RFID systems, however, this system's design is complex, the efficiency is low and the reusability is hard [3]. To get rid of this kind of complexity, there exists the RFID middleware to guarantee the fast developments of RFID applications. The RFID middleware is a software used to process and deal with the data from the reader. It is also the hub used to connect the low-level reader and the high-level enterprise applications. Middleware system which works at the independent level to take the processes to deal with the data shields the reader hardware and the high-level applications. Thus, the expansibility and applicability of the system have been improved greatly. During the early developments of RFID

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_21

applications, the RFID data were transported to the applications directly and the applications would deal with these original data. The RFID middleware is an application used to process and deal with all the information and flow of events. It will help to filtrate and format data the reader collected, then transport the processed data to the background applications [4]. The middleware shields the reader hardware and the high-level systems, shares the processes to deal with the data and takes the preprocessing to the original data. It decreases the difficulty to develop the applications, and the programs being developed do not need to face the low-level structure directly but to call the middleware directly [5]. It provides a team of universal applications interfaces (API), which can be connected to the RFID readers and read the RFID tag data. It is able to filtrate, divide and count the tag data, decreases the flow data sent to the background applications, reduces the effect from misreading, multi-reading and redundancy in RFID systems [6].

The impact of middleware in the RFID structure is shown as the Fig. 1.

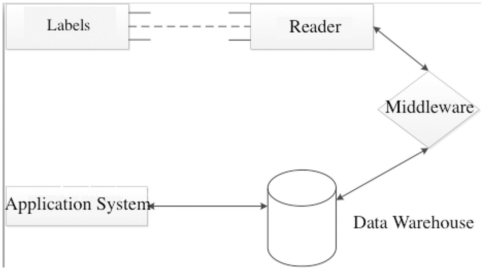


Fig. 1. RFID middleware system layer

The advantages of the RIFD middleware are [5]:

(1) Decreasing the difficulty to develop RFID applications. When companies take the extended developments, they do not need to consider about the complex RFID hardware, and they can focus on the events which they are cared at, which reduces the burden of software developers.

(2) Cutting down the development cycle. The basic developments of the software cost a lot of time. The development of RFID system is different from other common application developments. Since the simple software technology cannot deal with all RFID application problems where the most important issue is how to connect to exist RFID hardware. If we choose the mature RFID middleware, we can save 50%–75% time of the development cycle.

(3) Increasing the quality of the development. The mature RFID middleware is clear and ruled on interfaces which can efficiently prove the quality of the applications and reduce the maintenance of system.

(4) Saving the development costs. Using the mature RFID middleware can save 25%–60% money of the development. The logical structure of the RFID middleware is shown as Fig. 2. It needs to deal with three main problems: Data filtering (Data cleaning), Data integration and Data transferring. This paper mainly designed a new architecture in the RFID system middleware for data filtering and transferring.

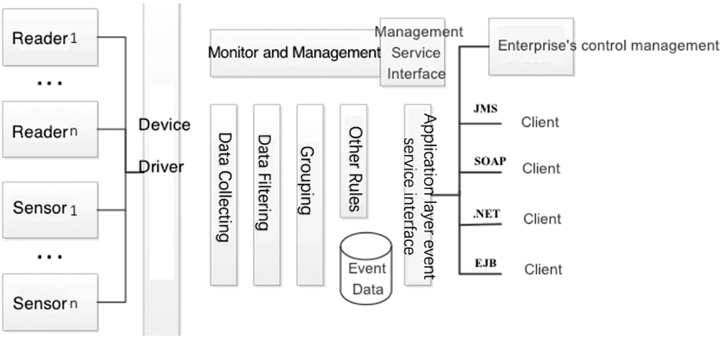


Fig. 2. Middleware logical structure

2 RFID Middleware Structure

This paper uses Java technology as the basis of the middleware design. The typical PC application system environment is shown as Fig. 3. It can be divided to three parts: Operating System, RFID application system and RFID middleware.

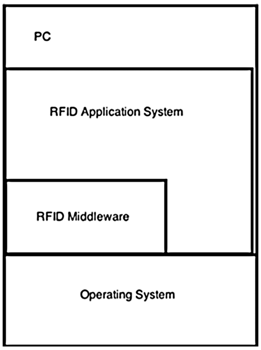


Fig. 3. RFID application software environment

In these parts, RFID middleware is the key hub to connect high-level applications and low-level RFID readers. The high-level application programs do not need to care about the low-level specific machines' (RFID readers and RFID tags) physic features, and the middleware wraps the details of the tag data using filtering. As long as the definition of the interfaces of the RFID middleware do not change, the high-level applications program will not need to change, which means it improves the scalability and the practical of the system. Role of RFID middleware in the whole RFID application system is shown in the Fig. 3. This system is realized through the Java language, because of the definite superiority of Java on the cross-platform [7] which is able to run under Windows and Linux. Thus, it is very flexible.

This paper study the RFID middleware system structure based on Java. We do not consider the other specific application scenes, and just put forward the middleware system structure based on the middleware function and the mobile devices features. The design structure is shown in Fig. 4:

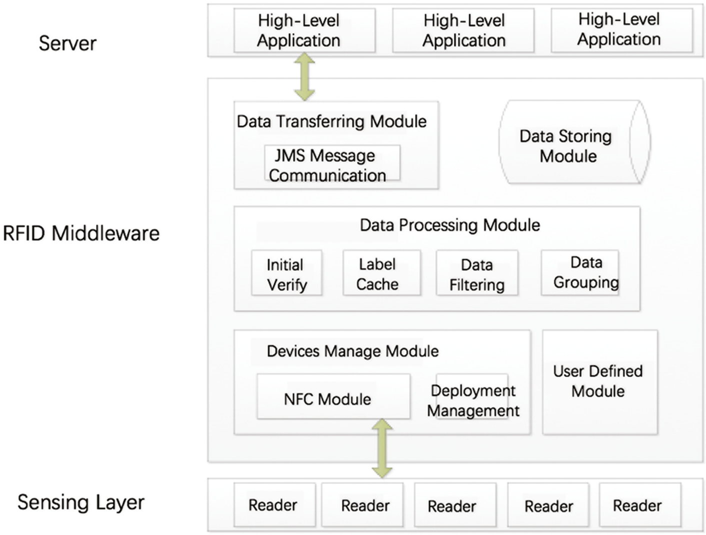


Fig. 4. Middleware system structure

(1) Devices Manage Module: It supplies the interfaces which satisfy to manage different and compatible RFID reader devices. As long as the definition of the interface does not change, the background application program will not need any change, and the function of the middleware makes the RFID system has a high scalability and practical application, which reduces the cost of RFID devices that the company needs to maintain.

(2) Data Processing Module: This module includes the preliminary verification, tag cache, redundancy data process, loss read process and data dividing. The preliminary verify uses specific protocol to verify the data. The tag cache takes the Blocking Queue [8] to memorize the data which have been preliminary verified. The redundancy data processing takes the L-NCD algorithm [9] to wipe off the redundancy data. The loss read processing takes the SMURF algorithm [10] to fill up the loss reading data. The filter function of the middleware is the significant data cleaning technology. When the device reads a lot of tags, there exists redundancy. The more reading will cause more misreading. So how to transfer real and useful data to the background applications becomes a hot study.

(3) Data Storing Module: This module takes the MySQL database to store RFID data.

(4) Data transferring Module: This module is used in the data transference between high-level applications and the middleware. It uses JMS message delivery mechanisms [11]. JMS cannot only store messages efficiently but also prove the

correct and accurate transference of the messages. JMS affords a kind of “Proving Transferring” mechanism.

(5) User Defined Module: It includes work diary management, system status query, task managements and so on.

3 Data Processing Module

3.1 Data Processing Module

The Data Processing module’s flow chart is shown as Fig. 5. Firstly, the system takes the preliminary verification to the collected tag data, which takes the Blocking Queue to cache the processed data to delete the tag data which does not belong to this system.

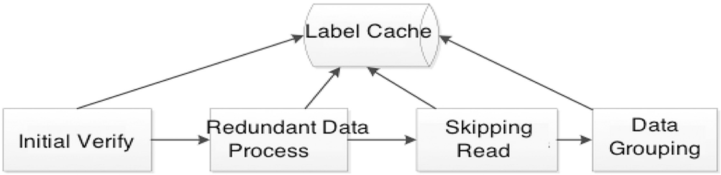


Fig. 5. Data processing module flow chart

The tags’ format is shown as Table 1, where TagID is the only tag’s marker. The first 7 bits of the verified data is the original check code, and the server generates the verified code through the CRC16 algorithm. The 8th bit is the RFID tag marker, TagID and the check code data can be transformed into PC’s MySQL database through JSON version conversation. When the middleware get the tag data, it can judge whether this tag belongs to this system through the 8th bit marker, then take the same CRC16 algorithm to generate the first 7 bits check code and compare with the check code in the database and judge whether the tag data has been changed. Checking data will be deleted after the preliminary verification.

Table 1. RFID tag data format

TagID	Source data	Data
EF4BF52A	000006D0	00 0 53 32 60 23 00 55 23 65 77 33 65 22 00 ED 4A 42

3.2 Design of Redundancy Data Elimination

The unreliability of the RFID devices result in the large number of tag data’s redundancy data, where these redundancy data can be divided into two kinds: reader redundancy and tag data redundancy. This is because several readers read one tag data or one reader read one tag data repeatedly. If we do not deal with these kinds of redundancy data, quite a lot of redundancy data rush into background applications will result in greatly decrease of the system’s performance and reduce the efficiency of the system. Our RFID

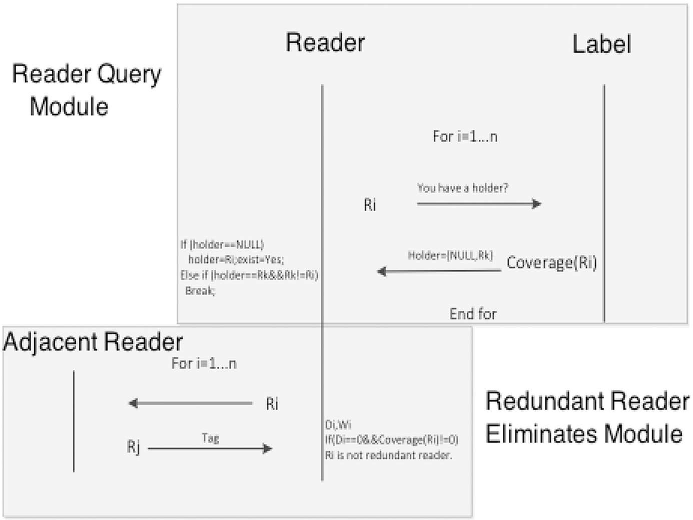


Fig. 6. Redundant data processing module

middleware system chooses L-NCD algorithm to process data redundancy. It can be divided into two specific modules: Reader query module and redundant reader elimination module. Redundant Data Processing Module is shown in Fig. 6.

The work steps of the reader query module are shown as follow:

Step1: The reader broadcasts a “query message” (you have a holder?).
Step2: Tags in the covering domain reply the query message, and notify the reader whether there is a holder.
Step3: If the holder == ‘NULL’, the reader writes its ID into this tag; if the tag has a holder and is different from the reader’s, then the reader ignores this reply; otherwise switch to the next step.
Step4: All the tags in the reader’s covering domain do not have ‘NULL’ in their replies, and this reader is a redundant reader, go to the next step.
Step5: Reader broadcasts a “query message” to the adjacent reader.
Step6: The adjacent reader replies the tag value (tag is the marker to record the number of tags in the covering domain).

After the whole query process finished, go to the redundant reader eliminate module. The steps run as follow:

Step1: Reader calculates the adjacent coverage density D and the weight W.
Step2: If the reader’s adjacent coverage density is 0 and covering domain’s tag’s number is not zero, this reader is not a redundant reader. Otherwise go to the next step.
Step3: List the readers’ weight comparing chart.
Step4: The tag chooses the largest weight reader as its holder.
Step5: Eliminate the readers which are not chosen by the tags.

3.3 Design of Skipping Reading Data Process

After the redundant data processing finished, it should go to the skipping reading data processing. We take the SMURF algorithm to filter the skipping reading data, the filtering processes are shown in Fig. 7.

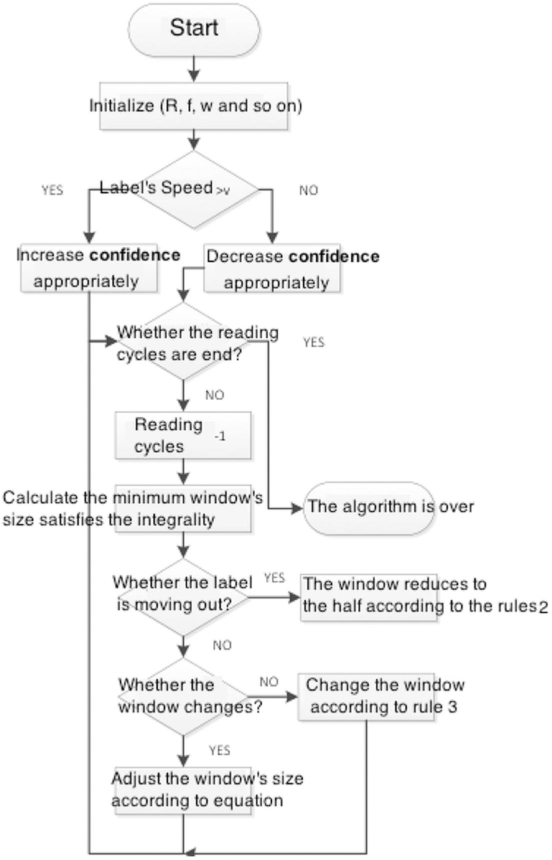

Start
Initialize (R, f, w and so on)
Label's Speed >v
YES
NO
Increase confidence appropriately
Decrease confidence appropriately
Whether the reading cycles are end?
YES
NO
Reading cycles -1
Calculate the minimum window's size satisfies the integrality
The algorithm is over
Whether the label is moving out?
YES
The window reduces to the half according to the rules2
NO
Whether the window changes?
NO
Change the window according to rule 3
YES
Adjust the window's size according to equation

Fig. 7. SMURF algorithm filtering flow

3.4 Data Transferring Module

RFID middleware is a Message-Oriented Middleware (MOM) [12]. Information is transferring from one program to another program through messages. Information can be transferred using asynchronous method, so the submitter does not need to wait for the reply. The middleware aiming at the message contains the functions not only the passing message but also includes translating data, security, data broadcasting, error recovering and so on. This module takes the responsibility of transferring data between RFID middleware and the sever through JMS message transferring mechanism. JMS can not only store messages efficiently but also prove the correct and accurate

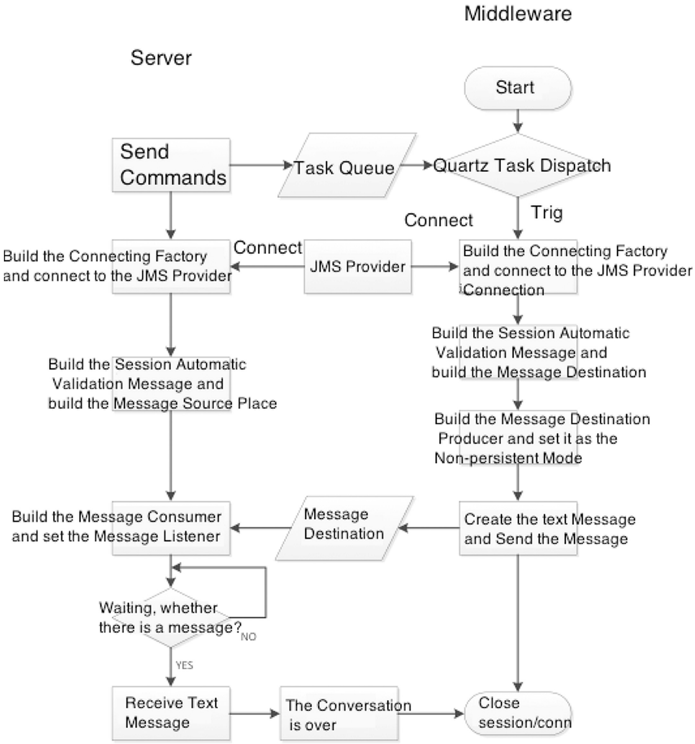

Middleware
Server
Start
Send Commands
Task Queue
Quartz Task Dispatch
Connect
Trig
Build the Connecting Factory and connect to the JMS Provider
Connect
JMS Provider
Build the Connecting Factory and connect to the JMS Provider Connection
Build the Session Automatic Validation Message and build the Message Destination
Build the Session Automatic Validation Message and build the Message Source Place
Build the Message Destination Producer and set it as the Non-persistent Mode
Build the Message Consumer and set the Message Listener
Message Destination
Create the text Message and Send the Message
Waiting, whether there is a message?
NO
YES
Receive Text Message
The Conversation is over
Close session/conn

Fig. 8. Data transferring module flow chart

transference. It also provides a kind of "proving transferring mechanism". This module's flow chart is shown in Fig. 8.

Data transferring between middleware and sever is realized through Java. It mainly uses class library which includes ConnectionFactory, Connection, Session, Destination, MessageProducer, MessageConsumer, TextMessage and MessageListener.

3.5 Other Modules' Design

Devices Manage Module is responsible for managing and monitoring low-level reader devices. This module mainly includes NFC module and other related modules. The NFC module contains reader management and electric tag management which is shown in Fig. 9.

Reader manage module can take NFC with sensing layer to manage readers, the electric tag manage module is responsible for reading and writing on the RFID electric tag. This module can be realized through libnfc_1.5.0. Libnfc [13] was the first free bottom of the NFC development kit and programming API released under the GNU public license.

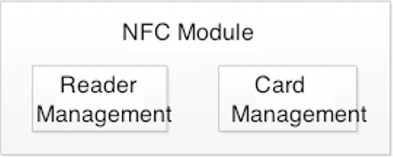
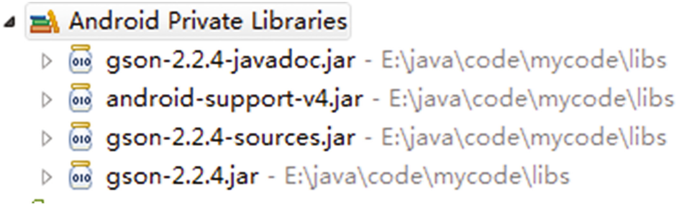


Fig. 9. NFC module

The data storage module is in charge of storing and back up of data. It generates the JSON file from the related Java subjects which is sent to the server and stores to the local MYSQL database. JSON and java subjects can be shown as Fig. 10.



Android Private Libraries
gson-2.2.4-javadoc.jar - E:\java\code\mycode\libs
android-support-v4.jar - E:\java\code\mycode\libs
gson-2.2.4-sources.jar - E:\java\code\mycode\libs
gson-2.2.4.jar - E:\java\code\mycode\libs

Fig. 10. GSON component library

The related code is as follow:

```
Gson gson = new Gson();
String jsonStr = gson.toJson(mMyDraw);
MyDraw mFormJson = gson.fronJson(jsonStr, MyDraw.class);
```

4 Conclusion

This paper designed a kind of RFID middle ware system based on java and the existed middleware technology; it shows the system structure and the flow chart of each modules; introduces the data processing module includes the skipping reading processing module and the redundant reader elimination module, which use the L-NCD and VSMURF algorithms; it also introduces the data transferring module which is in charge of transferring between RFID middleware and the sever, it uses the JMS message transferring mechanism; at last it shows the results and the analysis the experiment.

Acknowledgments. This work is financially supported by the National Natural Science Foundation of P. R. China (No. 61373017, No. 61572260, No. 61572261, No. 61672296, No. 61602261), the Natural Science Foundation of Jiangsu Province (No. BK20140886, No. BK20140888), Scientific & Technological Support Project of Jiangsu Province (No. BE2015702, BE2016185, No. BE2016777), Natural Science Key Fund for Colleges and Universities in Jiangsu Province (No. 12KJA520002), China Postdoctoral Science Foundation (No. 2014M551636, No. 2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1302090B, No. 1401005B), Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 14KJB520030), Jiangsu Postgraduate Scientific Research and Innovation Projects (SJLX15_0381, SJLX16_0326), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks (WSNLBZY201509), NUPTSF(Grant No. NY214060, No. NY214061) and the STITP projects of NUPT (No. XZD2016032 and No. SZD2016030).

References

1. Sakurai, S.: Prediction of sales volume based on the RFID data collected from apparel shops. Int. J. Space-Based Situated Comput. **1**(2–3), 174–182 (2011)
2. Chiou, J.C., Hsu, S.H., Kuei, C.K., et al.: An addressable UHF EPCGlobal class1 gen2 sensor IC for wireless IOP monitoring on contact lens. In: 2015 International Symposium on VLSI Design, Automation and Test (VLSI-DAT), pp. 1–4. IEEE (2015)
3. Chen, J.C., Cheng, C.H., Huang, P.T.B.: Supply chain management with lean production and RFID application: a case study. Expert Syst. Appl. **40**(9), 3389–3397 (2013)
4. Aazam, M., Huh, E.N.: Fog computing: the Cloud-IoT/IoE middleware paradigm. IEEE Potentials **35**(3), 40–44 (2016)
5. Bouhouche, T., Boulmalf, M., Bouya, M., et al.: A new middleware architecture for RFID systems. In: 2014 14th Mediterranean on Microwave Symposium (MMS), pp. 1–7. IEEE (2014)
6. Zhu, J., Huang, J., He, D., et al.: The design of RFID middleware data filtering based on coal mine safety. In: The Proceedings of the Second International Conference on Communications, Signal Processing and Systems, pp. 859–867. Springer (2014)
7. Tian, M., Wan, M., Song, Y., et al.: An improvement on software of zeeman experiment based on Java cross-platform technology. J. Pingdingshan Univ. **5**(2), 17–25 (2015)
8. Upadhyaya, G., Rajan, H.: An automatic actors to threads mapping technique for JVM-based actor frameworks. In: Proceedings of the 4th International Workshop on Programming based on Actors Agents & Decentralized Control, pp. 29–41. ACM (2014)
9. Xu, H., Shen, W., Li, P., et al.: A novel algorithm L-NCD for redundant reader elimination in P2P-RFID network. J. Algorithms Comput. Technol. (2017). doi:10.1177/1748301816688020
10. Hongsheng, Z., Jie, T., Zhiyuan, Z.: Limitation of RFID data cleaning method—SMURF. In: 2013 IEEE International Conference on RFID-Technologies and Applications (RFID-TA), pp. 1–4. IEEE (2013)
11. Li, J.: Design of RFID middleware based on SOA with JMX and JMS. Appl. Electron. Tech. **36**(4), 119–122 (2010)
12. Herron, D., Castillo, O., Lewis, R.: Systems and methods for individualized customer retail services using RFID wristbands. U.S. Patent Application 14/034, 395, 23 September 2013
13. Madlmayr, G., Kantner, C., Grechenig, T.: Near field communication. In: Secure Smart Embedded Devices, Platforms and Applications, pp. 351–367. Springer New York (2014)

Multidimensional Zero-Correlation Linear Cryptanalysis on PRINCE

Lu Cheng(✉), Xiaozhong Pan, Yuechuan Wei, and Liqun Lv

Department of Electronic Technology,
Engineering University of Chinese People's Armed Police, Xi'an Shaanxi, China
18302972151@163.com

Abstract. The PRINCE is a light-weight block cipher with the 64-bit block size and 128-bit key size. It is characterized by low power-consumption and low latency. PRINCEcore is the PRINCE cipher without key-whiting. For evaluating its security, a statistical testing on linear transformation is performed, and a statistical character matrix is given. By using the "miss-in-the-middle" technique, we construct 5-round zero-correlation linear approximations. Based on the 5-round distinguisher, a 9-round attack on the PRINCEcore is performed. The data complexity is $2^{62.9}$ known plaintexts and the time complexity is $2^{55.14}$ 9-round encryptions. The testing result shows that the PRINCEcore reduced to 9 rounds is not immune to multidimensional zero-correlation linear analysis.

1 Introduction

In recent years, the term "Internet of things" [15–22] has become more and more important in people's field of vision. The development of Internet of Things has been put on the strategic level by many countries and the scale of pertinent industry has been expanding. Due to manufacturing costs and portability limitations, many of the devices in the Internet of Things are computationally weak microprocessors. In order to protect the radio frequency identification tags and smart card and other equipment communication security, while adapting to resource-constrained environment, many lightweight block ciphers has been designed, such as PRESENT, LBlock, LED, KATAN, KTANTAN and so on [1–4]. On Asiacrypt2012, Borghoff Proposed a lightweight block cipher-PRINCE [5] which with 64-bit block size and 128-bit key size. PRINCE is a cryptographic algorithm based on the FX-structure which is an interesting property with a slight change in the round key. And it makes encryption and decryption consistent, so that the decryption costs will be negligible. And the property is called a reflection feature.

The zero-correlation linear cryptanalysis method was first proposed by Bogdanov and Rijmen in [6]. The biggest drawback is the data complexity for attack. In order to further reduce the data complexity, Bogdanov proposed multiple zero-correlation linear analysis using a number of zero-correlation linear approximations to distinguish statistical distributions at FSE2012 in [7], but required multiple zero-correlation linear approximation is independent of each other, and this assumption is difficult to satisfy. Then Bogdanov proposed a multidimensional zero-correlation linear analysis model at

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_22

ASIACRYPT2012 in [8] to overcome the strong assumptions required for multiple zero correlation linear analysis. The data complexity is substantially equivalent of multiple zero correlation linear analysis. The validity of the zero correlation linear analysis is applied on a series of important cryptographic algorithms for AES, LBlock, TWINE, E_2, SMS4, FOX and MIBS [8–13]. It is a focus in the current block cipher analysis to evaluate the security of partial block cipher algorithm by using zero correlation linear analysis.

The paper is organized as follows: Sect. 2 give an introduction of PRINCE and introduce some properties of the matrix M. In Sect. 3, we briefly introduce multidimensional zero-correlation linear cryptanalysis. A 5-round zero correlation linear approximations of PRINCE is discussed in Sect. 4. Section 5 gives the 9-round zero correlation linear attack on PRINCE. Section 6 concludes the paper and summarizes our results.

2 Backgrounds

In this section, we give an introduction of PRINCE and introduce some properties of the matrix M.

2.1 The Encryption Process of PRINCE

PRINCE is a block cipher with 64-bit block size and supports 128-bit key size. The key is split into two parts of 64 bits each,

$$k = k_0 \parallel k_1$$

and extended to 192 bits by the mapping

$$(k_0 \parallel k_1) \rightarrow (k_0 \parallel k_0' \parallel k_1) = (k_0 \parallel (k_0 > \; > \; > 1) \oplus (k_0 > \; > 63) \parallel k_1)$$

The first two sub-keys k_0 and $k_0^{'}$ are used as whitening keys, while the key k_1 is the 64-bit key for a 12-round block cipher we refer to as PRINCEcore, and the middle transformation is recorded as two rounds. The whole encryption process of PRINCE is shown in Fig. 1.

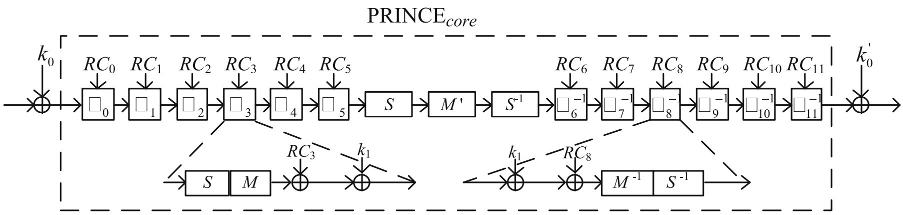


Fig. 1. Structure of PRINCE

PRINCEcore has two kinds of round transformation R_i and R_i^{-1}, and they are inverse transforms. And both iterate 6 rounds, and the middle transformation is recorded as two rounds.

$$PRINCE_{core} = R_{11}^{-1} \circ R_{10}^{-1} \circ R_9^{-1} \circ R_8^{-1} \circ R_7^{-1} \circ R_6^{-1} \circ M_{mid} \circ R_5 \circ R_4 \circ R_3 \circ R_2 \circ R_1 \circ R_0$$

And the middle transformation M_{mid} is defined as $M_{mid} = S \circ M' \circ S^{-1}$.

The 64-bit block is regarded as a 4 × 4 state matrix, each element is 4 bits (nibbles), and the round transformation is composed of four basic transformations, namely, S-box, a linear permutation M, round constant addition, and key addition, $R = SB \circ M \circ AC \circ AK$.

Key-add (AK): 64-bit state is xored with the 64-bit round key.
RC-add (AC): The 64-bit round constant is xored with the 64-bit state.
S-box (SB): A nibble uses a 4-bit S-box and there are 16 S-boxes. The S-box mapping values are given in Table 1.

Table 1. S-box mapping (in hexadecimal)

Input	0	1	2	3	4	5	6	7
Output	B	F	3	2	A	C	9	1
Input	8	9	A	B	C	D	E	F
Output	6	7	8	0	E	5	D	4

The Matrices (M): The linear transformation M is applied to a 4 × 4 state matrix, $M = SR \circ M'$,and SR is a row shift to the matrix. $M' = diag(\hat{M}^0, \hat{M}^1, \hat{M}^1, \hat{M}^0)$ is a 64 × 64 diagonal matrix, which is composed of two transformations $\hat{M}^0$ and $\hat{M}^1$, $\hat{M}^0$ and $\hat{M}^1$ are both 16 × 16 binary matrices, defined as follows.

$$\hat{M}^0 = \begin{pmatrix} M_0 & M_1 & M_2 & M_3 \\ M_1 & M_2 & M_3 & M_0 \\ M_2 & M_3 & M_0 & M_1 \\ M_3 & M_0 & M_1 & M_2 \end{pmatrix} \qquad \hat{M}^0 = \begin{pmatrix} M1 & M2 & M3 & M0 \\ M2 & M3 & M0 & M1 \\ M3 & M0 & M1 & M2 \\ M0 & M1 & M2 & M3 \end{pmatrix}$$

$$M_0 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} M_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} M_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} M_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

2.2 Properties of the Matrix M

It is easy to verify that the linear transformation M' of the PRINCE cipher is a diagonal matrix, and when the diagonal matrix is applied to a 4×4 state matrix, it is equivalent to applying each sub-matrix to each column of the state matrix. For further analysis, the following properties of linear transformation M' can be obtained. We agree that the 16 nibbles of the state matrix are mapped to 64-bit words in the following order.

0	4	8	12
1	5	9	13
2	6	10	14
3	7	11	15

Property 1 [14]: The state matrix is regard as 16 columns of 4-bit data $i = 0, \ldots, 15$, and the linear transformation M' is equivalent to 16 independent linear transformations were applied to each column, the transformation is as follows.

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \rightarrow \begin{pmatrix} wt(x) \bmod 2 \\ wt(x) \bmod 2 \\ wt(x) \bmod 2 \\ wt(x) \bmod 2 \end{pmatrix} + Rot_{n_i} \begin{pmatrix} x_4 \\ x_3 \\ x_2 \\ x_1 \end{pmatrix}$$

Where $wt(x)$ is hamming weight of x and Rot_{n_i} denotes some rotations by n_i positions to the top.

Based on the Property 1, we use the computer program to test the linear transformation M', and then get the Property 2.

Property 2: The 16 nibble state is divided into 4 groups, and each column in the state matrix is a group. We find that the confusion property of the linear transformation M' makes that if a group has a nonzero nibble state, then only the same group has nonzero nibble state after the transformation M'. n_{ij} denotes the number of group in which $i (0 \leq i \leq 4)$ nonzero nibble state with fixed position transforms to $j (0 \leq j \leq 4)$ nonzero nibble state with fixed position by the transformation M'. The values of n_{ij} is showed in Table 2 which are obtained by computer search, so that the probability of that linear

Table 2. n_{ij} values

i	j				
	0	1	2	3	4
0	1	0	0	0	0
1	0	0	0	16	44
2	0	0	28	256	1066
3	0	16	256	2848	10380
4	0	44	1066	10380	39135

transformation M' transforms i nonzero nibble masks with a fixed position into j nonzero nibble masks with a fixed position is $p_{i,j} = \frac{n_{ij}/C_4^i C_4^j}{(2^4-1)^i}$.

3 Multidimensional Zero-Correlation Linear Cryptanalysis

A linear approximation of a vectorial function f based on F_2^n is determined by n bit input mask α and output mask β. And we define the corresponding linear approximation $(\alpha \xrightarrow{f} \beta)$, it can be abbreviated as $(\alpha \rightarrow \beta)$. The correlation coefficient of the linear approximation is defined as:

$$\begin{aligned} C_f(\alpha, \beta) &= C\mathrm{or}_x(\beta \cdot f(x) \oplus \alpha \cdot x) \\ &= 2\mathrm{Pr}_x(\beta \cdot f(x) \oplus \alpha \cdot x = 0) - 1 \end{aligned}$$

If $C_f(\alpha, \beta) = 0$, we call that the linear approximation is zero-correlation linear approximation.

In multidimensional zero-correlation linear cryptanalysis,we treat the $\ell = 2^m$ zero-correlation linear approximations available as a linear space spanned by m base zero-correlation linear approximations, and these linear approximations can be defined as $(a_i \rightarrow b_i)_{i=0,1,\cdots m-1}$. The attacker chooses N pairs of plain-cipher and initializes a counter $N[\mathbf{z}]$, where $\mathbf{z}$ is a m bit vector. Then extending some rounds before and after the linear approximation, guessing the keys and partially encrypting and decrypting the selected plain-cipher pairs, then the corresponding intermediate state of the linear mask could be obtained, and is recorded as (p', c'). Then, for each of the selected plain-cipher pair calculated the (p', c') by partially encrypting and decrypting, and the value of the vector z can be obtained by Eq. (1):

$$\begin{aligned} \mathbf{z} &= (\mathbf{z}[0], \mathbf{z}[1], \cdots \mathbf{z}[m-1]) \\ &= (a_0 \cdot p' \oplus b_0 \cdot c', a_1 \cdot p' \oplus b_1 \cdot c' \cdots a_{m-1} \cdot p' \oplus b_{m-1} \cdot c') \end{aligned} \tag{1}$$

Update the corresponding counter $N[z]$. Then, calculate the statistics T:

$$T = \sum_{z=0}^{2^m-1} \frac{(N[z] - N2^{-\mathrm{m}})^2}{N2^{-\mathrm{m}}(1 - 2^{-m})} \approx N2^m \sum_{z=0}^{2^m-1} (\frac{N[z]}{N} - \frac{1}{2^m}) \tag{2}$$

If the guessed key is right key, the statistic T follows a normal distribution with the mean $\mu_0 = (\ell - 1)\frac{2^n-N}{2^n-1}$ and the variance $\sigma_0^2 = 2(\ell - 1)(\frac{2^n-N}{2^n-1})^2$. If the guessed key is wrong key, the statistic T follows a normal distribution with the mean $\mu_1 = \ell - 1$ and the variance $\sigma_1^2 = 2(\ell - 1)$. If we denote the probability that the right key guess is wrongfully discard as α, the probability that the wrong key guess is regarded as the right key as β, and the decision threshold as $\tau = \mu_0 + \sigma_0 \mathbf{z}_{1-\alpha} = \mu_1 - \sigma_0 \mathbf{z}_{1-\beta}$, then the number of known plaintexts for the attack should be approximately:

$$N = \frac{(2^n - 1)(\mathbf{z}_{1-\alpha} + \mathbf{z}_{1-\beta})}{\sqrt{(\ell - 1)/2} + \mathbf{z}_{1-\alpha}} + 1 \tag{3}$$

Where n is the block length; $\mathbf{z}_{1-\alpha}, \mathbf{z}_{1-\beta}$ are the respective quantiles for the standard normal distribution. For a more complete and detailed description of this method, please refer to [8].

4 5-Round Zero Correlation Linear Approximations of PRINCE

In this section, by using the "miss-in-the-middle" technique, we construct 5-round zero correlation linear approximations of PRINCE. x_i denotes the intermediate state of each round in the encryption process $x_3; x_4; x_5; x_8 \rightarrow R_3; R_4; R_5; R_6^{-1}$;$x_i^S$, $x_i^{M'}$ and $x_i^{SR^{-1}}$ denotes the output of the S-box, the linear transformation M', and the SR^{-1} shift respectively. We agree that the 16 nibbles of the state matrix are mapped to 64-bit words in the following order.

0	4	8	12
1	5	9	13
2	6	10	14
3	7	11	15

Theorem 1: If the input mask Γx_3^S of the linear transformation M' in the 3rd round R_3 is $(1000, 1000, 1000, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and the output mask $\Gamma x_8^{SR^{-1}}$ of the transformation SR^{-1} in the 8th round R_6^{-1} is in the form of (0,?,?,0,0,0,?,?,?, 0,0,?,?,?,0,0), then there is a contradiction, so a 5-round zero correlation linear approximation is constructed. Where in each of the nibble positions, 0 represents 4 bits are all 0 mask, and ? represents the random mask value.

Proof: Referring to Fig. 2 for proof. "•" represents a nonzero nibble mask. If the output mask Γx_3^S of S-box transformation in the 3rd round R_3 is $(1000, 1000, 1000, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, after a linear transformation M', the mask is $(1000, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, and the line shift SR does not affect the mask form. In the 4th round R_4, the S-box maps a nonzero nibble mask to a nonzero nibble mask, and the linear transformation M' extends at least one nonzero mask to three (as shown in Fig. 2a, b, the linear transformation M' extends one nonzero mask to four), the transformation SR transforms three nonzero nibble mask in the same column to a different column, the linear transformation M' in the 5th round R_5 extends three nonzero nibble masks to at least nine (as shown in Fig. 2a, b, after the linear transformation M' in the 5th round R_5, the number of nonzero mask can be up to 16), in middle transformation, after the S-box transformation, there are at least nine nonzero nibble masks in the state matrix. In the decryption direction, considering the output mask $\Gamma x_6^{SR^{-1}}$ of transformation SR^{-1} in the 8th round R_6^{-1} is (0,?,?,0,0,0,?,?,?,0,0,?,?,?,0,0), and ? represents the random mask value, there are at least eight nonzero nibble masks of the output mask of M' in middle transformation, which is contradict to the conclusion that there is at least nine nonzero masks in the state in the encryption direction. Thus, 5-round zero correlation linear approximations of PRINCE can be constructed. The following corollary is easily obtained by the proof of Theorem 1.

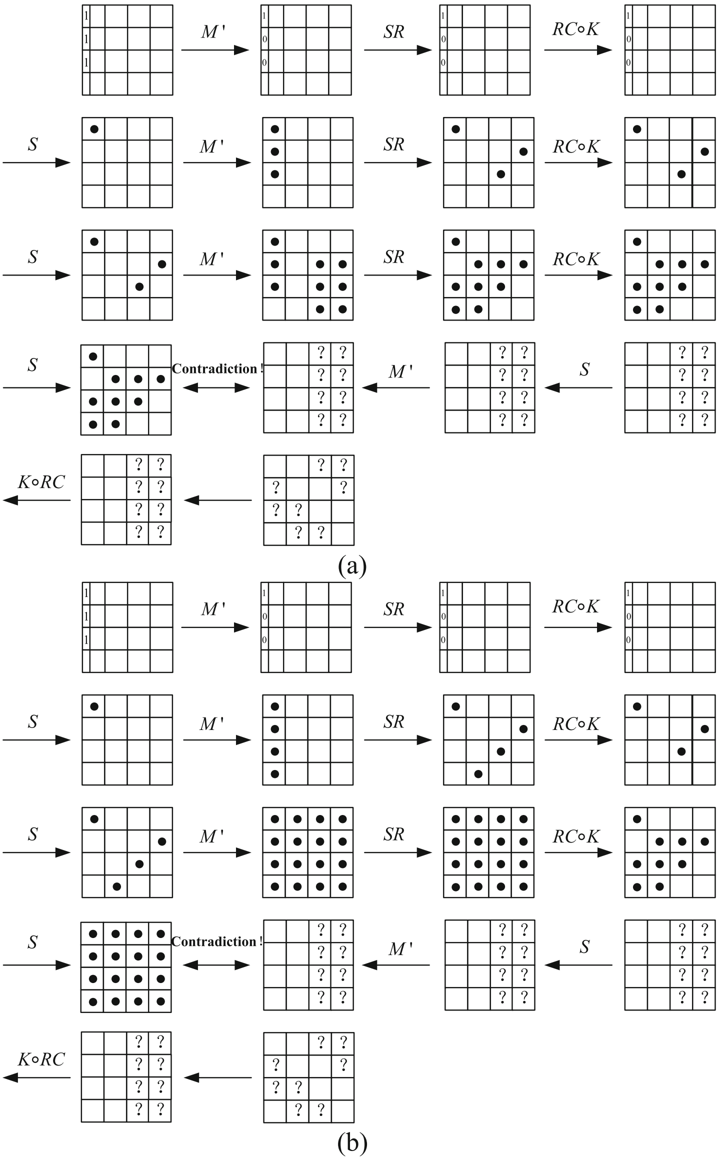


Fig. 2. 5-round zero correlation linear approximations of PRINCE

Corollary 1: If the input mask Γx_3^S of the linear transformation M' in the 3rd round R_3 is in the form of $(\bullet,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$, and the output mask $\Gamma x_8^{SR^{-1}}$ of the transformation SR^{-1} in the 8th round R_6^{-1} is in the form of (0,?,?,0,0,0,?,?,?,0,0, ?,?,?,0,0), then there is a contradiction, so a 5-round zero correlation linear approximation is constructed. Where in each of the nibble positions, 0 represents 4 bits are all 0 mask, and ? represents the random mask value.

5 9-Round Zero Correlation Linear Attack on PRINCE

The attack process in this section is based on the 5-round zero-correlation linear approximations in Corollary 1, and the 5-round zero-correlation linear approximations (the 3rd round - the 8th round) for the attack are selected $(a,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0) \not\rightarrow (0,b,b,0,0,0,0,0,0,0,0,0,0,0,0,0)$.Based on the special properties of the middle transformation of PRINCE, the attack path chosen is two rounds extending from both sides of the middle transformation, that is, the attack path is from the 2nd round to the 10th round.

Let (P, C) denotes plain-cipher pairs; x_i denotes the intermediate state in the attack process $x_2; x_3; x_8; x_9; x_{10} \rightarrow R_2; R_3; R_6^{-1}; R_7^{-1}; R_8^{-1}$; x_i^S, $x_i^{M'}$ and x_i^{SR} denotes the output of the S-box, the linear transformation M', and the SR shift respectively; $x_i(j)$ denotes the $(j+1)^{th}$ nibble of the intermediate state; $x_{i,col(l)} (1 \leq l \leq 4)$ denotes the l^{th} column of x_i; $k_1(j) (0 \leq j \leq 15)$ denotes the $(j+1)^{th}$ nibble of k_1.

Here is the 9-round zero-correlation linear attack on PRINCE in Fig. 3. And it shows the attack path and the probability of each step, $''*''$ denotes the involved state nibbles in the attack, and the attack steps are as follows.

Pre-computing: Allocate 16-bit counters $N_0[y_0]$ for 2^{112} possible values of $y_0 = x_2(0,1,2,3,4,5,6,7,8,9,10,11)||x_{10}(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)$ where $x_2; x_{10}$ denotes the initial state of the 2nd and 10th round in the attack process, and initialize the counters to zero. Collect N PC pairs, and extract y_0 from each of N PC pairs and increment the corresponding counter $N_0[y_0]$ by one. No more than 2^{64} PC pairs are divided into 2^{112}, so 16-bit counter is sufficient.

Step 1: Allocate 16-bit counters $N_1[y_1]$ for 2^{44} possible values of $y_1 = x_2^{SR}(0,1,2)|| x_{10}^{SR}(0,1,2,3,4,5,6,7)$ and initialize the counters to zero. Compute $x_2^{SR} = SR \circ M' \circ S(x_2)$; $x_{10}^{SR} = SR \circ M' \circ S(x_{10})$, and update the corresponding counter $N_1[y_1] += N_0[y_0]$, the time complexity of this step is about N S-box computing.

Step 2: Allocate 16-bit counters $N_2[y_2]$ for 2^{36} possible values of $y_2 = x_3^{M'}(0)||x_9^S(0,1,2)||x_{10}^{SR}(3,4,5,6,7)$ and initialize the counters to zero. Guess 12-bit $K_1(0,1,2)$, and compute $x_3^{M'}(0) = M' \circ S \circ RC \circ K[x_2^{SR}(0,1,2)]$; $x_9^S(0,1,2) = S \circ RC \circ K[x_{10}^{SR}(0,1,2)]$, and update the corresponding counter $N_2[y_2] += N_1[y_1]$, the time complexity of this step is about $2^{12} \times 2^{44} = 2^{56}$$S$-box computing.

Step 3: Allocate 16-bit counters $N_3[y_3]$ for 2^{24} possible values of $y_3 = x_3^{M'}(0)||x_9^{M'}(0)||x_{10}^{SR}(4,5,6,7)$ and initialize the counters to zero. Guess 4-bit $K_1(3)$, and compute $x_9^S(3) = S \circ RC \circ K[x_{10}^{SR}(3)]$; $x_9^{M'}(0) = M'[x_9^S(0,1,2,3)]$, and update the corresponding counter $N_3[y_3] + = N_2[y_2]$, the time complexity of this step is about $2^4 \times 2^{12} \times 2^{36} = 2^{52}$ S-box computing.

Step 4: Allocate 16-bit counters $N_4[y_4]$ for 2^{12} possible values of $y_4 = x_3^{M'}(0)||x_9^{M'}(0,5)$ and initialize the counters to zero. Guess 16-bit $K_1(4,5,6,7)$, and compute $x_9^S(4,5,6,7) = S \circ RC \circ K[x_{10}^{SR}(4,5,6,7)]$; $x_9^{M'}(5) = M'[x_9^S(4,5,6,7)]$, and update the corresponding counter $N_4[y_4] + = N_3[y_3]$, the time complexity of this step is about $2^{16} \times 2^{16} \times 2^{24} = 2^{56}S$-box computing.

Step 5: Allocate 16-bit counters $N_5[y_5]$ for 2^{12} possible values of $y_5 = x_3^{M'}(0)||x_8^{M'}(1,2)$ and initialize the counters to zero. Guess 8-bit $K_1(0,1)$, and compute $x_8^{M'}(1,2) = M' \circ S \circ RC \circ K \circ SR[x_9^{M'}(0,5)]$, and update the corresponding counter $N_5[y_5] + = N_4[y_4]$, the time complexity of this step is about $2^8 \times 2^{32} \times 2^{12} = 2^{52}S$-box computing.

Step 6: Allocate 16-bit counters $N_6[y_6]$ for 2^8 possible values of $y_6 = I_0||I_1$ and initialize the counters to zero. Compute $I_0 = x_3^{M'}(0); I_1 = x_8^{M'}(1) \oplus x_8^{M'}(2)$, and update the corresponding counter $N_6[y_6] + = N_5[y_5]$.

Step 7: $\mathbf{z}$ is an 8bit vector, for each possible $\mathbf{z}$, allocate an 8bit counter $N[\mathbf{z}]$, and all initialized to zero. For the eight 8-bit basic vectors, that is, the $i + 1^{th}$ bit is 1 and the other bits are 0. Compute $\mathbf{z}[i] = \mathbf{z}_i \cdot y_6, 0 \leq i \leq 7$, and compute $\mathbf{z}$, and then update the corresponding counter $N[\mathbf{z}] + = N_6[y_6]$. According to Eq. (2), compute the

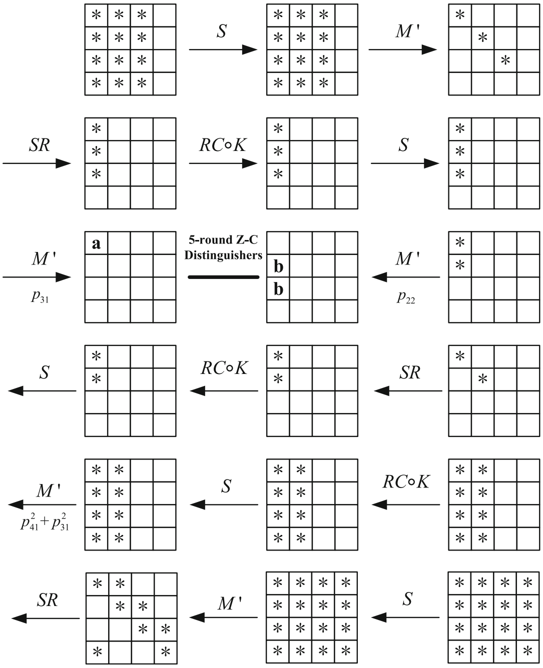


Fig. 3. 9-round zero correlation linear attack on PRINCE

statistics T. If $T \le \tau$, then the guessed key may be the right key, exhaustive search for all possible right keys.

In the attack process, if $\alpha = 2^{-2.7}$,$\beta = 2^{-20}$, then $\mathbf{z}_{1-\alpha} \approx 1$,$\mathbf{z}_{1-\beta} \approx 4.76$. And because of $n = 64$ and $\ell = 2^8$, according to the Eq. (3), the PC pairs required is generally $2^{62.9}$ and the decision threshold $\tau \approx 2^{7.2}$. Therefore, the attack requires a total of $2^{48.6} + 2^{62.2} + 2^{54.6} + 2^{58.1} + 2^{44.6} \simeq 2^{62.3}$ S-box computing, equivalent of $2^{62.9} \times \frac{1}{16} \times \frac{1}{9} \simeq 2^{55.7}$ 9-round encryption operations.

6 Summary

This paper first analyzes the immunity of the PRINCE cipher for the multidimensional zero-correlation linear cryptanalysis. For evaluating its security, a statistical testing on linear transformation is performed, and a statistical character matrix is given. By using the "miss-in-the-middle" technique, a 5-round zero-correlation linear approximation is constructed. Based on the 5-round distinguisher, a 9-round attack on the PRINCEcore is performed. The data complexity is $2^{62.9}$ known plaintexts and the time complexity is $2^{55.7}$ 9-round encryptions. The testing result shows that the PRINCEcore reduced to 9 rounds is not immune to multidimensional zero-correlation linear attact.

Acknowledgments. The authors would like to thank the anonymous referees for their valuable remarks. This work is partially supported by the National Natural Science Foundation of China (61202492,61572521), the Foundation of Science and Technology on Information Assurance Laboratory (KJ-15-010), the Natural Science Foundation of Shanxi Province (2016JQ6030).

References

1. Bogdanov, A., Knudsen, L.R., Leander, G., Paar, C., Poschmann, A., Robshaw, M.J.B., Seurin, Y., Vikkelsoe, C.: PRESENT: an ultra-lightweight block cipher. In: Paillier, P., Verbauwhede, I. (eds.) CHES 2007. LNCS, vol. 4727, pp. 450–466. Springer, Heidelberg (2007). doi:10.1007/978-3-540-74735-2_31
2. Wu, W., Zhang, L.: LBlock: a lightweight block cipher. In: Lopez, J., Tsudik, G. (eds.) ACNS 2011. LNCS, vol. 6715, pp. 327–344. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21554-4_19
3. Guo, J., Peyrin, T., Poschmann, A., Robshaw, M.: The LED block cipher. In: Preneel, B., Takagi, T. (eds.) CHES 2011. LNCS, vol. 6917, pp. 326–341. Springer, Heidelberg (2011). doi:10.1007/978-3-642-23951-9_22
4. Cannière, C., Dunkelman, O., Knežević, M.: KATAN and KTANTAN — a family of small and efficient hardware-oriented block ciphers. In: Clavier, C., Gaj, K. (eds.) CHES 2009. LNCS, vol. 5747, pp. 272–288. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04138-9_20
5. Borghoff, J., et al.: PRINCE – a low-latency block cipher for pervasive computing applications. In: Wang, X., Sako, K. (eds.) ASIACRYPT 2012. LNCS, vol. 7658, pp. 208–225. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34961-4_14

6. Bogdanov, A., Rijmen, V.: Linear hulls with correlation zero and linear cryptanalysis of block ciphers. Des. Codes Crypt. **70**(3), 369–383 (2014)
7. Bogdanov, A., Wang, M.: Zero correlation linear cryptanalysis with reduced data complexity. In: Canteaut, A. (ed.) FSE 2012. LNCS, vol. 7549, pp. 29–48. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34047-5_3
8. Bogdanov, A., Leander, G., Nyberg, K., Wang, M.: Integral and multidimensional linear distinguishers with correlation zero. In: Wang, X., Sako, K. (eds.) ASIACRYPT 2012. LNCS, vol. 7658, pp. 244–261. Springer, Heidelberg (2012). doi:10.1007/978-3-642-34961-4_16
9. Wang, Y., Wu, W.: Improved multidimensional zero-correlation linear cryptanalysis and applications to LBlock and TWINE. In: Susilo, W., Mu, Y. (eds.) ACISP 2014. LNCS, vol. 8544, pp. 1–16. Springer, Cham (2014). doi:10.1007/978-3-319-08344-5_1
10. Wen, L., Wang, M., Bogdanov, A.: Multidimensional zero-correlation linear cryptanalysis of E2. In: Pointcheval, D., Vergnaud, D. (eds.) AFRICACRYPT 2014. LNCS, vol. 8469, pp. 147–164. Springer, Cham (2014). doi:10.1007/978-3-319-06734-6_10
11. Ma, M., Zhao, Y., Liu, Q., Liu, F.: Multidimensional zero-correlation linear cryptanalysis on SMS4 algorithm. J. Cryptol. Res. **2**(5), 458–466 (2015)
12. Yi, W., Chen, S.: Multidimensional zero-correlation linear attacks on FOX block cipher. J. Cryptol. Res. **2**(1), 27–39 (2015)
13. Yi, W., Lu, L., Chen, S.: Integral and zero-correlation linear cryptanalysis of lightweight block cipher MIBS. J. Electron. Inf. Technol. **38**(4), 819–826 (2016)
14. Canteaut, A., Fuhr, T., Gilbert, H., Naya-Plasencia, M., Reinhard, J.-R.: Multiple differential cryptanalysis of round-reduced PRINCE. In: Cid, C., Rechberger, C. (eds.) FSE 2014. LNCS, vol. 8540, pp. 591–610. Springer, Heidelberg (2015). doi:10.1007/978-3-662-46706-0_30
15. Sakurai, S.: Prediction of sales volume based on the RFID data collected from apparel shops. Int. J. Space-Based Situated Comput. **1**, 174–182 (2011)
16. Varaprasad, G., Murthy G, S., Jose, J., D'Souza, R.J.: Design and development of efficient algorithm for mobile ad hoc networks using cache. Int. J. Space-Based and Situated Comput. **1**, 183–188 (2011)
17. Yuechuan, W., Yisheng, R., Xu An, W.: Security analysis of cipher ICEBERG against bit-pattern based integral attack. Int. J. Technol. Hum. Interact. (IJTHI) **12**, 60–71 (2016)
18. Xiuguang, L., Yuanyuan, H., Ben, N., Kai, Y., Hui, L.: An exact and efficient privacy-preserving spatiotemporal matching in mobile social networks. Int. J. Technol. Hum. Interact. (IJTHI) **12**, 36–47 (2016)
19. Ivaylo, A., Anastas, N., Evelina, P., Rozalina, D., Martin, I.: An approach to data annotation for internet of things. Int. J. Inf. Technol. Web Eng. (IJITWE) **10**, 1–19 (2015)
20. Seghir, N.B., Kazar, O., Khaled, R.: A decentralized framework for semantic web services discovery using mobile agent. Int. J. Inf. Technol. Web Eng. (IJITWE) **10**, 20–43 (2015)
21. Barenghi, A., Gerardo, P., Federico, T.: Secure and efficient design of software block cipher implementations on microcontrollers. Int. J. Grid Utility Comput. **4**, 119–127 (2013)
22. Itishree, B., Chita, R.T.: Performance modelling and analysis of mobile grid computing systems. Int. J. Grid Utility Comput. **5**, 21–32 (2014)

Design and Implementation of Simulated DME/P Beaconing System Based on FPGA

Ye Sun[1,2](✉), Jingyi Zhang[2], Xin Xiang[2], and Kechu Yi[1]

[1] State Key Laboratory of Integrated Service Networks,
Xidian University, Xi'an 10071, China
[2] Institute of Aeronautics and Astronautics Engineering,
Air Force Engineering University, Xi'an 710038, China
ye-sun@163.com, Zhangiy_1991@163.com,
xxisdn2002@sina.com

Abstract. Distance Measuring Equipment/Precision (DME/P) provides precisely distance for the approaching plane. The Simulated DME/P Beaconing System is designed for testing the performances index of airborne facility and researching interference of white noise and pulse in laboratory environment. Particular design scheme is raised,and implementation based on FPGA is discussed in this paper. The simulation results indicate that this system can satisfy for testing requirement and also quantify the influence of interference.

1 Introduction

DME/P, compatible with DME/N, employs a two-way pulse ranging system. That is, airborne facilities measure the elapsed time t between interrogation pulse and transponder pulse, and further calculate distance for the approaching plane by formulation $R = C(t-T_0)/2$ (among which, C stands for speed of light and T_0 stands for fixed elapse). It is required by the International Civil Aviation Organization (ICAO) that in the mode of initial approach (IA, 41 km–13 km), path following error (PFE) shall be limited within ±30 m and control motion noise (CMN) within ±15 m, while in the mode of final approach (FA, 13 km–0 km), PFE shall be limited within ±15 m and CMN within ±10 m [1, 2]. Due to high requirement of DME/P system accuracy and high sensibility to interference of noise and pulse, more difficulties exist for DME/P to test performances index of airborne facility compared with DME/N.

Simulated DME/P beaconing system is designed for testing performances of the following index of airborne DME/P facilities [3, 4].

(a) Test and calibration of distance accuracy
Complete test and calibration of distance measuring accuracy of airborne DME/P, with the requirement of FPE within ±15 m and CMN within ±10 m in FA mode.

(b) Test of tracking velocity
Complete test of distance tracking rate of airborne DME/P with the accuracy requirement of ±150 m/s in FA and IA mode, and no less than ±2 km/s in DME/N mode.

(c) Test of RF index

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_23

Receive sensitivity: Where, in IA mode, the receive sensitivity shall be better than −81 dBm, in FA mode better than −71 dBm, and DME/N mode better than −89 dBm.

Transmission power: Transmission power shall be no less than 500 w for full waveband.

In order to research the anti-interference performance of airborne DME/P, the simulated DME/P beaconing system shall also have abilities as follows.

(d) Impact of White Gaussian noise on measuring performances

Noise simulation ranges in 0 A–100% A, among which A refers to peak intensity of receiving pulse signals.

(e) Impact of simulated asynchronous and synchronous interference on performances of airborne DME.

Asynchronous interference can simulate interference from other DME with the same type by applying spurious random interference pulse in the same format with receiving signal and the applied pulse number is 800–9999 pairs per second. Synchronous interference can simulate multi-path interference by applying the multi-path pulse with the min elapse 0.01 µs and the amplitude attenuation 0–10 dB, and the applied pulse-pair number is no less than 5 pairs.

2 Overall Scheme of Simulated DME/P Beaconing System

Considering system function extensibility, modularized design is applied in simulated DME/P beaconing system, as shown in Fig. 1, which consists of three parts of host computer, signal processing unit and RF front-end [5].

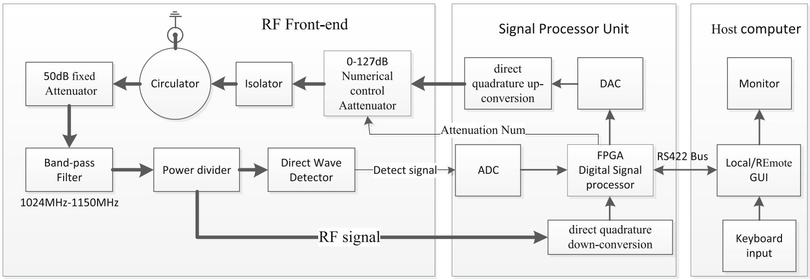


Fig. 1. Overall scheme of simulated DME/P beaconing system

2.1 Host Computer

It mainly carries out logic control, data display as well as such tasks with lower real-time requirements as auto-test process control, data processing and document editing. Industrial Personal Computer, PC104, or portable laptops may be up to requirements of the platform.

2.2 RF Front-End

This unit is usually customized in accordance with special requirements of certain system. This system apply RF front-end design as shown in Fig. 1, which consists of transmission path and receive path. In transmission path, signal is transmitted through numerical control attenuator, isolator and circulator, while signal in receiving path is divided into two pathways after treatment by circulator, fixed attenuator and band-pass filter, one for direct detector and the other for direct quadrature down-converter. Selection of the above-mentioned units shall base on comprehensive consideration of parameters of the whole system, such as frequency band, device power tolerance, signal transmission direction, accuracy level and redundancy.

2.3 Signal Processing Unit

It mainly carries out RF signal Processing, simulation signal acquisition, digital baseband (intermediate frequency) signal processing, and digital synthesis of analog waveform. The module basic architecture employs FPGA as its core and with peripheral baseband (intermediate frequency) signal processing board consisting of DAC and ADC and direct quadrature up/down conversion module. The schematic is as shown in Fig. 2.

The architecture features in strong adaptability and high flexibility so that different RF system such as communication, navigation or radar may be realized with certain baseband signal processing software and RF chip for proper frequency band.

2.3.1 RF Receiving

RF signal receiving employs AD8347 direct quadrature down-conversion chipset (work at 800 MHz–2700 MHz). The signal is mixed to zero intermediate frequency with the signal output by local oscillator ADF4351 (35 MHz–4400 MHz). Then the voltage is controlled by auto-gain with digital analog conversion chipset AD5621 so as to perform gain regulating (−30 dB–+39.5 dB) of the received signal. Output of AD8347 is acquired as digital signal by AD9218 dual 10-bit dual-channel high speed (105 MHz) ADC.

2.3.2 RF Transmission

In transmission, with digital analog conversion by AD9763 dual-channel 10-bit high-speed (125 MHz) DAC chipset, the simulation signal is modulated to RF spectrum (100 MHz–2400 MHz) through direct quadrature up-conversion by ADRF6755. And then it goes through filter network and numerical control attenuator to the antenna port (RF connector).

2.3.3 FPGA Signal Processing [6]

(a) Signal processing of distance measuring accuracy

Receiving signal acquired by quadrature sampling and input to FPGA is tested for pulseamplitude. Since receiving signal is the narrow pulse signal, elapse comparison is

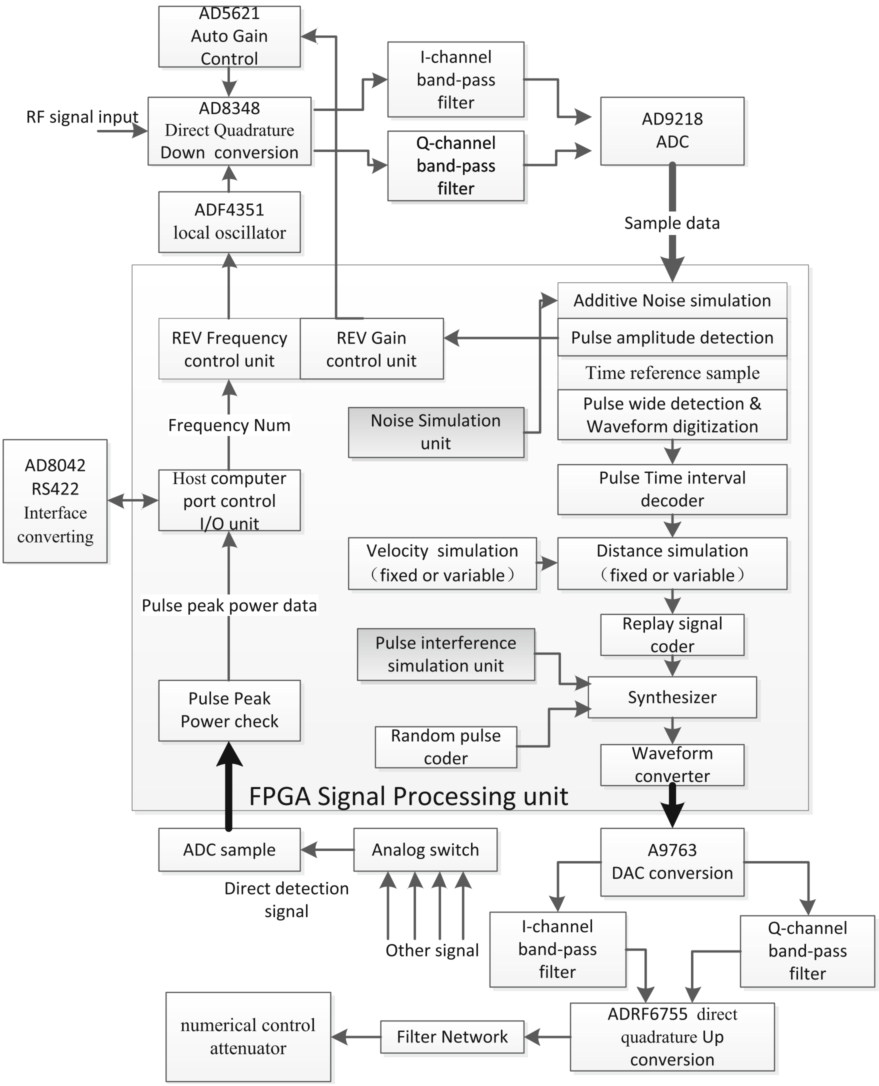


Fig. 2. Schematic diagram of signal processing unit

applied to determine the peak value of pulse signal, and the signal is used to control AGC circuit gain so that the waveform of sampling pulse keeps in rational range of amplitude. Then determine the time reference with half-amplitude power detection [7] or DAC (delay and attenuate compare) technique according to interrogation pulse waveform, and transform the quasi-Gaussian pulse or cos-cos^2 pulse form into square-wave pulse in order to pulse interval test, which is decoding. Refer to Bibliography [8] for decoding methods.

The triggering signal after decoding is treated by the fixed distance elapse and encoded into transponder signal. Finally, the signal is transformed into quasi-Gaussian or *cos-cos*^{2} pulse digital waveform and is sent to DAC convertor for analog signal. The elapse time *t* between transponder signal and interrogation signal is the corresponding simulation of distance of the airborne facility.

(b) Signal processing of distance tracking velocity

Signal processing of distance tracking velocity, based on the above-mentioned signal processing of distance measuring precision, is realized by manual control of change rate of elapse time between interrogation and transponder signals. When the change rate of elapse time is a constant value, it simulates a plane in uniform motion. Also a plane at a certain acceleration can be simulated.

(c) Test of RF index

Test requirements of RF index of airborne facility includes test of pulse peak power and receive sensitivity.

Test of pulse peak power

The pulse peak power is the key index to measure the generating performance of the airborne facility. To ensure the accuracy, the test end shall as close to the receiving RF port as possible. Therefore, the design applies sampling and quantifying of direct detecting signal. It is important to note that because the direct detector is connected after power divider, in order to increase the accuracy, an overall calibration shall be carried out to the fixed attenuator, pass-band filter, power divider and detector, and compensations shall be made each frequency band.

Test of receive sensitivity

The receive sensitivity is the key index to measure the receive performance of the airborne facility. The test method is to have the RF signal attenuate with numerical control attenuator (0–127 dB) until the airborne facility cannot measure the distance or the distance measured is out of tolerance. Strength of the radiation signal at this moment is the receive sensitivity. As in the above case, in order to make the test more accurate, an overall calibration shall be carried out to pass-band fluctuation of RF chipset output, filter network, numerical control attenuator and the circulator, and compensations shall be made each frequency band.

(d) Simulation of impact white Gaussian noise put on distance measuring performance

In order to enhance the understanding of distance measuring system, the noise simulation module is added to FPGA, which employs pseudo random sequence generator to generate pseudo random noise with controllable intensity and to superpose with the received signal. The impact white Gaussian noise put on distance measuring performance researched in this way.

(e) Simulation of synchronous and asynchronous interference

Pulse interference is one of the major interference sources of distance measuring system, which can be classified into synchronous interference and asynchronous interference in terms of consistency with repetition frequency of the interrogation signal.

Asynchronous interference, of which the repetition frequency is irrelevant to the interrogation signal, can be generated by random pulse encoder. Synchronous interference, of which the repetition frequency is in consistency with the interrogation signal but the elapse time and signal amplitude is different with the local transponder signal, can be regarded as multi-path interference. Simulation of this part is not carried out in traditional beacon imitation and the module here is designed to enhance understanding of synchronous interference.

3 Implementation and Simulation Result of Baseband Signal Processing

3.1 Pulse Amplitude Test and Time Reference Sampling

Pulse amplitude test is designed to perform real-time test of the peak value of receive pulse, which is realized by elapse comparison circuit. Simulated waveform in Fig. 3 indicates that data in the peak value test is the real-time peak value of each pulse, which will be taken as the reference amplitude value in time reference sampling of half-amplitude power and as time reference in DME/N signal. For DME/P signal, DAC technique is applicable and the equal point of delayed pulse and attenuated pulse can be taken as time reference sampling. The pulse peak value is transformed into AGC control signal after latch treatment to ensure the integration of the waveform of original receive pulse. In Fig. 3, due to noise interference, a series of narrow pulse interference appears in the pulse front-end of DAC time sampling and half-amplitude power sampling. As a result, in post circuit treatment, it is proposed to take into consideration filtering removal of the interference and the error in distance measuring introduced by noises.

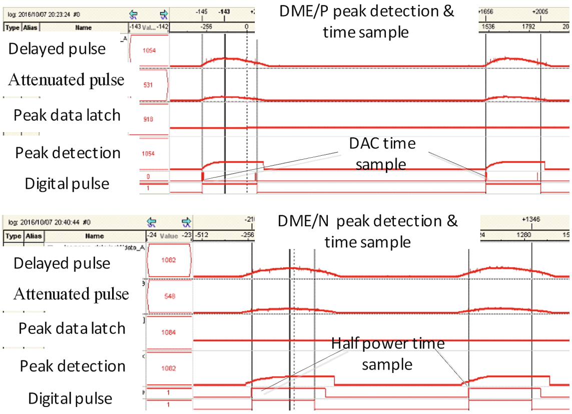


Fig. 3. Simulated amplitude test and time reference sample

3.2 Simulation on Fixed Distance and Fixed Rate

According to the principle of DME/P, the time delay and distance do exist between the interrogation pulse and the transponder pulse. In the formula $R = C(t{-}T_0)/2$, C refers to speed of light while T_0 refers to the fixed elapse required by the system [9].

The distance simulation module delays the generating time of transponder signal for a fixed time t = constant according to the distance value that the user inputs in the simulation driver, and then the fixed distance R can be simulated. Figure 4 presents the simulated waveform of the distance setting at 1030 m. In the figure, with the sample clock 100 MHz and the corresponding delay 1.5 m, we get the distant delay R = 686*1.5 = 1029 m. Delay between encoding pulse and triggering pulse is the system delay T_0, which should be 50 µs, but it is adjusted to 10 µs for facilitating observation.

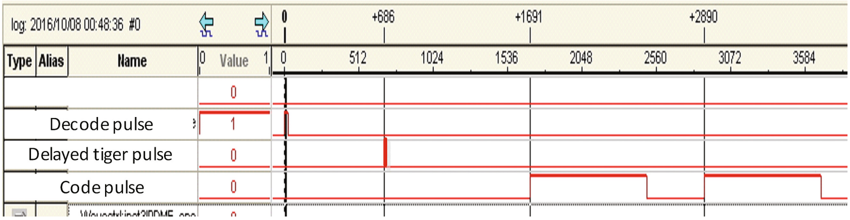


Fig. 4. Simulated waveform of fixed distance delay

For simulation of plane in uniform motion, use the counter to calculate the frequency needed according to velocity input, change the distance delay periodically and then the time changes according to $f(t) = at_{clk} + B(constant)$, where a may be negative.

For simulation of acceleration, similar to the above-mentioned method, only change the formula as $f(t) = \beta t_{clk}^2 + B(constant)$ and the delay data will obtained by square operation of output from the counter.

3.3 Simulation of Pulse Interference and Random Pulse Encoder

The pulse interference herein is simulation mainly of multi-path interference and synchronous interference. The trigger pulse is generated in accordance with the interrogation pulse received and then the encoder generates interference in conformance with multi-path interference and synchronous interference.

Figure 5 shows the synchronous interference of two groups of delay under different attenuations. According to the figure, the major impact of synchronous interference lies in signal distortion. the synchronous interference simulation module works effectively.

Random pulse encoder simulates impact that asynchronous interference has on DME/P. Firstly, pseudo-random encoder generates the delay data. And then add the asynchronous interference to the delay unit according to the number needed so as to form the random pulse of which the format is in conformance with the transponder pulse. It is important to note that the number of random pulse is 800 pairs/s in

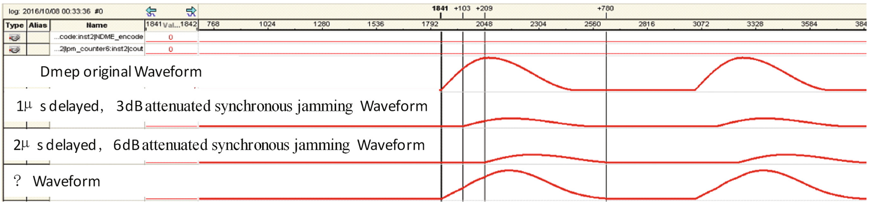

log: 2016/10/08 00:33:36 #0
Type
Alias
Name
1841 +103 +209
+780
...code:inst2|NDME_encode
...2|lpm_counter6:inst2|cout
Dmep original Waveform
1μ s delayed, 3dB attenuated synchronous jamming Waveform
2μ s delayed, 6dB attenuated synchronous jamming Waveform
? Waveform

Fig. 5. Simulated waveform of synchronous interference

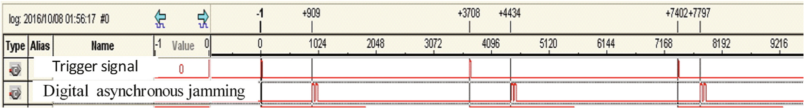


Fig. 6. Simulated wave form of asynchronous interference

minimum and zero is not allowed, or else AGC cannot be formed for the airborne facility. In Fig. 6, the simulation waveform shows that delay of the encoding pulse is note equal to that of the trigger pulse in each cycle, which is up to requirement of asynchronous interference simulation.

3.4 Noise Simulation

In pulse distance measuring system, noise leads to pulse distortion and then impact selection of time reference sample. As a result, noise simulation module mainly simulates the impact that poor Signal to Noise Ratio (SNR) brings to pulse distance measuring. The design applies pseudo random encoder to generate White Gaussian Noise, adjust its amplitude according to SNR, and sends the superposition of the noise and the interrogation pulse received to pulse amplitude test and time reference sample. The same principle may put the module to the sending end of simulation driver so as to simulate the anti-interference performance of airborne receiver under the simulation of White Gaussian Noise channel. In Fig. 7, it simulates the impact that noise put on receiving and transmitting signals at SNR 10 dB.

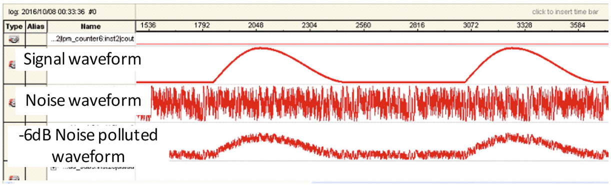


Fig. 7. Waveform simulation of noise polluted signal

Table 1. Test result of distance measurement accuracy of airborne facility with simulated beaconing system

Test item	DME/P setting 2000 (m)	DME/N setting 30000 (m)	Conclusion
No addition of noise or interference, with smoothing	2000 ± 5	30000 ± 18	Up to test accuracy requirement with no noise addition
SNR 10 dB at receiving end, without smoothing	2000 ± 60	30000 ± 125	Without smoothing, the noise at the receiver impacted the accuracy of time reference sample, and the accuracy decreased
−10 dB noise added at the receiving end, with smoothing	2000 ± 10	30000 ± 35	The measuring accuracy improved with smoothing, but it still decreased with the level of noise
−10 dB noise added to the transmission end	2000 ± 8	30000 ± 25	Smoothing is designed at the airborne facility, which decreased the impact of noise interference
Synchronous jamming of 1 μs delay, without noise.	2000 ± 5	30000 ± 12	Impact of multi-path interference varied from different delay condition. The comparison test indicates that the sampling of airborne DME/P lies around 0.1 μs. Therefore, multi-path over 0.3 μs doesn't have much impact
Synchronous jamming of 0.1 μs delay, without noise	2000 ± 12	30000 ± 28	
With 800–9000 pairs of asynchronous jamming, without noise	2000 ± 5 m	30000 ± 18 m	The airborne facility employs tracking test of local pulse, thus the asynchronous jamming didn't impact the accuracy of measurement basically

4 Conclusion

Apart from quantifying the performance index of airborne facility, the other major function of the simulated system is to carry out simulation analysis on factors impacting precision of distance measurement and the according results in terms of hardware. In another, the impact of building existing in channel is considered. [10, 11] The simulated test result is as shown in Table 1.

(a) Range resolution up to 1.5 m with error ±5 m, with no manual additive noise and synchronous interference, up to calibration accuracy requirement of airborne facility.

(b) Velocity test range −3 km/s–3 km/s, up to requirement of airborne tracking rate range ±2 km. Speed resolution 1.5 m/s, and add simulated test capability of acceleration compared with traditional imitation devices.
(c) It is capable to implement test of related RF parameters such as receive sensitivity and pulse peak power of generator.
(d) Simulation analysis capability of synchronous jamming and asynchronous interference is added compared with traditional beaconing imitation devices.
(e) Simulation analysis capability of noise impact is added, compared with traditional beaconing imitation devices.

References

1. Liu, J.: Research on digital DME/P technology. Mod. Electron. Tech. **35**, 161–163 (2012)
2. Zhu, Y.: Research on DME/P technology. Navigation **4**, 73–77 (2002)
3. ICAO: International Standards and Recommended Practices Annex 10 Aeronautical Telecommunications. ICAO (1992)
4. Technology and Industry for National Defense: GJB 914-90. Signal Standard and Test Method of TACAN, Beijing (1991)
5. Ye, H.: Research and development of digital up/down-convertor in software radio. Thesis of master, Guilin University of Electronic Technology (2010)
6. Wang, D.: The Research and implementation of multimode signal generat based on FPGA. Master degree thesis, Graduate School of National University of Defense Technology (2010)
7. Wang, D., Zhang, X., Luan, B.: Analysis of effect of half-amplitude detection circuit on measuring accuracy in TACAN beacon. Modern Navigation **2**, 117–1 (2014)
8. Sun, Y., Yi, K., Xiang, X. (eds.): The effect and elimination of multipath interference in pulse distance measuring system. In: Microelectronic Technology. Application of Electronic Technique, vol. 42, pp. 28–31 (2016)
9. Li, K., Pelgrum, W.: Optimal time-of-arrival estimation for enhanced DME. In: Proceedings of the 24th International Technical Meeting of the Satellite Division of the Institute of Navigation, ION GNSS 2011, Portland, pp. 3493–3502 (2011)
10. Iwashige, J., Barolli, L., Hayashi, K.: Analysis and characteristics of diffracted waves going over building rooftop. Int. J. Space-Based Situated Comput. **1**, 197–203 (2011). doi:10.1504/IJSSC
11. Sakamoto, S., Kulla, E., Oda, T., Ikeda, M., Barolli, L., Xhafa, F.: A simulation system for WMN based on SA: performance evaluation for different instances and starting temperature values. Int. J. Space-Based Situated Comput. **4**, 217–232 (2014). doi:10.1504/IJSSC

LDPC Codes Estimation Model of Decoding Parameter and Realization

Li Xiaobei[1(✉)], Zhang Jingyi[2], Wang Rui[2], and Zou Yazhou[2]

[1] Institute of Information and Navigation,
Air Force Engineering University, Xi'an 10077, China
624702291@qq.com

[2] Institute of Aeronautics and Astronautics Engineering,
Air Force Engineering University, Xi'an 10038, China
zhangjy_1991@163.com, 15353724115@qq.com,
m18789448338@163.com

Abstract. It is necessary to use the channel parameters to initialize the iterative decoding in sum-product decoding algorithm of LDPC codes. The channel parameters of the receiver are unknown, so it is essential to estimate the channel parameters or the signal-to-noise ratio (SNR). In this paper, two estimation models are established by the characteristics of the Gaussian channel, and their precision are analyzed; then we use the moment estimation, maximum likelihood estimation and Bayesian estimation respectively to estimate the models mentioned above; finally use the estimated signal-to-noise ratio to decode the LDPC codes by sum-product decoding. Simulation results show that: the result of the estimation model 2 is more accurate than that of the estimation model 1. Using the result of the estimation model 2 to decode the LDPC codes by sum-product decoding, the curve of the decoding results and that of using true value to decode are almost overlapped.

1 Introduction

Low density parity check (LDPC) codes have the excellent properties of approaching Shannon limit [1, 2]. Their studies on practical application have important significance. It is necessary to use the channel parameters σ^2 [3] to initialize the iterative decoding [4] in sum-product decoding algorithm of LDPC codes. And the channel parameters σ^2 can also be presented as SNR (the signal to noise ratio) information. While the channel parameters on the receiver are unknown, it is essential to estimate the channel parameters or the signal-to-noise ratio (SNR). According to the knowledge of statistics, we can get the conclusion that if we use the received random signal to estimate the channel parameters directly, the error of estimation is usually large, which can affect the iterative decoding convergence speed and performance of LDPC codes obviously [5]. In order to improve the accuracy of estimation and enhance the decoding accuracy, two estimation models are established in this paper. And a proper estimation model is obtained by analyzing and simulating the established estimation models

Section 2 establishes two estimation models; Sect. 3 the estimated accuracy of two established models is analyzed in detail; Sect. 4 deduces the formulas of moment

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_24

estimation, maximum likelihood estimation and Bayesian estimation. Section 5 simulates and analyses two established models; and the final section is the conclusion.

2 Estimation Model

The output bit sequence of the channel encoder is inputted the bipolar input channel after the bipolar mapping. It is assumed that a bit is $u \in \{0, 1\}$, and the bipolar mapping is $x = 1 - 2u$, we can get $x = \{+1, -1\}$ after the bipolar mapping [6].

In the smooth memoryless additive white Gaussian noise (AWGN) channel, the transmission signal $a \in \{+1, -1\}$ is the mapping of the transmitted bit under BPSK modulation. Since the number of the codeword "+1"and that of the codeword "−1" are substantially equal in the actual communication system, we can assume that the appearance probability of codeword "+1" and that of codeword "−1" are 1/2 respectively. Suppose that the information emitted by the source is X and the initial messages received by the receiver is Y, the quantized output is Z. Then the likelihood function is:

$$f(y|x = a) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left\{ -\frac{(y-a)^2}{2\sigma_n^2} \right\} \tag{1}$$

where $a \in \{+1, -1\}$, and σ^2 is the variance of Gaussian white noise. Based on the aforementioned hypothesis, we can get that the probability that a is "+1" or "−1" is equal. Thus

$$\begin{aligned} f(y) &= \sum_a f(a) \times f(y|a) \\ &= \frac{1}{2\sqrt{2\pi}\sigma_n} \left\{ \exp\left\{ -\frac{(y-1)^2}{2\sigma_n^2} \right\} + \exp\left\{ -\frac{(y+1)^2}{2\sigma_n^2} \right\} \right\} \end{aligned} \tag{2}$$

Based on the above equation, we can know that after the signal is transmitted in the AWGN channel, the initial information y_j received by the receiver obeys weighted superimposed distribution of two normal distributions, whose mathematical expectations are "+1" and "−1", respectively, as shown in Fig. 1.

Traditional SNR estimation algorithm is carried out based on the distribution of direct estimation of the initial information [3]. When the moment estimation is adopted directly, we can obtain the equations as follows:

$$\begin{cases} E(X) = 0 \\ E(X^2) = \mathrm{Var}(X) + (E(X))^2 = \sigma^2 \end{cases} \tag{3}$$

Thus

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \tag{4}$$

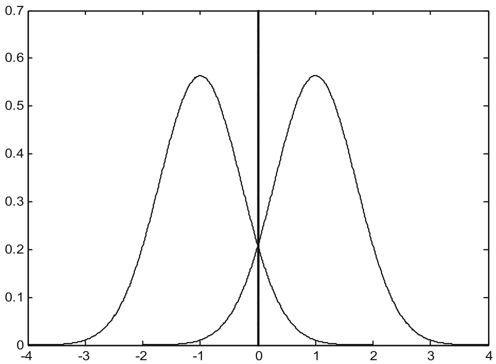


Fig. 1. Initial information distribution.

It is clear that variance estimated through the Eq. (4) is not the variance estimated through the Eq. (2) but the overall variance. Thus the error of the estimated variances through between Eqs. (4) and (2) is proportionable large. For example, when SNR is 1 dB, the estimated SNR is −2.5887 dB. In other words, when the variance σ^2 in Eq. (2) is 0.7943 dB, we can obtain that the estimated variance is 1.815 dB. Based on the statistical properties of output soft information after iterative decoding, the joint timing synchronization and SNR estimation algorithm which is assisted by the symbol soft information of iterative decoding can more accurately estimate the received signal or SNR through several iterations. However, when the initial value is not accurate, this algorithm requires more iteration, which will consume more resources, and make the precision of estimation results reduce possibly.

Based on the probability theory and mathematical statistics knowledge, we can get that when the distribution of the estimation object is known, it is best to adopt the maximum likelihood estimation and Bayesian estimation to estimate its parameters. However, adopting the maximum likelihood estimation and Bayesian estimation to estimate the variance σ^2 will lead to the situation that the formula is too complex and difficult to obtain the analytical solution. Therefore, we have to ignore some factors which less affect the estimation results and establish an ideal channel estimation model to achieve a higher estimation accuracy by a simplified estimation formula.

Since the curve graph of formula (2) is weighted superimposed by two Gaussian probability density curves, we assume that Gaussian probability density curve on one side less affects that curve on the other side. And because the mean is known, we can use the statistics sample X to estimate the variance σ^2. Thus, we can obtain two estimation models: based on Fig. 1, when the mean of 1 or −1 is known, we can use the statistics sample X when $x > 0$ or $x < 0$ to estimate the variance σ^2 of formula (2) in the model 1; Based on Fig. 1, when the mean of 1 or −1 is known, we can use the statistics sample X when $x > 1$ or $x < -1$ to estimate the variance σ^2 of formula (2) in the model 1. This shows that: the above model is suitable for the situation using the BPSK modulation and the AWGN channel.

3 Precision Analysis of Model

For Model 1, we only employ the sample X when $x > 0$ or $x < 0$ to estimate the variance σ^2 of formula (2). And the Gaussian distribution on the other side has a certain influence on the estimation result. Table 1 shows the impact situations that the Gaussian distribution on the left side affect the estimation result by employing the sample X when $x > 0$, where the relative impact is the ratio between the left Gaussian distribution and the right Gaussian distribution. Similarly, we can get the impact situations when $x < 0$.

Table 1. The probability of Gaussian distribution on the left side when x > 0

SNR	0 dB	1 dB	2 dB	3 dB	4 dB
σ^2	1	0.7943	0.631	0.5012	0.3981
Redundancy	0.1587	0.1309	0.1040	0.0789	0.0565
Relative impact	0.1886	0.1506	0.1161	0.0857	0.0599

As shown in Table 1, when the SNR is smaller, the left Gaussian distribution has a greater impact on the estimation result; when the SNR is larger, the left Gaussian distribution has a smaller impact on the estimation result.

For Model 1, we only employ the sample X when $x > 1$ or $x < -1$ to estimate the variance σ^2 of formula (2). Table 2 shows the impact situations that the Gaussian distribution on the left side affect the estimation result by employing the sample X when $x > 1$, where the relative impact is the ratio between the left Gaussian distribution and the right Gaussian distribution. Similarly, we can get the impact situations when $x < -1$. Because the probability of the right Gaussian distribution is 0.5, the relative impact is twice as much as the redundancy.

Table 2. The probability of Gaussian distribution on the left side when x > 1

SNR	0 dB	1 dB	2 dB	3 dB	4 dB
σ^2	1	0.7943	0.631	0.5012	0.3981
Redundancy	0.0228	0.0124	0.0059	0.0024	0.00076
Relative impact	0.0456	0.0248	0.0118	0.0048	0.0015

As shown in Table 2, when the SNR is smaller, the left Gaussian distribution has a smaller impact on the estimation result; and the estimation accuracy increases with the SNR becoming higher. Compared with the model 1, the estimation accuracy using model 1 is lower than that of model 2. The estimation accuracy using model 2 when the SNR is 0 dB is better than that of model 1 when the SNR is 4 dB. Besides, when the SNR is greater than 4 dB, the relative impact of model 2 is much less than that of model 1. Based on the Sect. 4, although the statistical data between the Model 2 and Model 1 is different, the estimation formulas which use these statistical data are same. So at the same statistical data, the computational complexities of the two models are the same. Besides, both of them are lower than the computational complexity of the algorithm in [3].

In general, the estimation precision that is estimated by using the model 2 can meet with the most applications. When we need higher estimation accuracy, the joint timing synchronization and SNR estimation algorithm [7] can more accurately estimate the received signal or SNR through several iterations, which are assisted by the symbol soft information of iterative decoding and are based on the statistical properties of output soft information after iterative decoding.

4 Estimation Method

Both these two estimation models are used to estimate variance when the mean is known, and the difference between two estimation models is only the ranges of the statistical samples. Thus, we only derive the estimation methods of variance to obtain the estimated SNR when the mean is known.

4.1 Moment Estimation and Maximum Likelihood Estimation

In this section, the estimation method is derived when the mean value is known. We assume that $X_1, X_2, \cdots, X_{\mathrm{n}}$ is a simple random sample of X, and its mean value is 1. Then we can get

$$\begin{cases} E(X) = 1 \\ E(X - E(X))^2 = \sigma^2 \end{cases} \tag{5}$$

Since the mean value is 1, we can get

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - 1)^2 = \sigma^2 \tag{6}$$

Thus the moment estimator of σ^2 can be obtained as follow

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - 1)^2 \tag{7}$$

Maximum likelihood estimation is a general important estimation method. This estimation method has many excellent properties, so when the type of overall distribution is known, it is the best to employ the maximum likelihood estimation method to estimate the unknown parameters of the overall distribution.

We assume that $X_1, X_2, \cdots, X_{\mathrm{n}}$ is a simple random sample of X, whose mean value is 1, and observed values of its sample are $(x_1, x_2, \cdots, x_n)$, then we can get

$$L(\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \tag{8}$$

$$\ln L(\theta) = -n \ln \sqrt{2\pi} - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{9}$$

Therefore the likelihood equation can be obtained as follow

$$\left.\frac{\partial \ln L(\theta)}{\partial \sigma^2}\right|_{\substack{\mu = 1 \\ \sigma^2 = \hat{\sigma}^2}} = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i - 1)^2 = 0 \tag{10}$$

The solution can be obtained as follow

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - 1)^2 \tag{11}$$

So its likelihood estimator is equal to the moment estimator.

4.2 Bayesian Estimation

Based on the previous research, we can find that the Bayesian estimation can achieve a better estimation effect by employing the priori information of the parameters.

Based on the estimation model in Sect. 2, we can get that $X_1, \cdots, X_N$ are independent and identically distributed, which obey the normal distribution. Each of their mean values is 1, and each of their variance is σ^2. We set $Y_i = X_i - 1\ (i = 1, \cdots, N)$. Then we can get that $Y_1, \cdots, Y_N$ are also independent and identically distributed, which obey the normal distribution. Each of their mean values is 1, and each of their variance is σ^2. Thus we can get that the joint distribution density of Y_i is $C\tau^r \exp(-\tau \sum_{i=1,\cdots,N} Y_i^2)$, where $r = n/2, \tau = 1/(2\sigma^2)$, the prior distribution of the conjugate τ is $\Gamma(g, 1/\alpha)$ [8], thus

$$\begin{array}{cc} E(\tau) = g/\alpha; & E(\tau^2) = g(g+1)/\alpha^2; \\ E(1/\tau) = \frac{\alpha}{g-1}, g > 1; & E(1/\tau^2) = \frac{\alpha^2}{(g-1)(g-2)}, g > 2. \end{array} \tag{12}$$

We set $z = \sum_{i=1,\cdots,\mathrm{N}} y_i^2$, then the posterior density of τ is $C(y)\tau^{r+g-1}e^{-\tau(\alpha+y)}$ when the y_i is given, which obeys the distributed of $\Gamma(r+g, 1/(\alpha\ +\ \mathrm{y}))$. Supposed that the loss is the squared error, the Bayesian estimation of $2\sigma^2 = 1/\tau$ is the posterior mean of $1/\tau$. We can get the Bayesian estimation of $\sigma^2 = 1/(2\tau)$ from Eq. (12) as follow

$$(\alpha + Y)/(n + 2g - 2),\ n + 2g > 2. \tag{13}$$

Because the information content of σ is proportional to $1/\sigma^2$ [8], the Jeffreys prior density should be proportional to non-genuine density of $1/\sigma$ in the above example. This density can derive that the density of τ is $1/\tau d\tau$, which equals to the limiting case that $\alpha = 0$ and $g = 0$.

Thus, by the formula (13), we can obtain that the Bayesian estimation of σ^2 can be $Y/(n+2)$ under the situation of the square error loss.

5 Performance Simulation and Analysis

In this section, we simulate the performance of LDPC under a binary phase-shifting keying (BPSK) modulated in an Additive white Gaussian noise (AWNG) channel. The rate of the adopted LDPC is 1/2, and its code length is 2048. The LLR decoding algorithm is used to decode. The maximum number of decoding iterations is set to 50. We get 10 million packets statistics at each SNR, Figs. 2 and 3 are both the BER simulation results.

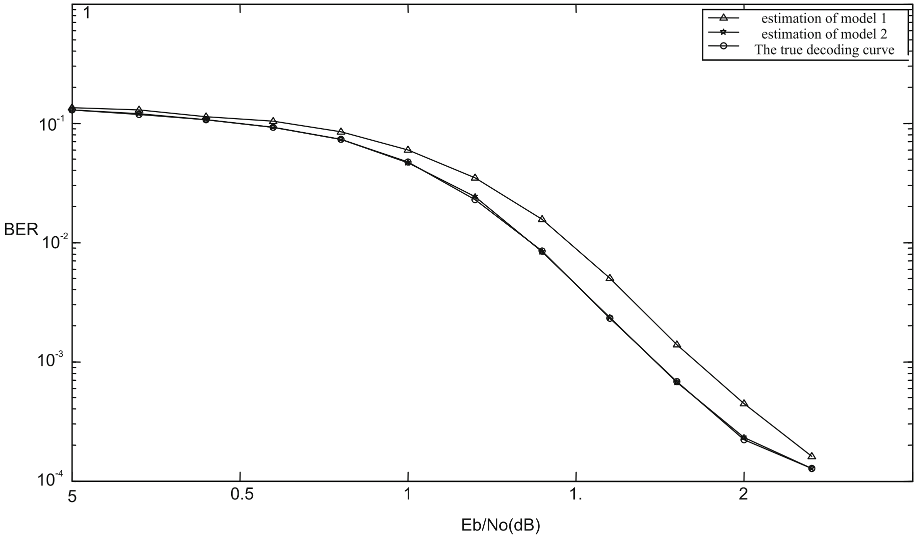


Fig. 2. The performance curve of moment estimation and maximum likelihood estimation

Figures 2 and 3 show the performance curves of different estimation methods. As shown in Figs. 2 and 3, the decoding result curves by using the model 2 to estimate the channel parameters can coincide with the true decoding curve; while the decoding result curves by using the model 1 to estimate the channel parameters have a certain gap with true decoding curve. Besides, using three estimation methods to estimate the channel parameters can get essentially the same decoding decodes curve under the different models.

As shown in Table 3, we can obtain that the estimation error by using model 1 is quite large. With the SNR increasing, the estimation error decreases. It can be obtained from Table 4 that the estimation results by using model 2 are basically equal to the true SNR. The mean square error is 0.0055 by using moment estimation or maximum likelihood estimation. While the mean square error is 0.0050 by using Bayesian estimation. Thus, we can conclude that the performance of using Bayesian estimation is the best.

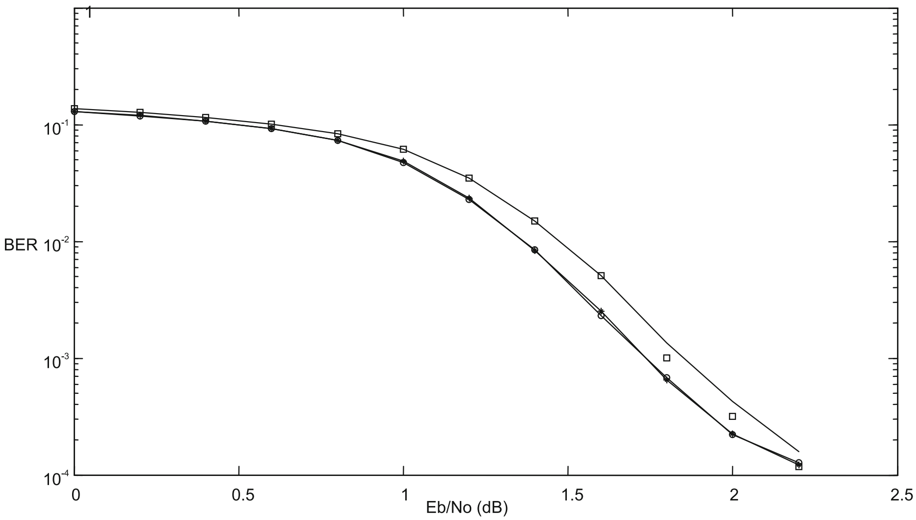


Fig. 3. The performance curve of Bayesian estimation.

Table 3. The different SNR estimation results by using model 1 (unit: dB)

Estimation method	SNR (dB)											
	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2
Moment estimation (maximum likelihood estimation)	1.7295	1.9039	2.0770	2.2482	2.4016	2.5299	2.6209	2.7833	2.9538	3.1080	3.2059	3.2974
Bayesian estimation	1.7286	1.9030	2.0761	2.2473	2.4008	2.5290	2.6200	1.7286	1.9030	2.0761	2.2473	2.4008

Table 4. The different SNR estimation results by using model 2 (unit: dB)

Estimation method	SNR (dB)											
	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2
Moment estimation (maximum likelihood estimation)	0.0268	0.2010	0.4442	0.6058	0.8287	1.0153	1.2215	1.3896	1.6246	1.8209	1.9893	2.2046
Bayesian estimation	0.0252	0.1994	0.4426	0.6042	0.8270	1.0137	1.2198	1.3879	1.6229	1.8192	1.9876	2.2030

6 Conclusion

In this paper, two simple and practical estimation models of channel parameter σ^2 are established based on the characteristics of Gaussian channel. Based on the specific characteristics of the established models, the estimation expressions of moment estimation, maximum likelihood estimation and Bayesian estimation are derived under the qualification situation of the models. Then the channel parameter σ^2 is estimated by using the moment estimation method, maximum likelihood estimation method and Bayesian estimation method. Since the Gaussian distribution on the other side of the model 1 has great influence on estimation, its estimation performance is poor. While the Gaussian distribution on the other side of the model 2 has little influence on estimation, its estimation performance can achieve a good effect. Besides, the estimation performance by using moment estimation, maximum likelihood estimation and Bayesian estimation has made a good estimation effect. With the SNR increasing, the estimation error decreases generally. The estimation performance of Bayesian estimation is better than those of both moment estimation and maximum likelihood estimation under the criterion of minimum mean square error. The sum-product decoding of LDPC is using the estimation results of the established model 1 and model 2. Simulation results show that the curve of the decoding results by using model 2 and that of using true value to decode are almost overlapped. Though the decoding results by using model 1 has a certain estimation error, it has little effect on the decoding accuracy due to the excellent decoding performance of LDPC codes. Using the proposed estimation methods has several advantages. For example, the statistical calculations are low, the calculation expressions are simple, and the estimation results are very accurate. Thus, the proposed estimation methods have a strong practical.

References

1. Chung, S.-Y., Forney Jr., G.D., Richardson, T.J., et al.: On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit. IEEE Commun. Lett. **5**(2), 58–60 (2001)
2. Richardson, T.J., Urbanke, R.L.: The capacity of low-density parity-check codes under message-passing decoding. IEEE Trans. Inf. Theory **47**(2), 599–618 (2001)
3. Li, S., Wang, J., Ma, L.: Convergence of BP decoding for LDPC codes under SNR mismatch. Syst. Eng. Electron. **32**(3), 491–494 (2010)
4. Saeedi, H., Banihashemi, A.H.: Performance of belief propagation for decoding LDPC codes in the presence of channel estimation error. IEEE Trans. Inf. Theory **55**(1), 84–86 (2007)
5. Saeedi, H., Banihashemi, A.: A design of irregular LDPC codes for BIAWGN channels with SNR mismatch. IEEE Trans. Commun. **57**(1), 6–11 (2009)
6. Li, L., Chen, Z., Fan, J., et al.: Implementation of LDPC codes decoding based on maximum average mutual information quantization. Int. Rev. Comput. Softw. (I.RE.CO.S.) **6**(6), 1135–1139 (2011)
7. Pan, X., Liu, A., Zhang, B.: A joint timing synchronization and SNR estimation algorithm for LDPC-coded systems. J. Electron. Inf. Technol. **30**(1), 125–130 (2008)

8. Lehmann, E.L., Casella, G.: Theory of Point Estimation, 2nd edn. Springer, New York, Inc (1998)
9. Chen, Z., Zhang, H., Li, L., et al.: Application of maximum average mutual information quantization in LDPC decoding under QPSK modulation. J. Syst. Eng. Electron. **34**(3), 598–602 (2012)
10. Huang, Q., Diao, Q., Lin, S., Abdel-Ghaffar, K., et al.: Cyclic and quasi-cyclic LDPC codes on constrained parity-check matrices and their trapping sets. IEEE Trans. Inf. Theory **58**(5), 2648–2671 (2012)

A Novel Query Extension Method Based on LDA

Yongjun Zhang[1(✉)], Jialin Ma[1], Zhijian Wang[2], and Bolun Chen[1]

[1] Huaiyin Institute of Technology, Huaian and College of Computer
and Information, Hohai University, Nanjing, China
13511543380@139.com

[2] College of Computer and Information, Hohai University, Nanjing, China

Abstract. *Information retrieval* (*IR*) is a major technology helping people to retrieve the information they are interesting in. One of challenge in *IR* is that the input query consists of very few words so that *IR* can't catch the user's intention. The *pseudo correlation query extension* (*PCQE*) is a power technology in *IR* aim to solve this challenge. In this paper, we propose a *PCQE* method which based on *LDA*, it apply the LDA model to fit the document set, then the latent topics are exploited and each document is represented as a multinomial distribution over topics. We calculate the probability of the document generating the query to measure the correlation between them, then the documents are ranked in terms of the correlation and top documents are extracted to seek informative words to extend the original query. Our experiment on the *Ohsumed* data set shows our method outperforms the other state-of-art *PCQE* methods.

1 Introduction

With the explosive growth of data, the network has become a major way for people to browse information, they resort to the popular search engine tools, such as Google, Baidu, to retrieve the information they are interesting in. One of key technologies that build these search engine tools is the *Information Retrieval* (*IR*) technology. The task of *IR* is retrieving the documents relevant to the input query from a large document set, then returns the rank order of them in terms of how they are relevant to the query. When the user inputs a query, the search engine often returns too many results and many of which are not or weak related so that the results are not desirable. One of the reasons affecting the user search experience is that the user often input short queries, which makes the search engine can't understand the user's intention. Another reason is that different users often use diverse words to represent the same retrieve intention, for example, to retrieve the information about the American history, some users may input "America history", whereas other users may input "US history" since US and America are nearly equal, it may confuse the search engineer to understand the input query.

Since the user often inputs queries with a few keywords, one of research hotpots in the field of *IR* is extending the input query with additional words, which is called q*uery expansion* (*QE*). The *QE* technology extends the original query with words chosen from first retrieval top documents, it has been proved can greatly improve the retrieval results, and has become a major way to promote the accuracy and recall of search

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies* ,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_25

results. The *QE* can be divided into two kinds of models in terms of whether exploiting the user's browsing information or not: *correlation query extension* (*CQE*) and *pseudo correlation query extension* (*PCQE*). *CQE* extends the input query with the history input queries of users so it need to configure a user profile for each user to save the user's query history, which limits the application of *CQE*. While *PCQE* is an automatic analysis method and is transparent to users. it refers to the top documents among retrieving results and then choose the extension words from them. These words are added to the original query to improve the ranking result of the relevant documents. One of the main problems of *PCQE* is that the assumption of the top *k* documents are strongly related to the query is often inconsistent with the fact, especially in the case of the very short query.

Since the *Latent Dirichlet Allocation* (*LDA*) [1] is proposed by David M. Blei et al. it has been paid more and more attention. *LDA* also has applied to *IR*, which leads to another research direction. In this paper, a *PCQE* method based on *LDA* topic model is proposed, it firstly uses the *LDA* to represent each document in the topic space and calculates how each document is relevant to the input query by the probability of the document generating the query. The experiment shows that our method can improve the retrieval results.

2 Related Works

Marton firstly proposed the *QE* technique in 1960 [2], he exploits the features relevant to the original query to return more relevant literature, the author uses expansion for query. Rocchio applied the *QE* in *SMART* retrieval system [3], he used a set of documents as the feedback documents, then ranked every word in the document set in terms of *TF-IDF* weights. In the following decades, *QE* has a rapid advancement in the *IR* field. Most query expansion methods are based on the framework of the Rocchio extension technology. Most *QE* methods are based on the framework of the technology proposed by Rocchio [4, 5], they can be divided into three categories: explicit, implicit and pseudo relevance feedback extension. Carterette et al. made use of the information of users clicking documents to determine whether the document is strongly relevant to the query [6]. Rajaram et al. uses the manual judgement of users that whether the result document is relevant to the query to choose the feedback documents [7]. However, this method requires the participation of users, in most cases, the user's feedback on the search results of the document is less, which limits the application of user feedback to expand the query word.

The *QE* technology based on feedback has experienced a period of vigorous development, in which *PCQE* has been widely concerned owning to its advantages of user transparency. Since few people are willing to judge the relevance of the retrieved documents, it is difficult to make use of the relevance of the query results. Therefore, as an alternative solution, the *PCQE* technology comes into play to extend the query without the interaction of users. It makes full use of the relevant information of the document to expand the query word. Some extended words extracted from the first top retrieval documents are combined with the original query to build a new query, then the new query is issued to the search engineer to perform the second search. *PCQE* has be proved by some research works that it can improve retrieval performance [8–10]. For example, Iwayama [11] used a method of pseudo feedback to extend query, his work clusters the documents in terms of the property that the retrieval documents relevant to the query are very similar but other documents have a diverse distribution.

In addition, with the development of *language model* (*LM*), some new *QE* techniques are used in the framework of language model [12]. The *PCQE* based on the language model framework often uses the feedback information (the ranked top *k* documents) to expand the query to obtain a more accurate query language model, its essence is to update the probability of query words in the framework of language model in terms of feedback information. For example, Zhai et al. [13] proposed a feedback model based on *LM*, their experiments showed that the model has an excellent performance in their experiments. Lv and Zhai [14] did a comparative experiment to compare the performance of the most representative five methods in the ad hoc retrieval task of *IR*, they found *Relevance Model 3* (*RM3*) and *Simple Mixture Model* (*SMM*) is the more *PCQE* methods and *RM3* is more robust, so *RM3* is widely used as a benchmark algorithm [15] for its excellent performance and good robustness.

On the other advancement line of our related works is the application of topic model to *IR* field. Wei firstly applied *LDA* to *IR* field for the ad hoc retrieval task [16], the topics extracted by *LDA* been used as document smoothing model under the framework of language model retrieval, their experiments show that this smoothing method can improve retrieval performance greatly. Xing et al. [17] evaluated the performance of the retrieval based on the topic model. They also analyze the search results of three topic models for *IR* [18], the result shows that the topic models for document smoothing of language model can improve the retrieval performance, especially the performance of *LDA* model is the most excellent. There are some research works finding that the topic model can improve the retrieval results when it is applied to the feedback documents. Ye [19] proposed an extension method based on topic model to find one or more related to the query topic, and proved by experiments, the selected and query related topics in query expansion based on performance obviously superior to the language model under the framework of some representative query expansion method.

3 Our Method

3.1 Fit Document Set with LDA

LDA is a powerful generative probabilistic model to model text corpora, it can be regarded as a mixture component model, in which each component is represented as a multinomial distribution over terms. The components, also called topics, are shared by all text documents. Each text document is associated with a multinomial distribution over topics, in this way *LDA* maps text documents to lower dimensional topic space. The probability generate process of *LDA* is shown in Fig. 1:

```
(1) For each topic $k \in \{1,2,\ldots,K\}$
    (1.1) Draw a topic-term distribution $\beta_k \sim Dir(\eta)$
(2) For each document $d \in D$
    (2.1) Draw a document-topic distribution $\theta_d \in Dir(\alpha)$
    (2.2) For each word index $n \in \{1,\ldots,N_d\}$
        (2.2.1) Draw a topic $z_{dn} \sim Multi(\theta_d)$
        (2.2.2) Draw a word $w_{dn} \sim Multi(\beta_{z_{dn}})$
```

Fig. 1. The probability generate process of LDA

Figure 2 illustrates the graph model representation of *LDA*:

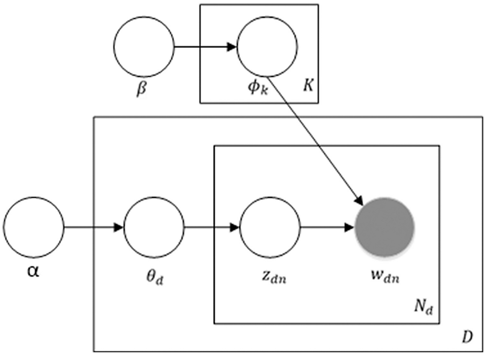


Fig. 2. The graph model representation of *LDA*

The notations listed in Figs. 1 and 2 are illustrated in Table 1:

Table 1. Notations of *LDA*

Notation	Meaning
K	The topic number
β	The parameter of *Dirichlet* prior distribution of topic-term distribution
ϕ_k	The topic-term distribution of topic k, where ϕ_{kv} is the probability of term v in topic k
D	The text document corpus
α	The parameter of *Dirichlet* prior distribution of topic-term distribution
θ_d	The document-topic distribution of document d, where θ_{dk} is the proportion of topic k in document d
N_d	The term count of document d
z_{dn}	The topic of the nth position in document d, it is drawn from the multinomial distribution $Multi(\theta_d)$, with a value range of from 1 to K
w_{dn}	The nth term in document d

Thomas L. Griffiths and Mark Steyvers [19] develops a *collapse Gibbs Sample* algorithm to inference θ_d, z_{dn} and ϕ_k:

$$P(z_{dn} = k|\mathbf{z}_{-dn}, \boldsymbol{w}) \propto \frac{n^{w_{dn}}_{-dn,k} + \beta}{n^{(\cdot)}_{-dn,k} + V\beta} \frac{n^{(d)}_{-dn,k} + \alpha}{n^{(d)}_{-dn,\cdot} + K\alpha} \tag{1}$$

$$\widehat{\phi}_{kv} = \frac{n^{(v)}_k + \beta}{n^{(\cdot)}_k + V\beta} \tag{2}$$

$$\widehat{\theta}_{dk} = \frac{n^{(d)}_k + \alpha}{n^{(d)}_{\cdot} + K\alpha} \tag{3}$$

Table 2 lists the notations and their illustrations:

Table 2. Notations in formula (1), (2) and (3)

Notation	Illustration
$\mathbf{z}_{-dn}$	All the topic assignments for each term in the corpus, but not including the *n*th topic assignment in document *d*
$\boldsymbol{w}$	All the tokens occurred in the corpus
$n^{w_{dn}}_{-dn,k}$	The count of term w_{dn} which has a topic assignment *k* in the corpus excluding the *n*th term of document *d*
$n^{(\cdot)}_{-dn,k}$	The count of all the terms that have a topic assignment *k* in the corpus excluding the *n*th term of document *d*
$n^{(d)}_{-dn,k}$	The count of topic *k* in document *d* excluding the *n*th position of document *d*
$n^{(d)}_{-dn,\cdot}$	The term count of document *d* excluding the *n*th position
$\hat{\phi}^{(v)}_k$	The estimated value of ϕ_{kv}, which is the probability of term *v* for given topic *k*
$n^{(v)}_k$	The count of term *v* which has a topic assignment *k*
$n^{(\cdot)}_k$	The count of topic *k* in the corpus
V	The size of term table
$\hat{\theta}_{dk}$	The estimated value of θ_{dk}, which is the proportion of topic *k* in document *d*
$n^{(d)}_k$	The count of topic *k* in document *d*
$n^{(d)}_{\cdot}$	The term count of document *d*

We use the *collapse Gibbs Sample* algorithm to fit the document set to be retrieved, then each topic ϕ_k and document-topic distribution θ_d are estimated in terms of the formulas (2) and (3). Moreover, a topic assignment z_{dn} for each token in the document set is generated for each iteration of the sample process in terms of the formula (1), which can be regarded as a sample of distribution $p(\mathbf{z}|\boldsymbol{w})$. We can obtain a mean estimate of the posterior distribution of topic $p(\mathbf{z}|\boldsymbol{w})$ by averaging the samples of numerous iterations. To make the samples to be average independent as far as possible, we gather a sample of $p(\mathbf{z}|\boldsymbol{w})$ every 50 iterations after the burn-in period. If we represent the topic assignment $\mathrm{z}_{dn} = k$ with a 1-of-K vector $\mathrm{z}'_{dn} = [\mathrm{z}'_{dn,1}, \ldots, \mathrm{z}'_{dn,K}]^T$, Where

$$\mathrm{z}'_{dn,k} = \begin{cases} 1 & \mathit{if}\ \ \mathrm{z}_{dn} = k \\ 0 & \mathit{ohterwise} \end{cases} \tag{4}$$

Then we can define the estimation for $p(\mathbf{z}|\boldsymbol{w})$ as:

$$\hat{E}(\mathbf{z}'|\boldsymbol{w}) = \frac{\sum_i \mathbf{z}'^{(i)}}{T} \tag{5}$$

Where $\mathbf{z}'^{(i)} = \{\mathrm{z}'^{(i)}_{11}, \ldots, \mathrm{z}'^{(i)}_{dn}, \ldots\}$ and $\mathrm{z}'^{(i)}_{dn}$ is the topic assignment for the ith sample of $p(\mathbf{z}|\boldsymbol{w})$ in terms of the formula (1), T is the number of gathered samples used for averaging. Then the topic proportion of each topic in the document set can be calculated by:

$$p(\mathrm{z} = k|\boldsymbol{w}) = \sum\nolimits_d \sum\nolimits_n \hat{E}(\mathbf{z}'|\boldsymbol{w}) \tag{6}$$

The formula (6) indicates that there may exist an unbalanced topic distribution in the given document set. For example, if $p(\mathrm{z} = 1|\boldsymbol{w}) = 0.6$ then in the document set $\boldsymbol{w}$ the topic #1 is very popular and many documents are focused on this topic.

3.2 Retrieve the First Relevant Documents for Input Query

After we fit the document set to be retrieved by *LDA*, then two multinomial distributions can be estimated in terms of the formulas (2) and (3), i.e. we can obtain the probability distribution over topics for each document as well as the probability distribution over terms for each topic. For a given input query q, it can be regarded as a token sequence $< w_1, \ldots, w_Q >$, then we regard the relevance between q and the document d as:

$$p(q|d) = p(q|\theta_d) = \sum\nolimits_{\mathrm{z}=1}^{K} p(q, \mathrm{z}|\theta_d) = \sum\nolimits_{z=1}^{K} p(q|\mathrm{z})p(\mathrm{z}|\theta_d) \tag{7}$$

It can be viewed as the probability of the document d generating q, the higher probability means the document d is more relevant to the query q. We rank the documents in terms of the generate probability calculated by the formula (7), then the top $N = 50$ retrieved documents are regarded as the first relevant documents for input query q.

3.3 Extend the Original Query

From the first relevant documents, we can get the candidate words to extend the original query. Using the notation W' to denote the candidate words, extending the original query then can be considered to choose some words $w \in W' - q$ to extend the original query q with the aim of w has high relevance to q. We use the posterior probability $p(w|q) = p(w|w_1, \ldots, w_Q)$ to measure how w is relevant to q:

$$\begin{aligned} p(w|w_1, \ldots, w_Q) &= \frac{p(w, w_1, \ldots, w_Q)}{p(w_1, \ldots, w_Q)} \\ &\propto p(w, w_1, \ldots, w_Q) \\ &= \sum\nolimits_{\mathrm{z}} p(w, w_1, \ldots, w_Q|\mathrm{z})p(\mathrm{z}) \end{aligned} \tag{8}$$

Where $p(\mathrm{z})$ can be calculated by the formula (6), $p(w, w_1, \ldots, w_Q|\mathrm{z})$ is the generate probability of words $w, w_1, \ldots, w_Q$, it can be calculated by the formula (2):

$$p(w, w_1, \ldots, w_Q|\mathrm{z}) = p(w|\mathrm{z}) \prod\nolimits_{i=1}^{Q} p(w_1|\mathrm{z}) = \phi_{\mathrm{z}w} \prod\nolimits_{i=1}^{Q} \phi_{\mathrm{z}w_i} \tag{9}$$

Where $\phi_{\mathrm{z}w}$ can be estimated by the formula (2).

4 Experiments

We use the *Ohsumed* data set to test our method, which was originally created to improve the performance of information retrieval. The OHSUMED data set is an objective oriented medline data set, comprising 348,566 documents from 270 medical journals between 1987–1991. It also provides a total of 106 queries by the doctors, the query consists of two parts of the patient's condition description and demand information.

We remove the stop words with a standard stop word list, and then stem the rest words with the *Potter* stem algorithm. To evaluate the effect of our method (named *LDAE*), we introduce two benchmark query expansion algorithms, namely *RM3* and *LCA*. The *RM3* extended algorithm is stable and robust as well as excellent performance, so many research works use it as a query benchmark algorithm. On the other hand, LCA is a widely-used query expansion algorithm based on the conventional word co-occurrence technology, so it is also introduced to compare with *LDAE*. We compare *LDAE* with *RM3* and *LCA* in both *MAP* and *P@10* metrics. The result is shown in Table 3:

Table 3. The experiment results of *LDAE*, *RM3* and *LCA*

Algorithm	*MAP*	*P@10*
LDAE	41.2%	41.9%
RM3	39.7%	39.2%
LCA	34.5%	33.6%

It can be seen that *LCA* has a poor performance and *RM3* has a better performance than it, whereas *LDAE* outperforms both *MAP* and *P@10*.

Acknowledgments. This work was supported by the Chinese National Natural Science Foundation (Grant No.: 61602202), the Natural Science Foundation of Jiangsu Province, China (Grant No.: BK20160428), the Social Key Research and Development Project of Huaian, Jiangsu, China (Grant No.: HAS2015020) and the Graduate Student Scientific Research and Innovation Project of Jiangsu Province, China (Grant No.: 2015B38314).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Maron, M.E., Kuhns, J.L.: On relevance, probabilistic indexing and information retrieval. J. ACM **7**(3), 216–244 (1960)
3. Rocchio, J.: Relevance Feedback in Information Retrieval. The SMART Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice-Hall, Englewood Cliffs (1971)
4. Szekeres, A., et al.: A keyword search algorithm for structured peer-to-peer networks. Int. J. Grid Util. Comput. **2**(3), 204–214 (2011)
5. Lai, C., Moulin, C.: Semantic indexing modelling of resources within a distributed system. Int. J. Grid Util. Comput. **4**(4), 21–39 (2013)
6. Carterette, B., Bah, A., Zengin, M.: Dynamic test collections for retrieval evaluation. In: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, Amsterdam, The Netherlands, July DBLP, pp. 55–62 (2007)
7. Rajaram, S., et al.: Classification approach towards ranking and sorting problems. In: Machine Learning: European Conference on Machine Learning, ECML 2003 Cavtat-Dubrovnik, Croatia, 22–26 September 2003, Proceedings DBLP, pp. 301–312 (2003)
8. Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 111–119. ACM (2001)
9. Carpineto, C., et al.: An information-theoretic approach to automatic query expansion. ACM Trans. Inf. Syst. **19**(1), 1–27 (2001)
10. Iwayama, M.: Relevance feedback with a small number of relevance judgements: incremental relevance feedback vs. document clustering. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 10–16 (2000)
11. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. University of Massachusetts, pp. 275–281 (1998)
12. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. ACM, pp. 403–410 (2001)
13. Lv, Y., Zhai, C.X.: A comparative study of methods for estimating query language models with pseudo feedback. In: ACM Conference on Information and Knowledge Management, pp. 1895–1898. ACM (2009)

14. Cao, G., et al.: Selecting good expansion terms for pseudo-relevance feedback. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2008)
15. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2006)
16. Yi, X., Allan, J.: Evaluating topic models for information retrieval. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM (2008)
17. Yi, X., Allan, J.: A comparative study of utilizing topic models for information retrieval. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 29–41. Springer, Heidelberg (2009)
18. Wang, X., et al.: LDA based pseudo relevance feedback for cross language information retrieval. In: 2012 IEEE 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS), vol. 3. IEEE (2012)
19. Ye, Z., Huang, J.X., Lin, H.: Finding a good query-related topic for boosting pseudo-relevance feedback. J. Am. Soc. Inform. Sci. Technol. **62**(4), 748–760 (2011)

Target Recognition Method Based on Multi-class SVM and Evidence Theory

Wen Quan[1(✉)], Jian Wang[2], Lei Lei[2], and Maolin Gao[1]

[1] Air Traffic Control and Navigation College, Air Force Engineering University,
Xi'an 710051, China
937182228@qq.com, gml2003@gmail.com
[2] Air and Missile Defense College, Air Force Engineering University,
Xi'an 710051, China
wendyandpaopao@163.com, 26471375@qq.com

Abstract. In order to conquer the hard outputs defect of Support Vector Machine (SVM) and extend its application, an improved target recognition method based on Multi-class Support Vector Machine (MSVM) is proposed. Firstly, the typical Probability Modeling methodologies of MSVM were deeply analyzed. Secondly, the structure of one-against-one multi-class method which matches with Basic Probability Assignment (BPA) outputs of evidence theory by coincide, so a special Multi-class BPA output method is derived, and multi-sensor target recognition model based on MSVM and two-layer evidence theory is constructed. Finally, the results of experiments show that the proposed approach can not only conquer the overlap area of one-against-one multi-class method, but also improve classification accuracy.

1 Introduction

Automatic target recognition (ATR) is an important development direction of current and future weapon system. In order to overcome the limitation of single sensor target recognition system. It is imminent to integrate the information from different sensors to complete the target identification. However, due to the information source of uncertainty and randomness, complexity of multi-sensor target recognition problem itself, as well as a variety of complex interference the electromagnetic environment, the organic combination of the solution of the problem depends on a variety of technology and comprehensive application [1, 2], such as methods based on evidence theory and neural network [3, 4], evidence theory method combined with fuzzy set theory [5], evidence theory combined with rough set theory [6], have been paid more attention in the field of multi-sensor target recognition, it has become a hot spot and trend of research at home and abroad [7].

One of the key interesting Data fusion methods of ATR is to fuse the data from SVM by Evidence theory. But unfortunately, the outputs of SVM are hard data, which cannot be well disposed by Evidence theory, so the main work of this paper is to explore how to combine the Evidence theory and the MSVM.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_26

The rest of this paper is organized as following. Section 2 presents basic definitions about evidence theory and SVM. The combination method of MSVM and evidence theory is explored in Sect. 3 and the Multi-sensor Target Recognition Structure Model is constructed. Section 4 is dedicated to validate the performance of proposed method. Finally, conclusions are drawn in Sect. 5.

2 SVM and D-S Theory

2.1 SVM Theory

Support vector machine (SVM) is a classification method based on statistical learning theory. It shows the advantages of high dimensional pattern classification and regression estimation, and there is no local optimal problem [8].

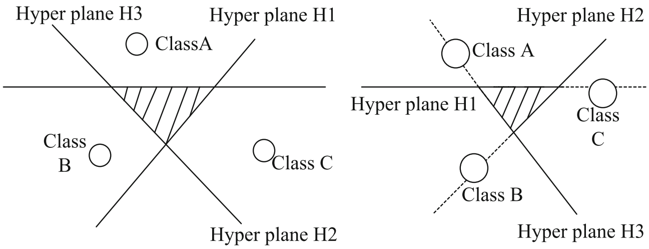
The principle of classification is to minimize the expected risk $R[f] = \iint L(y, f(x)) dP(y|x) dP(x)$ by finding the function $f(x) = w * x + b$, so as to minimize the normalized risk $R_{reg}[f]$, $R_{reg}[f] = 1/2 \cdot ||w||^2 + C \sum |1 - y_i f(x_i)|$ of the empirical risk of the data $(x_i, y_i)_{i=1 \cdots n}$. This is an Optimization problem, which can be solved by the following formula.

$$-\frac{1}{2}\sum_{i.,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i * x_j + \sum_{i=1}^{n} \alpha_i \to \min,\ w.r.t \quad \sum_{i=1}^{n} \alpha_i y_i = 0,\ \forall i : 0 \le \alpha_i \le C \tag{1}$$

2.2 Multi Classification SVM

A single SVM can only solve two classification problems, multiple targets classification problem [9, 10] can be solved by combining multiple SVM classifiers [11], including "one to many (One-Against-All, OAA)", or "one to one (One-Against-One, OAO)" method, or directed loop diagram method and method based on decision tree, all these methods are directly using the output decision function as criteria for classification, we called the discriminant principle as "functional distance" classification.

In OAA, the difference of the total samples in each two-class classifier between the numbers in two kinds of sample is very big; the training time is very large, at the same time, the classification results in favor of class that has more samples, namely "data skew" phenomenon. While OAO overcomes the shortcomings of the OAA method, avoided the "skewed data set", "inseparable" phenomenon (Fig. 1(a) is shown in the shaded part). But due to the OAO decision stage using the voting method, it may produce strange result that the votes are identical from totally different classes, so that the unknown samples are misjudged to belong to many different categories at the same time, namely "classification overlap" phenomenon (Fig. 1(b) shown in the shaded part), which severely impair the classification accuracy.



(a) OAA MSVM (b) OAO MSVM

Fig. 1. The comparison between OAA and OAO

2.3 D-S Evidence Theory

There are many kinds of methods such as the combination of multiple classifiers, based on Bayesian theory fusion strategy [12] (including average Bayes method, median Bayes method), Voting fusion strategy [13] (including absolute majority voting, weighted voting method), Fuzzy integral fusion strategy [14], Decision Template fusion strategy, the expert product of algorithm fusion strategy [15], D-S theory fusion strategy [16]. Especially, Literature [12] pointed out that the performance of D-S theory fusion strategy is the most robust and solid through comprehensive comparative analysis [17]. The introduction of D-S theory may eliminate the “inseparable” and “overlap” phenomenon in Multi class classification. So we tried to uses D-S theory to combined each classifier output decision value.

D-S theory of evidence [18] was modeled based on a finite set of mutually exclusive elements, called the frame of discernment denoted by Θ. The power set of Θ, denoted by 2^{Θ}, containing all possible unions of the sets in Θ including Θ itself. Singleton sets in a frame of discernment Θ will be called atomic sets because they do not contain nonempty subsets.

Definition 1: Let $\Theta = \{\mathrm{A}_1, \mathrm{A}_2, \mathrm{A}_3, \ldots, \mathrm{A}_\mathrm{n}\}$ be the frame of discernment. A basic probability assignment is a function m: $2^{\Theta} \rightarrow [0,1]$, satisfying the two following conditions:

$$m(\phi) = 0, \sum m(A) = 1, A \subseteq \Theta$$

Definition 2: Given two belief structures $\mathrm{m}_1$and m_2 on Θ, they are exclusive with each other, the combined BPA is $m(.) = [m_1 \oplus m_2](.)$, it is given by

$$\begin{cases} m(\phi)=0 \\ m(A)=\dfrac{\sum\limits_{\substack{X,Y\subset 2^{\Theta} \\ X\cap Y=A}} m_1(X)m_2(Y)}{1-\sum\limits_{\substack{X,Y\subset 2^{\Theta} \\ X\cap Y=\phi}} m_1(X)m_2(Y)} \end{cases} \quad \forall (A \neq \phi) \in 2^{\Theta}. \tag{2}$$

In the formula, the degree of conflict between the two evidences is $\sum m_1(X)m_2(Y),\ X\ and\ Y \subset 2^{\Theta}, X \cap Y = \phi$, if $k_{12} \neq 1$, then a basic probability assignment function is determined; if $k_{12} = 1$, it is considered that m_1 and m_2 are contradict with each other. The combination rules satisfy the combination law and the commutative law, and it is suitable for the combination of multiple evidences.

3 The Combination of MSVM and Evidence Theory

3.1 SVM Probability Output

There are two typical kinds of methods to determine the posterior probability [19]: one is the theoretical framework of Bayes which calculates all the class conditional probability density, and then calculates the posterior probability based on the Bayes theory [20–22]. Another kind of method is not to calculate probability density, fitting the posterior probability of the posterior probability directly, as Vapnik and cosine function [23], Platt treats posterior probability as Sigmoid function in the form of [24] and on the deformation of Sigmoid function [19, 25, 26]. At present, the commonly used method is put forward by Platt, he mapped hard decision output f(x) to [0, 1]. The posterior probability output form is as follows:

$$p(y = 1|f) = 1/(1 + \exp(Af + B)) \tag{3}$$

$$\min - \left(\sum t_i \log(p_i) + (1 - t_i \log(1 - p_i)\right). \tag{4}$$

$$p_i = 1/(1 + \exp(Af + B)),\ t_i = (y_i + 1)/2 \tag{5}$$

The literature [27] presented the comparative analysis of probability function of the output on several monotone functions, the probability output of SVM.

3.2 MSVM Soft Output

Upper bound for the test sample classification error rate is the proportion of the average number of support vectors in the training sample to the total number of training samples:

$$E(P(error)) \leq \frac{E(\text{number of support vectors})}{\text{total number of training samples} - 1} \tag{6}$$

The literature [28] considers that the upper bound of the error exactly reflects the uncertainty of the SVM for the sample, the output of the MSVM BPA function is defined as:

Given the total number of class is $T(T \geq 2)$, the identification framework is $\{A_1, A_2, \ldots, A_T, \Theta\}$, the corresponding set of n of the SVM of the two category, $n = T(T-1)/2$, which n SVM constitute a MSVM.

For the sample x of the test sample, assuming that the two types of SVM A_i, A_j, then according to the formula (3), formula (4), formula (5) and formula (6), for the allocation of a two SVM:

$$m(A_i) = p(x)(1 - Nsv/(l-1)). \tag{7}$$

$$m(A_j) = [1 - p(x)](1 - Nsv/(l-1)). \tag{8}$$

$$m(\Theta) = Nsv/(l-1). \tag{9}$$

Among them, $p(x) = 1/(1 + \exp[Af(x) + B])$, Nsv is the number of support vectors corresponding to the two types of SVM, l is the total number of training samples, Θ is the uncertainty of the results.

For the sample points x, according to the OAO rule, we give the BPAs of all two sub classifiers MSVM with evidence combination rule form:

$$\begin{array}{llllllll}
 & A_1 & A_2 & A_3 & \cdots & A_{T-1} & A_T & \Theta \\
m_1 & m_1(A_1) & m_1(A_2) & 0 & \cdots & 0 & 0 & m_1(\Theta) \\
m_2 & 0 & m_2(A_2) & m_2(A_3) & \cdots & 0 & 0 & m_2(\Theta) \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\
m_n & 0 & 0 & 0 & \cdots & m_n(A_{T-1}) & m_n(A_T) & m_n(\Theta)
\end{array}$$

However, the synthesis of this type BPA structure, only to and classification of experts on classification and assignment; expert m_1 only assign believe to A_1 and A_2, expert m_2 only assign believe to A_2 and A_3, if we combine the judgments of these two experts with typical evidence theory, there will be high conflict results. For example, obviously, expert m_1 is more inclined to class $A_1 (m_1(A_1) > 0.5)$, the experts m_2 tend to class $A_2 (m_2(A_2) > 0.5)$, to synthesis expert opinions with evidence theory, we first calculate the conflict indicator K value:

$$\begin{aligned} K_{12} &= m_1(A_1)m_2(A_2) + m_1(A_1)m_2(A_3) + m_1(A_2)m_2(A_3) \\ &= 1 + m_1(A_1) - m_1(\Theta)m_2(A_3) - m_2(\Theta) \end{aligned}$$

As $m_i(A_1) + m_i(A_2) + m_i(\Theta) = 1,\ 1 + m_1(A_1) > 1.5,\ m_2(\Theta) < 0.5,\ m_1(\Theta)m_2(A_3) < 0.25$, so $K_{12} > 0.75$;

Then calculate the pignistic probability function as follows:

$$\begin{gathered} BetP_{m1}(A_1) = \frac{m_1(A_1)}{1} + \frac{m_1(\Theta)}{n}, BetP_{m1}(A_2) = \frac{m_1(A_2)}{1} + \frac{m_1(\Theta)}{n}, BetP_{m1}(\Theta) = \frac{m_1(\Theta)}{n}; \\ BetP_{m2}(A_2) = \frac{m_2(A_2)}{1} + \frac{m_2(\Theta)}{n}, BetP_{m2}(A_3) = \frac{m_2(A_3)}{1} + \frac{m_2(\Theta)}{n}, BetP_{m2}(\Theta) = \frac{m_2(\Theta)}{n} \\ difBetP = \max \Sigma_{A \subseteq \Omega} |BetP_{m1}(A) - BetP_{m2}(A)| \end{gathered}$$

That means to calculate $\max[X_1, X_2, X_3, X_4]$,

$$\begin{aligned} &X_1 = |\tfrac{m_1(A_1)}{1} + \tfrac{m_1(\Theta)}{n}|, \quad X_2 = |m_1(A_2) - m_2(A_2) + \tfrac{m_1(\Theta)}{n} - \tfrac{m_2(\Theta)}{n}| \\ &X_3 = |m_2(A_3) - \tfrac{m_2(\Theta)}{n}|, \quad X_4 = |\tfrac{m_1(\Theta)}{n} - \tfrac{m_2(\Theta)}{n}| \end{aligned}$$

As $m_1(A_1) > 0.5$, $X_1 = |\frac{m_1(A_1)}{1} + \frac{m_1(\Theta)}{n}| \geq 0.5$

By the literature [29], $cf(m_1, m_2) = \, < K, difBetP > \, \geq \, < 0.75, 0.5 >$

It can be seen that is a high conflicts problem. For this paper is a study of multi-sensor and multi-target recognition problem, in order to avoid the loss of information in the fusion process. However, using the evidence theory of multi-sensor measurement fusion two layer evidence reasoning definitely increased system processing steps, it increase the system time consuming. This problem has been solved very well in the next section.

3.3 Fast Murphy Combination Rule

Murphy's improved combination rule [30] is put forward specifically for D-S inference rules that cannot solve the problem of high conflict information. Through in-depth research, we conclude the special rule of the same evidence combination, and propose a Fast Murphy Combination Rule (FMCR). The FMCR formula is given below. The framework is $\{A_1, A_2, \ldots, A_T, \Theta\}$, assuming the mean value of the uncertainty that derived from all evidences is Q.

$$\begin{array}{ccccccc} & A_1 & A_2 & A_3 & \cdots & A_T & \Theta \\ m_1 & x_1 & x_2 & x_3 & \cdots & x_T & Q \\ m_2 & x_1 & x_2 & x_3 & \cdots & x_T & Q \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ m_n & x_1 & x_2 & x_3 & \cdots & x_T & Q \end{array}$$

And the FMCR is:

$$m_{12\cdots i}(A_i) = \frac{(x_i + Q)^n - Q^n}{1 - 2\sum\limits_{1 \leq i,j \leq T, i \neq j} x_i x_j - \sum\limits_{k=2}^{n-1} [\sum\limits_{i=1}^{T} x_i (\sum\limits_{1 \leq i,j \leq T, i \neq j} [(x_j + Q)^k - Q^k])]} \tag{10}$$

Proof:

(1) For example, $n = 2$

Synthesis of evidence m_1 and m_2 with D-S combination rules:

$$
\begin{aligned}
M_{12}(A_1) &= \frac{(x_1+Q)^2 - Q^2}{1 - x_1(x_2+x_3+\cdots+x_T) - \cdots - x_T(x_1+x_3+\cdots+x_{T-1})} \\
&= \frac{(x_1+Q)^2 - Q^2}{1 - 2\sum\limits_{1\le i,j\le T, i\ne j} x_i x_j} = m_{12}(A)
\end{aligned}
$$

So the equation is correct;

(2) Assuming $n = \lambda$,

$$
\begin{aligned}
M_{12\cdots\lambda}(A_1) &= \frac{(x_1+Q)^\lambda - Q^\lambda}{1 - 2\sum\limits_{\substack{1\le i,j\le T\\ i\ne j}} x_i x_j - \sum\limits_{k=2}^{\lambda-1}[\sum\limits_{i=1}^{T} x_i(\sum\limits_{\substack{1\le i,j\le T\\ i\ne j}}[(x_j+Q)^k - Q^k])]} \\
&= m_{12\cdots\lambda}(A)
\end{aligned}
$$

If $n = \lambda + 1$, Synthesis $\lambda + 1$ evidence by D-S evidence combination rule:

$$
M_{12\cdots(\lambda+1)}(A_1) = \frac{x_1[(x_1+Q)^\lambda - Q^\lambda] + x_1 Q^\lambda + Q[(x_1+Q)^n - Q^n]}{\Delta - [\sum\limits_{i=1}^{T} x_i(\sum\limits_{j\ne i}(x_i+Q)^n - Q^n]} = m_{12\cdots(\lambda+1)}(A_1)
$$

Among them, Δ is $1 - 2\sum x_i x_j,\ 1 \le t, j \le T, i \ne j$, the equation $n = \lambda + 1$ is correct. According to mathematical induction rule, the equation is correct in any situation.

3.4 Multi-sensor Target Recognition Structure Model

In 1990, Hansen and Salamon [31] proposed a novel method, which can improve the classification accuracy of the system by training a number of neural networks and synthesizing the results [11].

By this idea, we construct a multi-sensor target recognition model, With m MSVM in target recognition, the same target x, by measuring m sensor to obtain m measurement data, in the inner MSVM, n probability output is produced by n two class classifier (by Eqs. 3, 4, 5), and then according to the Eqs. 7, 8, 9 we obtain the BPA, using FMCR evidence reasoning output value $BPA_1, BPA_2 \cdots, BPA_m$ of the sensors, and then using D-S for final decision fusion results, the structure model is shown in Fig. 2.

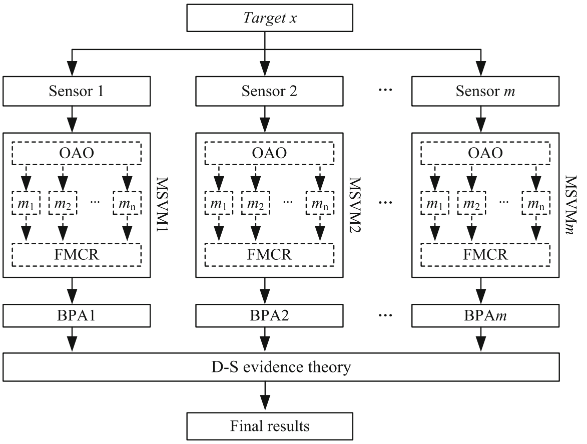


Fig. 2. Multi sensor target recognition structure model based on MSVM and evidence theory

4 Simulation Experiments

The simulation experiments are carried out in the MATLAB2010b platform, the experimental data using the actual levels of Polari metric HRRP data of 5 kinds of aircraft (J600, F1601, F1700, FY200, and B200). The original data are the horizontal polarization of X band amplitude and phase data of each target: the frequency is 8–12 GHz, a total of 101 frequency points; angle is 0°–180°, a total of 181 point of angel; each frequency point measuring 181 angle data. From the measured data by the transformation, we can obtain 128 dimensional angle intervals for HRRP data with 0°–180° angular range.

Because the original data is few, the corresponding attitude angle interval is large (1°), so the method of inserting artificial data is used to simulate the measurement data in different angles:

HRRP data can be expressed as $\mathbf{H} = \{H_{i,j}\}, 1 \le i \le m, 1 \le j \le n$, m is the number of sampling the attitude angle, n is the number of sampling frequency points.

(1) First calculate $\boldsymbol{\sigma} = \left\{\sigma_j\right\}_{1 \le j \le n}$, $\sigma_j = \text{mean}\big(\text{abs}(\mathbf{H}_{i,j} - \mathbf{H}_{i-1,j})\big)/2$

(2) $\mathbf{H}_{Ins}(i,j,k) = \mathbf{H}_{i,j} + \mathrm{x_v}$, $k = 1, \cdots, K_I$ are the index of the inserting data, $\mathrm{x_v} \sim \mathrm{N}(0, \sigma_j)$, the frequency function is $f(x_v) = 1/\sqrt{2\pi}\sigma_j \exp(x_v^2/2\sigma_j^2)$.

Synthesize the artificial data and raw data; we get a new data set $\mathbf{H}_{new} = \mathbf{H}_{Ins} \cup \mathbf{H}$. The experiment using 5 kinds of target HRRP data within 0°–30° angular variations, it produce azimuth interval 0.5°, 0.33°, 0.25° respectively, and the data of the three groups are HD1, HD2, HD3, including data (128 dimension) and the corresponding class labels, see Table 1.

Experiment was carried out on HRRP data, which was set 5 types of aircraft J600, F1601, F1700, FY200, and B200 in three different azimuth intervals under the (HD1, HD2, and HD3). The experiment set 1–3 MSVM, each MSVM randomly selected data sets of each type of 1/3 sample training as training set, set all samples as test set.

Table 1. Experiment data

Dataset	HD1	HD2	HD3
Azimuth interval	0.5°	0.33°	0.25°
Class number/dimension number	5/128	5/128	5/128
Sample number	60 × 5	90 × 5	120 × 5

Table 2. Comparison between the traditional method and the new method

Method		Dataset		
		HD1	HD2	HD3
OAO and Majority voting		72.97 ± 2.42	86.16 ± 1.30	91.47 ± 1.42
The new method	One MSVM	76.53 ± 3.37	89.53 ± 2.31	94.00 ± 1.40
	Two MSVMs	84.33 ± 2.53	93.71 ± 0.77	95.57 ± 0.70
	Three MSVMs	85.63 ± 1.83	96.04 ± 1.02	97.40 ± 0.79

By contrast, OAO traditional classification method and majority voting are adopted as "OAO majority plus majority voting" a method, comparing experimental results are using 10 repeated "mean variance" experiments, as shown in Table 2.

It can be seen from Table 2, testing on the HD1 data set, by the new method, recognition accuracy was 4.88% higher than "OAO majority voting", two MSVMs was 10.19% higher than that of the single sensor recognition accuracy, three MSVMs was 11.89% higher than that of single sensor precision. Testing on the data set HD2, by the new method, recognition accuracy was 3.91% higher than "OAO majority voting", two MSVMs was 4.67% higher than that of the single sensor identification accuracy, three MSVMs was 7.27% higher than that of single sensor precision; Testing on data set HD3, recognition accuracy was 2.77% more than the majority of "OAO the voting method", MSVMs was 1.67% higher than that of the single sensor recognition accuracy; Three MSVMs was 3.62% higher than that of single sensor precision.

The data analysis shows that in the new multi sensor target recognition model, with the increase in the number of sensors in target recognition, target recognition accuracy of the system increases, especially for the small sample, the effect is more obvious. Moreover, due to the introduction of evidence theory method in system, converting to the hard outputs to soft outputs, it retained the useful information as far as possible, overcame the "overlap" phenomenon that may occur with hard output and majority voting, target recognition accuracy is effectively improved.

5 Conclusion

This paper proposes Multi-sensor target recognition model of MSVM and the two layer architecture is given based on the evidence theory, fusing the data of the measure level and the decision level respectively, in the inner sensors we uses FMCR method, between each sensor we use evidence combination rule, which not only overcomes the problem of conflict and speed up the processing speed. The simulation data in a set of aircraft targets on the results have shown that this method is an effective method for multi-sensor target recognition, and has certain theoretical significance and application value. Further research will be focused on other possible soft output method of evidence theory to enhance the target recognition effect.

Acknowledgement. This work is supported by National Natural Science Foundation of China, Grant Nos 61273275, 60975026, 61503407.

References

1. Pllana, S., Benkner, S., Xhafa, F., Barolli, L.: A novel approach for hybrid performance modeling and prediction of large-scale computing systems. Int. J. Grid Util. Comput. **1**, 316–327 (2009)
2. Bouaziz, R., Krichen, F., Coulette, B.: C-SCRIP: collaborative security pattern integration process. Int. J. Inform. Technol. Web Eng. **10**, 31–46 (2015)
3. Qu, D.C., Meng, X.W., Huang, J., He, Y.: Research of artificial neural network intelligent recognition technology assisted by Dempster Shafer evidence combination theory. In: 7th International Conference on Signal Processing, vol. 1, pp. 46–49 (2004)
4. Wang, M., Li, S., Mao, S.: Method of using neural network combined with D-S theory to carry out HRR target recognition, vol. 4. Beijing University of Aeronautics and Astronautics (2001)
5. Guo, H.: Approach to evidence combination based on fuzzy theory and its applications. Control Decis. 2 (2008)
6. Wang, G.-Y., Yao, Y.-Y., Yu, H.: A survey on rough set theory and applications. Chin. J. Comput. **32**, 1229–1240 (2009)
7. Bssis, N., Asimakopoulou, E., Xhafa, F.: A next generation emerging technologies roadmap for enabling collective computational intelligence in disaster management. Int. J. Space-Based Situated Comput. **1**, 76–85 (2011)
8. Mu, S., Tian, S., Yin, C.: Multiple kernel learning based on cooperative clustering. J. Beijing Jiaotong Univ. **32**, 10–13 (2008)
9. Kun, W., Kang, J., Chi, K.: Research on fault diagnosis method using improved multi-class classification algorithm and relevance vector machine. Int. J. Inf. Technol. Web. Eng. **10**, 1–16 (2015)
10. Lin, H.-C.K., Su, S.-H., Wang, S.-C.: Influence of cognitive style and cooperative learning on application of augmented reality to natural science learning. Int. J. Technol. Human Interact. **11**, 41–66 (2015)
11. Bernardo, D.V., Hoang, D.B., Bernardo, D.V.: Multi-layer Security analysis and experimentation of high speed protocol data transfer for GRID. Int. J. Grid Util. Comput. **3**, 81–88 (2012)
12. Xu, L., Krzyzak, C., Suen, C.Y.: Methods of combining multiple classifiers and their applications to handwriting recognition. Trans. Syst. Man Cybern. **22**, 418–435 (1992)
13. Srihari, S.N.: Reliability analysis of majority vote systems. Inf. Sci. **26**, 243–256 (1982)
14. Sugeno, M.: Theory of Fuzzy Integrals and its Applications. Tokyo Institute of Technology, Tokyo, Japan (1974)
15. Hinton, G.E.: Products of experts. In: Proceedings of the Ninth International Conference on Artificial Neural Networks, Edinburgh, Scotland, pp. 1–6 (1999)

16. Li, Y., Cai, Y., Yin, R., Xu, X.: Support vector machine ensemble based on evidence theory for multi-class classification. J. Comput. Res. Dev., **45**, 571–578 (2008)
17. Li, Z., O'Brien, L., Zhang, H.: Circumstantial-evidence-based effort judgement for web service composition-based SOA implementations. Int. J. Space-Based Situated Comput. **2**, 31–44 (2012)
18. Yang, K., Liu, S., Li, X., Wang, X.A.: D-S evidence theory based trust detection scheme in wireless sensor networks. Int. J. Technol. Hum. Interact. **12**, 48–59 (2016)
19. Zhang, X., Xiao, X., Xu, G.-Y.: Probabilistic outputs for support vector machines based on the maximum entropy estimation. Control Decis. **7**, 767–770 (2006)
20. Sollich, P.: Bayesian methods for support vector machines, evidence and predictive class probabilities. Mach. Learn. **46**, 21–52 (2002)
21. Kwok, J.T.Y.: Moderating the outputs of support vector machine classifiers. IEEE Trans. Neural Netw. **10**, 1018–1031 (1999)
22. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. Ann. Stat. **26**, 451–471 (1998)
23. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
24. Platt John, C.: Probabilistic output for support vector machine and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifier, pp. 1–11. MIT Press, Cambridge (1999)
25. Lin, H.T, Lin, C.J., Weng, R.C.: A Note on Platt's Probabilistic Outputs for Support Vector Machines. National Taiwan University, Taipei (2003)
26. Madevska-Bogdanova, A., Nikolik, D., Curfs, L.: Probabilistic SVM outputs for pattern recognition using analytical geometry. Neurocomputing **62**, 293–303 (2004)
27. Wang, Z., Zhao, Z., Weng, S., Zhang, C.: Solving one-class problem with outlier examples by SVM. Neurocomputing **149**, 100–105 (2015)
28. Zhou, H., Li, S.: Combination of support vector machine and evidence theory in information fusion. Chinese J. Sens. Actuators **21**, 1566–1570 (2008)
29. Liu, W.: Analyzing the degree of conflict among belief functions. Artif. Intell. **170**, 909–924 (2006)
30. Murphy, C.K.: Combining belief functions when evidence conflicts. Decis. Support Syst. **29**, 1–9 (2000). Elsevier Publisher
31. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 993–1001 (1990)

Selective Ensemble Based on Probability PSO Algorithm

Wen Quan[1(✉)], Jian Wang[2], Zhongmin He[1], and Jiaofeng Zuo[1]

[1] Air Traffic Control and Navigation College,
Air Force Engineering University, Xi'an 710051, China
937182228@qq.com, martiansi@163.com, 2018763723@qq.com
[2] Air and Missile Defense College,
Air Force Engineering University, Xi'an 710051, China
26471375@qq.com

Abstract. In order to overcome the disadvantages of high complexity, low speed and accuracy of selective ensemble algorithm based on genetic algorithm. We proposed a new selective ensemble algorithm based on probability PSO algorithm. First, in order to tackle the converging slowly and easily to partial minimum problems of simplified PSO, we introduce the cloud model and the method of complex; Second, we present the definition of probability PSO and the formula that converting the particle position vector to base learner selection problem, that make the transformation from continuous space to discrete space become true. Finally, we choose integration model generalization error as the adaptive function of PPSOSEN. The numerical results show that, compared with discrete PSO, PPSOSEN improved the recognition precision with the same time consumption, and it is an efficient selective ensemble algorithm.

1 Introduction

Ensemble learning is a kind of effective machine learning technology, and its most representative algorithms are Bagging [1] and Boosting [2]. Selective ensemble learning methods are important research direction of the existing integration methods [3, 4], and most of them are built [3–5] based on the genetic algorithm. But unfortunately, the classical genetic algorithm is only suitable for solving discrete problems and there are problems such as Hamming cliff, which make the crossover and mutation of the genetic algorithm are difficult to overcome, one of the most serious result is "premature" [6], another defect is the selective ensemble based on genetic algorithm (GASEN) algorithm has a high computational complexity and slow speed in the practical application [7]. However, many research show that the particle swarm optimization algorithm (PSO) is adapt for solving real problems [8, 9], which is better than the genetic algorithm, and the PSO algorithm has been introduced into the Ensemble learning [10, 11].

However, their study was set based on the discrete binary PSO algorithm [12]. The PSO algorithm is a simulation of bird foraging behavior [13], so it has the continuity inborn [14]. While the search space is of discrete points in the discrete binary PSO (BPSO) algorithm, so it loss the advantages of continuous PSO algorithm,

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_27

moreover, the research results of the continuous PSO algorithm is rich, and the discrete PSO algorithm only have slow progress by many restrictions, therefore, study of the introduction of the continuous PSO algorithm to selective integration problems will be more promising [15]. Based on this, we tried to present the improved PSO algorithm, which has fast continuous convergence speed and avoid falling into local extreme value. It can be used to enhance the selectivity of integrated performance.

The rest of this paper is organized as following. Section 2 presents basic definitions about PSO algorithm. An improved continues PSO algorithm is explored in Sect. 3. Section 4 is dedicated to validate the performance of proposed method. Finally, conclusions are drawn in Sect. 5.

2 Particle Swarm Optimization Algorithm and Its Improvement

The basic PSO algorithm updates the position of particle by velocity indirectly, but the particle movement speed does not mean particles can effectively reach the optimal solution, it may make the particles deviates from the right direction, that calls "particle divergence" phenomenon [16–19]. So the literature [20] proposed simplified particle swarm optimization algorithm (SPSO), based on the fact that PSO evolutionary process is independent of particle velocity, controlling the evolutionary process only by the position of the particle, which avoided the particle "slow convergence" that caused by the particle velocity.

Although SPSO resolves the problem of "slow convergence", but the evolution trajectory of the particles is determined by two parameters, the individual extreme value and swarm's extreme value. Any one of these two kinds of extreme local value will affect further optimization. If the swarm no longer evolves, it either falls into the extreme local value, the swarm need to jump out of the local minimum by variation of the swarm optimal particle, or it has been positioned within the neighborhood of the global optimal position, we need to search for the best particle neighborhood, in order to find the swarm optimal particle. To overcome this defect, this paper introduces the cloud model into SPSO [21].

The cloud model, as a uncertainty representation conversion model between qualitative concept and quantitative data, representing in natural language. Using it to improve the PSO is a very good method, because the cloud model has the characteristics of randomness and stable tendency. The random search can avoid getting into the local extreme value, and stable tendency can guarantee good location of the global extreme value. Therefore, if the cloud model is introduced, optimal particle that no longer evolution, can using the parameters of the cloud model entropy *En* and hyper entropy *He* to adapt adjustment, which makes the algorithm not only avoid sink into the local extreme value, but also and improve search accuracy and convergence speed in the later time.

If a particle extreme value x_r continuous evolution t' generation still remain unchanged, and the particle is not the global optimal particle, the particle may sink into a local minimum. It is the time for the particles to escape. However, even if the particle can escape, it also always tries to optimize itself. We use Complex method [22, 23], the escape mechanism is as follows:

Particle escape algorithm:
Input: training set is X, the current particle swarm is x, the particles to be escaped is x_r, reflection times is α, threshold is ε.
Output: particles after escaping is x_r.

Step1 : find the center of the particle x_0 that is better than the current particle x_r (assuming there are k particles)

$$x_0 = \sum_{i=1}^{k} x_i/k. \tag{1}$$

Step2 : calculate the α reflection point x_0.

$$x_\alpha = x_0 + \alpha(x_0 - x_r). \tag{2}$$

Step3 : judge whether x_α is in the feasible domain, if it is then turn to Step4; otherwise, $\alpha = \alpha/2$ judge $\alpha > \varepsilon$? if it is then turn to Step4; otherwise, randomly produce a particle p, $x_r = p$, then output.

Step4 : if $f(x_\alpha) > f(x_r)$, then $x_r = x_\alpha$, then output;
Otherwise $\alpha = \alpha/2$, judge $\alpha > \varepsilon$? if it is then turn to Step2, otherwise, randomly generated a particle, then output.

3 Selective Integration Based on Probabilistic PSO

3.1 Basic Ideas

The learning device selection problem represented by vector $L = \{l_1, l_2, \cdots, l_N\}$, which is composed of only 0 or 1, $l_i = 1$ stands for the study will be selected, otherwise it means that the learning device will be eliminated. Thus, selective integration problem becomes a typical 0-1 combination optimization problem. In order to use the continuous PSO algorithm for individual learners, we use each dimension of the particle position vector to represent select probability of the learning device, limiting the search area in the range of [0,1], we call the probability PSO (PPSO) algorithm, that is to say, the PPSO algorithm is a special case of the continuous PSO algorithm. Therefore, the updating formulas of position and speed of PPSO algorithm is completely follow the continuous PSO algorithm, inherits all features of continuous PSO algorithm. In order to solve transfer problem that from continuous spatial domain to the discrete domain, so the formula for the conversion from particle position vector to the learner selection problem is as follows

$$l_i = \begin{cases} 1 & x_i \geq q \\ 0 & else \end{cases}. \tag{3}$$

The random number q in the interval (0.2, 0.8) is used to evaluate the subset of the learner as the evaluation of the corresponding particle, and the best learning subset vector is used to replace the worst particle vector for the next round of evolution.

Therefore, the continuous improvement PSO algorithm can be used to enhance the performance of a selective ensemble, so a selective ensemble probability PSO algorithm is presented.

3.2 Selection of Fitness Function

Judging whether a integrated model is better than the other models better, often depends whether on the generalization error on the validation set V is small. So we take the generalization error on the validation set V as the fitness function of PPSOSEN algorithm.

Definition 1 $L = \{l_1, l_2, \cdots, l_N\}$ are independent N training learner, $\bar{l} = \sum_{i=1}^{N} w_i l_i$ are integrated classifier outputs. In the validation set V $= \{x_k, y_k\}$, $k = 1, \cdots, M$, M is the number of samples of the validation set, the generalization error of the ensemble classifier (the fitness function of the algorithm) is expressed as:

$$\begin{aligned} E &= \sum_{k=1}^{M} \left(\bar{l}(x_k) - y_k\right)^2 = \sum_{i=1}^{N} w_i (E_i - A_i) \\ &= \overline{E} - \overline{A}. \end{aligned} \tag{4}$$

In the formula, the generalization error of individual members is $E_i = \sum_{k=1}^{M} (l_i(x_k) - y_k)^2$.

The difference between the members and the integration is $A_i = \sum_{k=1}^{M} (l_i(x_k) - w_i \sum_{j=1}^{N} l_j(x_k))^2$, $\overline{E}$ is the average generalization error of the members and $\overline{A}$ is the mean difference degree.

Thus, in the PPSOSEN algorithm, the smaller the fitness of the particle, the smaller the generalization error of the ensemble model, the higher the classification accuracy, and the stronger the generalization ability.

3.3 Algorithm Flow

Select SVM as the PPSOSEN base learner, the algorithm flow is as follows:

Input: training set S, maximum number of iterations is K.
Output: integration model

Step1 : using the Bootstrap method [24] to generate M training subsets from the training set S, and get a learner in each training set, and there are M learner $SVM_t(t = 1, \cdots, M)$;

Step2 : the generation of M-dimensional n particles, each dimension of each particle is a random number on the [0,1] interval, constitute the initial particle swarm $x(1) = (x_1(1), \cdots, x_n(1))$, and set the parameter $\omega, c_1, c_2, c_3, c_4, \alpha, \varepsilon, t'$, T', U[15], the number of alteration is $k = 1$;

Step3 : Convert each particle to subset of learners vectors by formula (5), and then calculate the fitness value of each particle by formula (6), $f_i(1) = f(x_i(1))$, $i = 1, 2, \cdots, n$, each particle position pb_i is as the best position of the particle, $pb_i = x_i(1)$. Replace the worst particles in the swarm by the corresponding best position particle of swarm subset vector, an define the particle as the best position gb, that is $i = \arg\min_j(f_j(1)), gb = x_i(1)$.

Step4 : to determine whether the algorithm convergence, if it is, then turned to Step8; otherwise, turn to Step5;

Step5 : $k = k + 1$, perform the following operations for all particles:

(1) update the particle position using the SPSO algorithm;

(2) make the following operations for each particle $x_i(k)$:

① if $f_i(k) < f(pb_i)$, $pb_i = x_i$, $t_i = 0$, otherwise $t_i = t_i + 1$;

② $t_i \geq t'$,$gb \neq pb_i$, implement the particle escape algorithm,$pb_i = x_i$;

(3) if $i = \arg\min_j(f_j(k))$, so $f_i(k) < f(gb)$,$gb = x_i(k)$,$T = 0$, turn Step4; otherwise, $T = T + 1$, perform Step6;

Step6 : if $T \geq T'$, then execute Step7; otherwise turn Step4;

Step7 : adopt the cloud models, make the best particle gb of the current swarm to variate:

(1) $E = gb$;

(2) S = variable search range/c_3;

(3) $H = S/c_4$;

(4) after running the positive cloud generation algorithm, the particle is gb';

(5) if $f(gb') < f(gb)$, then $gb = gb'$, turn Step4, otherwise, turn to (6);

(6) if $u \geq U$, turn to Step8, otherwise, $u = u + 1$, return to (4);

Step8 : use (5) to convert gb to 0,1 vector, to verify a group of learners that selected by every particle in the validation set, select the smallest fitness of the output, the algorithm runs to an end.

4 Numerical Experiments and Analysis

The experiment was carried out in the Matlab2010b platform, using SVM as the base classifier, the experimental data was collected from UCI standard database, select three data sets, IRIS, WINE, and GLASS, Table 1 lists the number of categories of experimental data set, the number of features and the number of samples.

Table 1. Characteristics of experimental data set

Test data	Categories	Feature number	Sample number
IRIS	3	4	150
WINE	3	13	178
GLASS	6	9	214

The experimental procedure is as follows:

First, uses Bootstrap to train 20 OAA MSVM classifiers, then using PPSOSEN method to select the 20 MSVM classifier, so we got the integration model. The validation set is selected randomly in the training set, the integrated model for optimal selection of optimal integration. The selected model on the test set for testing.

The experimental parameters are as follows:

The learning factor are $c_1 = c_2 = 1.4962$, swarm size is $N = 40$, other parameters are set up in reference [15], the maximum number of iterations is $K = 40$. The stop conditions are the optimal fitness value has not improved after continuous iteration 20 times.

In order to reduce the randomness, we used the [4] method (GASEN), [6] (DPSOSEN) and the method of this paper (PPSOSEN) to test 50 times on each data set, the average correct recognition rate and the time consumed as shown in Tables 2. and 3.

Table 2. Comparison of the recognition rates of the three methods on the standard data set

Dataset	Average recognition rates(%)		
	GASEN	DPSOSEN	PPSOSEN
IRIS	95.33	96.00	96.67
WINE	88.32	88.79	89.72
GLASS	71.96	71.96	72.43

Table 3. Comparison of the time consumption of the three methods on the standard data set

Dataset	Time consumption(s)		
	GASEN	DPSOSEN	PPSOSEN
IRIS	3.11	1.34	1.37
WINE	4.24	2.12	2.14
GLASS	5.67	2.60	2.64

From the comparison of the three methods in Table 2 can be seen, DPSOSEN obtains as good recognition accuracy as GASEN, and the recognition accuracy of PPSOSEN in the three data sets is higher than that of DPSOSEN and GASEN, thus, DPSOSEN is proved better than DPSOSEN and GASEN in optimization.

From comparing the time consuming on the three methods in Table 3, it can be seen that the convergence rate of the system of PPSOSEN and DPSOSEN is faster than GASEN, the execution time of PPSOSEN and DPSOSEN were much larger than that of GASEN and gained high efficiency. Since PPSOSEN made judgment whether into the local extreme value in the search stage of particle detection, and let the local minim of particle to escape, so the execution time was less than DPSOSEN. Therefore, with the overall consideration of speed and accuracy, PPSOSEN is best.

5 Conclusion

In order to improve the speed and accuracy of selective integration problems, the continuous PSO algorithm is introduced to the selective ensemble, adopting the cloud model and the method of complex combination, overcame the low convergence rate of the continuous PSO algorithm and avoided fall into local extreme value easily. Numerical experiments show that, the method proposed in this paper is an effect and promising selective integration method.

Acknowledgment. This work is supported by National Natural Science Foundation of China, Grant Nos 61273275, 60975026, 61503407.

References

1. Schapire, R.E.: The strength of weak learnability. Mach. Learn. **5**, 197–227 (1990)
2. Breiman, L.: Bagging predicators. Mach. Learn. **24**(5), 123–140 (1996)
3. Zhou, Z.H., Wu, J.X., Tang, W.: Ensembling neural networks: many could be better than all. Artif. Intell. **137**, 239–263 (2002)
4. Bssis, N., Asimakopoulou, E., Xhafa, F.: A next generation emerging technologies roadmap for enabling collective computational intelligence in disaster management. Int. J. Space-Based Situated Comput. **1**, 76–85 (2011)
5. Zhou, Z.H., Wu, J.X., Tang, W., et al.: Combing regression estimators: GA-based selectively neural network ensemble. Int. J. Computat. Intell. Appl. **1**, 341–356 (2001)
6. Zhuming, W., Lin, T., Tang, N.: Explore the use of handwriting information and machine learning techniques in evaluating mental workload. Int. J. Technol. Hum. Interact. **12**, 18–32 (2016)
7. Yang, H.-C., Wang, X.A.: A study on components and Features in Face Detection. Int. J. Inf. Technol. Web Eng. **10**, 33–45 (2015)
8. Li, Z., O'Brien, L., Zhang, H.: Circumstantial-evidence-based effort judgement for web service composition-based SOA implementations. Int. J. Space-Based Situated Comput. **2**, 31–44 (2012)

9. Uchida, K., Takematsu, M., Lee, J.H., Honda, J.: A particle Swarm optimization algorithm to generate inhomogeneous triangular cells for allocating base stations in urban and suburban areas. Int. J. Space-Based Situated Comput. **3**, 207–214 (2013)
10. Zhang, H.D., Wang, X.D., Wu, C.M., et al.: Selective SVM ensembles based on modified BPSO. In: Proceedings of PACIIA 2008, pp. 243–246 (2008)
11. Li, Z.-R., Deng, Y.-H., Zhang, Q.: Research and implement of selective SVM ensemble classifier. Comput. Sci. **37** (2010)
12. Kennedy, J., Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: IEEE Conference on Systems, Man and Cybernetics, Orlando, FA, pp. 4104–4109 (1997)
13. Lin, H.-C.K., Su, S.-H., Wang, S.-T., Tsai, S.-C.: Influence of cognitive style and cooperative learning on application of augmented reality to natural science learning. Int. J. Technol. Hum. Interact. **11**, 41–66 (2015)
14. Mathiyalagan, P., Suriya, S., Sivanandam, S.N.: Hibrid enhanced ant colony algorithm and enhanced bee colony algorithm for grid scheduling. Int. J. Grid Util. Comput. **2**, 45–58 (2011)
15. RahmaBouaziz, F.K., Coulette, B.: C-SCRIP: Collaborative Security Pattern Integration Process. Int. J. Inf. Technol. Web. Eng. **10**, 31–46 (2015)
16. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 1942–1948. IEEE Service Center, Piscataway (1995)
17. Quan-de, Q., Rong-jun, L.: Two-population particle swarm optimization algorithm based on bioparasitic behavior. Control Decis. **26**, 32–36 (2011)
18. Wang, W.-B., Feng, Q.-Y.: Chaotic particle swarm optimization algorithm based on hierarchical multi-subpopulation. Control Decis. 25, pp. 26–30 (2010)
19. Zhao, S.Z., Liang, J.J., Suganthan, P.N.: Dynamic multi-swarm particle swarm optimizer with local search for large scale global optimization. In: Proceedings of IEEE Swarm Intelligence Symposium, Hong Kong, pp. 3845–3852 (2008)
20. Wang, H.U., Zhishu, L.I.: A simpler and more effective particle swarm optimization algorithm. J. Softw. **18**, 861–868 (2007)
21. Deyi, L., Jiawei, H., Xuemei, S.: Knowledge representation and discovery based on linguistic atoms. Knowl.-Based Syst. **15**, 431–440 (1998)
22. Mo, Y.-B., Chen, D.-Z., Hu, S.-X.: A complex particle swarm optimization for solving system of nonlinear equations. Inf. Control **35**, 423–427 (2006)
23. Zheng, C., Wang, X.-D., Zheng, Q.-D.: Self-adaptive escape simple particle swarm optimization algorithm based on cloud theory. J. Chin Comput. Syst. **31**, 1457–1460 (2010)
24. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap, pp. 1–4. Chapman & Hall, New York (1993)

The Research of QoS Monitoring-Based Cloud Service Selection

Ling Li[✉], Feng Ye, and Qian Huang

College of Computer and Information, Hohai University, Nanjing, China
43041260@qq.com

Abstract. With the rapid development of cloud computing, more and more consumers adopt third-party cloud services to implement their critical business. Cloud services are provided by different service providers, and usually have different cost performance and quality of service. Service consumers have to make trade-offs on multiple factors in order to choose the appropriate cloud services. Therefore, an efficient cloud service selection mechanism makes sense for both cloud service providers and consumers. According to existing works, this research field still requires appropriate models, effective design paradigm, in-depth experimentations and practical implementations. Because agent can identify the paradigm of the clouds by learning algorithm, they can be trained to observe the differences and behave flexibly for cloud service selection. To rank different clouds, we propose and assign QoS factor for each cloud service and ranks it as whole. According to simulation experiment, we validate the approach, which emphasizes the need to rank cloud services of widely spreading and complex domains.

1 Introduction

With the rapid development of cloud computing, more and more companies adopt third-party cloud services to implement their critical business applications. The different cloud services provided include Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a service (SaaS) [1]. Also the deployment model of cloud computing is various, which can be classified into four categories: public cloud, private cloud, hybrid cloud and community cloud. Because of the diversity of cloud service providers (CSP) and complexity of the cloud computing environment, cloud services usually have different cost performance and quality of service (QoS) [2]. Therefore, the common focus of CSP and service consumers mainly reflects two aspects: (1) how to provide dependable, scalable, sustainable and high cost-effective services for consumers; (2) how to choose the "suitable" one from more and more cloud services. Aimed at this situation, an efficient cloud service selection mechanism makes sense for both CSPs and consumers. If service consumers rely on single service provider' services, they may suffer with unstable QoS. Therefore, there is pressing need of comprehensive techniques for cloud service selection [3]. Though there are some techniques [4] available to select a particular service based on requirements, this research field still requires effective models, in-depth experimentations and practical implementations.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_28

The agent [5, 6] is a type of software component that acts for a user or other program in a relationship of agency. The agent has many prominent features, such as autonomy, proactivity, reactivity, communication and cooperation, negotiation and learning. With certain intelligent, agent allows people to delegate numerous works to them, and they can employ repetitive tasks, learning, compute complex data and make decisions in various fields [7–9]. QoS and intelligence will become the important requirements for cloud services selection. The incorporation of cloud computing and agent will be useful for providing high quality services to the service consumers. To this end, agent-base cloud service provider (ACSP) is proposed, which provides cloud services for consumers' requirements with rank. The rank shows the QoS of the cloud services in respect of the feedback. QoS factor for cloud services (QFCS) maintains the history of the services provided along with their feedbacks, and it assigns the normal services to QoS by giving them a QoS factor which identifies its ability assessment for the usage. QFCS improves the overall services and rank of the CSPs as well. The QoS factor is based on the feedback calculation monitored and performed on different parameters of cloud services such as cost, reliability, availability. The objective is to enable cloud consumers select cloud services according to their own requirements.

The paper is organized as follows. The related works are introduced in Sect. 2. The proposed algorithm for cloud service selection using agent is discussed in Sect. 3. The architecture of the ACSP proposed is presented in Sect. 4. In Sect. 5, we utilize CloudSim [10] to simulate the scenario for verifying our solution. The conclusion and future works are described at last.

2 Related Works

Nowadays, as the usage of water, electricity and gas, cloud resources are utilized to handle demanding computations and provide huge volumes of data storage in an efficient manner [11]. Issues rise like standardization, reliability, cost effectiveness, security, fault tolerance, service selection in cloud computing domain.

A range of studies has been carried out to develop advanced techniques that will assist service consumers to choose appropriate services, such as performance analysis or recommender systems. However, most managed cloud service information on some functions and performance remains static once they are registered into broker agents by respective CSP, and has little possibility to update their abilities to fulfil the service selections. To provide the most suitable services to the clients, agents play a vital contribution to cloud service selection, and they have been considered to provide automatic selection of cloud services [12]. For example, an agent-based semantic search engine for cloud service discovery is proposed, which is called Cloudle [13]. In [14], the researchers further proposed a dynamic cloud service selection model and strategy, which employ intermediary service agent model to help users select and integrate their cloud service requirements.

With the help of machine learning techniques [15], some researchers have employed machine learning techniques to develop an adaptive deployment policy, providing an optimal match between the customer demands and the available cloud service offerings [16, 17]. In [16], the authors employed machine learning techniques to develop an adaptive deployment policy, providing an optimal match between the customer demands and the available cloud service offerings. A machine learning based approach [17] was proposed to evolutionarily learn user preferences according to their ratings on historical execution plans for business processes that adapt to user preferences. We can see that the machine learning techniques play a very important role. However, the works above do not consider the different paradigm of cloud computing, especially the cloud services selection becomes more complex when cloud services are deployed on open source platform, such as OpenStack, Apache CloudStack.

According to existing works, this research field still requires appropriate models, effective design paradigm, in-depth experimentations and practical implementations. Hence, we propose an agent-base mechanism which could help consumers to look for appropriate cloud services. Moreover, it actively discovers rank for the clients according to their preferred criteria.

3 QoS Factor for Cloud Services

For efficient cloud services selection, quality attributes as non-functional aspect, are very important, e.g. cost, usability, reliability, availability. The learning mechanism is helpful to make best decision while providing cloud services on demand. Especially, the decision making becomes more and more important when there are more multiple choices under consideration. Therefore, we should consider implementing a mechanism where QoS can be monitored and service consumers can acquire the QoS status of CSP's cloud services for usage. Here, we propose an unsupervised learning technique which is based on Q-Learning [18] named QFCS. Q-learning can be used to find an optimal action-selection policy for any given markov decision process. It acquires positive reward every time by using learning rate and discount factor to move ahead. In our solution, we collect all types of feedback from past and current service consumers. These feedbacks may be positive/negative. The QoS factor is calculated on the basis of feedback provided by its users. The proposed mechanism stores and manipulates the information along with the quality attributes available in repository.

QFCS uses parameters of cloud computing like reliability, cost effective, availability. "Parameter repository (PR)" is a repository that containing parameters and allows the extension as well. The component of ACSP takes parameters from PR one by one and then performs its calculation on it. It generates the QoS factor of all CSP's cloud services and this collective performance proceeds towards QoS factor of CSP as a whole. ACSP monitors the QoS factor of each CSP against the defined threshold value and provide services to their consumers accordingly.

The QoS of the CSP depends on the QoS of all services provided individually. It is calculated as:

$$QoS_{CSP} = \sum\nolimits_{k=1}^{n} \left(qos_k(s)\right) \tag{1}$$

In (1): n is total number of cloud services provided by CSP.

And qos_k is QoS factor of single service w.r.t all attributes of QoS in PR and is de-fined as:

$$qos_{PR}(s) = \sum\nolimits_{i=1}^{PR} (f_i - v_i) \tag{2}$$

In (2): ‘PR’ is the number of repository attribute which belongs to QoS; v_i represents the value assigned by the CSP as a threshold value to a service S.

f_s is the feedback against the particular quality attribute of the cloud service and is defined as:

$$f_s = (f_{c_1} + f_{c_2} + \cdots + f_{c_N})/N, \;\; 0 < f_s < 100 \tag{3}$$

Where N is the total number of consumers, which is the value assigned by the CSP as a company threshold value to a specific service *S*.

When a service consumer gives the preferred requirements, ACSP finds the service providers in the cloud community which matches the demands. Once the CSP is selected, it checks each cloud service of the CSP one by one and calculates the feedback *fs*. Based on *fs*, it calculates the QoS factor *qos* according to Eq. (2) and adds it to the CSP’s rank. After the completion of the process applied on every CSP, we will found the rank of each CSP.

4 The Architecture of Agent-Based Cloud Service Provider

The proposed model is shown in Fig. 1.

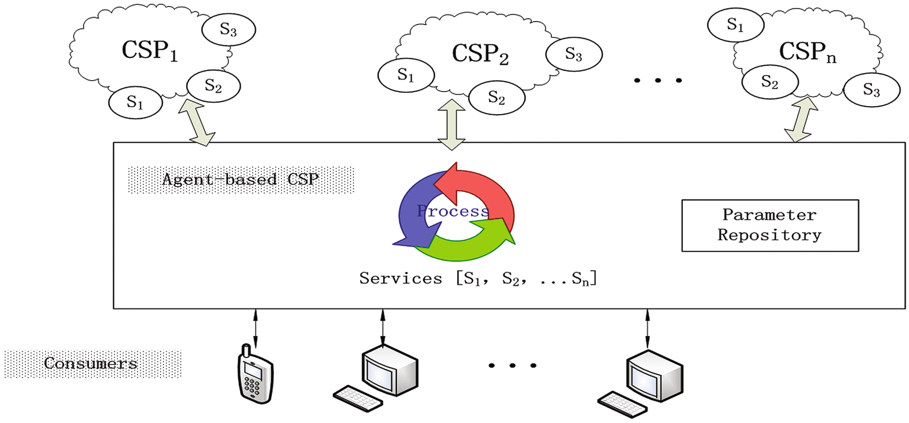


Fig. 1. The architecture of ACSP

There are many CSPs which can provide cloud services, and service consumers can utilize agent-based CSP to access to cloud services. The critical quality attributes are stored in parameter repository. When service consumers select the cloud services required, the results are shown to them along with service's rank and CSP's QoS factor. The Agent-based CSP interacts with different clouds according to the consumers' demands, and acquires cloud services from them, and then processes them as its internal feature. At last, it provides the list of cloud services along with QoS factor. The Agent-based CSP allows the service consumers to make best decision according to its business needs.

In order to demonstrate the details of Agent-based CSP, the components and interaction of Agent-based CSP is shown in Fig. 2. There are six critical components in Agent-based CSP, namely: Rank, Parameter Repository, Quality Facilitator, Service Selector, Feedback Unit and User Interface. User Interface is responsible for direct interaction with service consumers. Specifically, it can receive preferences and requirements of service from consumers, and send the results of service with rank and QoS factor to consumers.

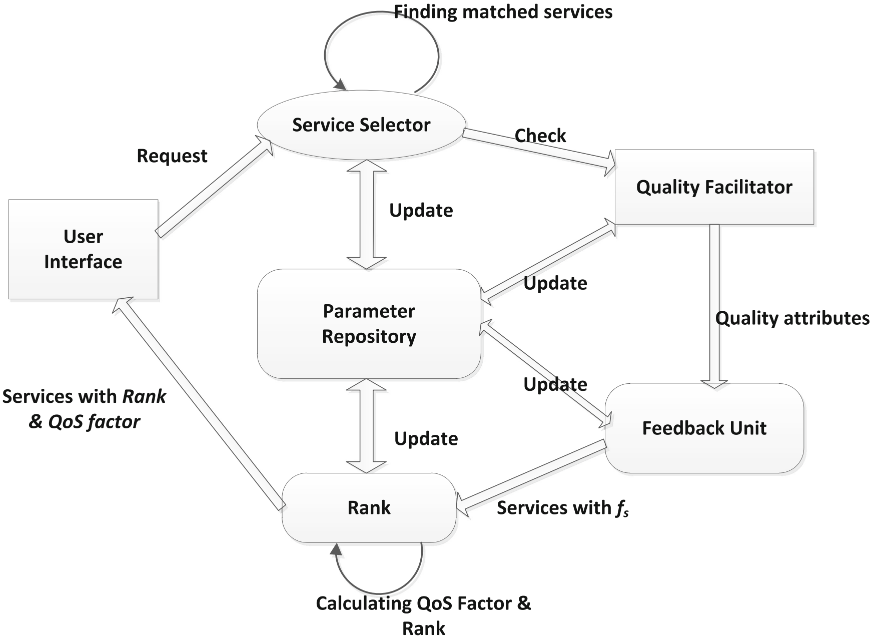


Fig. 2. The components of ACSP and interaction process

Service Selector is responsible for finding out the cloud services which can match with the service consumer's criterion from different CSP. It receives requirements from User Interface module and finds out the services and CSP in the cloud community. Service selection can be done with the help of multiple criteria techniques, such as [3].

Parameter Repository includes all the information regarding QoS attributes, and also maintains the feedback. It is core of the agent system and communicates with all parts of the CSP.

Quality Facilitator allows service consumers and providers to add quality attributes for service selection. Service Selector continually monitors and checks Quality

Facilitator. After selecting the appropriate attributes from Parameter Repository, it sends the array of QoS attributes to Feedback Unit.

Feedback Unit collects the feedback from the consumers found by the Service Selector and updates the Parameter Repository respectively. It contains the average values of the feedback of particular service based on QoS attributes.

Rank is used to calculate the rank/QoS factor of all the services and also performs the rank of the CSP. The rank utilizes feedback of each service that calculated on QoS attributes. It provides numeric value as QoS factor to visualize the rank of the service as well as CSP. The Rank and QoS factor of all its services is provided to the user interface.

5 Experiment and Results

We assume that there are three CSP's and every CSP have three services, namely S_1, S_2 and S_3. We use the proposed method on one of QoS attributes from the Parameter Repository. There are different columns in the Tables 1, 2 and 3, where C_1, C_2, C_3, C_4 and C_5 representing the feedbacks from different consumers. The column '*f*' presents the accumulative average feedback value according to Eq. (3). 'qos' is the QoS factor value of the service as per Eq. (2). The last row column 'Q' is the QoS factor/Rank of the CSP_i as whole. The positive values of 'Q' show that these service providers have the higher rank and their services are rated by their users.

Table 1. Cloud service provider CSP_1

Services	C_1	C_2	C_3	C_4	C_5	f	v	qos
S_1	65	70	70	75	65	69	65	4
S_2	60	80	75	65	70	70	70	0
S_3	65	80	70	70	70	70	70	2
Service	Q							6

Table 2. Cloud service provider CSP_2

Services	C_1	C_2	C_3	C_4	C_5	f	v	qos
S_1	80	90	75	80	75	79	80	0
S_2	90	70	65	75	70	74	80	−6
S_3	65	80	70	70	65	70	70	0
Service	Q							−6

Table 3. Cloud service provider CSP_3

Services	C_1	C_2	C_3	C_4	C_5	f	v	qos
S_1	70	80	85	65	70	74	80	4
S_2	75	75	70	75	70	72	80	2
S_3	80	85	90	75	65	82	70	7
Service	Q							13

Following tables represents CSP's services and their feedback of their services respectively. Consider the Table 2, where in first row, 80, 90, 75, 80 & 70 are the feedback values from five different clients/users. So, from Eq. (3), the f is calculated as:

$$\begin{aligned} &\mathrm{f} \ = \ (80+90+75+80+75)\,/5 = 80;\\ &\mathrm{v} \ = 80\,(\text{company value});\\ &\mathrm{qos} = \mathrm{f} - \mathrm{v} = 0;\\ &\text{QoS FactorQ} = \ ((0)\ +\ (-6)\ +0)\ =\ -6. \end{aligned}$$

Table 4 shows the form of QoS factor and ranks of all services of CSP's and as a whole. The first column presents services of the CSP's and remaining columns shows the calculated QoS factors of their respective services. The last row of the tables represents the rank of the different CSPs.

Table 4. CSP's QoS factor along with thier services

Service	CSP_1	CSP_2	CSP_3
S_1	0	4	4
S_2	−6	0	2
S_3	0	2	7
Rank	−6	6	13

From Table 4, we can easily see the service's status of different CSPs according to its requirements. In CloudSim, we further implement this scenario. The case study above verify that the methodology proposed works very well in cloud service environment where there is need to rank the services based on the requirements of service consumers.

6 Conclusion and Future Works

With an increasing number of cloud services, it is necessary to assist cloud consumers to choose the appropriate service. There may be some similar cloud services provided by different CSPs, and service consumers have to make trade-offs on multiple factors in order to choose the ideal cloud services. Therefore, we propose a solution based on QoS factor to select cloud service to its seekers which not only considers the user requirements but also quality of services and feedback of several other clients. Through the simulation, the method proposed enhances the level of the cloud service selection in an efficient manner.

In future, we need carry out more experiments in real cloud services environment, and do more comparison with other methods.

References

1. Chang, S.F.: A reference architecture for application marketplace service based on SaaS. Int. J. Grid Util. Comput. (IJGUC) **2**(4), 243–252 (2011)
2. Serhani, M.A., Atif, Y., Benharref, A.: Towards an adaptive QoS-driven monitoring of cloud SaaS. Int. J. Grid Util. Comput. (IJGUC) **5**(4), 263–277 (2014)
3. Rehman, Z.U., Hussain, F.K., Hussain, O.K.: Towards multi-criteria cloud service selection. In: Proceedings of 2011 5th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 44–48. IEEE Computer Society, Seoul (2011)
4. Afify, Y.M., Moawad, I.F., Badr, N.L., Tolba, M.F.: Cloud services discovery and selection: survey and new semantic-based system. In: Hassanien, A.E., Kim, T.H., Kacprzyk, J., Awad, A.I. (eds.) Bio-inspiring Cyber Security and Cloud Services: Trends and Innovations, pp. 449–477. Springer, Heidelberg (2014)
5. Cetnarowicz, K.: A Perspective on Agent Systems: Paradigm, Formalism, Examples. Springer, Cham (2015)
6. Jezic, G., Chen-Burger, J.Y.H., Howlett, R.J., Jain, L.C. (eds.): Agent and Multi-Agent Systems: Technology and Applications. Springer, Cham (2016)
7. Seghir, N.B., Okba, K., Rezeg, K.: A decentralized framework for semantic web services discovery using mobile agent. Int. J. Inf. Technol. Web Eng. (IJITWE) **10**(4), 20–43 (2015)
8. Dam, H.K., Ghose, A., Qasim, M.: An agent-mediated platform for business processes. Int. J. Inf. Technol. Web Eng. (IJITWE) **10**(2), 43–61 (2015)
9. Chen, H.P., Li, S.C.: SRC: A service registry on cloud providing behavior-aware and QoS-aware service discovery. In: Proceedings of 2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA), pp. 1–4. IEEE Computer Society, Perth (2010)
10. Wickremasinghe, B., Calheiros, R., Buyya, R.: Cloudanalyst: a cloudsim-based visual modeller for analysing cloud computing environments and applications. In: Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications (AINA 2010), pp. 20–23. IEEE Computer Society, Perth (2010)
11. Domenico, T.: Cloud computing and software agents: towards cloud intelligent services. In: Proceedings of the 12th Workshop on Objects and Agents, pp. 2–6. Sun SITE Central Europe CEUR-WS, Rende (2011)
12. Sun, L., Dong, H., Hussain, F.K., Hussain, O.K., Chang, E.: Cloud service selection: state-of-the-art and future research directions. J. Netw. Comput. Appl. **45**, 134–150 (2014)
13. Sim, K.M.: Agent-based cloud computing. IEEE Trans. Serv. Comput. **5**(4), 564–577 (2012)
14. Wang, X.: A dynamic cloud service selection strategy using adaptive learning agents. Int. J. High Perform. Comput. Netw. **9**(1/2), 70–81 (2016)
15. Kubat, M.: An Introduction to Machine Learning. Springer, Cham (2015)
16. Samreen, F., Elkhatib, Y. Rowe, M., Blair, G.S.: Daleel: Simplifying cloud instance selection using machine learning. In: Proceedings of 2016 IEEE/IFIP Network Operations and Management Symposium, pp. 557–563. IEEE, Istanbul (2016)
17. Kang, D.S., Liu, H., Singh, M.P., Sun, T.: Adaptive process execution in a service cloud: service selection and scheduling based on machine learning. In: Proceedings of 2013 IEEE 20th International Conference on Web Services, pp. 324–331. IEEE Computer Society, Santa Clara (2013)
18. Watkins, C.J.C.H.: Learning from delayed rewards. Ph.D thesis, University of Cambridge, England (1989)

Developing Cloud-Based Tools for Water Resources Data Analysis Using R and Shiny

Feng Ye[1(✉)], Yong Chen[2], Qian Huang[1], and Ling Li[1]

[1] College of Computer and Information, Hohai University,
Nanjing 211100, China
yefengl022@hhu.edu.cn
[2] Nanjing Longyuan Microelectronic Company, Nanjing 211106, China

Abstract. Aimed at developing and utilizing the water resource appropriately, it is critical to analyze, mine and present the valuable information and knowledge. Until recently, analyzing the big data in an online environment has not been an easy task especially in the eyes of data consumers in water conservancy domain. Moreover, there is no single tool or a one-size-fits-all solution for big data processing and data visualization in a special field. This barrier is now overcome by the availability of cloud computing, R and Shiny. In this paper, we propose to develop cloud-based tools for water resource data analysis using R and Shiny. Following the whole solution, the implementation using long-term hydrological data collected from Chu River is introduced as an example. The results show that these tools are valuable and practical resource for individuals with limited web development skills and offer opportunity for more dynamic and collaborative water resource management.

1 Introduction

Water resource is the important foundation of human survival and development. Along with the economic development and population growth, the global demand for water is increasing day by day. However, the processes of development are accompanied by severe water waste and pollution. As the largest developing country, China faces many water resource problems to be solved, including: (1) the water resource is unevenly distributed in area and time, so it directly affects the whole planning and regional development; (2) the extensive management mechanism restricts the efficient utilization of water resources; (3) the situation of water pollution is not optimistic, and the flood and drought control still remain a daunting task.

In order to develop and utilize water resources reasonably, it is necessary to quantify and compute water resource data from the space-time dimension, then acquire the valuable information and knowledge by data analyzing, mining and presentation. With the advent of the era of big data [1], business data is also being increasingly cheap to collect and maintain, and the users expect to acquire real-time, intuitive data analysis result for making decisions. However, there are three main technical challenges: (1) bulk of domain business data that are collected by the sensors or other open data sources need to be analyzed and presented in an efficient, transparent and integrated manner; (2) there is demand for more decision-making tools in water resources,

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_29

however, they are often difficult to come by because of the divergent skill sets of scientists, engineers and decision makers; (3) analyzing the big data in an online environment has not been an easy task especially in the eyes of data consumers in water conservancy domain. Existing studies suggest that how to implement data-intensive [2] knowledge discovery architecture in the specific domain is critical and there is no single tool or a one-size-fits-all solution for big data processing and visualization, but until now there are only some theoretical researches without any detailed use cases for water resource analysis and mining [3–5]. This barrier is now overcome by the availability of cloud computing, R [6–8] and Shiny [9], which can provide a novel solution for implementing web-based online data analysis. With the rapid development of R language and gaining popularity among big data and cloud computing [10–12] community, creating cloud-based tools for water resources data analysis becomes a new opportunity. Therefore, we propose to utilize cloud computing and open software related for provisioning resource and infrastructure for data processing, and use R and Shiny to create web-based data analysis and mining applications online.

The remainder of this paper is structured as follows. Section 2 introduces the related works; we introduce the key technologies in Sect. 3. In Sect. 4, the solution of developing cloud-based tools for water resources data analysis using R and Shiny are proposed. And then, the implementation using long-term hydrological data collected from Chu River is introduced as an example. Finally, the conclusion and the future works are described at last.

2 Related Works

In water conservancy domain, business data are collected by the sensors, samples or other open data sources, such as hydrological steaming data and weather data. How to process the large scale of data and how to implement online manner are the challenges. In [3], the authors discussed the concept “Smart Basin” with the whole architecture and possible application scenarios; similarly, Wang, Z.J. [4] introduced “The Internet of Water” from general framework, core functions to key technologies. Ai [5] put forward a framework for processing water resources big data, to process and analyze modern water resources data rapidly, and discussed the related application. However, all of them are still in the exploratory stage, and there are not specific use cases or tools.

For data analysis, the popularity of R language provides the new chance. In [6], the authors explain that why R is a better integrated solution for the data processing using a stylized example. Many applications [13–18] adopt R and Shiny to develop web-based data analysis application. For example, the authors introduce a web-based decision support tool (ViRTUE) [13] for performing climate risk evaluations of water supply systems. In agriculture domain, web-based data analysis tools for precision farming using R and Shiny are integrated. AMADA [18] is an interactive web application to analyze multidimensional datasets, which provides a number of clustering visualization diagnostics such as heatmaps, chord diagrams and graphs. However, the works above still cannot consider the problem of big data. Because of the complexity, there is no single tool or one-size-fits-all solution for deeply mining, analyzing and visualizing the big data of specific domain.

To sum up, the requirement of developing tools for water resources data analysis is necessary. Using R and Shiny are appropriate for data consumers of water conservancy domain, which can provide an efficient, transparent and integrated manner. Most of all, the tools need to have the ability of dealing with big data.

3 Key Technologies

In order to tackle the challenge of big data mining and analyzing, we consider to integrating many different key tools and technologies, from infrastructure resource management to rich statistical computation and graphic functions.

3.1 Cloud Computing and Apache CloudStack

Cloud computing [19] is defined as "A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [12]. Therefore, cloud computing can serve as effective platforms for addressing the computational and data storage needs of most big data analytics applications.

As a highly available, scalable open source Infrastructure-as-a-Service platform, Apache CloudStack [20] is designed to deploy and manage large networks of virtual machines. The purpose is to utilize Apache CloudStack to manage infrastructure resource by constructing a private cloud, and then integrate relational and NoSQL database for storing multiple datasets. Therefore, it can serve as an effective paradigm for addressing both the computation and data storage requirements of big data mining and analyzing applications.

3.2 Map-Reduce Model, Hadoop and Cascading

Map-Reduce [21, 22] is a programming model for expressing distributed computations on massive amounts of data and an execution framework for large-scale data processing on clusters of machines. It consists of two common built-in functions are Map and Reduce. The Map function receives a key/value pair as input and generates intermediate key/value pairs to be further processed. The Reduce function merges all the intermediate key/value pairs associated with the same (intermediate) key and then generates final output. Nowadays, it has been widely adopted to implement scalable data analysis algorithms and applications executed on multiple machines, to efficiently analyze big amounts of data.

Hadoop is open source, popular Map-Reduce implementation and runs well in the cloud. Based on this reason, Hadoop can be deployed on virtual machines cluster of the private cloud using Apache ClouadStack and adopted for providing distributed big data processing ability.

However, the skills required to write large Hadoop programs directly in Java are difficult to learn for most developers and far outside the norm of expectations for analysts. In order to solve this problem, we can integrate Cascading [23]. Using Cascading is an alternative approach: (1) It significantly simplifies and streamlines application development, job creation, and job scheduling, which provide a declarative way of specifying the high level logic of an application of Apache Hadoop and other big data frameworks, hiding the low level details. (2) It can integrate existing software modules, datasets, and services in complex compositions. (3) It supports nearly every type of data source.

3.3 R and Shiny

As a free software, R has a wealth of statistical computing and data presentation capabilities, and have become a fairly flexible and powerful tool for various data analysis, mining and graphical presentation.

Currently, the CRAN package repository features 10085 available R packages, including the time series analysis, graphic displays & dynamic graphics, analysis of ecological and environmental data, and so on. According to [24], we list the some R package related in Table 1, and more packages can be integrated into the system and became a new tool.

Among them, Shiny is the perfect companion to R, making it quick and simple to share analysis and graphics from R that users can interact with and query over the Web. With the shiny package, ordinary controllers or widgets are provided for ease of use of application programmers. Many of the procedures like uploading files and refreshing the page for drawing new plots and tables are provided automatically. These tasks are done based on the pre-built output widgets. The communication between the client and server is done over the normal TCP connection.

4 The Architecture Proposed

The architecture proposed is shown in Fig. 1, and there are four tiers: infrastructure layer, virtualization layer, dataset processing layer and web-based user interface. Infrastructure layer provides the hardware foundation for big data processing, such as PCs, various servers and network equipment. Various low-level infrastructure resources are abstracted into different resource pools, such as data resource pool, network resource pool.

In virtualization layer, Apache CloudStack is installed, configured and deployed to construct virtual machines cluster and then used to manage the infrastructure resource. Above the virtualization layer, it is dataset processing layer. In this layer, Hadoop, Cascading and R language runtime environment are installed and deployed. According to a variety of business requirements, different data management solutions are applicable for different data sizes or types, for example: if local resource of single virtual machine is sufficient for data processing, it is not necessary to use Hadoop. And it provides business logics for interactive manipulation between R and Cascading. At last, in order to hide the complexity of the R language, the architecture provides WYSIWYG Web-based user interface using Shiny.

Table 1. Some R packages related

Category	Packages	Functions
Hadoop-related	plyrmr	**plyrmr** enables the R user to perform common data manipulation operations on very large data sets stored on Hadoop
	rmr	**rmr** package allows an R user to perform statistical analysis via MapReduce on a Hadoop cluster
	rhdfs	**rhdfs** package provides basic connectivity to the Hadoop Distributed File System. R users can browse, read, write, and modify files stored in HDFS
	rhbase	**rhbase** package provides basic connectivity to HBASE. R users can browse, read, write, and modify tables stored in HBASE
	……	**more packages**
Hydrology-related	hyfo	**hyfo** package provides hydrology and climate forecasting
	dynatopmodel	**dynatopmodel** package provides implementation of the dynamic TOPMODEL hydrological model
	getMet	**getMet** package provides getting meteorological data for hydrological modeling.
	hydrostats	**hydrostats** package calculates a suite of hydrologic indices for daily time series data that are widely used in hydrology and stream ecology.
	rivervis	**rivervis** package is a flexible and efficient tool to visualise both quantitative and qualitative data from river surveys. It can be used to produce diagrams with the topological structure of the river network
	waterData	**waterData** package is used for retrieval, analysis, and anomaly calculation of daily hydrologic time series data
	……	**more packages**
Graphics and Web-based UI for hydrological data	shinydashboard	**shinydashboard** package provides a theme on top of ‘Shiny’, making it easy to create attractive dashboards
	miniUI	**miniUI** package provides UI widget and layout functions for writing Shiny apps that work well on small screens
	shiny	**shiny** package makes it incredibly easy to build interactive web applications with R
	……	**more packages**

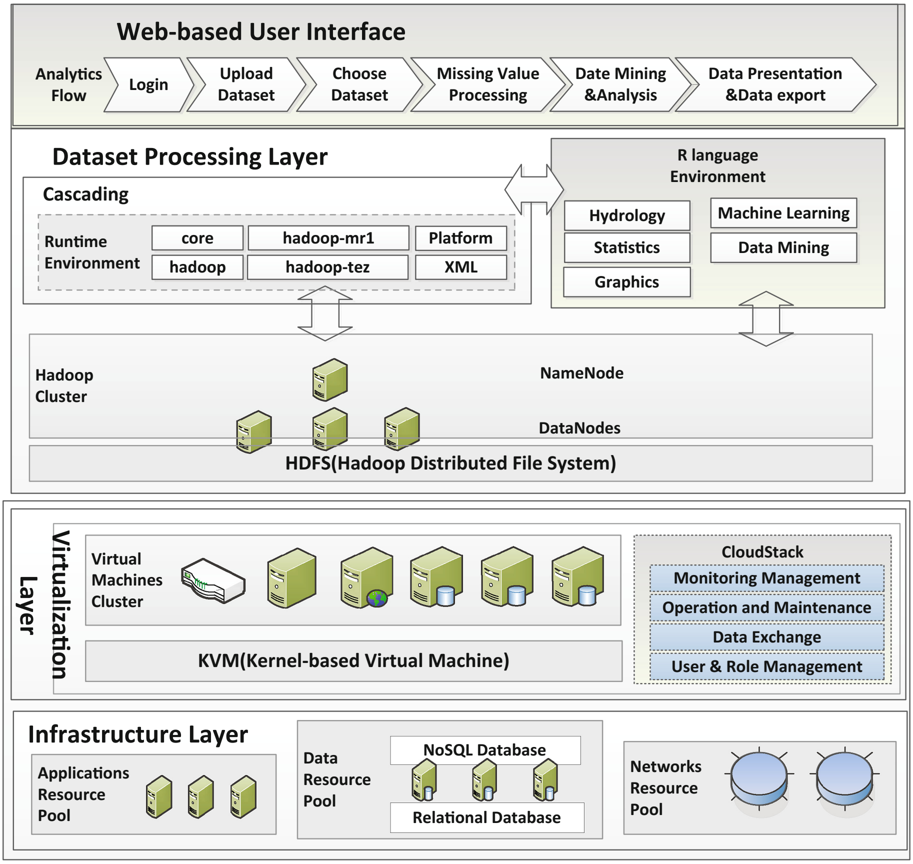


Fig. 1. The infrastructure of cloud-based tools for water resources data analysis using R and Shiny

The basic idea of data visualization technology is the database for each data item as a single pixel element represents, then a large number of data sets constitute image of data. Meanwhile using multi-dimensional data to represent each attribute value of data, the data can be observed from different dimensions and used more in-depth observation and analysis. The main purpose of data visualization is to convey information by using graphical tools and communicating clearly and effectively. Due to the different degrees of data, data visualization [25] must implement the zoom feature. At the same time, users can browse or specific knowledge about the data set using the dynamic response graphics.

5 Use Case

Chu river is an important tributary, which is the focus of flood and drought control of Nanjing. Currently, the administrative department of Chu river has accumulated a long sequence of hydrological data since 1960, and more than 70 sensors for rainfall and water lever value continuously collect data. The management departments want to find out similar sequence in order to use efficacious and rational emergent plan before. Similarity mining of hydrological time series is an important aspect of hydrological time series mining. It will be of great important in flood forecasting and flood control scheduling.

Therefore, based the methodology introduced in Sect. 4, we try to develop similarity mining of hydrological time series. According to the characteristics of hydrological data, a DTW [26] clustering-based similarity mining method over hydrological time series is implemented. Firstly, on the premise of wavelet de-noising, feature point segmentation and semantic classification, hierarchical cluster analysis were used to the classification of sub-sequences based on DTW distance and the sub-sequences were symbolized. Then, we filtered candidate sets of time series according to the edit distance between symbol sequences. Finally, we got the similar hydrological time series precisely from the candidate sets by DTW exact matching. In this process, the methods and functions in dtwclust and dtw package are integrated. The experimental results show that the method proposed can narrow the candidate sets effectively and improve the efficiency of searching for semantic similarity of hydrological time series in Fig. 2.

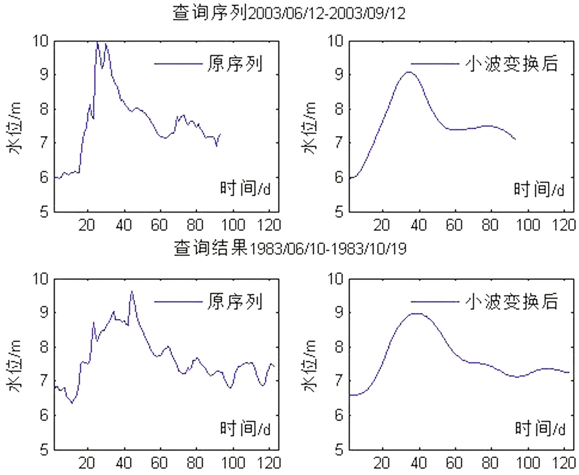


Fig. 2. Similarity analysis result of hydrological time series of Chu river

6 The Conclusion and Future Work

In this paper, we report our own experiences in developing cloud-based tools for water resources data analysis. Following the whole solution, the implementation using long-term hydrological data collected from Chu River is introduced as an example. The results show that these tools are valuable and practical resource for individuals with limited web development skills and offer opportunity for more dynamic and collaborative water resource management.

The better management and analysis of big data in special domain will become the next frontier of innovation, competition and productivity. And multidisciplinary and transdisciplinary effort continues to deliver new techniques and tools for the analysis of very large collections of data. The future directions of our work will focus on two aspects: (1) combine more business requirements and utilize the tools to analyze the dataset related; (2) develop more data statistical and analysis functions which is based on Map-Reduce programming model.

Acknowledgments. The research is supported by (1) National Natural Science Foundation of China (61300122); (2) A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions; (3) Water Science and Technology Project of Jiangsu Province (2013025); (4) Central University of basic scientific research and business fee of Hohai University (2009B21614).

References

1. Kudyba, S.: Big Data, Mining, and Analytics. CRC Press, Boca Raton (2014)
2. Hey, T., Tansley, S., Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Washington (2009)
3. Jiang, Y.Z., Ye, Y.T., Wang, H.: Smart basin and its prospects for application. Syst. Eng. Theor. Pract. **31**(6), 1174–1181 (2011)
4. Wang, Z.J., Wang, G.Q., Wang, J.H.: Developing the internet of water to prompt water utilization efficiency. Water Resour. Hydropower Eng. **44**(1), 1–6 (2013)
5. Ai, P., Yue, Z.X.: A framework for processing water resources big data and application. Appl. Mech. Mater. **519–520**, 3–8 (2014)
6. Munzert, S., Rubba, C., Meißner, P., Nyhuis, D.: Automated Data Collection with R. Wiley, Chichester (2015)
7. Kabacoff, R.I.: R in Action: Data Analysis and Graphics with R. Manning, New York (2011)
8. Gohil, A.: R Data Visualization Cookbook. Packet Publishing, Birmingham (2015)
9. Beeley, C.: Web Application Development with R Using Shiny. Packet Publishing, Birmingham (2013)
10. Oril, A.: R for Cloud Computing: An Approach for Data Scientists. Springer, New York (2014)
11. Buyya, R., Broberg, J., Goscinski, A.: Cloud Computing: Principles and Paradigms. Wiley, Hoboken (2011)
12. Wang, L., Ranjan, R., Chen, J., Benatallah, B.: Cloud Computing: Methodology, Systems and Applications. CRC Press, Boca Raton (2012)
13. Whateley, S., Walker, J.D., Brown, C.: A web-based screening model for climate risk to water supply systems in the northeastern united states. Environ. Model Softw. **73**, 64–75 (2015)
14. Boelaert, J., Bendhaiba, L., Olteanu, M., Vialaneix, N.V.: SOMbrero: an R package for numeric and non-numeric self-organizing maps. In: Villmann, T., Schleif, F.M., Kaden, M., Lang, M. (eds.) WSOM 2014. Advances in Intelligent Systems and Computing, vol. 295, pp. 219–228. Springer, Heidelberg (2014)
15. Wojciechowshi, J., Hopkins, A.M., Upton, R.N.: Interactive pharmacometric applications using R and the Shiny package. Pharmacometrics Syst. Pharmacol. **4**, 2–3 (2015)

16. Jahanshiri, E., Shariff, A.R.M.: Developing web-based data analysis tools for precision farming using R and Shiny. In: 7th IGRSM International Remote Sensing & GIS Conference and Exhibition, pp. 1–6. IOP Publishing, Kuala Lumpur (2014)
17. Nisa, K.K., Andrianto, H.A., Mardhiyyah, R.: Hotspot clustering using DBSCAN algorithm and shiny web framework. In: Proceedings of 2014 International Conference on Advanced Computer Science and Information Systems, pp. 129–132. IEEE (2014)
18. de Souza, R.S., Ciardi, B.: AMADA—analysis of multidimensional astronomical datasets. Astron. Comput. **12**, 100–108 (2015)
19. Furht, B., Escalante, A.: Handbook of Cloud Computing. Springer, New York (2010)
20. Apache Cloudstack. http://cloudstack.apache.org/
21. Lin, J., Dyer, C.: Data-Intensive Text Processing with MapReduce. Morgan & Claypool Publishers, College Park (2010). 1−64
22. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Commun. ACM **51**(1), 107–113 (2008)
23. Nathan, P.: Enterprise Data Workflows with Cascading. O'Reilly Media, Sebastopol (2013)
24. Prajapati, V.: Big Data Analytics with R and Hadoop. Packt Publishing, Birmingham (2013)
25. Zabukovec, A., Jaklic, J.: The impact of information visualization on the quality of information in business decision-making. Int. J. Technol. Human Interact. (IJTHI) **11**(2), 61–79 (2015)
26. Zhu, Y.L., Wang, Y.M., Wan, D., et al.: Similarity mining of hydrological time series based on semantic similarity messures. J. China Hydrol. **31**(1), 35–40 (2011)

Perception Mining of Network Protocol's Dormant Behavior

Yan-Jing Hu[1,2](✉)

[1] Key Laboratory of Cryptology & Information Security
Under the Chinese PLA, Engineering University of the Armed Police Force,
Xi'an 710086, China
huyanjing2007@126.com

[2] State Key Laboratory of Integrated Services Networks,
Xidian University, Xi'an 710071, China

Abstract. Unknown network protocol's dormant behavior is becoming a new type of stealth attack, which greatly harms the cyber space security, and seriously affects the credibility of network protocols. By studying the characteristics of the dormant behavior, we discovered that protocol's dormant behavior is different from the normal one in instruction level. The protocol's behaviors can be represented by the labeled instruction sequences. The similar instruction sequences means the similar protocol behavior, but the dormant behaviors are the ones that can only be triggered under particular conditions. The main frame of perception mining network protocol's dormant behavior, and sensitive information automatic generation are proposed in this paper. Using our proposed instruction clustering algorithm, all kinds of executed or unexecuted instruction sequences can automatic clustering. On the basis of this, we can effectively distinguish and mine the protocols' potential dormant behaviors. A virtual protocol behavior analysis platform HiddenDisc has been developed, and the protocol execution security evaluation scheme is proposed and implemented. Through the analysis and evaluation of the 1297 protocol samples, the experimental results show that the present method can accurately and efficiently perception mining unknown protocol's dormant behaviors, and can evaluate the protocol's execution security.

Keywords: Perception mining · Dormant behavior · Stealth attack

1 Introduction

The protocol reverse technologies have evolved phenomenally over the last few years. As the rapid development of dormant protocol flourishes, there is an exponential increase in the number and diversity of protocol's dormant behaviors [1]. A crucial question in protocol reverse analysis research is how to percept and mine protocol's dormant behavior instructions or signatures that faithfully describe similar behavior intent and also clearly stand out from other behaviors. Network protocol's dormant behaviors can be considered as any activities or phenomenon which don't behave as expected. Such behaviors are usually caused by protocol design bugs, misconfigurations, network attacks, and etc. It is considered that such a behavior can be

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_30

either benign or malicious, but the effects may be calamitous. Some important confidential information can be stolen by attackers through protocol's dormant behavior, even secretly controlling the target hosts. Try our best to find out the principles of dormant behavior, intensive study the methods and techniques to prevent such behavior is our main motivation. Most previous works focus on identifying malicious behaviors in protocols, further practice is automatically classifying malware into known families [2–4], but the biggest flaw is such methods are usually difficult to mine dormant behaviors quickly and effectively. The main reason for such works are lack of research on protocol's dormant behavior perception, we believe, is that dormant behaviors do not always appear, and the growing invisibility, robustness and survivability makes it difficult to distinguish between good and evil. Although the traditional protocol reverse approach based on message or instruction signature analysis are efficient, they can be easily defeated by various obfuscation techniques. Instruction clustering analysis technologies opened the door for implementing innovative, smarter and interactive analysis applications and architectures.

The closest study to ours is protocol reverse engineering. Related work in such area can be divided in two categories: static analysis and dynamic binary analysis. Recent works [5, 6] have shown that static analysis is cumbersome or even impossible to some extent due to the advanced code obfuscation techniques, and they are always applied in dormant behavior obfuscation. The dynamic analysis method executes the executable program for a limited time, thus monitoring and examining the instruction sequences. Unfortunately, dynamic analysis is limited by scalability issues, dedicated hardware and software is used to analyze a single protocol at one time [7–9]. Therefore, neither single static nor dynamic analysis method is able to deal with such huge number of protocols in a timely manner, and also cannot expose all the dormant behaviors.

The goal of the proposed work is to provide the methods to percept and mine the network-wide dormant behaviors using a protocol's machine-level instruction features. We have adapted and extended several existing binary code analysis methods to render them scalable for processing the protocols' dormant behaviors. We have developed the behavior distance-driven instruction clustering technologies for automatically exposing protocol's dormant behaviors. We run the proposed algorithm to analyze the protocol's dormant behaviors, and also developed a virtual analysis platform HiddenDisc. Many types of real-world protocol samples were analyzed in the analysis platform. The platform can assist security analysts or network administrators to identify potential protocol behavior anomalies. Experimental results show that using the analysis platform to mine dormant behaviors and their variants from unknown protocols is favorable and effective.

This paper is organized as follows. In Sect. 2 we discuss several related works that have been recently proposed in protocol reverse. In Sect. 3, we describe the architecture and our proposed analysis approach. The experiments and evaluation are demonstrated in Sect. 4. Finally, We conclude the paper in Sect. 5.

2 Related Work

The variety of protocol reverse analysis techniques can be divided into two layers according to the protocol specification, such as format, semantic mining in packet and behavior mining among packets [7, 10–15]. Most protocol reverse research focused on extracting protocol's format information, such as work [7], which can extract command-and-control (C&C) protocol's message structure by dynamic binary analysis. The limitation is it can only extract the monitored instruction sequences, and infer the message format from syntax section, cannot solve the protocol's behavior analysis issue, more importantly, the protocol's dormant behaviors cannot be explored from protocol's message format. Ying Wang [14] improved it with dynamic combined with static binary analysis. Experimental results show that the prototype tool can not only extract the message format, but also can speculate the state machine through relevant field attributes. But the limitation is this approach depends on the captured instruction sequences, the dormant behavior which never appears in the captured instruction sequences cannot be discovered. Although the process is not a once-and-done effort, we can still obtain rich information about closed (undocumented) protocols by protocol reverse engineering [16]. The protocol reverse analysis results have been applied to many network security applications and managements [17–21]. Most previous works focus on trying to translate the undocumented protocol to a readable formant [12, 15, 22, 23], but they are usually difficult to percept and discover the protocol's behavior information. Due to the limited behavior knowledge, we are always difficult to understand protocol's behavior, not to mention the detection of protocol's dormant behaviors. If there is no such important information, we cannot evaluate the protocol's execution security.

The protocols always have two types of resources available: one is protocol message [24, 25] and the other is binary program that implemented the protocol [26, 27]. With such resources, the protocol reverse approaches can be categorized into protocol message-based, program binary-based and hybrid analysis. [10]

Works [24, 28] mainly focus on the analysis of protocol message. The protocol fields, separators and several special information are identified by capturing and analyzing a huge number of network flows. The sequence alignment algorithm that has been used in bioinformatics for pattern discovery are also widely used in protocol reverse today. Although these techniques can obtain the message fields information, what the protocol's behavior like, furthermore, whether it has dormant behavior cannot be discovered from the message flow analysis. Work [23] got protocol's format information by clustering network flows. The limitation is the lack of semantic information in network flows. Such methods are powerless for encryption protocols, and the protocol's dormant behaviors still cannot be discovered from message information. Work [22] focuses on identifying and clustering different types of messages not only based on their structures, but also depending on the impact of each message on server behaviors. A conception of input similarity is related to execution similarity is presented in the work. That means, the program codes which process the same structure of network flows should also be similar, such as the same code blocks, libraries and system calls. Unfortunately, such approach does not consider the protocol's dormant behavior too.

Protocol's dormant behaviors are usually hidden in the program codes. The dormant behavior can be triggered only when special conditions are satisfied, and the network message it processed is the same structure. For example, when a specific time in a network message is captured, only depend on the message information, we are difficult to determine whether it is a normal time stamp or a trigger condition that can trigger certain behavior. The fact is that the message format can be the same, but the executed instructions can be quite different. This means that the same message structure does not always correspond to the same behavior. The main task of protocol dormant behavior analysis is to discover and explore all the dormant behavior instruction sequences.

The flaw of message-based analysis approaches are lack of semantic information, so the researchers began to focus on the binary code analysis methods. The binary code that implemented the protocol contains rich information. It is usually the most detailed description of a protocol. Polyglot [29] is able to extract the received message format by monitoring how the protocol program processes the message data, but it cannot deal with the sent message, and it does not consider the protocol behavior analysis, too. In order to compensate for Polyglot, Juan Caballero proposed a new tool Dispatcher [30], the enhanced version of Polyglot. There are two advantages in Dispatcher, one is it can extract the protocol format from both sender and receiver when only one endpoint's implementation is available. The other advantage is it has the ability to process encrypted protocols. Although the protocol's binary codes are used in Dispatcher, it can only infer the protocol's format, and it still cannot analyze the protocol's dormant behaviors.

Protocol's dormant behavior is certainly different from the normal one in machine-level instructions. However, such dormant behavior information is usually difficult to mine. Protocol's dormant behaviors are always implemented by special instructions, which can be hidden in either network message or protocol binary code, but the message data usually has no obvious abnormal. The former idea has been applied in more and more malware protocols [31], but the latter one seem to be rarely studied from published literatures. In order to resist reverse analysis, the protocol message data and program codes are always protected by special technologies such as control flow obfuscation or encryption. Encryption can be conquered by searching and analyzing the decryption process, such approaches are proposed in literatures [7, 28, 30], but existing methods are usually powerless to obfuscation. The seemingly "normal" protocol messages can guide the normal communication functions execute, and also can trigger dormant behaviors under certain circumstances. So only analyze the captured protocol message or binary code alone is not effective for protocol's dormant behavior analysis.

If a protocol contains dormant behaviors, the process of protocol message parsing can be divided into normal part and dormant part. A great quantity of case studies show that the dormant behavior is different from a normal behavior in instruction type and execution frequency. Our key technology is locate, judge and analyze dormant instructions by clustering all the instruction sequences no matter whether they are captured or hidden. This means a virtual protocol behavior analysis platform and hybrid analysis solution is required. With such an analysis platform, we can monitor the detail of message parsing process in a controlled environment, analyze the behavior. At the same time, all the functional instruction sequences are recorded and tracked in the platform.

3 Design of the Protocol Behavior Analysis System

As shown in Fig. 1, it is the description of our proposed protocol dormant behavior virtual analysis system. The system is designed to ensure the protocol's execution security.

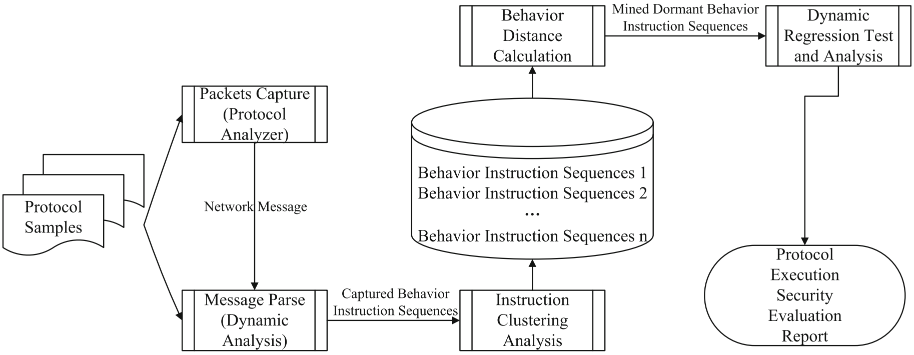

Behavior Distance Calculation
Mined Dormant Behavior Instruction Sequences
Dynamic Regression Test and Analysis
Packets Capture (Protocol Analyzer)
Protocol Samples
Network Message
Behavior Instruction Sequences 1
Behavior Instruction Sequences 2
...
Behavior Instruction Sequences n
Protocol Execution Security Evaluation Report
Message Parse (Dynamic Analysis)
Captured Behavior Instruction Sequences
Instruction Clustering Analysis

Fig. 1. The protocol dormant behavior analysis system

A protocol's behaviors are considered as a set of all the instruction sequences. A behavior can be expressed as one or more instruction sequences. A novel method to analyze the protocol behaviors based on clustering all the instruction sequences are proposed in this section. Our proposed protocol dormant behavior analysis system is consisted of six major functional modules, which are shown in Fig. 1.

(1) Network packets capture and analyze
The "RADCOM" hardware analyzer and "Wire shark" software protocol analyzer are both used as the network packets capture tools. This is more conducive to capture the active raw data generated by the protocol. The traffic raw data is captured appropriately in order to be reverse analyzed.

(2) Dynamic analysis of the message parsing process
The protocol's main behaviors are related to the message parsing process, and the instruction sequences are considered as a predictor for protocol's behaviors from the microscopic point of view. The dynamic program binary analysis techniques are used in this module to monitor how the protocol program parses the network message. Then the behavior instruction sequences (the machine level operation codes) are captured and analyzed.

(3) Clustering analysis of the instructions
The captured behavior instruction sequences can be considered as a part of the protocol's behaviors. With the help of instruction clustering algorithm, protocol's similar behavior patterns can generate a same cluster according to the known captured behaviors. The dormant behavior instruction sequences which are different from the captured behaviors can be mined in the clustering process.

(4) Calculate the behavior distance
Calculate the distance between the captured behavior and each mined behavior. If the captured behavior is normal, the behavior of most distance from which can be considered as a potential dormant behavior.

(5) Dynamic regression test and analysis of dormant behaviors
All the mined potential dormant behavior instruction sequences are triggered to execute. Through regression testing, we can monitor and analyze the dormant behavior's characteristic from the protocol's execution.

(6) Evaluate the protocol's execution security
According to the result of regression test and analyze, the protocol's execution security report is generated.

The potential dormant behaviors of unknown protocols can be mined by the instruction clustering algorithm. However, whether the behaviors are certainly to be dormant behaviors, and what is the behavior's specific content still needs to further analyze and validate. The mined potential dormant behaviors are executed and analyzed in the whole-system simulation platform HiddenDisc, which are developed by ourselves independently. HiddenDisc is built on the basis of TEMU virtual platform. The platform can simulate various hardware and software resources realistically. The dynamic taint analysis technique is added on HiddenDisc. HiddenDisc platform provides a user-defined-behavior interface (UDB-API). The user-defined-behavior interface facilitates the development of standard plug-ins, customizes and extends the functionality of the platform. Under the help of HiddenDisc, protocol's potential dormant behaviors can be monitored and analyzed in detail. We can understand the details of dormant behaviors in machine-level instruction by monitoring the value of registers (memory) in HiddenDisc. Due to the space, the detail of dormant behavior analysis issue is described in another paper.

4 Protocol Dormant Behavior Analysis

4.1 Experimental Scheme

In order to evaluate the effectiveness of the method, we have collected and analyzed 1297 protocol samples in our platform HiddenDisc. The experimental platform HiddenDisc is based on the open source project TEMU. HiddenDisc's analysis engine consists of 2 subsystems. One subsystem collect and record instruction sequences of the protocol samples, and the other implement instruction clustering for dormant behavior analysis. The second subsystem is achieved with our own custom functions. These functions can be easily changed and extended for different intentions. HiddenDisc can be deployed on many type of platforms to implement protocol behavior analysis. In our designed experimental scheme, HiddenDisc is installed on 6 host computers. The hardware configurations of the 6 host machines are all Intel Core i7 6700 k@4.20 GHz CPU, 16 GB of memory. The host operating systems and 2 virtual experimental environments are all installed with Ubuntu V16.10 Linux; the other 2 virtual operating systems are both Windows 7 SP1. The control server and the analysis server are both installed with Windows server 2008 operating systems.

4.2 Experimental Results

At the beginning of the experiment, we have already prepared 25 extracted instruction sequences, representing 25 different protocol behaviors. In order to effectively achieve cluster analysis, 25 typical behavior instruction sequences of public protocols have been captured, such as http, ftp, DNS and SMB etc. These known behaviors is the foundation of protocol instruction clustering analysis. Using HTTP as an example, we capture and record the instruction sequences about how the Apache server processes a http get request for the file "index.html", and the reply generated by the server. In order to understand the content of the FTP protocol, we have analyzed and extracted the instruction sequences of the messages which sent by the FileZilla server in response to a connection, as well as the sent messages when the username and password are received. An example for SMB, we have analyzed and extracted the behavior about a Negotiate Protocol Request received by the Sambad open source server. In addition, we have also analyzed some common and typical dormant behavior instruction sequences, such as keyboard sniffing, password sniffing, backdoor accessing, rootkiting and spying etc. Within 3 min, the 1297 unknown protocol samples are executed on HiddenDisc one after another automatically. All between the two behavior distance is computed according to the instruction clustering algorithm. Finally, by contrasting with the 25 original basic behaviors, 193 behavior clusters are automatically generated.

Under the help of our developed analysis platform HiddenDisc, all the normal and dormant behavior instruction sequences are generated. Some typical behaviors of protocol samples are shown in Table 1.

Table 1. Examples of typical behavior instruction sequences

Behavior category	Genetic instruction sequences	Whether dormant
Behavior1	DFFDDDF DFDDF CCDFD DD FDFD F DDDFDD	No
Behavior2	DFFDDDF FDDDF CC DFD DD FFFF CC DFDFDDF FDDDFD FDCC DFD F DDDFDD	No
Behavior3	FDFDDDCD FFFFFFFF FDFDFDDCD	Yes

As illustrated in Table 1, behaivor1 and behavior2 are both captured instruction sequences by dynamic analysis in HiddenDisc. While behavior3 is not the captured behavior, but mined by our developed instruction clustering algorithm. Using the instruction clustering algorithm, the experimental results show that, the genetic instruction characteristics for the three behaviors are $B_1 = (freq\ _1(F, D,C), distrib_1(F, D,C)) = (0.16,0.12,0.22,8,4,1)$; $B_2 = (freq\ _2(F, D,C), distrib_2(F, D,C)) = (0.26,0.18,0.61,8,6,2)$; $B_3 = (freq\ _3(F, D,C), distrib_3(F, D,C)) = (0.11,0.13,0.21,9,9,2)$. The behavior distances are $D\ (B_1, B_2) = | B_1–B_2 | = (B_1 –B_2)^2 = 2$; $D(B_1, B_3) = | B_1–B_3 | = (B_1–B_3)^2 = 25$; $D\ (B_2, B_3) = | B_2–B_3 | = (B_2–B_3)^2 = 15$. Actually, the three behaviors are different, because the behavior distance can reveal the secret. A large number of case studies show that if the behavior distance $D(B_1, B_2) < 3$, although the function of $B_1$and B_2 seem to be different, the nature of the two behaviors are almost the same. This means behavior B_1 and B_2 can form a same cluster. As a result, if B_1 is a normal behavior, B_2 can also be a normal one. Now suppose that B_1 is a normal behavior, if the behavior distances $D(B_1, B_2) > 3$, B_2 is considered to be a potential dormant behavior. Furthermore, if the behavior distance $D(B_1, B_2) > 9$, B_2 is considered to be a potential malicious behavior. The mined dormant behavior instruction sequences still need further analysis to understand their content. Once again, the potential dormant behavior instruction sequences are analyzed on HiddenDisc for regression test. Although behavior1 and behavior2 are different, we are sure that they are both normal behaviors. While behavior3 is a dormant behavior, because it secretly copies a suspicious file to the system directory. With the help of instruction clustering, we have extracted 193 potential dormant behavior instruction sequences from 1297 protocol samples. Finally, we discovered 187 malicious behaviors from the 193 instruction sequences by regression test.

5 Conclusions

In this research, a system solution for perception mining the protocol's dormant behavior is proposed. Different from previous works, our approach first combines dynamic taint analysis with Instruction Clustering to effectively integrate dynamic and static analysis method. The dynamic taint analysis is used to monitor protocol's public behavior and instruction clustering is used to mine the dormant behavior which has not executed. We have extended the function of instruction clustering algorithm to our own platform HiddenDisc. Our platform can find a partitioning of a given protocol samples, so that the same cluster exhibit similar behaviors and dormant behaviors are belonged to different clusters. Our research begins with monitoring each protocol sample in our dynamic taint analysis environment HiddenDisc. Next, All the behavior instruction sequences are extracted by instruction clustering. With the help of instruction clustering analysis, the dormant behavior instruction sequences are mined automatically. Finally, we use the new generated sensitive message to trigger the dormant behavior execute as regression test for extensive study. Until now we only have a preliminary knowledge about protocol's execution security. What is the relationship between dormant behavior and execution security still needs further research. Our future work will focus on improving the accuracy and versatility of the instruction clustering algorithm. We also intend to improve the ability to perceive various dormant behaviors.

Acknowledgments. This work was supported by the Natinal Natural Science Foundation of China under Grant no. 61103178, 61373170, 61402530, 61309022 and 61309008.

References

1. Ming, J., Xin, Z., Lan, P., Wu, D., Liu, P., Mao, B.: Impeding behavior-based malware analysis via replacement attacks to malware specifications. J. Comput. Virol Hack Tech., 1–15 (2016)
2. Han, K., Lim, J.H., Im, E.G.: Malware analysis method using visualization of binary files. In: Proceedings of the 2013 Research in Adaptive and Convergent Systems, Montreal, Quebec, Canada (2013)
3. Hu, X., Shin, K.G.: DUET: integration of dynamic and static analyses for malware clustering with cluster ensembles. In: Proceedings of the 29th Annual Computer Security Applications Conference, New Orleans, Louisiana (2013)
4. Ye, Y., Li, T., Chen, Y., Jiang, Q.: Automatic malware categorization using cluster ensemble. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA (2010)
5. Anderson, B., Storlie, C., Lane, T.: Improving malware classification: bridging the static/dynamic gap. In: Proceedings of the 5th ACM workshop on Security and Artificial Intelligence, Raleigh, North Carolina, USA (2012)
6. Egele, M., Scholte, T., Kirda, E., Kruegel, C.: A survey on automated dynamic malware-analysis techniques and tools. ACM Comput. Surv. **44**(2), 1–42 (2012)
7. Caballero, J., Song, D.: Automatic protocol reverse-engineering: message format extraction and field semantics inference. Comput. Netw. **57**(2), 451–474 (2013)
8. Meng, F.M., Liu, Y., Zhang, C., Li, T.: Inferring protocol state machine for binary communication protocol. In: Advanced Research and Technology in Industry Applications (WARTIA), 2014, pp. 870–874 (2014)
9. Sedaghat, L., Duerling, B., Huang, X., Tang, Z.: Exploring data communication at system level through reverse engineering: a case study on USB device driver. In: Wong, W.E., Zhu, T. (eds.): Computer Engineering and Networking, pp. 329–336. Springer (2014)
10. Li Xiang-Dong, L.C.: A survey on methods of automatic protocol reverse engineering. In: Proceedings of the 2011 Seventh International Conference on Computational Intelligence and Security, pp. 685–689 (2011)
11. Luo, J.-Z., Yu, S.-Z.: Position-based automatic reverse engineering of network protocols. J. Netw. Comput. Appl. **36**(3), 1070–1077 (2013)
12. Zhang Zhao, W.Q.-Y., Wen, T.: Survey of mining protocol specifications. Comput. Eng. Appl. **49**, 1–9 (2013)
13. Wei Lin, J.F., Zhu, Y., Shi, X.: A method of multiple encryption and sectional encryption protocol reverse engineering. In: 2014 Tenth International Conference Computational Intelligence and Security (CIS), pp. 420–424 (2014)
14. Wang, Y., Gu, L.-z., Li, Z.-x., Yang, Y.-x.: Protocol reverse engineering through dynamic and static binary analysis. J. China Univ. Posts Telecommun. **20**, 75–79 (2013)
15. Wondracek, G., Comparetti, P.M., Kruegel, C., Comparetti, P., Kirda, E.: Automatic network protocol analysis. In: Proceedings of the 15th Annual Network & Distributed System Security Symposium (NDSS 2008) (2008)
16. Rahbarinia, B., Perdisci, R., Lanzi, A., Li, K.: PeerRush: mining for unwanted P2P traffic. In: Rieck, K., Stewin, P., Seifert, J.-P. (eds.): Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 62–82. Springer (2013)
17. Cui, B., Wang, F., Hao, Y., Wang, L.: A taint based approach for automatic reverse engineering of gray-box file formats. Soft Comput., pp. 1–16 (2015)

18. Polino, M., Scorti, A., Maggi, F., Zanero, S.: Jackdaw: towards automatic reverse engineering of large datasets of binaries. In: Almgren, M., Gulisano, V., Maggi, F. (eds.): Detection of Intrusions and Malware, and Vulnerability Assessment. pp. 121–143. Springer (2015)
19. Rahimian, A., Ziarati, R., Preda, S., Debbabi, M.: On the reverse engineering of the citadel botnet. In: Danger, J.L., Debbabi, M., Marion, J.-Y., Garcia-Alfaro, J., Zincir Heywood, N. (eds.): Foundations and Practice of Security, pp. 408–425. Springer (2014)
20. Rostami, M., Majzoobi, M., Koushanfar, F., Wallach, D., Devadas, S.: Robust and reverse-engineering resilient PUF authentication and key-exchange by substring matching. IEEE Trans. Emerg. Top. Comput., 1 (2014)
21. Wang, Y., Xiang, Y., Zhou, W., Yu, S.: Generating regular expression signatures for network traffic classification in trusted network management. J. Netw. Comput. Appl. **35**(3), 992–1000 (2012)
22. Comparetti, P.M., Wondracek, G., Kruegel, C., Kirda, E.: Prospex: protocol specification extraction. In: Proceedings of the 30th IEEE Symposium on Security & Privacy, pp. 110–125 (2009)
23. Cui, W., Peinado, M., Chen, K.: Tupni: automatic reverse engineering of input formats. In: Proceedings of the 15th ACM Conferences on Computer and Communication Security, pp. 391–402 (2008)
24. João Antunes, N.N.: Automatically complementing protocol specifications from network traces (2014). http://www.di.fc.ul.pt/~nuno/Papers/ewdc11.pdf
25. João Antunes, N.F.N., Verissimo, P.: ReverX: reverse engineering of protocols (2011). http://hdl.handle.net/10455/6699
26. Deyoung, M.E.: Dynamic protocol reverse engineering - a grammatical inference approach. Ohio: Air Force Institute of Technology (2008)
27. Dreger, H., Feldmann, A., Mai, M. et al.: Dynamic application layer protocol analysis for network intrusion detection. In: Proceedings of the 15th USENIX Security Symposium, pp. 257–272 (2006)
28. Juan Caballero, D.S.: Automatic protocol reverse-engineering: message format extraction and field semantics inference. Comput. Netw. **54**(2), 451–474 (2012)
29. Caballero, J., Yin, H., Liang, Z., Dawn, S.: Polyglot: automatic extraction of protocol message format using dynamic binary analysis. In: Proceedings of the 14th ACM Conference on Computer and Communications Security, pp. 317–329 (2007)
30. Caballero, J., Poosankam, P., Kreibich, C., Song, D.: Dispatcher: enabling active botnet infiltration using automatic protocol reverse-engineering. In: Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 621–634 (2009)
31. Zhi, W., Jiang, X., Cui, W.-D., Wang, X.-Y., Grace, M.: ReFormat: automatic reverse engineering of encrypted messages. In: Proceedings of the 14th European Symposium on Research in Computer Security, pp. 200–215 (2009). 5789

Video Stabilization Algorithm Based on Kalman Filter and Homography Transformation

Cong Liu[1,2], Xiang Li[1,2], and Minghu Wu[1,2](✉)

[1] Hubei Key Laboratory for High-Efficiency Utilization
of Solar Energy and Operation Control of Energy Storage System,
Hubei University of Technology, Wuhan, People's Republic of China
wuxx1005@mail.hbut.edu.cn

[2] Hubei Collaborative Innovation Center for High-Efficiency Utilization
of Solar Energy, Hubei University of Technology,
Wuhan 430068, People's Republic of China

Abstract. The camera systems are usually suffered from random jitter. In this paper, a new method based on Kalman filter and homography transformation is proposed to stabilize the unstable video. Firstly, the SURF (Speed-Up Robust Feature) point-feature matching algorithm is employed to find the corresponding matching points between two consecutive frames, and the bidirectional nearest neighbor distance ratio method is used to clear false matches. Secondly, motion estimation is computed by homography model and least square method. Then, Kalman filter are applied to separate the global and local motion. Finally, the unstable video frames is compensated by global motion vector. The experiment result shows that proposed method can effectively eliminate the random jitter.

1 Introduction

The current maneuver technology and image stabilization technology can be divided into three categories: mechanical, optical, electronic. Mechanical and optical due to equipment manufacturing difficult, high cost, large size and other shortcomings, in the application is limited. Electronic can effectively reduce the size and design cost of the system, which is the hotspot of current shaking technology and image stabilization [1, 2]. The traditional electronic image stabilization algorithm mainly includes three steps: motion estimation, motion filtering, image compensation [3]. In [4], a good image stabilization effect is achieved by using the method of constrained L1 norm optimization. However, since the original trajectory is not hardened, the trajectory after smoothing is easy to deviate from the main trajectory. The traditional method of motion estimation tends to bring about cumulative error, which affects the effect of image stabilization. Recently, feature-based video stabilization approaches locate a sparse set of reliable features in adjacent frames for camera motion estimation. These features can be obtained from Harris coner [5], SIFT (Scale Invariant Feature Transform) [6] and SURF (Speeded Up Robust Features) [7]. In [8, 9], motion-compensated stable video frame sequence can be applied to image mosaic technology and intelligent video surveillance system,

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_31

but the motion trajectory curve of the video sequence is not smooth enough. As for this problem, in [10, 11], the motion estimation of motion estimation is carried out by combining the SURF and Kalman filter, and the local motion in the video is separated, and the image stabilization effect is obtained. In order to further improve the real-time performance of the image stabilization technique, [12] uses the value based on the feature window gradient matrix to design the feature point detection method, and uses the least squares method and the matching point based on the optical flow to solve the motion parameter equation, The calculation time is less than the frame sampling time, and the Kalman filter is effective for the robot visual jitter compensation.

In this paper, the algorithm based on Kalman is proposed to solve the problem of rotating image stabilization and smooth motion trajectory at the same time. The SURF algorithm is used to extract the SURF feature points of two adjacent frames. The homography transformation matrix and the motion estimation corresponding to the two adjacent frames are calculated by the transformation model and the least squares method. Then, Kalman filtering of the motion estimation is carried out horizontally, vertically and in the direction of rotation. The global motion vector and the local motion vector are used. Finally, the global motion vector is used to reverse the video sequence to obtain a stable video trajectory.

2 Research Methods

This paper presents the algorithm flow shown in Fig. 1, the system consists of three major components: motion estimation, motion separation, motion compensation [1]. The algorithms in the block diagram are described below.

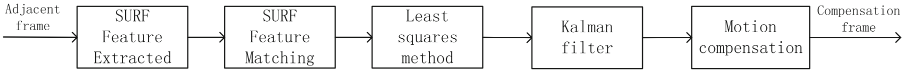


Fig. 1. Block diagram of proposed process

2.1 SURF Feature Point Extraction

At present, the feature point matching method has been widely applied to the field of video stabilization, the feature points can truly reflect the image of local information, such as corner, texture and other features. Harris corner has a good robustness, the rotation of the image, the effect of translation on its detection is very small, when the video scale transformation, Harris corner of the number of detection will increase with the scale decreases. SIFT (Scale-Invariant Feature Transform) and SURF (Speed-Up Robust Feature) are in the scale space. According to the neighborhood information of the feature points, the feature descriptor is established, which is a scale invariant feature descriptor, degeneration and anti - noise. The SURF feature point matching method proposed in this paper is an improved method of SIFT. SIFT algorithm uses the method of fixed Gaussian filter to reduce the size of the original image when constructing the image pyramid, and SURF proposes the concept of integral image and box filter, In the construction of the pyramid, to keep the image unchanged, change the size of the filter, the use of three addition and subtraction to replace the complex convolution operation, the speed relative to SIFT increased by about 3 times. Given a point X in image I (x, y), at this point, the Hessian matrix with dimension space σ is:

$$H(x,\sigma) = \begin{vmatrix} D_{xx}(x,\sigma) & D_{xy}(x,\sigma) \\ D_{xy}(x,\sigma) & D_{yy}(x,\sigma) \end{vmatrix} \tag{1}$$

$D_{xx}(x,\sigma)$, $D_{xy}(x,\sigma)$, $D_{yy}(x,\sigma)$ are Gaussian second derivative $\frac{\partial^2}{\partial x^2}g(\sigma)$, $\frac{\partial^2}{\partial x \partial y}g(\sigma)$, $\frac{\partial^2}{\partial y^2}g(\sigma)$. The approximate value of the convolution of the image I at point X, in order to preserve the energy of the original Gaussian kernel and the approximate Gaussian kernel, Adjust the parameter w so that the Hessian matrix has an approximate expression:

$$Det(H) = D_{xx}(x,\sigma)D_{yy}(x,\sigma) - wD_{xy}(x,\sigma)^2 \tag{2}$$

In theory, the value of w is related to σ. In practical calculation, w can be constant, and w ≈0.9 is obtained according to the experimental data of [7].

In order to obtain the rotation invariance feature, the SURF algorithm first calculates the repeatability direction using the horizontal and vertical Haar wavelet responses. Haar wavelet can be obtained quickly by the integral image method, and then a rectangular region is constructed in the main direction of the feature point. The extraction of the feature descriptor in the window. The window is divided into 4 × 4 sub-regions, each sub-region corresponds to a 4-dimensional vector, $V = (\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$, Where d_x represents the horizontal Haar wavelet response, d_x represents the vertical Haar wavelet response, and the same person $|d_x|$, $|d_y|$ is the absolute value of the response.

2.2 Nearest Neighbor Algorithm

In the case of matching the feature points with 64-dimensional feature descriptors, it is inevitable that there will be some cases of mis-matching. This error is divided into two categories: one is the inaccurate position of the feature points, the other is due to the video frame Of the local movement of the pseudo-feature points. Mistaken matching points will increase the error of adjacent inter-frame motion estimation, which will affect the subsequent motion compensation. Therefore, it is necessary to check the matching point, eliminate the mismatch point and improve the matching precision.

In this paper, we use a two-nearest neighbor algorithm to define a nested container. By calculating the Euclidean distance of the matching point pair, two matching points closest to the feature points in the reference frame are calculated and stored in the nested container. If the matching distance is less than the number of times the matching distance is multiplied by a factor, the matching point of the minimum distance is considered as the excellent feature matching point in the current frame, and the

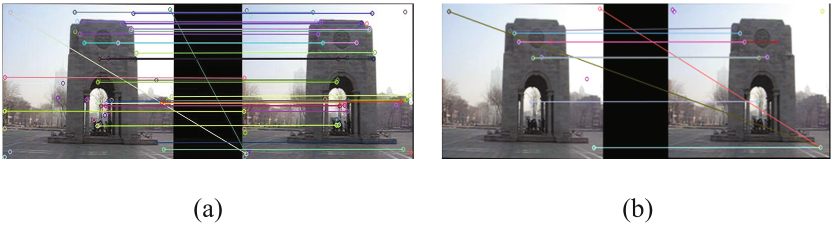


Fig. 2. Eliminate the matching of two adjacent points before and after mismatch

matching point which does not satisfy the condition is eliminated. According to the number of test coefficients k range of 0.2–0.6. Figure 2(a) is to eliminate the mismatch before the adjacent two-frame match point pairs, Fig. 2(b) is to eliminate the mismatch after the adjacent two-frame matching point pairs, by comparison can be found, it is clear that the wrong match point was removed, Reduced matching pairs, increased the number of correct pairs, and accelerated the calculation of motion estimates.

The set of feature descriptors of the reference frame is D1 and the set of feature descriptors of the reference frame is D2, the set of feature descriptors of adjacent two frames can be expressed by M as formula (3).

$$M = \{(d_{1,i}, d_{2,j}) | d_{1,i} \in D1, d_{2,j}, d_{2,k} \in D2, \frac{dist(d_{1,i} - d_{2,j})}{dist(d_{1,i} - d_{2,k})} < threshold\} \tag{3}$$

The least matching method is used to find the most suitable affine transformation matrix by using the nearest neighbor distance ratio method.

2.3 Motion Compensation

Motion compensation is the last process of the video image stabilization system. After the affine transformation matrix is obtained by the above method, a new transformation matrix is obtained after the Kalman filtering in the horizontal and vertical directions, and the filtered matrix is used to convert the current frame And the compensated frame is output to a video write stream. During the next iteration, the current frame is used as the next reference frame for the next motion estimation, so that after all the frame compensation is completed, Like after the video sequence.

3 Experimental Results

In order to verify the effectiveness of this algorithm, the experimental hardware platform: CPU frequency 2.8 GHz, RAM for the 4G, Windows7 operating system PC, software platform: Visual Studio 2013, Opencv2.4.9 library, Matlab2010b.

3.1 Video Rotation Compensation

In this experiment, three sets of videos were collected in a bumpy environment. Each set of videos contains horizontal, vertical vector, jitter and rotation, but each video has its own distinctive features, and unstable video contains a large rotation, and accompanied by a scale change. By observing the 21, 27, 33, and 39 frame images of the three sets of video images in Fig. 3, the stabilized video compensates for a lot of rotation. The experimental results show that the SURF algorithm can effectively remove the horizontal, vertical vector, jitter and rotation in the camera video.

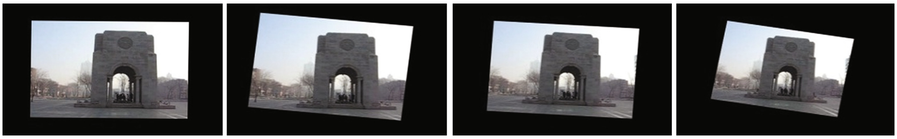


Fig. 3. Frames after video stabilization using proposed algorithm, 21, 27, 33, and 39 frames of stabilization video

3.2 Evaluation of Image Stabilization

The quality of the post-stabilization video depends on the inter-frame fidelity (ITF) and motion-filtered video trajectory, and the smoother motion trajectory in the horizontal and vertical directions is better. ITF is based on the PSNR (peak signal to noise ratio) [2] obtained, a group of continuous frame sequence ITF value of the larger, to prove that the inter-frame image misalignment is less, and its high degree of sequence composition The degree of video image stabilization, ITF calculation method for the formula (4).

$$ITF = \frac{1}{N}\sum_{i=1}^{N-1} PSNR(f_i) \tag{4}$$

Where N is the total number of video frames, the larger the PSNR value in Eq. (4), the larger the ITF value, and Table 1 is the PSNR value of the 21, 27, 33, 39 frames of a sequence of successive video frames. The peak signal to noise ratio of the video after image stabilization is improved by about 5 dB.

Table 1. Peak signal-to-noise ratio before and after image stabilization

Frames	Unstable video	Method using [7]	Proposed method
21	14.08	15.33	17.54
27	17.84	20.16	22.43
33	15.44	18.67	19.49
39	19.67	21.48	24.16

4 Conclusion

In this paper, a new algorithm based on Kalman filter and homography is proposed. The traditional image stabilization algorithm can only eliminate the compensation translation motion vector. In this paper, the feature points of two adjacent frames are found by SURF feature matching method. And then the least squares method is used to find the most suitable homography transformation matrix, and the motion estimation between the video frames is calculated accurately. Finally, the Kalman filter method is used to calculate the motion of the video frame. Experiments show that the motion trajectory can be smoother than Kalman filter. The proposed algorithm can be applied to a variety of video with rotation and random jitter, in the case of video resolution is not high, better real-time, to meet the requirements of modern enterprises.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (No. 60972016, No. 61231010), and Funds of Distinguished Young Scientists of China (No. 2009CDA150).

References

1. Qian, L., Wang, S., Zhang, J., et al.: A real-time despinning method for onboard video image. J. Projectiles Rockets Missiles Guidance **29**(3), 20–22 (2009)
2. Zeng, X.P., Yang, T.: Electronic system for real-time canceling image rotations. Opto-Electron. Eng. **32**(10), 27–30 (2005)
3. Chen, B.H., Kopylov, A., Huang, S.C., et al.: Improved global motion estimation via motion vector clustering for video stabilization. Eng. Appl. Artif. Intell. **54**, 39–48 (2016)
4. Grundmann, M., Kwatra, V., Essa, I.: Auto-directed video stabilization with robust L1 optimal camera paths. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 225–232. IEEE Computer Society (2011)
5. Yin, X., Kim, D.H., Hong, C.P., et al.: Advanced feature point transformation of corner points for mobile object recognition. Multimedia Tools Appl. **74**(16), 6541–6556 (2015)
6. Su, Y., Sun, M.T., Hsu, Y.F.: System and method for non-iterative global motion estimation: US, US 7684628 B2 (2010)
7. Bay, H., Ess, A., Tuytelaars, T., et al.: Speeded-up robust features. Comput. Vis. Image Underst. **110**(3), 404–417 (2008)
8. Pinto, B., Anurenjan, P.R.: Video stabilization using speeded up robust features. In: International Conference on Communications and Signal Processing, pp. 527–531. IEEE (2011)
9. Pradidtong-Ngam, C., Natwichai, J.: Content-based video search on peer-to-peer networks. Int. J. Grid Utility Comput. **2**(3), 234–242 (2011)
10. Yu, Y.C., You, S.C.D., Tsai, D.R.: A video-based portal system for remote appliance control. Int. J. Space-Based Situated Comput. **1**(2/3), 122–129 (2011)
11. Cheng, X., Hao, Q., Xie, M.: A comprehensive motion estimation technique for the improvement of EIS methods based on the SURF algorithm and Kalman filter. Sensors **16** (4), 486 (2016)
12. Wang, B.R., Jin, Y.L., Shao, D.L., et al.: Design of jitter compensation algorithm for robot vision based on optical flow and Kalman filter. Sci. World J. **2014**(4), 130806 (2014)

Towards a Web-Based Teaching Tool to Measure and Represent the Emotional Climate of Virtual Classrooms

Modesta Pousada, Santi Caballé(✉), Jordi Conesa, Antoni Bertrán,
Beni Gómez-Zúñiga, Eulàlia Hernández, Manuel Armayones,
and Joaquim Moré

Universitat Oberta de Catalunya, Barcelona, Spain
{mpousada, scaballe, jconesac, abertranb, bgomezz,
ehernandez, marmayones, jmore}@uoc.edu

Abstract. This paper presents the first results of a teaching innovation project named "Emotional Thermometer for Teaching" (ETT) carried out at the Universitat Oberta de Catalunya. The ETT project intersects the scopes of eLearning and Affective Computing with the aim of collecting and managing emotional information of online students during their learning process. Such information allows lecturers to monitor the overall emotional climate of their virtual classrooms whilst detecting critical moments for timely interventions, such as assisting in certain learning tasks that generate negative emotions (anxiety, fear, etc.). To this end, an innovative teaching tool named ETT was developed as a functional indicator to measure and represent the classroom emotional climate, which is dynamically evolving as the course goes by. In this paper, the technical development of the ETT tool is described that meets the challenging requirement of correctly identifying the overall emotional climate of virtual classrooms from the posts sent by students to in-class forums. First, a machine learning approach combined with Natural Language Processing techniques is described to automatically classify posts in terms of positive, neutral and negative emotions. Then, a web-based graphical tool is presented to visualize the calculated emotional climate of the classroom and its evolution over time. Finally, the post classification approach is technically tested and the initial results are discussed.

1 Introduction

Emotions and affective factors, such as confusion, frustration, shame and pride, are acknowledged as major influences in education. The identification of learners' emotions becomes a key aspect to promote a deeper and more enthusiastic learning experience and to regulate the interaction between learners and teachers [1]. Indeed, the transformation of negative emotions into positive ones, impacts on learners' motivation and engagement, self-regulation and academic achievement [2]. However, despite major advancements in fields such as artificial intelligence and human-computer interaction, e-learning environments are still struggling with incorporating emotional-aware tools. The limited-to-null adoption of emotional analysis tools and

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_32

affective feedback prevents both learners and teachers from reaping the benefits of emotion-aware Learning Management Systems (LMSs) [3].

On one hand, the teaching professions now face rapidly changing demands. Helping all learners to develop the skills they need in a rapidly evolving society, and a global labor market based on ever higher skill levels, requires new sets of competences [4]. In this sense, teachers are expected to adopt a new role as intellectual and educational "coaches", which establishes a new paradigm for teaching [5]. This coaching role is virtually impossible to achieve with a lack of emotional awareness; this problem intensifies even more in online teaching. To coach successfully, teachers do need a cognitive, social and emotional understanding of their learners and the ability to respond to these neuro-physiological states [6].

On the other hand, students learn by exploiting their cognitive abilities but also their emotional abilities, which have an influence in the teaching and learning processes. Literature provides solid links between emotions and student's motivation, self-regulation and academic performance [2, 6]. The latest reports of the Centre for Educational Research and Innovation of the OCDE (Organisation for Economic Co-operation and Development) also highlight emotions as one of the principles driving the development of the next learning scenarios of the 21st century [7].

In this context, Affective Computing comprises an emerging set of innovations that allow systems to recognize, interpret and simulate human emotions [8]. While current applications mainly focus on capturing and analyzing emotional reactions of individual users to improve the efficiency and effectiveness of product or media testing, this technology holds great promise for developing social awareness, communication, collaboration and emotional skills, such as greater empathy, improved self-awareness and stronger relationships [9]. However, very few studies are trying to represent the emotional state of learners leading to a lack of emotional feedback for learners in current educational environments [10].

Learning situations give rise to many emotions which influence how people learn and how they interact with each other [11]. It is thus fundamental to (i) better understand how emotions affect learning and social interactions and (ii) detect emotions and social signals in order to build a clear map of a socio-affective situation and regulate the learning process [12]. To this end, developers and designers are struggling to empower educational environments with usable interfaces, in an attempt to trace learner emotions in an unobtrusive and non-invasive way, in parallel with their tasks, without extra cost or equipment and expertise, and without language barriers [13, 14].

This new research area is propelled by the advancements attained in affective learning addressing a set of interrelated conceptual, technological and application problems [3]. Though previous research initiatives have indicated that taking emotions into account can make the learning offering more effective, the jury is still out on lowering dropout rates and raising the engagement of potentially bored cohorts of students [10, 14]. In addition, sign posting the learner's feelings in LMSs is still in its infancy and the answers to many technological questions remain open-ended. In particular, adequate user interfaces to represent emotions have yet to be integrated in seamless, non-intrusive and effective ways into LMSs [13]. Finally, the application of latest learning technologies and experiences gained in previous research projects, such as machine learning [15], have not been conveniently leveraged for augmenting learner's engagement in the learning process. Potential use of these outcomes for affective learning has not been fully explored yet.

The motivation of this research work is to provide some answers to the above challenges from the results of a teaching innovation project named "Emotional Thermometer for Teaching" (ETT) currently undertaken at the Universitat Oberta de Catalunya[1]. The aim of the project is to develop and test an emotion-aware teaching tool for lecturers to measure and represent the classroom emotional climate [6] in the context of online higher education. The ultimate goal is to allow lecturers to use the ETT tool as a supportive teaching instrument to monitor and make decisions in relation to teaching and learning in the short and medium term, with the purpose of (a) analyzing the correlation between emotional dynamics over time and learning units according to the schedule of the classroom (satisfaction and/or difficulty of the content of different learning units); (b) detecting critical moments for a timely intervention of the lecturer in the virtual classroom, thus also optimizing the lecturer's effort and time by intervening only when and where appropriate.

In the current stage of the project, a first prototype of the ETT tool has been developed and technically tested in a controlled situation. The aim of this paper is to report the main issues and challenges faced during the development and prototyping of the tool as well as discuss on the first results obtained so as to input the next steps of the project.

The remainder of the paper is structured as follows: Sect. 2 reviews literature and some previous works on affective learning models as well as technological approaches to incorporate emotion-aware technology into LMSs in the context of our project. Section 3 shows the research methodology, which includes the main decisions made for the development of an automatic classification system and its corresponding graphical representation as the main requirements of the ETT tool. Section 4 shows the technical testing of our automatic classification system and the evaluation results. Finally, Sect. 5 summarizes the main ideas of the paper and outlines the following steps in the development and evaluation of the tool in the context of the ETT project.

2 Background

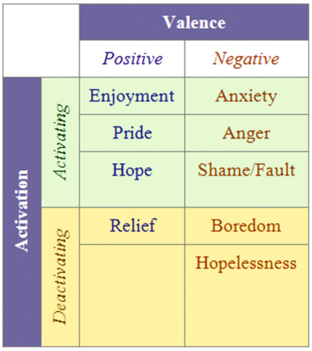
In this section, we review different emotion theories and models that classify emotions. For the sake of our ETT project, the emotion dimensions are also examined, especially the positive and negative valence with regard to students' engagement and motivation. Then, we review methods and existing tools that are applied in detecting (affect detection) and responding (affective feedback) to emotion. This information will serve as a basis for proposing models, methods and tools in the context of our project.

[1] The Universitat Oberta de Catalunya (UOC) is located in Barcelona, Spain. The UOC offers distance higher education fully over the Internet since 1995. Currently, about 54,000 students and 3,700 lecturers are involved in 6,400 online classrooms from about 300 graduate, post-graduate and doctorate programs in a wide range of academic disciplines. The UOC is found at http://www.uoc.edu.

.

2.1 Models for Emotion Recognition and Affective Feedback

Over the last three decades, there has been an emerging focus of research on the interplay between emotions and learning. Scherer [16] has distinguished three major schools of emotion recognition: the basic emotion (patterns are equivalent with basic emotions that can be easily recognized universally), the emotional dimension (acknowledges various emotion dimensions, e.g. arousal, valence, intensity, etc.), and the eclectic approach (verbal labels that seem appropriate to the aims of a particular study). There exists an extensive emotion taxonomy (see [14] for a full review):



		Valence	
		Positive	*Negative*
Activation	*Activating*	Enjoyment	Anxiety
		Pride	Anger
		Hope	Shame/Fault
	Deactivating	Relief	Boredom
			Hopelessness

Fig. 1. Academic emotions according to Pekrun et al. [24]

For the sake of our ETT project, we review broader the emotion model by Pekrun and his colleagues (see [17] and Fig. 1) who examined the impact of the so-called academic emotions (four positive: joy, hope, pride, relief and five negative: boredom, anger, anxiety, shame, hopelessness). According to their findings, positive mood fosters holistic, creative ways of thinking. Harmful effects can only be expected in situations, where students are in a good mood and the learning topics are of less importance to them. In this case, the positive emotion might detach them from learning. Negative emotions, on the other hand, direct students' attention to themselves, in most of the cases. Necessary attention for learning and task solving is lacking, because they try to find ways to get rid of the bad feeling. When negative emotions create a pessimistic perceptual attitude, they divert the learner's attention to aspects irrelevant to the task, which activate intrusive thoughts that give priority to a concern for a well-being rather than for learning [2, 6].

Although some work has been done to examine the relationship between the affective components of feedback and performance, a systematic examination is still pertinent since there is still controversy over the effects of certain feedback interventions [18]. Further works to be systematically investigated from the literature examine what sort of affective response is stimulated by "good feedback" [19]. Although more work has been done, there still remains some basic examples of feedback that merit further affective investigation, which include: (i) Elaborated feedback in manageable units; (ii) Specific clear feedback; (iii) Keep it simple; and (iv) Promote a learning goal objective [10]. Understanding the affective response and what correlates with lack of understanding of feedback will assist with automatic feedback systems in the future [20].

2.2 Emotion-Aware Methods and Techniques for ELearning

The literature on affective learning is offering successful affect detection and affective feedback methods and techniques. Next we review the most relevant contributions found, also considering our needs of the ETT project.

2.2.1 Emotional Detection

In [20] emotion measurement and detection tools are reviewed grouped into 3 areas:

- Psychological: subjective report using verbal or pictorial scales or questionnaires.
- Physiological: use of sensors to capture biometric signals, e.g., electromyogram (EMG), electrodermal activity (EDA), electrocardiogram (EKG or ECG), electrooculogram (EOG), blood volume pulse (BVP), etc.
- Behavioural: observation or capturing of motor-behavioural activity, e.g., facial expressions, voice intonation, body posture, sentiment analysis of text input, mouse and keyboard logs, etc.

For the sake of our ETT project, we add a fourth emotion detection group: artificial intelligence methods, made up of techniques from the fields of Natural Processing Language (sentiment analysis and opinion mining) [21] and machine learning [22], which combined can analyze and understand the students' emotional states from the text they produce in different forms during learning activities. Within the context of learning, a very limited number of proposals filed in scientific literature of affective learning have tried to take advantage of artificial intelligence methodologies to analyze the content of in-class reports, and comments and contents of social media (messages sent to discussion forums, blogs and social networks) [10].

2.2.2 Affective Feedback

Although not extensive, literature exhibits remarkable studies that evaluate computer mediated affective feedback strategies, and their impact on users. A rough classification includes dialogue moves (hints, prompts, assertions, and summaries), immersive simulations or serious games, facial expressions and speech modulations, images, imagery, cartoon avatars, caricatures or short video-audio clips [23].

Affective feedback design is aiming at producing effective rules that are sensitive to learners' emotional state. Agents can respond to student affect with either parallel-empathetic (exhibits an emotion similar to that of the target), reactive- empathetic (focuses on the target's affective state, in addition to his/her situation) or task-based (supplementary to empathetic strategies) [15]. Moreover, from a correct data analysis, the constant evolution in software development for visualization of data [24] can help to graphically interpret the analysis results in terms of what is happening in a classroom, which positively impacts on participant's motivation, emotional state and problem-solving abilities [23].

However, [19] concludes that we can never be entirely certain that the dynamic affect-adaptive tutoring systems are delivering useful affective feedback as there will always be some risk of unintentional negative consequences when attempting to intervene to modify student affect. In the context of our ETT project, we consider this direction and feedback is delivered by lecturers manually, thus minimizing this risk.

3 Research Methodology

The ETT project proposes a multidisciplinary mixed methods approach for the R&D process that will provide online teachers and lecturers with innovative teaching tools by combining affective learning theories, technological approaches and use cases for validation. These tools will become highly relevant for improving e-Learning. Next, the innovative aspects of the project are first summarized and then described in detail from a methodological and development perspective:

- *Affective learning theories* to model both the emotional behavior at individual level in collaborative and social learning situations (e.g. in-class discussion forums) as well as the provision of affective feedback at group level targeting students' motivation and engagement (see Sect. 2).
- *Learning technologies* from artificial intelligence methodologies [21, 22] to analyze the students' data generated inside social and collaborative forums, also to evaluate and predict the emotional polarity of their interactions and comments with respect to the courses that are being studied (see the application and evaluation of these technologies in Sects. 3 and 4).
- *Learning scenarios* to evaluate students' emotional attitude towards relevant educational situations (i.e. performing learning exercises and study units). In particular, use cases for student's affective profile are designed in order to obtain the academic modalities most appropriate to the characteristic of an individual, such as personal introduction to the group, collaborative activities and self-evaluation exercises (see ongoing and future work of our project in Sect. 5).

3.1 Conceptual Approach and Requirements

In the context of virtual classrooms, online students usually share and discuss their ideas and comments through in-class forums by posting messages. These posts may virtually contain any type of cognitive and emotional exchange move [25], from social support moves (greetings, encouragement, etc.) to request and informative moves (elaborate, justify, etc.). However, for the purpose of our ETT project, we are only interested in classifying the posts into 3 types of emotional moves or valence, namely positive, negative and neutral (the latter is a balance between positive and negative moves). This follows the model of Pekrun et al. [17] discussed in Sect. 2.

By this basic emotional classification approach, teachers can efficiently monitor the overall emotional climate of the classroom while identifying areas of improvement (activities that generate negative emotions, etc.). As a result, the teacher acts as a "coach" (see [5]) by making the right decisions in relation to teaching and learning in the short and medium term by both (a) analyzing the correlation between emotional dynamics over time and learning units according to the course schedule (satisfaction/difficulty of the content of different learning units), and (b) detecting critical moments for a timely intervention of the teacher in the classroom.

In terms of software system, the ETT tool is to provide functionality to constantly monitor and measure the classroom emotional climate. Text mining and sentiment analysis combined with machine learning techniques [21, 22] are used to detect and predict student's individual emotional state from the posts sent to in-class forums while aggregated data measure the classroom emotional climate. This information is sent to the lecturer in an effective visual format in order to perform the appropriate actions (i.e. feedback) addressed to the whole class, such as hints and advices or any other type of assessment to both emotionally balance the classroom and iteratively observe the consequences of these actions overtime. Following [25], we consider manual feedback from lecturers minimizing the unintentional negative consequences when attempting to intervene to modify student emotional state.

To sum up, the ETT tool is to meet the following functional requirements, which are next technically developed and prototyped:

- machine learning and sentiment analysis techniques for classifying the forum posts sent by individual students into positive, neutral and negative emotions;
- a graphical representation that visualizes the classroom emotional climate (number of positive, neutral and negative emotions) and its evolution over time.

3.2 Development and Prototyping

In this section, we report on the technical issues to meet the previous requirements of the ETT tool in order to both classify forum posts and show the classification results.

3.2.1 Post Classification

The goal of the classification component of the ETT tool is to correctly classify forum posts into 3 types of emotions, namely positive, negative and neutral. Next we provide our machine learning approach to automatically classify posts [15, 22].

Method

The input of our approach is a list of posts sent by students to regular in-class forums of the same real learning context of UOC where the ETT tool will be tested. We first build the training set from the list of posts which are classified manually by experts. Then, a set of indicators based on sentiment analysis are extracted from the training set. These indicators bear useful information about the posts that input our algorithm to classify new posts according to the emotional moves found in the text.

Training set

Following a similar approach to [16], we first classify the content of a set of posts manually. This training set was a pool of 700 posts selected randomly from in-class forums. The classification task of this training set into 3 tags (positive, negative and neutral emotion) was performed by a group of researchers from the Psychology and Computer Science faculties of UOC.

The procedure to tag each post consisted of considering the entire text of the post, many times containing both positive and negative moves in the same post, thus uniquely tagging the post according to the majority of exchanges moves in either direction (positive or negative) or balanced (neutral). Each and every of the 700 posts were manually tagged by at least 3 researchers from different faculties and separately. These researchers performed the tagging task according to their experience and knowledge. Then, the tagged posts were discussed collaboratively in several project group meetings in order for each post to issue an agreed tag, eventually obtaining a fairly amount of correctly tagged posts to train the machine learning method.

Sentiment analysis indicators

For the purpose to implement our classification algorithm, groups of indicators are considered following sentiment analysis techniques [15], as follows:

- *Word relevance:* An emotional wording list is created with all the relevant words appearing in the posts. To this end, we use first a tag set[2] for the language sources (Catalan and Spanish) to identify the grammatical type of each word for all the posts and only consider those types with emotional connotation (i.e. nouns, adjectives and verbs), thus removing the rest of words, such as articles and adverbs. Then, the posts are analyzed again by the statistic Term Frequency – Inverse Document Frequency (TF-IDF)[3], which takes those words of the created wording list and calculate their relevance according to the number of appearances (term frequency) in each post.
- *Emotional positivity and negativity*: For each word in the list, we calculate its relative expressiveness in terms of positive and negative emotional move according to the tag of the post in the training set where the word is found.
- *Text particles*: We calculate the number of other text particles in the posts that may have emotional meaning, such as multiple punctuation marks and negations (e.g., the text particle "……" may suggest an erratic speech, thus transmitting stress). Direct emotional particles (emoticons) may also appear in the text.

Eventually we have a list of concrete indicators that characterize and define the posts:

– Number of positive words (positivity > negativity)
– Number of negative words (positivity < negativity)
– Number of negations (Freeling tag = RN = adverb negative)
– Number of multiple punctuation marks ("??", "????", "!!!!", "……", etc.)
– Sum of positivity weighted as the relevance calculated by TF-IDF of each word
– Sum of negativity weighted as the relevance calculated by TF-IDF of each word
– Number of words indicating etcetera ("etc.", "and so on", "…", etc.)
– Number of positive emoticons ("☺", ":D", etc.)
– Number of negative emoticons ("☹", ";(", etc.)

[2] https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-ca.html.

[3] http://www.tfidf.com.

The values obtained by these indicators are then normalized by this operation:

$$\frac{(value - M)}{SD}$$

where M = mean of all the values and SD = standard deviation of M. The purpose of this standardization process is for our implemented classification algorithm to compare, operate and combine all these indicators in order to create the internal rules to automatically classify (predict) the emotional polarity of new posts.

Validation

For validation purposes, we used the k-fold cross validation technique (see [27] and Fig. 3) by the following process (k = 10):

- Divide the training dataset into 10 equal parts (folds)
- One fold is set aside in each iteration for testing, so each fold is used once for testing, nine times for training (see Fig. 2)
- Average the testing scores

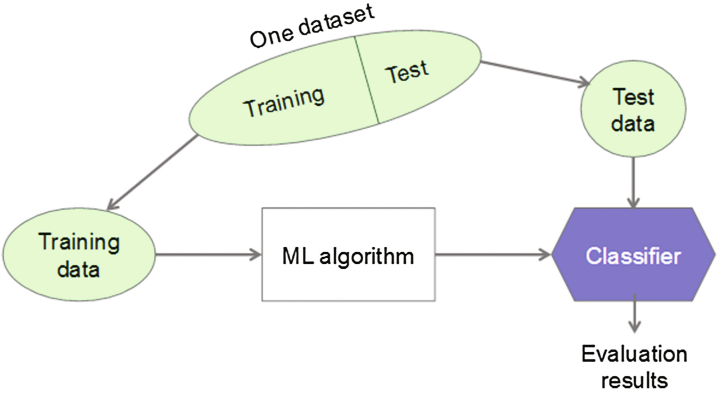


Fig. 2. Training and testing our machine learning approach

For optimization purposes, this process is repeated for different machine learning methods, such as Neural Network and Random Forest [22], and with different algorithm set-ups within each method. Then, we compare the average of the testing scores obtained by each method and set-ups. The purpose is to obtain the best machine learning algorithm optimized with the best set ups, and also at different data scales in order to improve the prediction capabilities of the ETT tool. Section 4 presents initial evaluation results.

3.2.2 Graphical Representation

In order to meet the second requirement of the ETT tool (see Sect. 3.1), we developed a web-based prototype to provide lecturers with a graphical representation of the emotional climate of the classroom and its evolution over time (see Fig. 3).

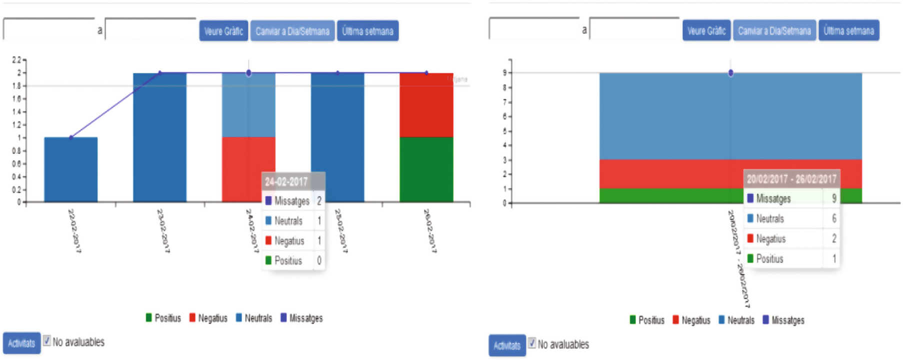
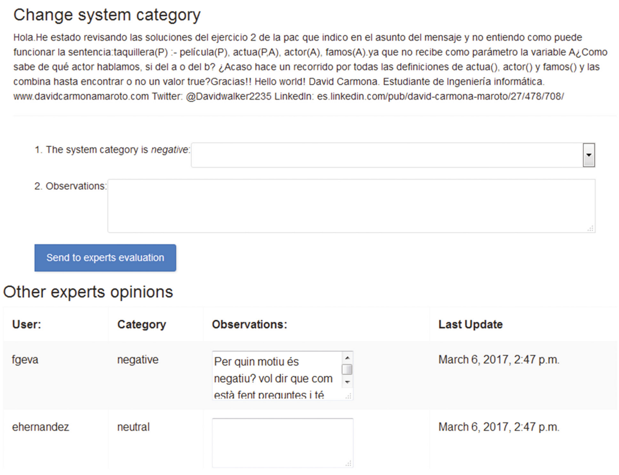


Fig. 3. Different time scales of the ETT tool report: daily (left) and weekly (right)

This web-based visualization system was developed by python-based Django with the charting package C3.js [26] to rendering dynamic and interactive graphics that shows the number of positive, negative and neutral messages from our post classification approach. This graphical information is sent to the lecturers who can use it to measure the emotional climate of their classrooms and see the day-to-day and weekly evolution. Hence lecturers can identify emotionally problematic moments in their classroom and perform timely actions to correct them. For instance, some suggested feedback from the emotional state detected for the whole classroom:

- Negative (e.g., during a complex learning activity): lecturer shares with the classroom some hints for solving the activity.
- Neutral: no specific action.
- Positive (e.g., after a feedback action performed to correct a negative emotion): encourage the classroom to keep up the good work.



Change system category

Hola.He estado revisando las soluciones del ejercicio 2 de la pac que indico en el asunto del mensaje y no entiendo como puede funcionar la sentencia:taquillera(P) :- película(P), actua(P,A), actor(A), famos(A).ya que no recibe como parámetro la variable A¿Como sabe de qué actor hablamos, si del a o del b? ¿Acaso hace un recorrido por todas las definiciones de actua(), actor() y famos() y las combina hasta encontrar o no un valor true?Gracias!! Hello world! David Carmona. Estudiante de Ingeniería informática. www.davidcarmonamaroto.com Twitter: @Davidwalker2235 LinkedIn: es.linkedin.com/pub/david-carmona-maroto/27/478/708/

1. The system category is *negative*:

2. Observations:

Other experts opinions

User:	Category	Observations:	Last Update
fgeva	negative	Per quin motiu és negatiu? vol dir que com està fent preguntes i té	March 6, 2017, 2:47 p.m.
ehernandez	neutral		March 6, 2017, 2:47 p.m.

Fig. 4. Experts manage and discuss on the automatic post classification

From the graphical representation shown in Fig. 3, lecturers and other experts registered in the system can gain access to the posts from their associated tags (see Fig. 4). Daily, new posts sent to the targeted forums and classified by the system are sent to the experts for tag review and discussion. This way, the automatic post classification system is constantly fed back and tuned for improving the training of our machine learning algorithm.

4 Evaluation Results

In this section, we present and discuss on the main results of testing our post classification approach through the cross validation technique, both described in Sect. 3.2.1. We only evaluate here the post classification system as the main challenging requirement for our ETT tool to perform correctly, thus eventually validating its potential as a valuable teaching resource.

We first introduce the evaluation measurement *confusion matrix* [27], which is a table layout visualization of the performance of machine learning algorithms in supervised learning (see [22] and Fig. 5). After executing the classification algorithm, each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. Hence the rows and columns of the confusion matrix table reports the number of false positives (fp), false negatives (fn), true positives (tp), and true negatives (tn). For instance, in our testing, a row contains the percentage of negative posts predicted out of the total of negative posts, while a column contains the percentage of the actual negative posts out of the total of negative posts.

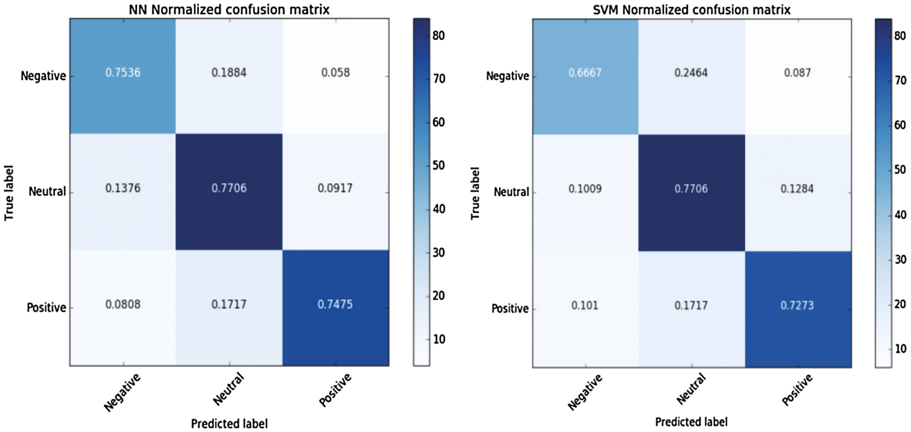


Fig. 5. Confusion matrix for neural network (left) and support vector machine (right)

Then, we consider two additional performance measures for evaluation [287]:

- *Precision*: measures the accuracy of the prediction, as follows:

$$\frac{tp}{tp + tn}$$

- *Recall*: measures the sensitivity of the prediction, as follows:

$$\frac{tp}{tp + fn}$$

Finally, we introduce the *F1 score* that combines the previous two measures by calculating their harmonic mean, as follows:

$$F1 = 2 \cdot \frac{(precision \cdot recall)}{(precision + recall)}$$

Based on the above evaluation measurements, we evaluated our post classification approach from testing it with three different machine learning methods, namely Neural Network, Random Forest and Support Vector Machine [22]. The best results were obtained by Neural Network, with *F1* = 0.76 and a *confusion matrix* (see Fig. 5-left) showing little confusion between positive and negative as well as acceptable *precision* of the three tags. As for the Random Forest and Support Vector Machine methods, we obtained similar but poorer results (*F1* = 0.74 and *F1* = 0.73 respectively). Therefore, Neural Network outperformed the other methods and obtained the best testing results.

5 Conclusions and Future Work

This paper presented the first stages of an innovative teaching project named ETT aiming at providing advance teaching tools that allow lecturers to identify and measure the classroom emotional climate in the context of online higher education. The main component of the ETT tool is a post classification system based on machine learning and sentiment analysis techniques, which is then reported to the lecturers by an interactive web-based graphical system. The initial evaluation results suggest the automatic post classification performs correctly, in terms of identifying the emotional polarity of new posts.

Ongoing work is to experiment with the ETT tool in real online learning. Different subjects from the Psychology and Computer Science degrees and about 2,000 students and 16 lecturers will participate either directly or indirectly in the experiment to evaluate the ETT tool as a valuable teaching resource.

Future work will take advantage of other Natural Processing Language techniques, such as academic analytics, to provide further cues on learners' behavior based on low-level interaction with the LMS stored in event logs, such as access frequency and session time, navigation patterns and time spent in learning activities.

Acknowledgements. This research was partially funded by both the Universitat Oberta de Catalunya through the teaching innovation project UOC-APLICA-2016 "Termomètre Emocional per la Docència" and the Spanish Government through the project: TIN2013-45303-P "ICT-FLAG".

References

1. Immordino-Yang, M.H., Damasio, A.: We feel, therefore we learn: the relevance of affective and social neuroscience to education. Mind Brain Educ. **1**(1), 3–10 (2007)
2. Hascher, T.: Learning and emotion: perspectives for theory and research. Eur. Educ. Res. J. **9**, 13–28 (2010)
3. Calvo, R., D'Mello, S.: Frontiers of affect-aware learning technologies. IEEE Intell. Syst. **27** (6), 86–89 (2012)
4. Commission Staff Working Document: Supporting the Teaching Professions for Better Learning Outcomes Accompanying the document Communication from the Commission Rethinking Education: Investing in skills for better socio-economic outcomes SWD/2012/0374 (2012)
5. Reinke, W.M., Stormont, M., Herman, K.C., Newcomer, L.: Using coaching to support teacher implementation of classroom-based interventions. J. Behav. Educ. **23**(1), 150–167 (2014)
6. Brackett, M., Reyes, M., Rivers, S., Elbertson, N., Salovey, P.: Classroom emotional climate, teacher affiliation, and student conduct. J. Classroom Interact. **46**(1), 27–36 (2011)
7. Dumont, H., Istance, D., Benavides, F. (eds.): The Nature of Learning: Using Research to Inspire Practice. OECD Publishing, Paris (2010)
8. Calvo, R., D'Mello, S., Gratch, J., Kappas, A.: The Oxford Handbook of Affective Computing. Oxford University Press, New York (2015)
9. Calvo, R.A., D'Mello, S.K.: Affect detection: an interdisciplinary review of models, methods, and their applications. IEEE Trans. Affect. Comput. **1**(1), 18–37 (2010)
10. Caballé, S.: Towards a Multi-modal emotion-awareness eLearnnng System. In: Proceedings of the Seventh IEEE International Conference on Intelligent Networking and Collaborative Systems, pp. 280–287 (2015)
11. Picard, R.W.: Emotion research by the people, for the people. Emot. Rev. **2**(3), 250–254 (2010)
12. Marsella, S., Gratch, J.: Computationally modeling human emotion. Commun. ACM **57**(12), 56–67 (2015)
13. Wong, M.: Emotion assessment in evaluation of affective interfaces. Neuron **65**(3), 293 (2006)
14. Feidakis, M., Caballé, S., Daradoumis, T., Gañán, D., Conesa, J.: Providing emotion awareness and affective feedback to virtualized collaborative learning scenarios. Int. J. Continuing Eng. Educ. Life-Long Learn. **24**(2), 141–167 (2014)
15. Caballé, S., Lapedriza, A., Masip, D., Xhafa, F., Abraham, A.: Enabling automatic just-in-time evaluation of in-class discussions in on-line collaborative learning practices. J. Digital Inf. Manage. **7**(5), 290–297 (2009)
16. Scherer, K.R.: Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? J. New Music Res. **33**(3), 239–251 (2005)
17. Pekrun, R., Goetz, T., Frenzel, A.C., Perry, R.P.: Measuring emotions in students' learning and performance: the achievement emotions questionnaire (AEQ). Contemp. Educ. Psychol. **36**(1), 36–48 (2011). Elsevier

18. Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2008). March 2007
19. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems. In: Proceedings of International Conference of Affective Feedback & Intelligent Interaction, pp. 37–42 (2009)
20. Feidakis, M., Daradoumis, T., Caballé, S.: Endowing e-Learning systems with emotion awareness. In: Proceedings of the Third International Conference on Intelligent Networking and Collaborative Systems, pp. 68–75 (2011)
21. Feldman, R.: Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013)
22. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
23. D'Mello, S.K., Craig, S.D., Witherspoon, A.M., McDaniel, B., Graesser, A.C.: Automatic detection of learner's affect from conversational cues. User Model. User-Adap. Inter. **18**(1–2), 45–80 (2008)
24. Leony, D., Muñoz-Merino, P.J., Pardo, A., Delgado-Kloos, C.: Provision of awareness of learners' emotions through visualizations in a computer interaction-based environment. Expert Syst. Appl. **40**(13), 5093–5100 (2013)
25. Caballé, S., Daradoumis, T., Xhafa, X., Juan, A.: Providing effective feedback, monitoring and evaluation to on-line collaborative learning discussions. Comput. Hum. Behav. **27**(4), 1372–1381 (2011)
26. Holovaty, A., Kaplan-Moss, J.: The Definitive Guide to Django: Web Development Done Right. Apress, New York (2009)
27. Powers, D.M.W.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. **2**(1), 37–63 (2011)

An Efficient and Secure Outsourcing Algorithm for Bilinear Pairing Computation

Xiaoshuang Luo[1,2], Xiaoyuan Yang[1,2](✉), and Xiangzhou Niu[1,2]

[1] Department of Electronic Technology, Engineering University of Chinese People's Armed Police Force, Xi'an, Shaanxi, China
xyyangwj@126.com

[2] Key Laboratory of Network and Information Security, Chinese People's Armed Police, Xi'an, Shaanxi, China

Abstract. Bilinear pairing computation is one of the most important cryptographic primitives, which is widely used in the public key encryption schemes. However, it has been considered the most expensive operation in the pairing-based cryptographic protocols. In this paper, we present an efficient and secure outsourcing algorithm for bilinear maps based on one untrusted servers. The client could outsource expensive computation to the cloud and perform simple operation to obtain the great efficiency. We analyze the security of this algorithm and compare it with prior works in efficiency. It is argued that our algorithm is more efficient and practical than the state of the art.

1 Introduction

Since the bilinear maps [1, 2] are designed by cryptographic researchers, it has been widely used in many aspects of cryptography, which promotes the huge development of cryptography. In 2000, Sakai et al. [3] proposed the identity-based key agreement scheme using bilinear pairings on elliptic curves. Especially in 2001, Boneh and Franklin [4, 5] designed the first practical identity-based encryption algorithm based on efficient computational bilinear maps constructions by exploiting the Weil pairing technology. No matter what the three party one round key agreement, identity-based encryption (IBE) [6], attribute-based encryption (ABE) [7, 8], predicate encryption (PE) [9], Function encryption (FE) [10], and searchable encryption (SE) [11] or short signatures [12] and other varieties of signatures, it plays an important role in cryptography.

With the advent of cloud computing era [13–17], bilinear maps as a cryptographic tool, plays a very significant role in the security field of cloud computing. Using the bilinear pairing operation, you can design a secure secret key sharing in the cloud and construct a searchable encryption scheme. However, bilinear pairing computation is considered to be one of the most common and expensive operations. The running time overhead on resource-constrained devices is still large, which gives a challenge to the efficiency of cryptographic algorithms. For example, some terminal equipment, including smart cards, RFID cards [18], tablets and mobile phones [19, 20], the computation power is weak. These resource-constrained devices will face the problem of limited computation. Outsourcing computation is an important way to solve this outstanding problem.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_33

In 2005, Hohenberger and Lysyanskaya [21] stated that how computation limited devices outsource complex computation task to another devices with a powerful but potential malicious property. Besides, they also defined the model of secure outsourcing computation and constructed two practical outsource-secure schemes. One is an outsource-secure exponentiation algorithm under two untrusted servers. Specially, they showed how to securely outsource modular exponentiation and presented the computation bottleneck in most public key cryptography computationally limited devices, which provide novel ideas and methods for solving complex operations. Due to the powerful computation ability of cloud, the amount of computation will reduce greatly for clients if expensive computations are outsourced to the cloud. In addition to this, it will benefit to the design and performance in devices. Therefore, how to decrease the computation complexity of the client by outsourcing computation to obtain the high efficiency has become a hot topic in the current academic research institutes.

Even though the outsourcing computation can solve many problems as the above, it inevitably brings about some security threats. There are mainly three aspects in the following required to be considered.

(1) *Secrecy.* The aim of the outsourcing computation is to reduce the cost of time rather than not guarantee the secrecy of sensitive information. The server only needs to perform computational operation as a tool regardless of what it computes. In other words, the secret information involving in the outsourcing algorithm must be kept secure and reliable.

(2) *Checkability.* What the cloud returns maybe not the trustworthy results entirely. Therefore, it must require the outsourcer to check the correctness of results with some certain probability. It means that the construction not only increase the efficiency, but also the checkability.

(3) *Assumptions.* At present, the outsourcing algorithms are based on three assumptions to be implemented [22]. A one-untrusted program (OUP) assumption means that one server implements an algorithm and the server could be malicious. A one-malicious version of a two-untrusted-program (OMTUP) assumption illustrates two servers to perform an algorithm and only one of them is malicious. A two-untrusted-program (TUP) assumption states that two servers carry out an algorithm and they could be malicious.

1.1 Related Works

Now that being this, bilinear pairing can be achieved naturally by outsourcing computation. The outsourcing computation of cryptographic operation has been researched for a long time. There are many achievements arising for secure outsourcing computation. As far as we know, the particular case of the delegated pairing computation has been first studied by Girault and Lefranc in [23]. For two points $A \in \mathbb{G}_1$ and $B \in \mathbb{G}_2$, the client tries to compute $e(A, B)$ and hide A from the server. The client can choose a random value $\mu \in \mathbb{Z}_q$ where q is the order of $\mathbb{G}_1$ and $\mathbb{G}_2$. Then the client delivers A and B to the server. The server computes $e(A, \mu B)$ and returns to the client. At last, the client could retrieve $e(A, B)$. However, except for the secrecy of the computation, they don't consider the verification of the results.

Chevallier-Mames et al. [24] proposed an elliptic curve pairing outsourcing algorithm based on one untrusted server for the first time. The card or the client hides A and B with randomly chosen numbers x and y, such that $\alpha = e(xA, yB)$ outsourced to a terminal or a computer to compute the result. Then the card obtains $e(A, B)$ by computing $\alpha^{(xy)^{-1}}$. It can also check the correctness of the results with a probability of 100% if the terminal is malicious and returns false.

Canard et al. [25] provided several delegated processes for a bilinear pairing gaining much efficiency and security than previous works. For the public value of A and B, the offline client needed one scalar multiplication in $\mathbb{G}_1$ and $\mathbb{G}_2$, one exponentiation in $\mathbb{G}_T$. Compared with the offline, the online client needed another test of membership in $\mathbb{G}_T$ or one exponentiation in $\left(\mathbb{F}_{p^k}\right)^*$. However, it only can keep verifiability, not secrecy.

Remarkably, Chen et al. [26] presented an outsourcing algorithm with high efficiency and security for bilinear pairing in the two untrusted program model (TUP). The algorithm didn't involve scalar multiplication or exponentiations. It only required the client to compute 5 point addition in $\mathbb{G}_1$ and $\mathbb{G}_2$ and 4 multiplication in $\mathbb{G}_T$, which improved the efficiency greatly. The only drawback is that the algorithm is checkable with a probability about $1/2$.

Aiming at the checkability and efficiency, Tian et al. [22] proposed two new outsourcing algorithms for bilinear pairings. One is a more efficient outsourcing algorithm under OMTUP model with the checkability about $1/2$. This algorithm required the client to operate 4 point addition in $\mathbb{G}_1$ and $\mathbb{G}_2$ and 3 multiplication in $\mathbb{G}_T$. The other is more flexible under TUP model with checkability $(1 - 1/3\text{s})^2$. Besides, they exploited the EBPV algorithm to reduce the quantity of pre-computation.

Recently, Arabaci et al. [27] put forward two new efficient algorithms for secure outsourcing of bilinear pairing. Compared with Tian et al.'s algorithms, their algorithms need less pre-computation, memory and queries to the servers. For example in communication overhead for 80-bit security, Tian et al.'s algorithm needs 0.117 KB whereas 0.078 KB or 0.098 KB needed for Arabaci et al.'s algorithms. The more details could be referred to [27].

However, except for [21] based on OUP assumption, other algorithms are based on two malicious cloud or servers. We put up with a new efficient secure bilinear pairing computation to the cloud under OUP assumption. Furthermore, we apply our outsourcing bilinear pairing algorithm as a subroutine to identity-based key-exchange protocol.

1.2 Paper Organization

In this section, the paper roadmap is given in the following. The Sect. 2 gives some preliminaries used in this paper. The security model of algorithm is introduced in Sect. 3. The proposed algorithm is described in Sect. 4 including efficiency analyses and security proof of it. Finally, we conclude this paper.

2 Preliminaries

2.1 Bilinear Pairing

Let $\mathbb{G}_1$ and $\mathbb{G}_2$ be two cyclic additive groups generated by P_1 and P_2 respectively, $\mathbb{G}_T$ be a multiplicative cyclic group. The order of $\mathbb{G}_1$, $\mathbb{G}_2$ and $\mathbb{G}_T$ is a large prime number and denoted by the symbol q. If a map $e : \mathbb{G}_1 \times \mathbb{G}_2 \to \mathbb{G}_T$ satisfies the following properties, it will be a bilinear pairing.

(1) Bilinear: For any $R \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$ and $a, b \in \mathbb{Z}_q^*$, $e(aR, bQ) = e(R, Q)^{ab}$.
(2) Non-degenerate: There are $R \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$ such that $e(R, Q) \neq 1$.
(3) Computable: There is an efficient algorithm to compute $e(R, Q)$ for any $R \in \mathbb{G}_1$, $Q \in \mathbb{G}_2$.

2.2 Pre-computation Algorithm

Particularly, the outsourcing algorithm needs a pre-computation algorithm *Rand* to produce random pairs. As the same as [22], we show an algorithm based on EBPV in the following to generate related parameters, which contains a static table *ST* and a dynamic table *DT*. The difference of them lies in the usage in algorithm. *ST* is used to store pre-computations and *DT* is used to maintain some pairs of elements.

Preprocessing Step: Generate n random integers $\alpha_1, \cdots, \alpha_n \in \mathbb{Z}_q$, where q is large prime. For $i = 1, \cdots, n$, compute $\beta_{i1} = \alpha_i P_1$ and $\beta_{i2} = \alpha_i P_2$, and store the values of a_i, β_{i1} and β_{i2} in the table.

Pair Preparation: It randomly generates $S \in \{1, \cdots, n\}$ such that $|S| = k$. For each $i \in S$, randomly choose $\chi_i \in \{1, \cdots, l-1\}$ where $l > 1$ is a small integer. Compute

$$x = \sum\nolimits_{i \in S} \alpha_i \chi_i \mod q \tag{1}$$

If $x = 0 \mod q$, start again. Otherwise compute

$$xP_1 = \sum\nolimits_{i \in S} \chi_i \beta_{i1} \tag{2}$$

and return the pair (x, xP_1). Following this way, it computes pairs (a, aP_1) and (b, bP_2) for preparation. Then randomly choose two secret values k_1, k_2 stored in the system, where $a, b, k_1, k_2 \in Z_p^*$ and compute

1. $ak_2^{-1}P_1$
2. $bk_1^{-1}P_2$
3. $e(P_1, P_2)^{-ab(k_1k_2)^{-1}}$

The entry

$$\left(A+aP_1, B+bP_2, k_1A, k_1A+ak_2^{-1}P_1, k_2B, k_2B+bk_1^{-1}P_2, e(P_1,P_2)^{-ab(k_1k_2)^{-1}}\right)$$

are stored in the *DT* table.

In addition, we compare our pre-computation algorithm with an algorithm A of [27] and an efficient algorithm 1 of [22] showed in the Table 1. The symbol of *SM*, *ME* and *PA* respectively represent scalar multiplication, modular exponentiation and point addition. According to [28], it requires about $k+h-3$ group operations for every computation of xP_1 or xP_2. Therefore, our pre-computation algorithm only needs to perform three times scalar multiplication, one time modular exponentiation and $2(k+h-3)$ times point addition operation. It improves the efficiency of our pre-computation algorithm to a large extent.

Table 1. Comparisons of pre-computation among [22, 27] and ours

Operation	Algorithm [22] A	Algorithm [27] 1	Our algorithm
SM	3	2	3
ME	2	2	1
PA	$5(k+h-3)$	$4(k+h-3)$	$2(k+h-3)$

2.3 Computational Indistinguishability

We review the notion of computational indistinguishability for secure outsourcing bilinear pairing [29].

Two distribution ensembles $X \overset{\text{def}}{=} \{X_n\}_{n\in I}$ and $Y \overset{\text{def}}{=} \{Y_n\}_{n\in I}$, where I is a countable index set, are said to be indistinguishable in polynomial time if for every probabilistic polynomial-time algorithm D, every positive polynomial $p(\cdot)$, and all sufficiently large n's,

$$|\Pr[D(X_n, 1^n) = 1] - \Pr[D(Y_n, 1^n) = 1]| < \frac{1}{p(n)} \tag{3}$$

3 Security Model

In 2005, Hohenberger and Lysyanskaya [21] defined the secure outsourcing model. Following this definition, Chen et al. and Tian et al. proposed their algorithms exactly. Our algorithm is also based on this security model. The definition and model of outsourcing computation are introduced in the following.

The algorithm includes a trusted party T and an untrusted party U. E represents an untrusted environment. T is a limited computation resourced party who tries to outsource its computation task to the party U. T^U represents a calculation of T invocating U who may be a program written by E and installed on a computer device. Therefore, we define U' as a software produced by E. Assume that two adversaries (E, U') can communicate before the execution of T^U and by means of T Otherwise. If U and E can't learn anything interesting about T's inputs and outputs, the algorithm is secure.

Definition 1 (Algorithm with outsource-*IO*): The algorithm *Alg* includes five inputs and three outputs. According to how much adversaries (E, U') knows about them, they can be classified with secret, protected and unprotected. The first input is the honest, secret input, which is unknown to E and U'. The second input is the honest, protected input, which is public for E, but is protected from U'. The third input is the honest unprotected input, which is known by both E and U'. These inputs are selected by the party T. Besides, there are two adversary-chosen inputs generated by E. One is the adversarial protected input that E know it and is protected by U'. The other is the adversarial unprotected input that are known by E and U'.

Definition 2 (Outsource-security): Let *Alg* be an algorithm with outsource-IO. The implementation T^U of *Alg* is secure if:

(1) Correctness. T^U is a correct implementation of *Alg*;
(2) Security. For all probabilistic polynomial time adversaries (E, U'), there exist probabilistic expected polynomial time simulators (S_1, S_2) such that the following pairs of random variables are computationally indistinguishable.

Pair One: $EVIEW_{real} = EVIEW_{ideal}$
The adversarial environment E can learn nothing about inputs or outputs during the execution of T^U. The real process and ideal process proceeds in rounds.

$$\begin{aligned} EVIEW^i_{real} = \Big\{ & \left(istate^i, x^i_{hs}, x^i_{hp}, x^i_{hu}\right) \leftarrow I\left(1^k, istate^{i-1}\right); \left(estate^i, j^i, x^i_{ap}, x^i_{au}, stop^i\right) \\ & \leftarrow E\left(1^k, EVIEW^{i-1}_{real}, x^i_{hp}, x^i_{hu}\right); \left(tstate^i, ustate^i, y^i_s, y^i_p, y^i_u\right) \\ & \leftarrow T^{U'\left(ustate^{i-1}\right)}\left(tstate^{i-1}, x^{j^i}_{hs}, x^{j^i}_{hp}, x^{j^i}_{hu}, x^{j^i}_{ap}, x^{j^i}_{au}\right) : \left(estate^i, y^i_p, y^i_u\right) \Big\} \end{aligned}$$

$EVIEW_{real} = EVIEW^i_{real}$ *if* $stop^i = TRUE.$

An honest and stateful process I inputs k which is a security parameter, and inputs its $i-1$ round internal states $istate^{i-1}$ to produce its i round honest state and honest inputs $x^i_{hs}, x^i_{hp}, x^i_{hu}$ for $T^{U'}$. In the same way, the adversarial environment E takes its $i-1$ round view $EVIEW^{i-1}_{real}$ and k and x^i_{hp}, x^i_{hu} to produce its i round internal state $estate^i$, the order of honest inputs j^i, the i round adversarial inputs x^i_{ap} and x^i_{au}, and a signal sign $stop^i$. The adversary U takes its $i-1$ round internal state $ustate^{i-1}$ to react with T in the round i stage. The implementation of T^U takes five inputs and the $i-1$ round internal state $tstate^{i-1}$ to produce i round internal states of T and U, and three i round outputs $\left(y^i_s, y^i_p, y^i_u\right)$. The view of the real process in round i consists of $estate^i$, and the values of y^i_p and y^i_u.

$$
\begin{aligned}
EVIEW_{ideal}^{i} = \Big\{ & \left(istate^{i}, x_{hs}^{i}, x_{hp}^{i}, x_{hu}^{i}\right) \leftarrow I\left(1^{k}, istate^{i-1}\right); \left(estate^{i}, j^{i}, x_{ap}^{i}, x_{au}^{i}, stop^{i}\right) \\
& \leftarrow E\left(1^{k}, EVIEW_{ideal}^{i-1}, x_{hp}^{i}, x_{hu}^{i}\right); \left(astate^{i}, y_{s}^{i}, y_{p}^{i}, y_{u}^{i}\right) \\
& \leftarrow Alg\left(astate^{i-1}, x_{hs}^{j^{i}}, x_{hp}^{j^{i}}, x_{hu}^{j^{i}}, x_{ap}^{i}, x_{au}^{i}\right); \left(sstate^{i}, ustate^{i}, Y_{p}^{i}, Y_{u}^{i}, replace^{i}\right) \\
& \leftarrow S_{1}^{U'\left(ustate^{i-1}\right)}\left(sstate^{i-1}, x_{hp}^{j^{i}}, x_{hu}^{j^{i}}, x_{ap}^{j^{i}}, x_{au}^{j^{i}}, y_{p}^{i}, y_{u}^{i}\right); \left(z_{p}^{i}, z_{u}^{i}\right) \\
& = replace^{i}\left(Y_{p}^{i}, Y_{u}^{i}\right) + \left(1 - replace^{i}\right)\left(y_{p}^{i}, y_{u}^{i}\right) : \left(estate^{i}, z_{p}^{i}, z_{u}^{i}\right)\Big\}
\end{aligned}
$$

$EVIEW_{ideal} = EVIEW_{ideal}^{i}$ *if* $stop^{i} = TRUE.$

In the ideal process, we use a stateful simulator S_1 to participate the algorithm. The algorithm Alg takes its $i-1$ round internal state $astate^{i-1}$ and five inputs to get i round internal state $astate^{i}$ and three outputs. The simulated implementation $S_1^{U'}$ inputs its $i-1$ round internal state $sstate^{i-1}$ and all the protected and unprotected inputs and outputs to produce the i round internal state of S_1 and U', the simulated protected and unprotected, and a signal $replace^{i} \in \{0, 1\}$. The response signal is used to determine i round $\left(z_{p}^{i}, z_{u}^{i}\right)$ for $EVIEW_{ideal}^{i}$.

Pair Two: $UVIEW_{real} = UVIEW_{ideal}$

The real view that the untrusted party U' obtains has been described in the $EVIEW_{real}^{i}$ of the above Pair One definition. So $UVIEW_{real}^{i} = ustate^{i}$ if $stop^{i} = TRUE.$

$$
\begin{aligned}
UVIEW_{real}^{i} = \Big\{ & \left(istate^{i}, x_{hs}^{i}, x_{hp}^{i}, x_{hu}^{i}\right)\Big\} \\
& \leftarrow I\left(1^{k}, istate^{i-1}\right); \left(estate^{i}, j^{i}, x_{ap}^{i}, x_{au}^{i}, stop^{i}\right) \\
& \leftarrow E\left(1^{k}, estate^{i-1}, x_{hp}^{i}, x_{hu}^{i}, y_{p}^{i-1}, y_{u}^{i-1}\right); \left(astate^{i}, y_{s}^{i}, y_{p}^{i}, y_{u}^{i}\right) \\
& \leftarrow Alg\left(astate^{i-1}, x_{hs}^{j^{i}}, x_{hp}^{j^{i}}, x_{hu}^{j^{i}}, x_{ap}^{i}, x_{au}^{i}\right); \left(sstate^{i}, ustate^{i}\right) \\
& \leftarrow S_{2}^{U'\left(ustate^{i-1}\right)}\left(sstate^{i-1}, x_{hu}^{j^{i}}, x_{au}^{i}\right) : \left(ustate^{i}\right)\Big\}
\end{aligned}
$$

$UVIEW_{ideal} = UVIEW_{real}^{i}$ if $stop^{i} = TRUE.$

The algorithm I and E are the same as those in the $EVIEW_{real}^{i}$ of the above Pair One definition. The algorithm Alg is also defined in the same way as that in the $EVIEW_{ideal}^{i}$ of the above Pair One definition. The simulated implementation $S_2^{U'}$ takes round i internal state $sstate^{i-1}$ and two unprotected inputs to produce the state of $sstate^{i}$ and $ustate^{i}$.

Assume that T^{U} is a correct execution of Alg, some definitions could be reached in the following.

Definition 3 (α-efficient, secure outsourcing): If for any input x, the running time of T is no more than an α-multiplicative factor of the running time of Alg, then the algorithm (T, U) is α-efficient secure outsourcing.

Definition 4 (β-checkable, secure outsourcing): If for any input x, T could detect any failure of U during the execution of T^U with a probability no less than β, then the algorithms of (T, U) is β-checkable secure outsourcing.

Definition 5 ((α, β)-outsource-security): If it is α-efficient and β-checkable secure outsourcing, then the algorithms (T, U) has the (α, β)-outsource security property.

4 The Efficient Secure Algorithm

This section introduces an efficient secure bilinear pairing computation algorithm and gives the proof of it in detail.

4.1 Construction

Given two cyclic addictive groups $\mathbb{G}_1$ and $\mathbb{G}_2$. There are two elements $A \in \mathbb{G}_1$ and $B \in \mathbb{G}_1$ generated by P_1 and P_2 respectively, which support bilinear pairing operation. The order of $\mathbb{G}_1$ and $\mathbb{G}_2$ is a large prime p. The notion of $U(R, Q) \to e(R, Q)$ used to denote the U taking (R, Q) as inputs to produce the output $e(R, Q)$. We use T to denote a computation limited device. The goal of this construction is to compute the result of $e(A, B)$. The detailed algorithm is in the following.

(1) *Initiation*: The pre-computation *Rand* function introduced in the Sect. 2.2 is invoked to generate random values

$$\left(A + aP_1, B + bP_2, k_1A, k_1A + ak_2^{-1}P_1, k_2B, k_2B + bk_1^{-1}P_2, e(P_1, P_2)^{-ab(k_1k_2)^{-1}}\right)$$

(2) *Query*: T queries U in a random order to get some values in the following.
 1. $U(A + aP_1, B + bP_2) \to V_1$
 2. $U\left(k_1A + ak_2^{-1}P_1, k_2B + bk_1^{-1}P_2\right) \to V_2$
 3. $U(A, B + bP_2) \to V_3$
 4. $U(A + aP_1, B) \to V_4$
 5. $U\left(k_1A, k_2B + bk_1^{-1}P_2\right) \to V_5$
 6. $U\left(k_1A + ak_2^{-1}P_1, k_2B\right) \to V_6$

(3) *Recover*: T checks whether $V_3 \cdot V_4^{-1} = V_5 \cdot V_6^{-1}$ or not. If equals, it computes

$$o = e(A, B) = \left(\frac{V_2 \cdot e(P_1, P_2)^{ab - \frac{ab}{k_1k_2}}}{V_1}\right)^{(k_1k_2-1)^{-1}}$$

$$= \left(\frac{V_2}{V_1}\right)^{(k_1k_2-1)^{-1}} e(P_1, P_2)^{-ab(k_1k_2)^{-1}}$$

and produces o as an input. Otherwise it rejects and produces “Error”.

4.2 Proofs

Theorem 1: The implementation algorithm of (T, U) for outsourcing bilinear pairing under OUP assumption is secure where the inputs are honest secret, honest protected and adversarial protected.

Proof: The correctness of this algorithm is shown in the following obviously.

$$\begin{aligned} o &= \left(\frac{V_2 \cdot e(P_1, P_2)^{ab - \frac{ab}{k_1 k_2}}}{V_1}\right)^{(k_1 k_2 - 1)^{-1}} \\ &= \left(\frac{e\left(k_1 A + a k_2^{-1} P_1, k_2 B + b k_1^{-1} P_2\right) \cdot e(P_1, P_2)^{ab - \frac{ab}{k_1 k_2}}}{e(A + aP_1, B + bP_2)}\right)^{(k_1 k_2 - 1)^{-1}} \\ &= \left(e(A, B)^{k_1 k_2 - 1}\right)^{(k_1 k_2 - 1)^{-1}} = e(A, B) \end{aligned}$$

Next, we will prove the security in detail. We first give the proof of $EVIEW_{real} = EVIEW_{ideal}$.

According to the inputs, we discuss three cases such that honest secret, honest protected and adversarial protected.

For round *i*, if the parameter (A, B) is an honest protected and adversarial protected input, the environment E always could know (A, B). Under this condition, the real environment is indistinguishable with the simulator S_1.

If the parameter (A, B) is an honest secret input, the simulator will randomly choose points $(t_1 P_1, t_2 P_2, t_3 P_1, t_4 P_2, t_5 P_1, t_6 P_2, t_7 P_1, t_8 P_2)$ for U to generate six groups.

1. $U(t_1 P_1, t_2 P_2) \rightarrow V_1'$
2. $U(t_3 P_1, t_4 P_2) \rightarrow V_2'$
3. $U(t_5 P_1, t_2 P_2) \rightarrow V_3'$
4. $U(t_1 P_1, t_6 P_2) \rightarrow V_4'$
5. $U(t_7 P_1, t_2 P_2) \rightarrow V_5'$
6. $U(t_1 P_1, t_8 P_2) \rightarrow V_6'$

On receiving the response from U, the simulator S_1 checks whether $V_3' \cdot V_4'^{-1} = V_5' \cdot V_6'^{-1}$ or not. If yes, S_1 sets $Y_p^i = \emptyset, Y_u^i = \emptyset$, and $replace^i = 0$. Otherwise, S_1 chooses a random value $o_r \in \mathbb{G}_T$ and sets $Y_p^i = o_r, Y_u^i = \emptyset$, and $replace^i = 1$.

For all cases, S_1 saves the appropriate states.

The distributions of input in the real and ideal environments are computationally indistinguishable for U. In the ideal environment, U chooses parameters uniformly at random. In the real environment, the inputs for U are also independently randomized.

If U is honest in the round i, the output of Alg in the ideal experiment is the same as the output of T^U in the real experiment. If U is dishonest, with a probability about 1 to produce “Error”. Note that the results returned by U are multiplied by random values only known by T or S_1 while others know nothing about honest secret inputs. In conclusion, $EVIEW^i_{real} = EVIEW^i_{ideal}$, and then $EVIEW_{real} = EVIEW_{ideal}$.

Theorem 2: The implementation algorithm of (T, U) for outsourcing bilinear pairing is $\left(\mathcal{O}\left(1/log_q\right), 1\right)$-outsource secure under OUP assumption.

Proof: The algorithm need a call to *Rand*, 4 point additions in $\mathbb{G}_1$ and $\mathbb{G}_2$, 2 modular multiplications, and 1 modular exponentiation. It takes roughly $1/log_q$ multiplication to compute a bilinear pair. So it is called $\mathcal{O}\left(1/log_q\right)$-outsource secure algorithm. If the server is dishonest, it fails with a probability 0. Thus, the client can check the correct of implementation with checkability 100% under OUP assumption.

4.3 Comparisons

In this subsection, we compare our algorithm with three efficient algorithms proposed by Canard [25], Chen et al. in [22] and Arabaci et al. in [27]. There are two tables shown in the following. The former is the comparison of efficiency among them and the latter is the properties. In the Table 2, the symbols of SM and ME are described in the Sect. 2.2. PA represents point addition, MM represents modular multiplication and MT represents a membership test operation.

Table 2. Comparisons of efficiency among [22, 25, 27] and ours

Operation	Algorithm [25]	Algorithm [22] A	Algorithm [27] 1	Our algorithm
SM	2	0	0	0
ME	1	0	0	1
MT	1	0	0	0
PA	2	4	4	4
MM	1	3	3	2

Compared with the state of the art, our algorithm needs to perform one modular exponentiation operation, which drops the efficiency in a way. However, we can invoke related outsourcing modular exponentiation algorithm in [30] as a subroutine to compute the results, which can solve the problem of computing one time modular exponentiation. What's more, under the same assumption, our algorithm don't need to perform scalar multiplication and membership test operation while it would need in [25]. From the Table 3, different from prior works, our algorithm is under OUP assumption and with about 100% checkability that is the great advantage among them. Even though the algorithm under OUP assumption could cost larger computational resource than OMTUP or TUP, it would be more practical in real life. To our interesting, our algorithm can reach the probability with 1.

Table 3. Comparisons of properties among [22, 25, 27] and ours

Operation	Algorithm [25]	Algorithm [22] A	Algorithm [27] 1	Our algorithm
Assumption	OUP	OMTUP	OMTUP	OUP
Secrecy	No	Yes	Yes	Yes
Checkability	1	$1/2$	$1/2$	1

5 Conclusions

In this paper, we provide an efficient and secure outsourcing algorithm for a bilinear pairing without any cryptographic assumption. Security modes are defined to guarantee the correct and secure implementation of our algorithm. It showed that pre-computation algorithm can compute random vectors efficiently by comparing with a previous algorithm. Furthermore, we proved the security and showed the efficiency.

Acknowledgments. This work was supported in part by the National Nature Science Foundation of China (Grant Nos. U1636114, 61402531, 61572521) and the Nature Science Basic Research Plan in Shaanxi Province of china (Grant Nos. 2014JM8300, 2014JQ8358, 2015JQ6231, 2016JQ6037).

References

1. Su, Z., Sun, C., Li, H., Ma, J.: A method for efficient parallel computation of Tate pairing. Int. J. Grid Util. Comput. **3**, 43–52 (2012)
2. Zhang, J., Zhang, F.: Linear threshold verifiable secret sharing in bilinear groups. Int. J. Grid Util. Comput. **4**, 212–218 (2013)
3. Sakai, R., Ohigishi, K., Kasahara, M.: Cryptosystems based on pairing. In: Symposium on Cryptography and Information Security, pp. 135–148 (2000)
4. Boneh, D., Franklin, M.: Identity-based encryption from the Weil pairing. In: Advances in Cryptology—CRYPTO 2001, pp. 213–229. Springer, Heidelberg (2001)
5. Luo, S., Chen, Z.: Hierarchical identity-based encryption without key delegation in decryption. Int. J. Grid Util. Comput. **5**, 71–79 (2014)
6. Sun, X., Jiang, Z., Zhou, M., Wang, Y.: Versatile identity-based signatures for authentication in multi-user settings. Int. J. Grid Util. Comput. **5**, 156–164 (2014)
7. Garg, S., Gentry, C., Halevi, S., Zhandry, M.: Fully secure attribute based encryption from multilinear maps. IACR Cryptology ePrint Archive 2014/622
8. Zhu, S., Yang, X.: Protecting data in cloud environment with attribute-based encryption. Int. J. Grid Util. Comput. **6**, 91–97 (2015)
9. Wee, H.: Dual system encryption via predicate encodings. In: Theory of Cryptography Conference, pp. 616–637. Springer, Heidelberg (2014)
10. Lewko, A., Okamoto, T., Sahai, A., Takashima, K., Waters, B.: Fully secure functional encryption: attribute-based encryption and (hierarchical) inner product encryption. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 62–91. Springer, Heidelberg (2010)
11. Boneh, D., Di Crescenzo, G., Ostrovsky, R., Persiano, G.: Public key encryption with keyword search. In: International Conference on the Theory and Applications of Cryptographic Techniques, pp. 506–522. Springer, Heidelberg (2004)

12. Boneh, D., Lynn, B., Shacham, H.: Short signatures from the Weil pairing. In: International Conference on the Theory and Application of Cryptology and Information Security, pp. 514–532. Springer, Heidelberg (2001)
13. Guo, S., Xu, H.: A secure delegation scheme of large polynomial computation in multi–party cloud. Int. J. Grid Util. Comput. **6**, 1–7 (2014)
14. Manoharan, M., Selvarajan, S.: An efficient methodology to improve service negotiation in cloud environment. Int. J. Grid Util. Comput. **6**, 150–158 (2015)
15. Khan, N., Al-Yasiri, A.: Cloud security threats and techniques to strengthen cloud computing adoption framework. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**, 50–64 (2016)
16. Yuriyama, M., Kushida, T.: Integrated cloud computing environment with IT resources and sensor devices. Int. J. Space-Based Situated Comput. **1**, 163–173 (2011)
17. Mezghani, K., Ayadi, F.: Factors explaining IS managers attitudes toward cloud computing adoption. Int. J. Technol. Hum. Interact. (IJTHI) **12**, 1–20 (2016)
18. Sakurai, S.: Prediction of sales volume based on the RFID data collected from apparel shops. Int. J. Space-Based Situated Comput. **1**, 174–182 (2011)
19. Varaprasad, G., Murthy, G.S., Jose, J., D'Souza, R.J.: Design and development of efficient algorithm for mobile ad hoc networks using cache. Int. J. Space-Based Situated Comput. **1**, 183–188 (2011)
20. Morreale, P., Goncalves, A., Silva, C.: Mobile ad hoc network communication for disaster recovery. Int. J. Space-Based Situated Comput. **5**, 178–186 (2015)
21. Hohenberger, S., Lysyanskaya, A.: How to securely outsource cryptographic computations. In: Proceedings of the 2nd International Conference on Theory of Cryptography, pp. 264–282. Springer, Berlin (2005)
22. Tian, H., Zhang, F., Ren, K.: Secure bilinear pairing outsourcing made more efficient and flexible. In: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security (2015)
23. Girault, M., Lefranc, D.: Server-aided verification: theory and practice. In: International Conference on the Theory and Application of Cryptology and Information Security, pp. 605–623. Springer, Heidelberg (2005)
24. Chevallier-Mames, B., Coron, J.S., McCullagh, N., Naccache, D., Scott, M.: Secure delegation of elliptic-curve pairing. In: International Conference on Smart Card Research and Advanced Applications, pp. 24–35. Springer, Heidelberg (2010)
25. Canard, S., Devigne, J., Sanders, O.: Delegating a pairing can be both secure and efficient. In: International Conference on Applied Cryptography and Network Security, pp. 549–565. Springer International Publishing (2014)
26. Chen, X., Susilo, W., Li, J., Wong, D.S., Ma, J., Tang, S., Tang, Q.: Efficient algorithms for secure outsourcing of bilinear pairings. Theor. Comput. Sci. **562**, 112–121 (2015)
27. Arabacı, O., Kiraz, M.S., Sertkaya, I., Uzunkol, O.: More efficient secure outsourcing methods for bilinear maps (2015)
28. Nguyen, P.Q., Shparlinski, I.E., Stern, J.: Distribution of modular sums and the security of the server aided exponentiation. In: Cryptography and Computational Number Theory, pp. 331–342. Birkhäuser, Basel (2001)
29. Goldreich, O.: The Foundation of Cryptography: Basic Applications, vol. 1, pp. 103–112. Cambridge University Press, Cambridge (2009)
30. Wang, Y., Wu, Q., Wong, D.S., Qin, B., Chow, S.S., Liu, Z., Tan, X.: Securely outsourcing exponentiations with single untrusted program for cloud storage. In: European Symposium on Research in Computer Security, pp. 326–343. Springer International Publishing (2014)

A New Broadcast Encryption Scheme for Multi Sets

Liqun Lv and Xiaoyuan Yang(✉)

Electronics Technology Department,
Engineering University of People's Armed Police, Xi'an, Shanxi, China
xyyangwj@126.com

Abstract. Broadcasters use broadcast encryption to broadcast confidential communications to arbitrary sets of users, and broadcasters individually send their corresponding ciphertext information for different sets of users. However, in the modern Internet, which is represented by cloud computing and complex networks, with the rapid increase of broadcast users and the rapid growth of the amount of broadcast information, the number of broadcast users is increasing. In order to solve this problem, a broadcast encryption scheme is proposed. In the environment of multi user collection, the new scheme has good communication and computation overhead, and the ciphertext length is only constant. The new scheme is flexible and efficient, and can be widely used in many fields, such as pay TV.

1 Introduction

Broadcast encryption is an encryption system that implements one-to-many secure communication in an insecure channel [1]. In the general broadcast encryption system, the broadcaster broadcasts the encrypted information to the users in the system, and any user can listen to the broadcast to obtain the encrypted information. Only the user in the authorized user set S can decrypt with the private key broadcast ciphertext, to restore the corresponding clear text information. If all unauthorized users collide with the broadcast information, the broadcast encryption system has complete anti-conspiracy characteristics. At present, broadcast encryption as a commonly used encryption means has been widely used in pay TV, digital rights management, satellite communications, video teleconferencing and wireless sensor networks [2].

The broadcaster can use broadcast encryption to broadcast secure communication to any user set. For different user sets, the broadcaster sends the corresponding ciphertext information separately. However, in the modern Internet, represented by cloud computing and complex networks, the number of broadcast users is increasing with the rapid increase of broadcast users and the rapid growth of broadcast information. In the current cloud environment, the cloud service provider can be divided according to the user to order business or payment costs, the cloud users will be divided into different sets of authorized users, different user collections to obtain the cloud service information the same. At present, the cloud service provider sends the corresponding ciphertext information separately for different user sets, which means that the ciphertext information sent by the cloud service provider is linearly related to the user's

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_34

collection. However, with the diversification of cloud services, the diversification of cloud users, the user's choice of ordering business is more and more complex and diverse, so the number of cloud users is also more and more, cloud service providers as broadcasters The broadcast of the information is also more and more, the burden of the broadcasting center is also increasingly heavy, performance bottlenecks also appeared, limiting the application of the broadcasting system. Therefore, the traditional simple one-to-many broadcast encryption can not meet the above application environment, and it is very important to design broadcast encryption for multi-user subsets.

Fiat and Naor first proposed the concept of broadcast encryption in 1994 [1], followed by a series of broadcast encryption schemes that have been proposed [3–6], but the ciphertext lengths of these schemes are related to users The number of linear relationship. In 2005, Boneh et al. Used the bilinear pairing BGW scheme [7], the ciphertext length and the user private key length were constant, Delerablee et al. Proposed identity-based dynamic broadcast encryption scheme DPP07 [8], Gentry And others constructed a broadcast security scheme GW09 [9] with adaptive security and short ciphertext length. But the public key length of these schemes is linearly related to the number of users. In order to reduce the public key overhead, Boneh et al. Constructed a low-overhead broadcast encryption scheme using multi-linear mapping [10]. Under the premise that the ciphertext and user private key length are constant, the public key length is only *O (log (N))*. Other such as identity-based broadcast encryption scheme, revocable broadcast encryption scheme, certificate-based broadcast encryption scheme, etc. have also been proposed [11–14]. BZ14 uses the indistinguishable confusion to construct the first low-cost broadcast encryption scheme with receiver privacy protection [15]. In the flexibility of the scheme, Ohtake et al. Proposed that the BEPM scheme achieves one-to-one private communication between broadcasters and users [16]. Xu et al. Constructed an identity-based BEPM scheme using multi-linear mapping [17] However, these schemes can cause large ciphertext and computational overhead in a multiuser subsurface environment, so it is worth further research to design low overhead for broadcast encryption in a multiuser subsets environment.

This paper combines the idea of broadcast encryption and group encryption, constructs a low overhead broadcast encryption scheme suitable for multiuser subsets environment, and proves the choice of plaintext security under the standard model. The user's private key and the broadcast ciphertext of the new scheme are composed of three groups of elements, which leads to lower storage and communication overhead.

In Sect. 2, we give the preliminary knowledge of this paper. We present our construction of key encapsulation mechanism scheme in Sect. 3. In Sect. 4, its security analysis is described respectively. In the next chapter, we make a conclusion.

2 Preliminaries

2.1 Multilinear Maps

Our scheme are implemented in multilinear mapping groups. We assume $\mathcal{G}$ is a group generator, which takes a security parameter 1^{λ} and the number of allowed pairing operation n as input. $\mathcal{G}(1^{\lambda}, n)$ outputs a sequence of groups $G = (\mathbb{G}_1, \ldots, \mathbb{G}_n)$ each of prime order $p > 2^{\lambda}$. We let $g = g_1$ and g_{i} be a canonical generator of $\mathbb{G}_{\mathrm{i}}$.

There is a sequence of bilinear maps:

$$\left\{e_{i,j} : G_i \times G_j \rightarrow G_{i+j} | i,j \geq 1; i+j \leq n\right\}$$

where $e_{i,j}$ satisfies: $e_{i,j}(g_i^a, g_j^a) = g_{i+j}^{ab} : \forall a, b \in \mathbb{Z}_p$.

2.2 The Diffie-Hellman Inversion Assumption

The Diffie-Hellman inversion assumption says that given $g, g^b \in G_1$, it is hard to calculate $e(g, \cdots g)^{1/b} \in G_n$. Define a randomized algorithm $\mathcal{A}$'s advantage in solving the problem to be the probability that $\mathcal{A}$ is able to compute $e(g, \cdots g)^{1/b}$ from g, g^b.

Definition 1. The multilinear map system satisfies the Diffie-Hellman inversion assumption if for all polynomial time algorithm $\mathcal{A}$, the probability is negligible.

2.3 Broadcast Encryption Scheme for Multi Sets

Combining the formal definition of broadcast encryption scheme and the general construction of group encryption, the formal definition of broadcast encryption for multiuser subsets are given below.

A broadcast encryption system for multiple user subsets can be described by the following four algorithms:

Setup: The algorithm take the system public key PK, the system master private key MSK and the security parameters as input, public the system public parameters and public key, and keep its system master private key.

KeyGen: The algorithm KeyGen take the public key PK, the master private key MSK and the user's information and broadcast user group i as input, the output of its corresponding private key SK_{ij}.

Enc: The broadcast encryption algorithm Enc take the system public key PK and the set S of the broadcast user group as an input, then it computes (Hdr, K, K_i), K is used to encrypt the common broadcast information and K_i is used for encrypting the information of each different group, and finally broadcasting $(Hdr, C, C_i(i \in S))$.

Dec: The broadcast decryption algorithm Dec takes the system public key PK, the broadcast group i where the user is located, the user private key, the broadcast header Hdr, and the broadcast group set S as input. If $i \in S$, the algorithm outputs (K, K_i), then decrypts the ciphertext and get the plaintext.

The broadcast encryption scheme for multi-user subsets needs to satisfy decryption consistency. For any

$$\begin{aligned}(PK, MSK) &\leftarrow \mathbf{Setup}(\lambda)\\ SK_{ij} &\leftarrow \mathbf{KeyGen}(PK, MSK, i, j)\\ (Hdr, K, K_i) &\leftarrow \mathbf{Enc}(PK, S)\end{aligned}$$

If $i \in S$, then there is $(K, K_i) \leftarrow \mathbf{Dec}(PK, i, j, SK_{ij}, Hdr, S)$。

2.4 Broadcast Encryption Scheme for Multi Sets Security Model

The chosen plaintext security model of broadcast encryption schemes for multiple user subsets is described by the attacker conducted by the attacker $\mathcal{A}$ and the challenger. The game process consists of the following six stages:

Initialization phase: The attacker $\mathcal{A}$ outputs the set of broadcast groups S^* to challenge to the challenger.

System build phase: Challenger running $(PK, MSK) \leftarrow \mathbf{Setup}(\lambda)$, keep the master private key MSK and sends the public key PK to the attacker.

Private key generation query: the attacker can ask any users in the group S^* for the private key, the challenger to run and then send the private key to the attacker $\mathcal{A}$.

Challenge: When an attacker $\mathcal{A}$ decides to finish the private key generation query, the challenger runs $(Hdr, K, K_i) \leftarrow \mathbf{Enc}(PK, S)$ to obtain the challenging broadcast header Hdr^* and the challenge symmetric key $(K_0^*, \{K_i^*\}_{i \in S^*})$. Challenger selects $\{b_i\}_{i \in S^*}, b \in \{0, 1\}$, if $b = 0, b_i = 0$, the challenger will send (Hdr^*, K_0^*, K_i^*) to the attacker. If $b = 1, b_i = 0$, the challenger randomly selects K_1^* and K_i^* sends to the attacker.

The attacker can continue to make private key generation queries, the challenger runs $SK_{ij} \leftarrow \mathbf{KeyGen}(PK, MSK, i, j)$ and sends the private key to the attacker.

Guess: finally, the attacker outputs guess $b', \{b_i'\}_{i \in S^*}$, if $b' = b, b_i' = b_i$, which means that attackers win the game. Let $|S^*|$ be the total number of elements in the groups, the attacker's advantage of winning the attack game can be expressed as

$$Adv^{BE}(\lambda) = |\Pr\left[(b' = b) \wedge (b_i' = b_i)_{\forall i \in S^*}\right] - \frac{1}{2} \times \frac{1}{2^{|S^*|}}|$$

Definition 2. For an attacker of any polynomial time, the attacker is given for the given multiple user subsets. If the attacker wins the advantage of the attack game, $Adv^{BE}(\lambda) \leq \epsilon$ where the probability associated with the security parameter is negligible, then the system is said to be statically chosen plaintext security.

3 Our Construction

In this part, we construct a key encapsulation mechanism from multilinear maps. Our scheme consists of the following four algorithms:

Setup $(1^\lambda, n)$: The algorithm takes as input the security parameter as well the number of broadcast groups. (p, G_1, G_n, e) is a multilinear map group system, let g, h be the generators of G_1. Then, it chooses randomly $g_1, \cdots g_n \in G_1$ and sets a function $f(i) : \{1, \cdots, n\} \to G_1$ as follows:

$$f(i) = \begin{cases} g_i \text{ if } i \in S \\ g \text{ otherwise} \end{cases}$$

where $S \subseteq \{1, \cdots, n\}$.. Next it randomly chooses $\alpha, \gamma \in \mathbb{Z}_p$ and sets $w = g^\gamma$. So the public key is $PK = \{w, g, g_1, \cdots g_n\}$. For each group i, $(i \in [1, n])$, it randomly selects $s_i \in \mathbb{Z}_p, i \in [1, n]$ and computes the public key of its group $PK_i = w^{s_i}$. For each user p_j in group i, the private key is calculated as follows:

- Randomly select $m_j \in \mathbb{Z}_P$ and compute $n_j \in \mathbb{Z}_P$ to satisfy $s_i = (m_j + n_j) \bmod p$。
- Compute $d_{ij1} = h^{m_j}, d_{ij2} = (h^\gamma)^{n_j}$。
- Compute $d_i = g_i^\alpha$.

So the private key for p_j in group i is $SK_{ij} = (d_i, d_{ij1}, d_{ij2})$.

Enc (PK, n): The algorithm takes the public key PK and the number of groups as input. Next, it randomly chooses $t \in \mathbb{Z}_p$ and computes:

$$C_0 = g^t, C_1 = w^t$$
$$K = e(f(1), \cdots, f(n))^\alpha$$

The encryption key for each group i is $K_i = e(h_2, PK_i)^t = e(g, h_2)^{\gamma t s_i}, i \in [1, n]$ So, the ciphertext is $Hdr = (C_0, C_1)$.

Dec (SK_{ij}, Hdr): The algorithm takes the ciphertext Hdr and secret key SK_{ij} as input, then it computes that

$$K = e(f(1), \cdots f(i-1), d_i, f(i+1), \cdots f(n))$$

The symmetric encryption key for each group i is:

$$K_i = e(C_1, d_{ij1}) \cdot e(C_0, d_{ij2})$$

4 Program Analysis

4.1 Correctness

$$\begin{aligned} K &= e(f(1),\cdots f(i-1),d_i,f(i+1),\cdots f(n)) \\ &= e(f(1),\cdots f(i-1),g_i^{\alpha},f(i+1),\cdots f(n)) \\ &= e(f(1),\cdots,f(n))^{\alpha} \\ K_i &= e(C_1,d_{ij1})\cdot e(C_0,d_{ij2}) = e(Y^t,h_2^{m_j})\cdot e(g^t,(h_2^{\gamma})^{n_j}) \\ &= e(g,h_2)^{\gamma t m_j}\cdot e(g,h_2)^{\gamma t n_j} \\ &= e(g,h)^{\gamma t s_i} \end{aligned}$$

4.2 Security Proof

In this section, we will prove that the security of the broadcast encryption scheme for multi sets constructed in this paper depends on the GDDHE assumption. The following proves that the theorem holds:

Theorem 1. Assuming MDHI (Multilinear Diffie-Hellman inversion assumption) is assumed to be true in the multilinear groups system (p, (G_1, G_n), e), then the constructed broadcast encryption scheme is CPA security.

Proof. Suppose that there is an attack algorithm $\mathcal{A}$ to break the CPA security of the constructed broadcast encryption scheme with a non-negligible advantage, then it is possible to construct a simulation algorithm to break the MDHI assumption with a non-negligible advantage.

Initialization: The simulation algorithm $\mathcal{B}$ constructs a multilinear group system (p, (G_1, G_n), e). The attacker $\mathcal{A}$ submits a broadcast set $S^*\subseteq[1,n]$ of challenges to the simulation algorithm.

Setup: $\mathcal{B}$ randomly chooses the generator $g \in G_1$ and $\alpha,\beta,r,t \in \mathbb{Z}_P$, sets $Y = g^{\beta}$, $W = g^t, g' = g^b, b \in [1, n-1]$. $\mathcal{B}$ also randomly chooses $r_1,\cdots r_n \in \{1,\cdots,n\}$. For $i \in S^*$ set $g_i = g^{r_i}$. For $i \notin S^*$ set $g_i = g'^{r_i}$. $\mathcal{B}$ sends $PK = \{g, g_1, \cdots, g_n, Y, W\}$ to $\mathcal{A}$.

Phase1: The attacker $\mathcal{A}$ can ask the algorithm $\mathcal{B}$ for the following private key generation query:

When the adversary ask for the private key generation query, for any user p_{ij} in $i \notin S^*$, $\mathcal{B}$ randomly chooses $b_1,\cdots,b_n \in \mathbb{Z}_P, h \in G$ and computes the private key that:

- $d_i = g_i = g'^{r_i} = g_i^{1/b}$, so $\alpha = 1/b$
- Randomly chooses $m_j \in \mathbb{Z}_P$ and computes $n_j \in \mathbb{Z}_P$ to satisfy $b_i = (m_j + n_j) \bmod p$
- $d_{ij1} = h^{m_i}, d_{ij2} = (h^{\beta})^{n_j}$

Challenge: The attacker $\mathcal{A}$ asked to challenge the ciphertext, $\mathcal{B}$ calculate the broadcast ciphertext header $Hdr = (W, W^{\beta}) = (g^t, g^{\beta t})$ and send it to $\mathcal{A}$. $\mathcal{B}$ randomly selects $b \in \{0,1\}$, if $b = 0$, $\mathcal{B}$ calculates $K^* = e(g_1, \cdots g_n)^{1/b}$, otherwise, $\mathcal{B}$ randomly selects $K^* \in G_n$. The same for each broadcast group $\mathcal{B}$ randomly selects $b_i \in \{0,1\}$, if $b_i = 0$,$\mathcal{B}$ calculates $K_i^* = e(W^{\beta}, h^{b_i}) = e(g,h)^{\beta t b_i}$, otherwise, $\mathcal{B}$ randomly selects K_i^*. The ultimate response to the challenge of the attacker is $(Hdr^*, K^*, \{K_i^*\}_{i \in S^*})$. Obviously the response is valid and effective, so the perfect simulation of the attacker on the structure of the broadcast encryption scheme attacks.

Phase 2: The attack algorithm $\mathcal{A}$ continues to ask for private key inquiries. The simulation algorithm $\mathcal{B}$ performs the same steps as the inquiry phase 1 to reply to both types of queries.

Guess: The attacker outputs the guess b' and $\left\{b_i'\right\}_{i \in S^*}$. When $\left\{b_i'\right\}_{i \in S^*} = \left\{b_i'\right\}_{i \in S^*}$ and $b' = b$, it means that the attacker $\mathcal{A}$ won the game. $\mathcal{A}_{win}$ indicates that the attacker can correctly guess b and $\left\{b_i'\right\}_{i \in S^*}$. $\mathcal{B}_{win}$ indicates that the simulation algorithm can solve the MDHI assumption. $|S^*|$ indicates the number of all elements in the collection. Therefore, if K^* and $\{K_i^*\}_{i \in S^*}$ are correct, then the probability of occurrence of the event is

$$\begin{aligned} Pr[\mathcal{B}_{win}] &= Pr[\mathcal{B}_{win}|\mathcal{A}_{win}] \cdot Pr[\mathcal{A}_{win}] + Pr[\mathcal{B}_{win}|\bar{\mathcal{A}}_{win}] \cdot Pr[\bar{\mathcal{A}}_{win}] \\ &= 1 \times (1/2 \times 1/2^{|S^*|} + \epsilon) + 1/2 \times (1 - (1/2 \times 1/2^{|S^*|} + \epsilon)) \\ &= 1/4 + 1/2^{|S^*|} + 1/2 + \epsilon/2 \end{aligned}$$

Similarly, if K^* and $\{K_i^*\}_{i \in S^*}$ are the random elements in group, that is, the attacker has no advantage to guess b' and $\{b_i'\}_{i \in S^*}$, then the probability of occurrence of the event is

$$\begin{aligned} Pr'[\mathcal{B}_{win}] &= Pr[\mathcal{B}_{win}|\mathcal{A}_{win}] \cdot Pr[\mathcal{A}_{win}] + Pr[\mathcal{B}_{win}|\bar{\mathcal{A}}_{win}] \cdot Pr[\bar{\mathcal{A}}_{win}] \\ &= 1 \times (1/2 \times 1/2^{|S^*|}) + 1/2 \times (1 - (1/2 \times 1/2^{|S^*|})) \\ &= 1/4 + 1/2^{|S^*|} + 1/2 \end{aligned}$$

Therefore, the advantages of the simulation algorithm $\mathcal{B}$ solves the MDHI assumption is $Pr[\mathcal{B}_{win}] - Pr'[\mathcal{B}_{win}] = \epsilon/2$.Therefore, if the attack algorithm $\mathcal{A}$ has an non-negligible advantage to break the constructed broadcast encryption scheme, then the simulation algorithm has an non-negligible to solve the MDHI assumption.

5 Performance Analysis

In this section, we analyze the performance of the new scheme through two aspects: computation overhead and communication overhead. In the performance comparison, n represents the number of users in the broadcast encryption scheme, p and e represent the pairing operations and exponential operations. Since the cost of the multiplication is much less than the cost of the pairing operations and exponential operations, Therefore, when calculating the overhead, the multiplication is ignored, and only the exponential

Table 1. Compares with the existing scheme performance

Scheme	Ciphertext length	Public key length	Private key length	Encryption overhead
BGW05	2 m	$O(n)$	1	$2me$
GW09	2 m	$O(n)$	n	$2me$
BWZ14	2 m	$O(log(n))$	1	$2me$
Our scheme	3	$O(m)$	3	$3e$

and pairing are considered. In addition, since the constructed scheme is for a multiple sets, the number of user sets in the scheme is represented by m. As shown in Table 1:

It can be seen from Table 1 that compared with the traditional broadcast encryption scheme, the new scheme has obvious advantages on the basis of a small increase in the length of the user's key, the length of the ciphertext and the amount of encryption. In the case of the m receivers In the environment of broadcast communication, the amount of secret text generated when encrypting is only three groups of elements, and the computational cost is only 3 times. In addition, the new scheme is well flexible, and when the number of users in each broadcast group is 1, the new scheme is transformed into a conventional broadcast encryption scheme; when all users are in a group, that is, m = 1, The new scheme is transformed into a broadcast encryption scheme with a specified user set, and the system's public key length, cipher text length and user private key length reach the constant level. In summary, the new scheme is superior in communication and computing overhead, and has good flexibility.

6 Summary

In this paper, we constructs a broadcast encryption for multiple sets, and proves the chosen plaintext security. Our scheme solves the problem that the traditional broadcast encryption traffic is large in the multiuser subsurface environment. The ciphertext length is only constant, and the encryption is only a few times. At the same time, the new program also has very good flexibility, can be converted into the traditional fixed ciphertext length of the broadcast encryption or designated user subset of the broadcast encryption. The analysis shows that the new scheme is flexible and efficient, and can be widely used in the current complex network communication environment.

Acknowledgments. This work was supported by National Natural Science Foundation of China (Grant Nos. U1636114,61572521), Natural Science Basic Research Plan in Shaanxi Province of China (Grant Nos. 2016JQ6037).

References

1. Fiat, A., Naor, M.: Broadcast encryption. In: Proceedings of Advances in Cryptology - CRYPTO 1993, International Cryptology Conference, Santa Barbara, California, USA, 22–26 August 1993, pp. 480–491 (1993)

2. Zou, X., Xiang, J.: Dynamic broadcast encryption scheme with revoking user. Wuhan Univ. J. Nat. Sci. **18**(6), 499–503 (2013)
3. Dodis, Y., Fazio, N.: Public key broadcast encryption for stateless receivers. In: DRM 2002, vol. 2696, pp. 61–80 (2002)
4. Dodis, Y., Fazio, N.: Public key trace and revoke scheme secure against adaptive chosen ciphertext attack. Lecture Notes in Computer Science, vol. 2567, pp. 100–115 (2003)
5. Goodrich, M.T., Sun, J.Z., Tamassia, R.: Efficient tree-based revocation in groups of low-state devices. In: Advances in Cryptology - CRYPTO 2004, International Cryptology Conference, pp. 511–527 (2004)
6. Halevy, D., Shamir, A.: The LSD broadcast encryption scheme. In: Advances in Cryptology - CRYPTO 2002, International Cryptology Conference, pp. 47–60 (2002)
7. Dan, B., Gentry, C., Waters, B.: Collusion resistant broadcast encryption with short ciphertexts and private keys. In: CRYPTO 2005, vol. 3621, pp. 258–275 (2005)
8. Delerablée, C.: Identity-based broadcast encryption with constant size ciphertexts and private keys. In: Advances in Crypotology, International Conference on Theory and Application of Cryptology and Information Security, pp. 200–215 (2007)
9. Gentry, C., Waters, B.: Adaptive security in broadcast encryption systems (with Short Ciphertexts). In: Advances in Cryptology - EUROCRYPT 2009 (2009)
10. Dan, B., Waters, B., Zhandry, M.: Low overhead broadcast encryption from multilinear maps. In: Advances in Cryptology – CRYPTO 2014. Springer, Heidelberg (2014)
11. Ren, Y., Wang, S., Zhang, X.: Non-interactive dynamic identity-based broadcast encryption without random oracles. In: Information and Communications Security. Springer, Heidelberg (2012)
12. Park, S., Lee, K., Dong, H.L.: New constructions of revocable identity-based encryption from multilinear maps. IEEE Trans. Inf. Forensics Secur. **10**(8), 1–1(2015)
13. Boneh, D., Waters, B.: A fully collusion resistant broadcast, trace, and revoke system. In: ACM Conference on Computer and Communications Security, pp. 211–220 (2006)
14. Gu, C.: An improved multilinear map and its applications. Int. J. Inf. Technol. Web. Eng. **10** (3), 64–81 (2015)
15. Dan, B., Zhandry, M.: Multiparty key exchange, efficient traitor tracing, and more from indistinguishability obfuscation. Algorithmica **8616**, 1–53 (2014)
16. Ohtake, G., Hanaoka, G., Ogawa, K.: Efficient broadcast encryption with personalized messages. In: International Conference on Provable Security, pp. 214–228 (2010)
17. Xu, K., Liao, Y., Qiao, L., Liu, Z., Yang, X.: An identity-based (IDB) broadcast encryption scheme with personalized messages (BEPM). PLoS ONE **10**(12), e0143975 (2015)
18. Wei, Z.: A pairing-based homomorphic encryption scheme for multi-user settings. Int. J. Technol. Hum. Interact. (IJTHI) **12**(2), 72–82 (2016)
19. Chen, Y., Chen, X., Li, H.: More dcca-secure public-key encryptions from kem + dem style hybrid paradigms and some observations on the 'inner-outer' structure. Int. J. Grid Util. Comput. **5**(1), 60–70 (2014)
20. Li, S., Zhang, F.: Leakage-resilient identity-based encryption scheme. Int. J. Grid Util. Comput. **4**(2/3), 187–196 (2013)
21. Chen, H., Hu, Y., Lian, Z., Jia, H., Wang, X.A.: An additively homomorphic encryption over large message space. Int. J. Inf. Technol. Web. Eng. **10**(3), 82–102 (2015)
22. Ma, J., Zhang, Y., Wang, Z., Yu, K.: A message topic model for multi-grain SMS spam filtering. Int. J. Technol. Hum. Interact. (IJTHI) **12**(2), 83–95 (2016)

Key Encapsulation Mechanism from Multilinear Maps

Liqun Lv[1], Wenjun Sun[1], Xiaoyuan Yang[1](✉), and Xuan Wang[1,2]

[1] Electronics Technology Department,
Engineering University of People's Armed Police, Xi'an, Shanxi, China
llq654@163.com
[2] Xidian University, Xi'an, Shanxi, China

Abstract. The key encapsulation mechanism (KEM) and the data encapsulation mechanism (DEM) form a hybrid encryption, which effectively solves the problem of low efficiency of public key cryptography and key distribution problems in symmetric encryption system. The security and efficiency of the key encapsulation mechanism directly affect the security and efficiency of hybrid encryption. In this paper, an identity-based key encapsulation scheme is constructed by using multilinear mapping. We proved that the scheme is under the standard model of adaptive chosen-ciphertext security. The scheme can be publicly verified and the key and ciphertext length are constant and have high efficiency.

1 Introduction

Identity-based encryption (IBE) has been a hot issue in public-key cryptography since it was introduced by Shamir et al. in 1984 [1]. In the identity-based cryptosystem, the user's public key can be any string, such as the user's e-mail address, telephone, etc. Identity-based encryption does not require a digital certificate, simplifying the user's public key management process. However, similar to other public key encryption schemes, the identity-based encryption system also has the drawback that the encryption and decryption speed is slow and the efficiency is not high, and the length of the plaintext is limited. The encryption and decryption operation of symmetric cryptosystem is speed, high efficiency, and there is no limit to the plaintext space, but there is a problem of key management distribution difficult. Hybrid encryption combines the advantages of two encryption systems, namely the use of key encapsulation machine KEM and data encapsulation mechanism DEM combination of the way [2]. Key encapsulation mechanism KEM is similar to the public key cryptosystem, except that the encryption of plaintext is transformed into encapsulation of keys. That is, KEM does not have a plaintext input, and the decryption result is a symmetric encrypted key. Combining the key encapsulation mechanism with the identity-based encryption system constitutes the identity-based key encapsulation mechanism [3]. Compared with the traditional key encapsulation mechanism, the identity-based key encapsulation mechanism does not need to use digital certificate, so it has a good application prospect.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_35

Key encapsulation mechanism is an open, insecure channel, for the key encapsulation mechanism of active attacks not only eavesdropping, intercept, and even tamper with the data in the channel, in order to obtain valuable information. Thus, for identity-based key encapsulation schemes, it is not enough to achieve the chosen plaintext security (CPA), and it is more desirable to be able to achieve chosen ciphertext security (CCA) to resist active attackers. According to Canetti, Halevi and Katz, we can get the chosen ciphertext security by using the identity based encryption system and one-time signature [4]. However, the one-time signature system itself will result in higher computational and storage overhead, and practicality is not strong. In order to improve efficiency, Boneh and Katz proposed the use of message authentication code (MAC) instead of a signature system [5]. However, the addition of message authentication code based on identity encryption scheme decryption requires the use of private key authentication, which limits the application of the scene. Therefore, whether it is a signature system or message authentication code there are some flaws. How to achieve the chosen ciphertext security without the use of a signature system and message authentication code worthy of further study. Boyen et al. proposed to improve the security of the scheme by using the hash function and the internal structure of the ciphertext [6].

Cramer et al. first defined the security model of hybrid encryption [2] in 2003, that is, the combination of KEM and DEM. In 2005, Bentahar first combined KEM with an identity-based encryption system and gave its general structure [3]. Subsequently, Chen et al. gave the specific structure of identity-based KEM [7]. So far, many scholars have proposed many safe and efficient identity-based key encapsulation schemes [8–10]. In addition, other key-based encapsulation schemes [14, 15], such as certificate-based key encapsulation schemes [11–13], have been proposed. Wang et al. constructed a identity-based key encapsulation scheme using multilinear mapping [16]. The ciphertext and private key length of the scheme are one group element. The decryption is only one operation, storage and operation efficiency of the scheme is higher, but the security of the scheme only achieves the chosen plaintext security. Therefore, it is worthy to study how to construct an efficient KEM with the chosen ciphertext security.

In this paper, we use the ideas proposed by Boyen et al. and multilinear mapping to construct an identity-based publicly verifiable key encapsulation scheme with adaptive ciphertext security. The ciphertext and private key length of the scheme is short and fixed length, without using a signature system or message authentication code. The scheme is computationally efficient and publicly verifiable.

In Sect. 2, we give the preliminary knowledge of this paper. We present our construction of key encapsulation mechanism scheme in Sect. 3. In Sect. 4, its security analysis is described respectively. In the next chapter, we make a conclusion.

2 Preliminaries

2.1 Multilinear Maps and Assumption

Our scheme are implemented in multilinear mapping groups. We assume $\mathcal{G}$ is a group generator, which takes a security parameter 1^λ and the number of allowed pairing operation n as input. $\mathcal{G}(1^{\lambda}, n)$ outputs a sequence of groups $G = (\mathbb{G}_1, \cdots, \mathbb{G}_n)$ each of prime order $p > 2^{\lambda}$. We let $g = g_1$ and g_i be a canonical generator of $\mathbb{G}_i$.

There is a sequence of bilinear maps:

$$\{e_{i,j} : G_i \times G_j \rightarrow G_{i+j} | i, j \geq 1; i + j \leq n\}$$

where $e_{i,j}$ satisfies: $e_{i,j}(g_i^a, g_j^b) = g_{i+j}^{ab} : \forall a, b \in \mathbb{Z}_p$.

2.2 Identity-Based Key Encapsulation Mechanism

An identity-based key encapsulation scheme can be described by the following four algorithms:

(PK, MSK) ← **Setup(** λ**):** The system establishes the algorithm **Setup** to output the public key PK and the master private key MSK with the security parameter λ as input.

SK_{ID} ← **KeyGen**(PK, MSK, ***ID***): The private key generation algorithm **KeyGen** outputs its corresponding private key SK_{ID} with the public key PK, the master private key MSK and the user identity $ID \in \mathcal{I}$ as input ($\mathcal{I}$ is user's identity space).

(C, K) ← **Encap**(PK, ID): The encapsulation algorithm **Encap** uses the public key PK and the user ID as input, and outputs the ciphertext C and the key $K \in \mathcal{K}$ where $\mathcal{K}$ is the key space.

K ← **Decap**(C, Sk_{ID}): The decapsing algorithm Decap uses ciphertext C and user private key SK_{ID} as input, output key K or invalid symbol ⊥.

The identity-based key encapsulation scheme must meet the decryption consistency. For any

$$(\text{PK}, \text{MSK}) \leftarrow \textbf{Setup}(\lambda)$$
$$SK_{ID} \leftarrow \textbf{KeyGen}(\text{PK}, \text{MSK}, ID)$$
$$(C, K) \leftarrow \textbf{Encap}(\text{PK}, ID)$$
$$\text{If } ID \in \mathcal{I}, \text{ then there is } K \leftarrow \textbf{Decap}(C, SK_{ID}).$$

2.3 Identity-Based Key Encapsulation Mechanism Security Model

The adaptive selection of the identity-based key encapsulation mechanism. The ciphertext security model is described by the attack game conducted by the attacker and the challenger. The game process consists of the following six stages:

Initialization phase: The attacker $\mathcal{A}$ outputs the identity to challenge the challenger ID^*

System build phase: Challenger run (PK, MSK) ← Setup (λ), keep the master private key, and sent public key PK to the attacker.

Query Phase 1: The attacker can adaptively asked the challenger for the following two types of inquiries:

(1) private key generation query: the attacker can conduct private key queries on any $ID \neq ID^*$, the challenger to run Sk_{ID} ← **KeyGen**(PK, MSK, ID), and then sent the private key SK_{ID} to the attacker.

(2) decaping query: The attacker can choose an identity $ID \neq ID^*$ to generate the corresponding ciphertext C_{ID}. The challenger runs the algorithm $K \leftarrow$ **Decap** (C, SK_{ID}) and sends the symmetric key K to the attacker $\mathcal{A}$.

Challenge stage: When the attacker decides to finish the end of the phase 1, the challenger runs $(C^*, K) \leftarrow$ **Encap**(PK, ID^*), obtains the ciphertext C^* and challenges the symmetric key $K \in \mathcal{K}$. The challenger then randomly selects another symmetric key $K' \overset{R}{\longleftarrow} \mathcal{K}$ in the symmetric key space and randomly selects b $\leftarrow$ {0, 1}. The challenger sets $K_b = K, K_{1-b} = K'$, and then sends (C^*, K_0, K_1) to the attacker.

Query Phase 2: The attacker continues to ask for private key generation query and decapsulation query. But in the decapsulation query, the attacker can not ask the challenge ciphertext C^*.

Guess: Finally, the attacker output a guess b', if $b' = b$ then the attacker wins the game.

q_e said the number of attackers to execute attack information in private key generation in the game, q_d said the attackers to execute the total number of solutions in the attack in the game package information. At this point, the attacker to win the offensive game advantage can be expressed as

$$Adv_{A,n,q_e,q_d}^{IB-KEM}(\lambda) = |\Pr[b' = b] - \frac{1}{2}|$$

Definition 1. For an IND-sID-CCA2 attacker for any polynomial time, the attack game is performed for a given identity-based key encapsulation mechanism. The private key generation ask for q_e time, decapsulation q_d time to ask. If an attacker $\mathcal{A}$ win the advantage to satisfy attacking game

$$Adv_{A,n,q_e,q_d}^{IB-KEM}(\lambda) \leq \varepsilon$$

Where ε is the probability negligible function associated with the safe parameter, then the system is said to be $(n, q_e, q_d, \varepsilon)$-IND-sID-CCA2.

3 Our Construction

In this part, we construct a key encapsulation mechanism from multilinear maps. Our scheme consists of the following four algorithms:

Setup $(1^\lambda, n)$: The algorithm takes as input the security parameter as well the bit-length n of identities. It first runs $\mathcal{G}(1^\lambda, n)$ and outputs a sequence of groups $(G_1, \cdots, G_n)$ of prime order p, with generators $g_1, \cdots g_n$, where we let $g = g_1$.

Then, it chooses random exponents $(b_{1,0}, b_{1,1}), \cdots (b_{n,0}, b_{n,1}) \in \mathbb{Z}_p^2$ and sets $B_{i,\beta} = g^{b_{i,\beta}}$ for $i \in [1, n], \beta \in \{0, 1\}$. Next it randomly chooses $y_1, y_2 \in \mathbb{Z}_p$ and computes $u_1 = g_{n-1}^{y_1}, u_2 = g_{n-1}^{y_2}$. We also random choose a hash function $H_0 : G_1 \rightarrow \mathbb{Z}_P$. So the public parameters is $PK = \{(B_{1,0}, B_{1,1}), \cdots (B_{n,0}, B_{n,1}), H_0, u_1, u_2\}$ and the master key $MSK = \{(b_{1,0}, b_{1,1}), \cdots (b_{n,0}, b_{n,1})\}$.

Keygen (PK, MSK, ID): The algorithm takes as input the master key MSK, the public key PK and the identity $ID = (id_1, \cdots, id_n)$ and outputs the secret key $SK_{ID} = g_{n-1}^{\prod_{i \in [1,n]} b_{i,id_i}}$.

Encap (PK, ID): The algorithm takes as input the public key PK and identity ID. Next, it randomly chooses $t \in \mathbb{Z}_p$ and computes:

$$C_0 = g^t, C_1 = u_1^t u_2^{tw}, \text{where } w = H_0(C_0)$$

$$K = e(B_{1,id_1}, \cdots, B_{1,id_1})^t = g_n^{t \prod_{i \in [1,n]} b_{i,id_i}}$$

So, the ciphertext is $C = (C_0, C_1)$.

Decap (SK_{ID}, C): The decapsulation algorithm takes as input the ciphertext C and secret key SK_{ID}, then it computes that

(1) Computes the hash value of the ciphertext $C_0 : w = H_0(C_0)$
(2) Verify the validity of the ciphertext C: Verify the equation $e(C_0, u_1 u_2^w) = e(C_1, g)$ is valid, if not equal, then the ciphertext C is invalid and outputs $\perp$.
(3) If the ciphertext C is valied, computes $K = e(SK_{ID}, C_0)$.

4 Program Analysis

4.1 Correctness

If the given ciphertext $C = (C_0, C_1)$ is a valid ciphertext, then $w = H_0(C_0)$,

$$\begin{aligned} e(C_0, u_1 u_2^w) &= e(g^t, u_1 u_2^w) = e(g, u_1^t u_2^{tw}) \\ &= e(C_1, g) \end{aligned}$$

Therefore, if the equation $e(C_0, u_1 u_2^w) = e(C_1, g)$ is satisfied, then

$$\begin{aligned} K &= e(SK_{ID}, C_0) = e(g_{n-1}^{\prod_{i \in [1,n]} b_{i,id_i}}, g^t) \\ &= g_n^{t \prod_{i \in [1,n]} b_{i,id_i}}. \end{aligned}$$

4.2 Security Proof

In this section, we will prove that the security of the key encapsulation scheme constructed in this paper depends on the MDDH assumption. The proof of thinking is similar to that of Wang et al.'s key encapsulation scheme, but should be modified slightly to allow the simulation algorithm to reply the decapsulation query. In the initial stage of the attack, when the simulation algorithm receives the challenge ID^* selected by the attacker, the simulation algorithm first calculates the C_0^* part of the challenge ciphertext, and then calculates $w^* = H_0(C_0^*)$ to generate the public key of scheme. When the attacker submits the ciphertext $C = (C_0, C_1)$ to the decryption query, the simulation algorithm first verifies the validity of the ciphertext C, and then the simulation algorithm calculates the encapsulated key to answer the attacker's decapsulation query.

The following proves that the theorem holds:

Theorem 1. Assuming MDDH is assumed to be true in the multilinear groups system $(p, (G_1, \cdots, G_n), e)$ then the constructed identity-based key encapsulation scheme is IND-sID-CCA security.

Proof: Suppose that there is an attack algorithm $\mathcal{A}$ to break the IND-sID-CCA2 security of the constructed key encapsulation scheme with a non-negligible advantage, then it is possible to construct a simulation algorithm to break the MDDH assumption with a non-negligible advantage. $\mathcal{B}$ takes $D = (g, g^{c_1}, \cdots, g^{c_{n+1}})$ and T as input, where $g = g_1, T = g_n^{\prod_{j\in[1,n+1]} c_j}$ or T is a random element in group G_n. The target of $\mathcal{B}$ is to output 1 when $T = g_n^{\prod_{j\in[1,n+1]} c_j}$, otherwise 0 is output.

Initialization: The attack algorithm $\mathcal{A}$ outputs an identity $ID^* = (id_1^*, \cdots, id_n^*)$ which it want to challenge.
Setup: $\mathcal{B}$ randomly chooses $b_1, \cdots, b_n \in \mathbb{Z}_p$, and sets $B_{i,id_i^*} = g^{c_i}, B_{i,1-id_i^*} = g^{b_i}$, $i \in [1, n]$. $\mathcal{B}$ randomly chooses a hash function $H_0 : G_1 \rightarrow \mathbb{Z}_P$ and computes $w^* = H_0(g^{c_{n+1}})$. $\mathcal{B}$ also randomly choose $\delta \in \mathbb{Z}_p$ and sets $u_1 = (g_{n-1}^{\prod_{j\in[1,n+1]} c_j})^{-w^*} g_{n-1}^{\delta}, u_2 = g_{n-1}^{\prod_{j\in[1,n-1]} c_j}$. $\mathcal{B}$ sends $\{(B_{1,0}, B_{1,1}), \cdots, (B_{n,0}, B_{n,1}), H_0, u_1, u_2\}$ to $\mathcal{A}$.

Phase1: The attacker $\mathcal{A}$ can adeptly ask the algorithm $\mathcal{B}$ for the following private key generation query and key decapsulation query:

(1) When the adversary ask for the private key generation query, for any $ID \neq ID^*$ and let $ID_i = (id_{i,1}, \cdots, id_{i,n})$ because of $ID \neq ID^*$, there must be at least one bit $id_{i,j} \neq id_j^*, j \in [1, n]$. $\mathcal{B}$ computes the private key $SK_{ID_i} = e(B_{i,id_{i,1}}, \cdots, B_{j-1,,id_{i,j-1}}, B_{j+1,,id_{i,j+1}}, B_{n,id_{i,n}})^{b_j}$ then the simulation algorithm will send the private key to attack algorithm $\mathcal{A}$.

(2) When the attack algorithm performs a key decapsulation query, for a query ciphertext $C = (C_0, C_1)$, let $w = H_0(C_0)$. The algorithm $\mathcal{B}$ first verifies the validity of the ciphertext: whether $e(C_0, u_1 u_2^w) = e(C_1, g)$ is true. If not true, the ciphertext C is not legal and $\mathcal{B}$ reply $\perp$. Otherwise, the ciphertext is legal, then ciphertext $C = (C_0, C_1) = (g^t, u_1^t u_2^{tw})$, where t is unknown.

If $w = w^*$, then the simulation algorithm can not reply to the query, and then the simulation algorithm $\mathcal{B}$ ends the simulation of the attack algorithm, randomly outputs of a bit $b \in \{0, 1\}$ finish the game.

If $w \neq w^*$, it should be calculated:

$$K = \frac{e(C_0, (g_{n-1}^{b_{n,id_n}})^{\frac{-\delta}{w-w^*}} u_1 u_2^w)}{e(C_1, (g^{b_{n,id_n}})^{\frac{-\delta}{w-w^*}} g)} \\ = g_n^{t \prod_{j \in [1,n]} b_{j,id_i,j}}$$

The algorithm $\mathcal{B}$ sends the encapsulated key K to $\mathcal{A}$.

Challenge: When the attack algorithm $\mathcal{A}$ decides to finish the query phase 1, the simulation algorithm $\mathcal{B}$ needs to output a valid ciphertext $C^* = (C_0^*, C_1^*) = (g^{c_{n+1}}, (g_{n-1}^{c_{n+1}})^{\delta})$ and the challenge encapsulation key $K^* = T \in G_n$.

We can see that only in $C^* = (C_0^*, C_1^*) = (g^{c_{n+1}}, u_1^{c_{n+1}} u_2^{c_{n+1} w^s})$, the challenge ciphertext is valid, where c_{n+1} is unknown to the challenger. When $T = g_n^{\prod_{j \in [1,n+1]} c_j}$, the challenge symmetric key K^* is valid. When $T \overset{R}{\leftarrow} G_n$, the challenge symmetric key K^* is not associated with the challenge ciphertext C^*.

Phase 2: The attack algorithm $\mathcal{A}$ continues to ask for private key inquiries and key decapsulation requests. The simulation algorithm $\mathcal{B}$ performs the same steps as the inquiry phase 1 to reply to both types of queries.

Guess: In the end, the attack algorithm $\mathcal{A}$ outputs a guess $b' \in \{0, 1\}$, and when $b' = 1$, it is assumed that K^* is the correct symmetric key. The simulation algorithm also uses b' as its output and let $b = b'$, when $b = 1$, the simulation algorithm is assumed to be $T = g_n^{\prod_{j \in [1,n+1]} c_j}$.

If $T = g_n^{\prod_{j \in [1,n+1]} c_j}$, then the challenge symmetric key K^* is the correct key associated with the challenge ciphertext C^*. If T is a random element in group G_n, then K^* is the random key in the key space, and the simulation process is perfect. Therefore, if the attack algorithm $\mathcal{A}$ has an non-negligible advantage to break the constructed identity-based key encapsulation scheme, then the simulation algorithm has an non-negligible advantage to solve the MDDH assumption.

5 Performance Analysis

In this section, we analyze the performance of the new scheme through two aspects: computation overhead and communication overhead, and the efficiency of the proposed scheme is compared with the existing identity based key encapsulation scheme. The computational cost is mainly composed of two parts: encryption and decryption. In addition, the length of the private key, public verifiability and security are compared, as shown in Table 1.

Table 1. Compares with the existing scheme performance

Scheme	Ciphertext length	Computation overhead	Private key length	Publicly verifiable	Security
Scheme [14]	$G_1 + 2G_n + S$	$8P + T_s$	G_{n-1}	Yes	CCA
Scheme [15]	G_1	$2P$	G_{n-1}	No	CPA
Our scheme	$G_1 + G_{n-1}$	$8P + T_s$	G_{n-1}	Yes	CCA2

In the performance comparison, only the time-consuming of multilinear pair operation is considered (using P for the time of a multilinear pair operation), the cost of one-time signature (using T_s for the calculated cost of the signature, S for the signature communication overhead) and the hash function computational overhead (denoted by T_h). In Table 1, g_i represents an element in a group G_i. As can be seen from Table 1, compared with [15], our scheme has not used a one-time signature on the basis of the IND-sID-CCA2 security, while reducing the ciphertext length and encryption and decryption of the calculation overhead. Compared with the scheme [16], our scheme still maintains a high efficiency on the basis of the standard model under the IND-sID-CCA2 security and publicly verifiable. In terms of communication overhead, the ciphertext length of the new scheme only increases the elements in a group G_{n-1}, and the private key length does not change. In terms of computational overhead, only one hash function operation and two multilinear pairs are added.

6 Summary

In this paper, we construct a publicly verifiable identity-based key encapsulation scheme, and proves the adaptive chosen ciphertext security under the standard model. Our scheme does not use one-time signature algorithm, only use the hash function, through the ciphertext internal verification mechanism to achieve a publicly verifiable and adaptive chosen ciphertext security. The analysis shows that the proposed scheme achieves better security at the expense of less cost, and is suitable for practical communication.

References

1. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Blakley, G.R., Chaum, D. (eds.) Advances in Cryptology. Springer, Heidelberg (1984). Lecture Notes in Computer Science, 21(2)
2. Cramer, R., Shoup, V.: Design and analysis of practical public-key encryption schemes secure against adaptive chosen ciphertext attack. SIAM J. Comput. **33**(1), 167–226 (2004)
3. Bentahar, K., Farshim, P., Malone-Lee, J., Smart, N.P.: Generic constructions of identity-based and certificateless kems. J. Cryptology **21**(2), 178–199 (2008)
4. Ran, C., Halevi, S., Katz, J.: Chosen-ciphertext security from identity-based encryption. SIAM J. Comput. **36**(5), 1301–1328 (2007)

5. Boneh, D., Katz, J.: Improved efficiency for CCA-secure cryptosystems built using identity-based encryption. In: Menezes, A. (ed.) Topics in Cryptology, vol. 261. Springer, Heidelberg (2005)
6. Boyen, X., Mei, Q., Waters, B.: Direct chosen ciphertext security from identity-based techniques. In: ACM Conference on Computer and Communications Security, pp. 320–329 (2005)
7. Cheng, Z., Malonelee, J., Chen, L., Smart, N.P.: Efficient id-kem based on the sakai–kasahara key construction. Inf. Secur. IEE Proc. **153**(1), 19–26 (2006)
8. Kiltz, E., Galindo, D.: Direct chosen-ciphertext secure identity-based key encapsulation without random oracles. In: Australasian Conference on Information Security and Privacy, pp. 336–347 (2006)
9. Long, Y., Chen, K.: Efficient chosen-ciphertext secure certificateless threshold key encapsulation mechanism. Inf. Sci. **180**(7), 1167–1181 (2010)
10. Lippold, G., Boyd, C., Nieto, J.M.G.: Efficient certificateless KEM in the standard model. In: International Conference on Information Security and Cryptology, 34–46 (2009)
11. Li, J., Huang, X., Mu, Y., Susilo, W., Wu, Q.: Constructions of certificate-based signature secure against key replacement attacks. J. Comput. Secur. **18**(3), 421–449 (2010)
12. Li, J., Huang, X., Mu, Y., Susilo, W., Wu, Q.: Certificate-based signature: security model and efficient construction. In: Public Key Infrastructure, European Pki Workshop: Theory and Practice, Europki 2007, pp. 110–125 (2007)
13. Li, J., Huang, X., Zhang, Y., Xu, L.: An efficient short certificate-based signature scheme. J. Syst. Softw. **85**(2), 314–322 (2012)
14. Zhang, M., Zhang, T., Wang, X.: Publicly Verifiable Encryption in Multilinear Maps[J]. J. Wuhan Univ. Nat. Sci. Ed. **2014**(6), 507–512 (2014)
15. Wang, H., Wu, L., Zheng, Z., Wang, Y.: Identity-based key-encapsulation mechanism from multilinear maps. IACR Cryptology ePrint Archive 2013/836
16. Wei, Z.: A Pairing-based homomorphic encryption scheme for multi-user settings. Int. J. Technol. Hum. Interact. (IJTHI) **12**(2), 72–82 (2016)
17. Chen, Y., Chen, X., Li, H.: More dcca-secure public-key encryptions from KEM + DEM style hybrid paradigms and some observations on the 'inner-outer' structure. Int. J. Grid Util. Comput. **5**(1), 60–70 (2014)
18. Li, S., Zhang, F.: Leakage-resilient identity-based encryption scheme. Int. J. Grid Util. Comput. **4**(2/3), 187–196 (2013)
19. Chen, H., Hu, Y., Lian, Z., Jia, H., Wang, X.A.: An additively homomorphic encryption over large message space. Int. J. Inf. Technol. Web. Eng. **10**(3), 82–102 (2015)
20. Ma, J., Zhang, Y., Wang, Z., Yu, K.: A message topic model for multi-grain SMS spam filtering. Int. J. Technol. Hum. Interact. (IJTHI) **12**(2), 83–95 (2016)

An Multi-hop Broadcast Protocol for VANETs

Li Yongqiang(✉), Wang Zhong, Fan Qinggang, Cai Yanning,
and Chen Baisong

Computer Staff Room, Xian Research Institute of High Technology,
Xian 710025, China
lyq200381@163.com

Abstract. Many applications in VANETs rely on reliable broadcast, while broadcasting can led to broadcast storm problem. Probabilistic broadcast is a kind of simple and effective way to suppress broadcast storm. However, a lot of existing probabilistic broadcast protocols for VANETs haven't taken network division problem in low traffic density into consideration. This paper design and implement an enhanced local connectivity-based broadcast protocol. Simulation results show that the new protocol can achieve better reliability with lower overhead in both dense and sparse traffic scenario.

1 Introduction

Many applications of VANETs, especially security applications, rely on reliable and efficient broadcasting. The important information will be broadcast within a certain range of the vehicle and the whole network node, and can effectively improve the driving safety and driving experience. The easiest way to broadcast is flooding, but flooding at high node density will cause serious competition and collision, causing the broadcast storm problem [1]. Probabilistic broadcast [2] is a simple and effective way to alleviate the broadcast storm.

Most of the existing vehicle network probabilistic broadcasting algorithm reduce the broadcast storm at the same time in different degrees, not considering the network vehicle network in the low node density of the segmentation problem, leading to low reliability in the low node density of the agreement. In the past, these Study on vehicle network broadcast probability, only Ozan et al. Considering the node density of vehicles caused by changes in network connectivity changes, put forward to work in dense and sparse scene node in DV-CAST protocol [3, 4]. However, DV-CAST also has some restricting its gain wider application, this paper will be based on these defects, put forward an enhanced multi hop broadcast protocol of vehicle network based on local connectivity.

2 ELCMBP Enhanced Broadcast Protocol

The probabilistic broadcast protocol is used to alleviate the broadcast storm, but it does not take into account the problem of network segmentation in low node density. The DV-CAST protocol proposed by Ozan et al. Makes up for this deficiency. DV-CAST gets

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_36

one hop node information through Beacon messages periodically, and combined with the node's location and driving direction to determine the 3 State markers: DFlag (Destination Flag), MDC (Message Direction Connectivity) and ODC (Opposite Direction Connectivity) value, and then informed the local connectivity information. When the vehicle is in the local scene identified full connectivity (MDC = 1), DV-CAST protocol adopts weighted-p stick method to suppress the broadcast storm, when that node is disconnected (MDC = 0 and ODC = 0) or local sparse connectivity (MDC = 0 and ODC = 1) scene, DV-CAST will broadcast messages into a cache queue or the buffer queue, when new the neighbors will broadcast the task will be forwarded to the new neighbors, to ensure that more nodes can receive the broadcast message.

Although the introduction of DV-CAST carry-and-forward attempts to increase the reliability of the protocol at low node density, the DV-CAST protocol has its inherent drawbacks, which will be introduced in section third.

2.1 Disadvantages of DV-CAST

First, DV-CAST is designed for safety message broadcast protocol, so DV-CAST only consider the broadcast message to the rear of the vehicle, and they can not be achieved multi direction broadcast extension in the condition of without adjustment protocol design [5]. But the vehicle network application for broadcast needs a variety of news broadcast only in the rear of the vehicle is not clearly enough, a simple example is: the vehicle nodes have access to the Internet and other vehicles (may not have Internet access) weather and traffic information sharing. Even for security applications, sometimes also need to spread the news to the direction of travel, such as the rear of the vehicle is out of control and risk warning to the front of the vehicle rear end broadcast in danger, vehicle traffic accidents occurred on the road are affected in each direction in the condition of no isolation belt.

Secondly, DV-CAST is not a good design of the forwarding conditions of cached messages, there is a situation that the cached messages are blindly forwarded, which seriously affects the efficiency of the protocol [6]. Figure 1 is an example: when the A initiates a broadcast, all nodes in the graph will go through a single hop or multi hop forwarding received the broadcast message, according to the DV-CAST protocol [7], the G node and the H node, the received broadcast packets into the cache queue in order to meet the new neighbor when traveling to different forwarding, so when C, D,

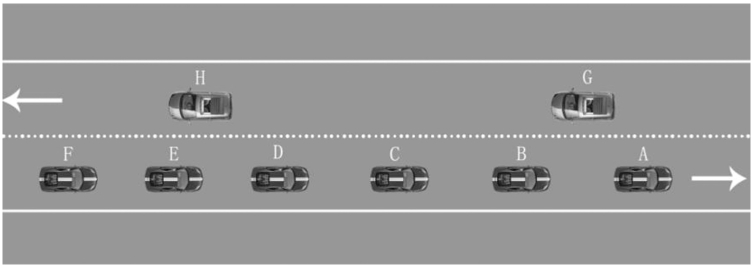


Fig. 1. Common traffic scenario in straight road

E, F and G were later met, although they have received a radio message from A to G, but they will still broadcast again.

In the scene and conditions of fourth section, we analyzed the DV-CAST will send a message in the buffer queue for the message transmission times and the total number of experiments has already received the news of the driving direction is different from the proportion of neighbors, as shown in Table 1:

Table 1. Redundancy sending ratio of messages in DV-CAST's buffer queue I

Scene	Vehicle density						
	2	4	8	12	16	20	24
Expressway	24.27%	49.09%	66.50%	59.07%	34.79%	22.45%	17.68%
City	34.24%	14.13%	44.01%	33.54%	51.22%	45.74%	42.85%

It can be seen that in the DV-CAST protocol, the redundancy ratio of I messages in the cache queue is very high, which will significantly increase the retransmission times of broadcast messages.

2.2 ELCMBP Protocol Design

First, the ELCMBP protocol uses a periodic Beacon message to obtain the information of a hop neighbor node, and the neighbors are divided into three groups: the forward neighbor, the backward neighbor and the opposite neighbor. Then according to the neighbor information to determine the status of 3 markers: PosFlag (Position Flag), DirFlag (Direction Flag) and ODC (Opposite Direction Connectivity) the value of the final agreement will be based on the value of PosFlag, DirFlag and ODC, and the news broadcast forwarding cache decision. The meanings of the three state flags are as follows: PosFlag: used to indicate the position of a vehicle in the same direction. When PosFlag is 0, it said that node does not travel in the same direction of neighbors; when PosFlag is 1 said the node for the first node of the vehicle group (no forward neighbors); when PosFlag is 2, it said that this node is the node of the vehicle tail of group (no backward neighbors); when PosFlag is 3, it said that in the vehicle group node edge. DirFlag: used to indicate the direction of broadcast message propagation. When DirFlag is 1, it said that the message is propagated forward; when DirFlag is 2, it said that the message is the back propagation; when DirFlag is 0, it is means that the direction of propagation is uncertain. the DirFlag Flag of 0 is used only when a new broadcast is initiated. ODC: used to indicate whether the node has a different direction to the neighbors. ODC takes 0 to indicate that there is no difference to the traveling neighbors; when ODC takes 1, it represents only the difference between the neighbors of the different directions; when ODC takes 2, it represents the number of different neighbors traveling in the neighborhood. Compared with DV-CAST, ELCMBP protocol, ODC not only represents the driving direction is different neighbors or not, also show the amount.

The ELCMBP protocol is introduced into DirFlag to realize the two-way propagation of broadcast messages, and the local connectivity of nodes is determined by PosFlag and ODC. The nodes of the DirFlag and PosFlag will be included in the broadcast message, it receives a broadcast message can be obtained on a node DirFlag and PosFlag information on a node. DirFlag information used to determine the node DirFlag, a node PosFlag will be used to assist the forwarding decision.

As the DV-CAST protocol, the ELCMBP protocol uses two cache queues: the cache queue I and the cache queue II, which store broadcast messages that need to be forwarded. when the node does not have a different neighbor, The cache queue I is used; when there is a different neighbor, the cache queue II is used.

2.2.1 Broadcast Message Initiation

When broadcasting, the DirFlag of the broadcast message is 0, because at this point in the direction of each the neighbors are involved in the message forwarding, used to inform the broadcast message received by the node at this time the message propagation direction undetermined. According to the neighbor information obtained PosFlag and ODC, Table 2 to determine the measures taken by the node:

Table 2. Originating rules of broadcast

PosFlag	ODC	Measures taken	
		Broadcast	Cache queue
0	0	No	DirFlag takes 0 into the cache queue I
1	0	Yes	DirFlag takes 1 into the cache queue I
2	0	Yes	DirFlag takes 2 into the cache queue II
0/1	1/2	Yes	DirFlag takes 1 into the cache queue II
Other possible combinations		Yes	No

Table 2 in addition to the “other possible combinations”, the nodes in local are poor connectivity in the scene, the message is to be broadcast into the buffer queue, when they met nodes which are appropriate messages spread to more distant nodes, they will retransmit the broadcast message.

2.2.2 Broadcast Message Forwarding

When receiving a new broadcast message, the ELCMBP protocol first determines the DirFlag and then selects the forwarding strategy according to the relationship between the previous hop node and its own. If the previous hop node is the same with their neighbors, when the last hop DirFlag is 0, this section will be based on the local connectivity to determine DirFlag; otherwise, this DirFlag and the last hop consistent. The specific details of the ELCMBP protocol are shown in Table 3:

If the previous hop node is different from the node to the neighbor, when the last hop DirFlag is 0, this section will be based on the local connectivity to determine DirFlag; otherwise, the hop message propagation direction is the opposite. The specific details of the ELCMBP protocol are shown in Table 4. In Tables 3 and 4, “broadcast suppression” refers to the use of a weighted p- to adhere to the probability of broadcasting, the other need to broadcast in the case of a probability of 1 broadcast. The different values of DirFlag, PosFlag and ODC indicate that the nodes are in different local connectivity environment, and assume different broadcast forwarding tasks.

Table 3. Forwarding rules in the same direction

Last DirFlag	PosFlag	ODC	This DirFlag	Measures taken	
				Broadcast	Cache queue
0	1	0	1	No	Into the cache queue I
	1	1/2	1	Yes	Into the cache queue II
	2	0	2	No	II
	2	1/2	2	Yes	No
	3	–	–	Broadcast suppression	No
1	1	0	1	No	Into the cache queue I
	1	1/2		Yes	Into the cache queue II
	3	–		Broadcast suppression	No
2	2	0	2	No	Into the cache queue I
	2	1/2		Yes	No
	3	–		Broadcast suppression	No

For the message of queue I: (1) when meting the neighbor that the driving direction is different and neighbor PosFlag is 0/1; (2) when PosFlag changes, it is because that this node catching up on the run in the same direction prior to the vehicle or traveling in the same direction after the vehicle to catch up, when any of these two conditions is satisfied, the node will broadcast the queue I message and the message is deleted from the queue. the message in the cache queue II, when meting the new neighbor which PosFlag is found to be 0/1, the node will broadcast the message in the cache queue II, the neighbor PosFlag will not be deleted from the queue.

Different from the DV-CAST protocol, the ELCMBP protocol will only retransmit the message in the cache queue when the new neighbor is different from the PosFlag protocol, which is 0/1. It is because of this improvement, the ELCMBP protocol is a good way to suppress the blind forwarding of cached messages, and the effect of this improvement will be described in the next section.

When the node receives the broadcast message is repeated, the protocol first check whether the message in the queue II, if there is, and d the received forwarding hops HOP1 queue II in forwarding hops HOP2 meet

$$\mathrm{HOP1}\ > = \mathrm{HOP2}\ +\ 2$$

The message in the cache queue II will be deleted. At this point, the node will be catch the repeat broadcast satisfied with the Tables 3 and 4 rules the conditions into the cache queue, but it will not broadcast the message.

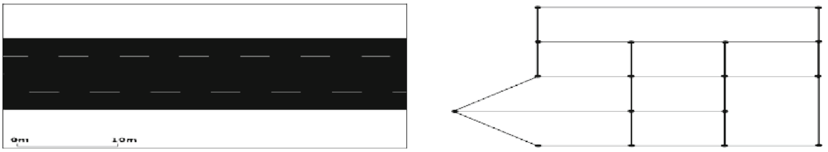
Table 4. Forwarding rules in the different direction

Last DirFlag	PosFlag	Last PosFlag	ODC	This DirFlag	ELCMBP taken measures	
					Broadcast	Cache queue
0	0	0	1	1	No	Enter the cache queue II
			2	0	Yes	DirFlag = 1 enter the cache queue II
		1	2	0	Yes	No
		2	1	1	No	Enter the cache queue II
			2	1	Yes	Enter the cache queue II
	1	0	–	2	Broadcast suppression	DirFlag = 1 enter the cache queue II
		1/3	–	2	Broadcast suppression	No
		2	–	1	No	Enter the cache queue II
	2	0/1	1	1	Broadcast suppression	No
			2	0	Broadcast suppression	No
		2/3	–	1	Broadcast suppression	No
	3	–	–	0	Broadcast suppression	No
1	0/1	0/2	–	2	–	Enter the cache queue II
		–	2		Yes	–
	2/3	–	–		Broadcast suppression	No
2	0/2	–	2	1	Yes	No
	1/3	–	–		Broadcast suppression	No

3 Simulation Experiment and Result Analysis

In order to evaluate the performance of the ELCMBP protocol, the network simulation tool NS-2 [12] is used to simulate the protocol, and the traffic flow simulation tool SUMO [13] is used to build the high speed and the urban scene. High speed scene for a long 20 km two-way 4 Lane straight highway, the local part is shown in Fig. 2a; urban scene road topology shown in Fig. 2b, (Table 5).

As a contrast, we also implemented DV-CAST, weighted p-persistence and Flooding protocol and simulation, and use the following metrics to measure the performance of the protocol:



(a) The scene of high speed local road map (b) Urban Scene road topology

Fig. 2. Roads in simulation, (a) The scene of high speed local road map, (b) Urban scene road topology

Table 5. NS-2 simulation parameters

Parameter	Value
Mac	802.11
Bandwith	2 Mbps
Communication radius	250 m
Speed in high speed	80–120 km/h
Speed in urban	0–50 km/h
Vehicle density	4, 8, 12, 16, 20, 24 Car/km
Beaconcycle	1 s
Simulation time	300 s

(1) reliability: the ratio of the number of nodes that receive a broadcast message to the number of nodes on the road.
(2) average cost: the ratio of the total number of packets sent to a broadcast to the total number of nodes. The lower the average cost, the higher the efficiency of each broadcast, the less likely to cause broadcast storm.
(3) the average overhead after correction: the ratio of the total number of packets sent by a broadcast to the number of nodes that receive the broadcast. The average overhead after correction can reflect the efficiency of broadcast protocol more objectively than average cost, especially in the case of low reliability.

From the simulation results, we can see that, compared with DV-CAST, ELCMBP protocol has better reliability, which is more obvious at low node density. The average cost of DV-CAST and the corrected average cost are very high, this is the message queue of the blind forwarding protocol caused; and the ELCMBP protocol, the average cost and corrected average cost is significantly reduced compared to DV-CAST, close to the weighted p- adhere to the protocol, the ELCMBP protocol can effectively inhibit blind forwarding cache message.

ELCMBP and DV-CAST are used to as weighted p-broadcast suppression method, but the introduction of store and forward mechanism in order to improve the reliability of broadcast, so compared to two weighted p- protocol has better reliability persist, especially in low node density; but the store and forward mechanism also increased the broadcast overhead, this is why the average overhead protocol and corrected average overhead than the weighted-p adhere.

The effect of ELCMBP on reducing the blind forwarding of cached messages can be seen in Table 6. Compared with Table 1, it is found that the ELCMBP protocol significantly reduces the redundant transmission of messages in I.

Table 6. Redundancy sending ratio of messages in ELCMBP's buffer queue I

Scene	Vehicle density						
	2	4	8	12	16	20	24
High speed	37.86%	35.66%	29.87%	9.01%	1.86%	0%	0.04%
City	33.15%	37.94%	29.43%	14.74%	1.31%	2.95%	1.53%

4 Conclusion

Based on the analysis and verification of probabilistic broadcast protocol, DV-CAST has a message can not be two-way communication, message forwarding cache easy to blindly disadvantages, we proposed an enhanced vehicle network broadcast protocol based on local connectivity. According to the local connectivity, the ELCMBP protocol determines the value of the 3 states of PosFlag, DirFlag and ODC, so as to realize the two-way broadcasting and forwarding decision. simulation results show that the ELCMBP can be fully connected and non connected, intermittent connectivity in the scene to achieve more reliable and more efficient (the average overhead and average cost adjusted measure) broadcasting, and can avoid caching in DV-CAST protocol message forwarding on the blind.

References

1. Tseng, Y.C., Ni, S.Y., Chen, Y.S., et al.: The broadcast storm problem in a mobile ad hoc network. Wirel. Netw. **8**(2–3), 153–167 (2002)
2. Reina, D.G., Toral, S.L., Johnson, P., et al.: A survey on probabilistic broadcast schemes for wireless ad hoc networks. Ad Hoc Netw. **25**, 263–292 (2015)
3. Tonguz, O.K., Wisitpongphan, N., Bai, F.: DV-CAST: a distributed vehicular broadcast protocol for vehicular ad hoc networks. IEEE Wirel. Commun. **17**(2), 47–57 (2010)
4. Wegener, A., Hellbrück, H., Fischer, S., et al.: AutoCast: an adaptive data dissemination protocol for traffic information systems. In: Proceedings of IEEE 66th Vehicular Technology Conference, pp. 1947–1951 (2007)
5. Alshaer, H., Horlait, E.: An optimized adaptive broadcast scheme for inter-vehicle communication. In: Proceedings of IEEE 61st Vehicular Technology Conference, pp. 2840–2844 (2005)
6. Wisitpongphan, N., Tonguz, O., Parikh, J.S., et al.: Broadcast storm mitigation techniques in vehicular ad hoc networks. IEEE Wirel. Commun. **14**(6), 84–94 (2007)
7. Zhou, L., Cui, G., Liu, H., et al.: NPPB: a broadcast scheme in dense VANETs. Inf. Technol. J. **9**(2), 247–256 (2010)
8. Panichpapiboon, S., Ferrari, G.: Irresponsible forwarding. In: Proceedings of the 8th International Conference on ITS Telecommunications, pp. 311–316 (2008)

9. Panichpapiboon, S.: Irresponsible forwarding under general inter-vehicle spacing distributions. In: Proceedings of the 8th International Conference on Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology, pp. 357–360 (2011)
10. Panichpapiboon, S., Cheng, L.: Irresponsible forwarding under real intervehicle spacing distributions. IEEE Trans. Veh. Technol. **62**(5), 2264–2272 (2013)
11. Mostafa, A., Vegni, A.M., Agrawal, D.P.: A probabilistic routing by using multi-hop retransmission forecast with packet collision-aware constraints in vehicular networks. Ad Hoc Netw. **14**, 118–129 (2014)
12. Miraldaa, C., Ilira, S., et al.: A simulation system based on ONE and SUMO simulators: performance evaluation of different vehicular DTN routing protocols. J. High Speed Netw. **23**(1), 59–66 (2017)
13. Prusty, A.R., Nayak, A.K.: A hybrid multi-hop mobility assisted heterogeneous energy efficient cluster routing protocol for wireless ad hoc sensor networks. J. High Speed Netw. **22**(4), 265–280 (2016)

DPHKMS: An Efficient Hybrid Clustering Preserving Differential Privacy in Spark

Zhi-Qiang Gao(✉) and Long-Jun Zhang

Department of Information Engineering, Engineering University of PAP,
Xi'an, Shaanxi, China
1090398464@qq.com, 15891741749@163.com

Abstract. *k*-means is one of the most widely used clustering algorithms by far. However, when faced with massive data clustering tasks, traditional data mining approaches, especially existing clustering mechanisms fail to deal with malicious attacks under arbitrary background knowledge. This could result in violation of individuals' privacy, as well as leaks through system resources and clustering outputs while untrusted codes are directly performed on the original data. To address this issue, this paper proposes a novel, effective hybrid *k*-means clustering preserving differential privacy in Spark, namely Differential Privacy Hybrid *k*-means (DPHKMS). We combined Particle Swarm Optimization and Cuckoo-search to initiate better cluster centroid selections in the framework of big data computing platform, Apache Spark. Furthermore, DPHKMS is implemented and theoretically proved to meet ε-differential privacy with determinative privacy budget allocation under Laplace mechanism. Finally, experimental results on challenging benchmark data sets demonstrated that DPHKMS, guaranteeing availability and scalability, significantly improves existing varieties of *k*-means and consistently outperforms the state-of-the-art ones in terms of privacy-preserving, verifying the effectiveness and advantages of incorporating heuristic swarm intelligence.

Keywords: Apache spark · Differential privacy · k-means · Particle Swarm Optimization · Cuckoo-search · Privacy-preserving

1 Introduction

The accessibility and abundance of big data today makes data mining a matter of considerable importance and necessity [1, 2]. However, the advantages of big data processing and analysis unfortunately come with a high cost in terms of new security and privacy exposures, especially, raising new challenges in the face of potentially malicious attackers [3–6]. For instance, the original users' data may be intercepted or damaged during the process of data mining by actively cheating attackers in the situation of social networks, where data privacy is seriously threatened [7]. Moreover, traditional data mining approaches can not satisfy the growing demand of the Internet and the growing complexity of cloud services [8]. Therefore, privacy-preserving data mining without violating privacy under arbitrary background knowledge has become a

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_37

new hot area of cross disciplinary research [9] and one of the ten extraordinarily challenging problems in data mining since it was introduced [10].

k-means [11], as one of the most widely used classical and effective clustering techniques in data mining, is friendly to implement and efficient with linear time complexity. Intriguingly, from the perspective of privacy-preserving, a growing body of literature is focused on differential privacy *k*-means and scholars have done a lot of fruitful research work in recent years. As demonstrated in [12], differential private data analysis can be divided into two categories: interactive approach which aims at customizing differential private algorithms and non-interactive approach aims at supporting various data mining tasks. Further, a hybrid *k*-means clustering [12] that combines interactive and non-interactive is analyzed and verified from extensive experiments by the error behavior. On the other hand, in terms of scaled big data, they could be distributed horizontally (each party owns some tuples), vertically (each party owns some attributes), or arbitrarily partitioned. Regarding this issue, privacy-preserving protocols are needed in these situations, where *k*-means can be developed and modified by applying secure multi-party computation (SMC) [13] on multi-sourced data distributions so that each party can securely produce *k* clusters and assign data in a coordinated manner. Meanwhile, protocols specialized for maintaining the privacy of *k*-means clustering when data is shared among two or more parties are highlighted in [14, 15].

Unfortunately, aimed at achieving privacy and security, numerous attempts in literature [13–15] are implemented in sacrifice of additional communication costs, such as Secure Multiparty Computation (SMC) or homomorphic encryption schemes [16]. While, many existing solutions explored for privacy-preserving *k*-means either suffer from computation overheads [13, 14] or have strict assumptions on the involved parties which are in conflict with the real-world requirements [15, 16].

To the best of our knowledge, performing differential privacy *k*-means clustering with heuristic swarm intelligence on Spark, which is a distributed program successfully deployed in production, has not been explicitly proposed in literature. Therefore, regarding the issue above, an effective differential privacy preserving *k*-means algorithm based on Spark framework (DPHKS) is proposed in this paper, which is implemented in the form of parallelization with swarm intelligence in order to improve the availability of data privacy protection and mining results. Even if attackers obtain the maximum data background knowledge at length, they are still unable to obtain sensitive information in detail. The contributions of this paper can be concluded as follows:

- We adopt the state-of-art efficient computing platform Apache Spark [17–19] to cope with the severe challenge of computational requirements in clustering very large data sets with several attributes of different types.
- We adopt differential privacy [20], a new paradigm for privacy preserving to guarantee the security and privacy of data mining in a mathematically rigorous way with regard to adversary's background knowledge.
- Notably, we implement bio-inspired swarm intelligence with Particle Swarm Optimization [21] and Cuckoo-search [22] in a parallelized form to initiate better cluster centroid selections in clustering problems. The novel mechanism can determine the optimal class number of data sets to be partitioned dynamically.

The reminder of the paper is organized as follows. Section 2 defines preliminary information of *k*-means clustering and differential privacy. In Sect. 3, we propose a hybrid privacy preserving *k*-means in Spark (DPHKMS) under a basic attack model and analyze the ability of DPHKMS regarding privacy preserving. Experimental results are presented in Sect. 4. Finally, Sect. 5 concludes the paper and points out the direction of future research.

2 Preliminaries

2.1 *k*-means Clustering

k-means clustering [11] is a powerful data mining tool, which is capable of grouping data into *k* clusters according to similarity measure, e.g. Euclidean distance. Pseudo Code of classical *k*-means can be summarized in Table 1.

Table 1. Pseudocode for classical *k*-means clustering

```
Algorithm 1. k-means clustering
Input: k, the number of centroid
Output: entities assignments to final clusters
        Select k entities as initial centroid randomly
        Repeat
            Assign each entity to the nearest centroid
            Recompute the centroid of each cluster
        Until termination criterion is achieved
```

Note that the process of *k*-means clustering is terminated when (1) the maximum number of iterations is satisfied, or when (2) no improvement can be made in centroid selection according to a specified prefixed parameter.

2.2 Differential Privacy

Differential privacy [20] is a recent privacy guarantee tailored to the problem of statistical disclosure control on how to publicly release statistical information about people without compromising the privacy of any individual. Moreover, differential privacy is accomplished by adding noise into data sets to achieve the effect of privacy protection and overcome the shortcoming of the traditional security model. From the view of statistics, differential privacy can be strictly defined and quantified as follows:

Condition: A statistical database is a set of rows or tuples. We say databases D and D_0 are adjacent or neighbors (namely identical or differ by only a record) when their Hamming distance is 1.

Definition 1. A randomized function $\boldsymbol{K}$ gives (ε, δ)-differential privacy if for all pairs of adjacent databases D and D_0 and all $S \subseteq Range(\boldsymbol{K})$,

$$Pr[\boldsymbol{K}(D_0) \subseteq Range(\boldsymbol{K})] \leqslant e^{\varepsilon} \cdot Pr[\boldsymbol{K}(D) \subseteq Range(\boldsymbol{K})] + \delta \tag{1}$$

In the special case that $\delta = 0$, we say the randomized function $\boldsymbol{K}$ obtains ε-differential privacy.

For the query function $\boldsymbol{F}$, global sensitivity is an important intrinsic attribute which measures the impact of change on a single record to the query output.

Definition 2. For query function $\boldsymbol{F}$: $D \rightarrow R^d$, the L_1-sensitivity of $\boldsymbol{F}$ is defined as follows, where D and D_0 are neighbors.

$$\Delta \boldsymbol{F} = \max_{D, D_0} \|\boldsymbol{F}(D) - \boldsymbol{F}(D_0)\|_1 \tag{2}$$

In this paper, we adopt Laplace mechanism to satisfy ε-differential privacy, which keeps the original data statistics characteristics better. The Laplace probability density function is given below:

$$\boldsymbol{Lap}(x) = \frac{1}{2\Delta \boldsymbol{F}/\varepsilon} exp(\frac{x|}{\Delta \boldsymbol{F}/\varepsilon}) \tag{3}$$

In addition, differential privacy is composable in the sense that combining ε_1, ε_2, ……, ε_m, that satisfy ε-differential privacy for $\varepsilon = \sum_{i=1}^{m} \varepsilon_i$, i.e. ε is referred to be the privacy budget in a privacy-preserving. Especially, when a data mining task involves multiple steps, each step can occupy part of ε so that the sum is allocated reasonably according to ε.

3 DPHKS Algorithm

3.1 Attack Model

In the era of big data, malicious attackers can use a variety of approaches to obtain the background knowledge of data privacy. Intuitively, when any records from data sets change, attackers, even with maximum background knowledge, can neither be able to infer any details about stored information nor learn details about query results in the DPHKS clustering process. The corresponding attack model is shown in Fig. 1.

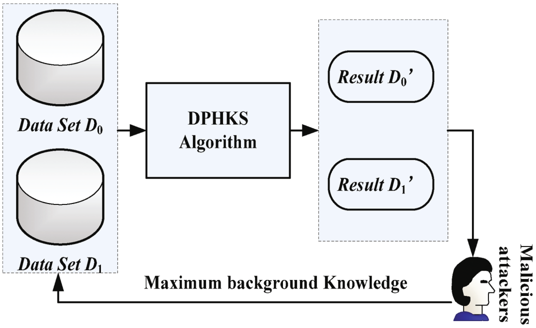


Fig. 1. The corresponding attack model

3.2 Design of DPHKS Algorithm

k-means is faced with two outstanding challenging: (1) The clustering quality highly relies on the selection of the initial centroid and easily falls into local minima. (2) The exposure of privacy is threatened during clustering. Regarding these issues, we propose DPHKS algorithm, which combines the Particle Swarm Optimization (PSO) [21] and Cuckoo-search (CS) [22] to optimize the initial position of center point which is implemented on Spark for achieving efficiency and scalability when faced with big data sets clustering problem. Additionally, differential privacy is utilized to preserve the clustering results. The pseudo code and the flowchart of DPHKS on Spark, are shown in Table 2 and Fig. 2 respectively.

Table 2. Pseudo Code for DPHKS

```
Algorithm 2. Our Proposed DPHKS
Step 1: Convert the initial data sets from HDFS into RDD data sets, and the initial cluster centers are encoded as particles;
Step 2: Map operation and train of the initial clustering centers with fitness value in form of (key, value) pairs; Then select the minimum cluster center distance and judge records belongings;
Step 3: Reduce and privacy preserving operation, where the calculation of the number of records in the cluster $num$ and its vector is processed. Then join Laplace noise disturbance to recalculate corresponding cluster center;
Step 4: Produce new RDD data sets, according to the key value of join operations.
    Step 4.1: Adopt the method of SaveAsTextFile to save clustering results and end iterations, if difference among clusters is smaller than preset value,;
    Step 4.2: Otherwise, loop Step 2~Step 4, and cache the intermediate results;
Step 5: Output clustering results.
```

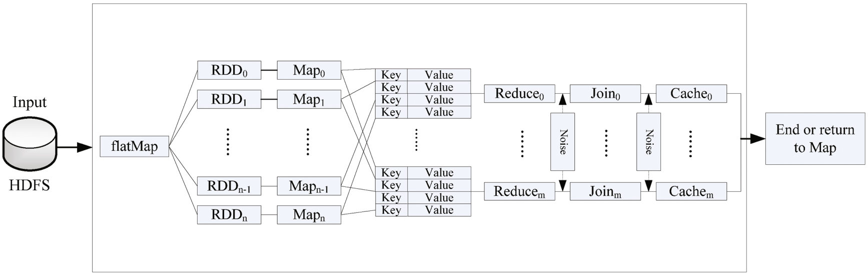


Fig. 2. Flowchart of DPHKS on spark

In **Step 1–2**, we utilize swarm intelligence optimization to initialize clustering center as particle (PSO and CS represents cluster centroid with unified name) to improve the performance of initial centers selection and accelerate the rate of clustering. Particle $\boldsymbol{X_i}$ represent centroids of k clusters, which is constructed as follows:

$$X_i = (Z_{i1}, \cdots\cdots, Z_{ik}) \tag{4}$$

where Z_{ik} refers to the k_{th} centroid of cluster C_j. Meanwhile, the updating operations of particles' velocity and position are modified in according with literature [21, 22]. The cluster centroid is described, using Eq. (5) as follows:

$$\boldsymbol{Z}_j = \frac{1}{N_{C_j}} \sum_{\forall \boldsymbol{X}_p \in C_j} \boldsymbol{X}_p \tag{5}$$

where N_{Cj} denotes the number of data cluster j. Also, particle fitness is evaluated by clustering function, which is used to measure the degree of compactness inside the clusters and be calculated according to Eq. (6).

$$J(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{K} \sum_{\boldsymbol{X}_j \in C_i} \left\| \boldsymbol{Z}_i - \boldsymbol{X}_j \right\| \tag{6}$$

In addition, the Mean square error (MSE) is the most widely used criterion, which can be defined mathematically as follows:

$$\boldsymbol{MSE} = \frac{1}{N} \sum_{j=1}^{K} \sum_{\boldsymbol{X}_i \in C_j} \left\| \mathbf{X}_i - \mathbf{Z}_j \right\| \tag{7}$$

Furthermore, Euclidean distance, which is treated as one of the most popular similarity metric in clustering can be calculated as follows:

$$d(\boldsymbol{X}_i,\boldsymbol{Z}_j) = \sqrt{\sum_{p=1}^{D} (x_{ip} - z_{jp})^2} \tag{8}$$

In **Step 3**, the variable *num* belongs to background knowledge of attackers, so we reduce the impact of excess noise on the clustering center by just adding Laplace noise to sum instead of variables of *sum* and *number*; In **Step 4**, by using key technique of spark framework, the RDD cache is adopted to avoid frequent I/O communication from HDFS.

According to Fig. 2, we calculate values in form of key-value pair from RDD in Mapper and then send them to Reducers for combining with additional calculation of Laplace noise. Due to the noise added to the results in DPHKS, the ε-differential privacy is achieved.

3.3 Privacy Analysis of DPHKS

In this paper, DPHKS is performed on the platform of Spark with all data in form of resilient distributed data set (RDD). Moreover, during the Reduce operation of clustering, data vector and variable sum are processed by adding Laplace noise disturbances to achieve privacy protection. We can conclude that DPHKS is executed in the mechanism of iteration calculation run in memory abstraction of data sets and eliminates significant amount of disk IOs due to Spark. In accordance with composable property of ε-differential privacy [23, 24], iteration of DPHKS satisfies ε privacy budget. In addition, the privacy budget allocation scheme we adopt is that each iteration t occupies privacy budget $\varepsilon_t = \varepsilon/2^t$.

On the other hand, the global sensitivity of variable *sum* in DPHKS is $d + 1$, where d is the dimension of the data sets. Therefore, in each iteration, we add $\boldsymbol{Lap}(d + 1)\ 2^{t+1}/\varepsilon$ into variable *sum* to guarantee DPHKS satisfying ε-differential privacy. Moreover, compared to traditional approaches, our DPHKS just needs to add noise disturbance to the intermediate variable *sum*, which can result in reducing the privacy preserving budget and improving the iteration efficiency with better clustering effect.

4 Experiment and Analysis

Our experiment platform has been performed on Hadoop HA, comprising 2 masters and 9 slaves. Each node features the configuration as follows: operating system CentOS 7.0 (Desktop), 3.30 GHz CPU, 16 GB of RAM and 500 GB hard disk, Ethernet 1000 Mb/s of network communication, Hadoop 3.0.0-Alpha and spark 2.0 on yarn cluster. All codes are programmed in Python 2.7.3.

To evaluate the performance of DPHKS, six benchmark data sets (Adult, Iris, Wine, Glass, Liver disorder, Vowel from the UC Irvine Machine Learning Repository on website address http://archive.ics.uci.edu), were selected. Moreover, the parameters of DPHKS are set empirically as follows: population size is equal to 80, maximum number of evaluations is set to 500. All experiments are performed 20 times.

The parameters for PSO and CS are set according to paper [21, 22]. The experimental results comparing our proposed DPHKS with several typical stochastic algorithms including *k*-means++ [25], *k*-means [11] and PSO-*k*-means [26] are presented in this section.

4.1 Clustering Efficiency

In this paper, the parallel efficiency of the algorithm is tested by using indicator of *Speedup* as testament. *Speedup* is a measure of the performance and effectiveness for parallel systems or procedures. As shown in Eq. (9), T_s denotes the single run time, and T_c denotes execution time consuming for clustering.

$$Speedup = T_s/T_c \tag{9}$$

The results from data sets Adult and Iris on *speedup* of DPHKS, *k*-means and *k*-means++ performed by 3, 5, 7, 9 nodes are shown in Table 3.

Table 3. Results on *Speedup*

Algorithm	Data set	*Speedup*			
		3 nodes	5 nodes	7 nodes	9 nodes
DPHKS	Adult	1.78	2.21	3.42	4.89
k-means		1.39	1.52	1.69	1.98
k-means++		1.55	1.74	2.05	2.76
DPHKS	Iris	1.95	3.87	4.73	6.09
k-means		1.78	2.07	2.56	3.32
k-means++		1.69	2.19	2.62	3.24

From Table 3, it shows that *Speedup*, namely acceleration rate of DPHKS is superior to comparison algorithms, *k*-means, *k*-means++ on the data sets of Adult and Iris. In addition, when the number of computing nodes increases, *Speedup* line of DPHKS is not linearly increasing. However, it can be also verified that the Spark in-memory computing model is more efficient in the iterative mode.

4.2 Computation Time

To investigate the effects of computation time, results from data set Wine on 11 nodes compared with *k*-means and *k*-means++ is shown in Fig. 3. Furthermore, the average computation time for each test benchmark data sets (Glass, Liver disorder) is calculated. The average computation time is computed as Eq. (10).

$$Avg = \frac{1}{N}\sum_{i=1}^{N} Time(i) \tag{10}$$

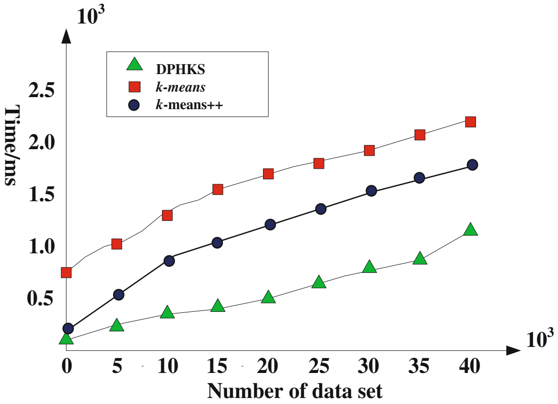


Fig. 3. Running time on 11 nodes

As is shown in Fig. 3, we can conclude that with the expansion of data sets, running time consuming also increases, but the ratio increases slow relatively. Notably, DPHKS consumes the least running time due to spark memory computing framework. Especially, the frequency of the I/O communication and operation rate have obvious advantage over compared algorithms, which is only about 1/2 of the running time of compared algorithms. In addition, DPHKS can not only reduce complexity of privacy protection, but improve the efficiency to some extent.

4.3 Clustering Results Usability

In this paper, the clustering feasibility index that combined with differential privacy is the weighted synthesis of accuracy and recall rate, and the formula is as follows:

$$F_{available} = \sum |U_i| F_i / N \tag{11}$$

From Eq. (11), F_i is the mean weighted harmonic of accuracy and recall rate, $|U_i|$ the collection of reference standard, and N is the total amount of record set. Meanwhile, we use $F_{available}$ to measure the availability of clustering results and similarity degree of the clustering results obtained by DPHKS. With the privacy budget changes, the experimental results from data set Vowel compared with PSO-k-means can be shown in Fig. 4.

As is shown in Fig. 4, when the level of privacy budget ε is low, results of clustering availability show very strong positive correlation with ε. When the value of ε reaches a certain high level, clustering availability tends to be gentle, which indirectly proves the balance effect of ε on privacy and usability. On the other hand, the collaborative effort from PSO and CS enhance the users' trust of the data privacy, and a key difference in comparison is verified by the excellent performance of DPHKS. We can infer that the small amount of noise introduced into the privacy preserving approaches is well within a tolerable error margin.

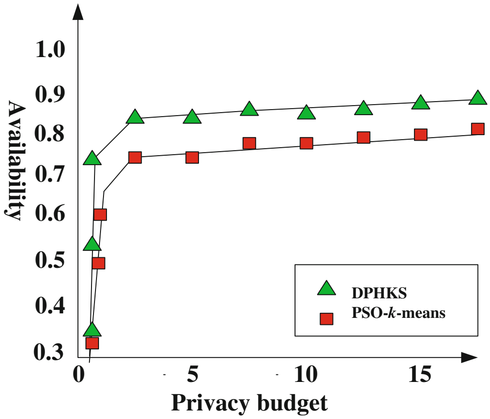


Fig. 4. Experimental results with different privacy budget

5 Conclusion

At present, Big Data has become another industrial growth point in the field of information technology after cloud computing [27], and its security issues have attracted great attention from academics and industry. In this paper, the spark parallel computing technique is adopted to support improved *k*-means by combining swarm intelligence optimization model to optimize the initial cluster center. Then ε-differential privacy is achieved by employing Laplace mechanism adding noise disturbance into clusters. At the same time, through convincing experiment results, the performance of our proposed DPHKS is obviously better than state-of-the-art clustering algorithms. In the future, the research work will mainly focus on the premise of ensuring the level of privacy protection. In addition, integrating swarm intelligence optimization into the framework of Spark platform still has a broad prospect.

Acknowledgments. We will thank for National Natural Science Foundation funded project 61309008 and Shaanxi Province Natural Science Funded Project 2014JQ8049. Also, we would also like to thank our partners at our Research Lab and their generous gifts in support of this research.

References

1. Chasaki, D., Mansour, C.: Security challenges in the internet of things. Int. J. Space-Based Situated Comput. **5**, 141–149 (2015)
2. Beldad, A.: Sealing one's online wall off from outsiders: determinants of the use of facebook's privacy settings among young dutch users. Int. J. Technol. Hum. Interact. (IJTHI) **12**, 21–34 (2016)
3. Barhamgi, M., Benslimane, D., Ghedira, C.: PPPDM–a privacy-preserving platform for data mashup. Int. J. Grid Utility Comput. **3**, 175–187 (2012)

4. Li, X., He, Y., Niu, B.: An exact and efficient privacy-preserving spatiotemporal matching in mobile social networks. Int. J. Technol. Hum. Interact. (IJTHI) **12**, 36–47 (2016)
5. Petrlic, R., Sekula, S., Sorge, C.: A privacy-friendly architecture for future cloud computing. Int. J. Grid Utility Comput. **4**, 265–277 (2013)
6. Duan, Y., Canny, J.: How to deal with malicious users in privacy-preserving distributed data mining. Stat. Anal. Data Mining **2**, 18–33 (2009)
7. Khan, N., Al-Yasiri, A.: Cloud security threats and techniques to strengthen cloud computing adoption framework. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**, 50–64 (2016)
8. Zhang, W., Jiang, S., Zhu, X.: Cooperative downloading with privacy preservation and access control for value-added services in VANETs. Int. J. Grid Utility Comput. **7**, 50–60 (2016)
9. Almiani, M., Razaque, A., Al, D.A.: Privacy preserving framework to support mobile government services. Int. J. Inf. Technol. Web Eng. (IJITWE) **11**, 65–78 (2016)
10. Yang, Q., Wu, X.: 10 challenging problems in data mining research. Int. J. Inf. Technol. Decis. Making **5**, 597–604 (2006)
11. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**, 129–137 (1982)
12. Su, D., Cao, J., Li, N.: Differentially private k-means clustering. In: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, pp. 26–37 (2016)
13. Samet, S., Miri, A., Orozco-Barbosa, L.: Privacy preserving k-means clustering in multi-party environment. In: SECRYPT, pp. 381–385 (2007)
14. Doganay, M.C., Pedersen, T.B., Saygin, Y.: Distributed privacy preserving k-means clustering with additive secret sharing. In: Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society, pp. 3–11 (2008)
15. Upmanyu, M., Namboodiri, A.M., Srinathan, K.: Efficient privacy preserving k-means clustering. In: Pacific-Asia Workshop on Intelligence and Security Informatics, pp. 154–166. Springer, Heidelberg (2010)
16. Chen, H., Hu, Y., Lian, Z.: An additively homomorphic encryption over large message space. Int. J. Inf. Technol. Web Eng. (IJITWE) **10**, 82–102 (2015)
17. Hadoop. http://hadoop.apache.org
18. Mllib. http://spark.apache.org/mllib
19. Spark. http://spark.apache.org
20. Dwork, C.: A firm foundation for private data analysis. Commun. ACM **54**, 86–95 (2011)
21. Kennedy, J.: Particle swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766. Springer, Heidelberg (2011)
22. Yang, X.S., Deb, S.: Engineering optimisation by cuckoo search. Int. J. Math. Model. Numer. Optimisation **1**, 330–343 (2010)
23. Zhou, M., Zhang, R., Xie, W.: Security and privacy in cloud computing: a survey. In: 2010 Sixth International Conference on Semantics Knowledge and Grid (SKG), pp. 105–112 (2010)
24. Roy, I., Setty, S.T., Kilzer, A.: Airavat: security and privacy for mapreduce. In: NSDI, pp. 297–312 (2010)
25. Bahmani, B., Moseley, B., Vattani, A., et al.: Scalable k-means++. Proc. VLDB Endow. **5** (7), 622–633 (2013)
26. Ahmadyfard, A., Modares, H.: Combining PSO and k-means to enhance data clustering. In: International Symposium on Telecommunications, IST 2008, pp. 688–691 (2008)
27. Kong, W., Lei, Y., Ma, J.: Virtual machine resource scheduling algorithm for cloud computing based on auction mechanism. Optik-Int. J. Light Electron Optics. **127**, 5099–5104 (2016)

Technique for Image Fusion Based on PCNN and Convolutional Neural Network

Weiwei Kong[1](✉), Yang Lei[2], and Jing Ma[3]

[1] School of Computer Science and Technology,
Xi'an University of Posts and Telecommunications, Xi'an 710121, China
wwkong_xupt@163.com

[2] Department of Electronics Technology,
Engineering University of Armed Police Force, Xi'an 710086, China
surina526@163.com

[3] Key Laboratory of Information Assurance Technology, Beijing 100072, China
mrsma919@163.com

Abstract. Image fusion has been a hotspot in the area of image processing. How to extract and fuse the main and detailed information as accurately as possible from the source images into the single one is the key to resolving the above problem. Convolutional neural network (CNN) has been proved to be an effective tool to cope with many issues of image processing, such as image classification. In this paper, a novel image fusion method based on pulse-coupled neural network (PCNN) and CNN is proposed. CNN is used to obtain a series of convolution and linear layers which represent the high-frequency and low-frequency information, respectively. The traditional PCNN is improved to be responsible for selecting the coefficients of the sub-images. Experimental results indicate that the proposed method has obvious superiorities over the current main-streamed ones in terms of fusion performance and computational complexity.

1 Introduction

Image fusion is a hotspot in the field of information processing, which has attracted a lot of attentions in both domestic and abroad. Due to the significant improvements of the visual performance, the fused image is very beneficial for the following computer processing so that the technology of image fusion has been widely used in a lot of areas, such as medical imaging [1], remote sensing [2], and so on.

So far, a lot of methods [3–30] have been presented to resolve the issue of image fusion which can be mainly classified into three categories, namely transform-domain-based ones, spatial-domain-based ones, and neural-network-based ones.

As for transform-domain-based methods, the core idea is composed of three steps as follows. (a) The source image is decomposed into a series of sub-images. (b) Certain fusion rules are adopted to complete the choice of coefficients in the sub-images. (c) The final image is reconstructed. As a result, the mechanism and effects of decompositions and reconstructions are very crucial in the whole course. Discrete wavelet transform (DWT) was regarded as an ideal pioneer before. However, further

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_38

researches indicate that DWT still has its inherent limitations. First, it is merely good at capturing point-wise singularities, but the edge expression performance is poor. Second, it captures limited directional information only along vertical, horizontal and diagonal directions [3]. Several improved versions have been proposed to resolve the disadvantages of DWT, such as quaternion wavelet transform (QWT) [4], ridgelet transform (RT) [5], curvelet directional transform (CDT) [6], quaternion curvelet transform (QCT) [7], contourlet transform (CT) [8] and shearlet transform (ST) [9]. However, the performance of the above models is severely limited because of the absence of the shift-invariance property introduced by the down-sampling procedure. The shift-invariance extension of CT, namely non-subsampled contourlet transform (NSCT) [1, 10] has been explored and used, but its computational complexity is rather higher compared with aforementioned transform-domain-based methods. Easley *et al.* proposed an improved version of ST called non-subsampled shearlet transform (NSST) [11] which not only has higher capability of capturing the feature information of the input images, but also costs much lower computational resources compared with NSCT. In spite of relatively good performance of preserving the details of the source images, transform-domain-based methods may produce brightness distortions since spatial consistency is not well considered in the fusion process.

Compared with the above ones, the spatial-domain-based methods are much easily implemented. A pioneer in SD is the weighted technique (WT) [12], and the final fused image is estimated as the weighted compromise among the pixels with the same spatial location in the corresponding inputs. WT is resistant to the existent noise in the inputs to a certain degree, but it always results in the decline of the contrast level of the fused image because it always treats all of the pixels without distinction. Furthermore, the theories of principal component analysis (PCA) [13] and independent component analysis (ICA) [14] have been used for image fusion as well. However, the methods based on PCA and ICA both put forward high requirements to the component selection.

Recently, the neural-network-based methods have become the research hotspot during the area of image fusion. As the third generation of the artificial neural networks (ANN), pulse coupled neural network (PCNN) [15] as well as its extensive versions, e.g., intersecting cortical model (ICM) [16] has been successfully proposed and widely used to deal with the issue of image fusion. In essence, ICM is the improved version of the traditional PCNN model. The above two models are able to simulate the process of biological pulse motivation to capture the inherent information of source images. Unfortunately, the traditional models of PCNN and ICM have so many parameters requiring setting. Furthermore, the mechanisms are not obvious enough so that it may bring troubles for further processing of the human or computers.

On the other hand, recently, deep learning (DP) [31] has been successfully used in the fields of image classification and natural language processing. Different from past neural network models, DP with multi-hidden layers has an remarkable feature-learning ability which is helpful to describe and represent the nature of data, and the difficulty during the training course can effectively decline via the mechanism of "gradually initialization". Although the great progress DP has made, its application in the area of image fusion has not been involved. As a typical branch of DP, convolution neural network (CNN) proves to be a good tool to deal with information processing.

Based on the content mentioned above, a novel technique for image fusion base on PCNN and CNN is proposed in this paper. The main idea consists of several following stages. On the one hand, CNN is responsible for decomposing the source images into several layers with different functions. On the other hand, PCNN is used to conduct the coefficients choice in each hidden layer.

The rest of this paper is organized as follows. The CNN model is briefly reviewed in Sect. 2. An improved PCNN model and the fusion framework based on the proposed technique are proposed in Sect. 3. Experimental results with relevant analysis are reported in Sect. 4. Conclusions are summarized in Sect. 5.

2 Convolution Neural Network

Convolutional neural network is a kind of artificial neural network, which has become a hot topic in the field of speech analysis and image recognition. The weight sharing network structure makes it more similar to the biological neural network, which reduces the complexity of the network model and the number of weights. The advantage is more obvious when the input of the network is a multi-dimensional image, so that the image can be directly used as the input of the network, so as to avoid the complex feature extraction and data reconstruction process in the traditional recognition algorithm. The convolutional network is a multilayer perceptron which is designed to recognize the shape of the two dimensions. The network structure is highly invariant to translation, scaling, inclination, or deformation.

CNN is affected by the early delayed neural network (TDNN). The time-delay neural network reduces the complexity of learning by sharing the weights on the time dimension, which is suitable for the processing of speech and time series signals. CNN is the first truly successful learning algorithm for multilayer network architecture. It reduces the number of parameters that need to be studied by the use of spatial relations in order to improve the training performance of general forward BP algorithm. CNN as a deep learning architecture is proposed to minimize the data preprocessing requirements. In CNN, a small part of the image (local receptive field) as the minimum level structure of the input information, and then transmitted to different layers, each layer by a digital filter to obtain the most significant features of observed data. This method can obtain the remarkable characteristics of the translation, zoom and rotation invariant observation data, because the characteristics of the local area of the image feel neurons or processing unit allowed access to the most basic, such as directional edge or corner.

CNN is a multilayer neural network, and each layer is composed of a number of two-dimensional planes each of which consists of a number of independent neurons. The concept demonstration of CNN is shown in Fig. 1.

The input image can be convoluted with three training filters and biases. After convolution, three feature maps appear in C1 layer, then four pixels in each group of the feature map is summed plus bias get. Three feature maps in S2 layer can be obtained through a sigmoid function. These maps are then filtered into the C3 layer. Similar to S2 layer, S4 layer can be obtained via the above hierarchical structure. Ultimately, these pixel values are rasterization and connected into a vector input to be the traditional neural network.

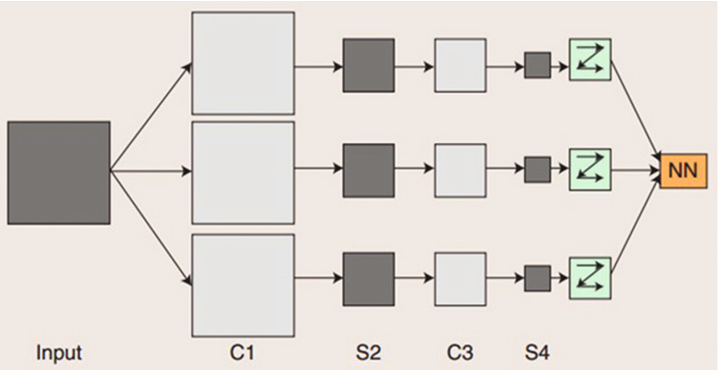


Fig. 1. Concept demonstration of convolutional neural network

In general, C layer is responsible for feature extractions. The inputs of each neuron are connected to the local receptive field of the previous layer to extract the local features. Once the local feature is extracted, its position relationships to other characteristics are determined. S layer is the feature mapping layer, and each computing layer of the network is composed of multiple feature maps. Each feature is mapped to a plane where the weights of neurons are equal. The sigmoid function is used as the activation function of the convolution network, so that the feature mapping has the invariance of the displacement.

In addition, since the neurons within a map share the weights, the number of free parameters of the network and the complexity of the network parameter selection both decrease. Each feature extracting layer (C layer) in CNN is followed by a computing layer (S layer) for computing the local average and twice extraction. This unique two feature extraction structure allows the network to identify the input sample with a higher tolerance to distortion.

3 Improved Pulse Coupled Neural Network

3.1 IPCNN and Its Time Matrix

The basic PCNN offers us with a new angle to resolve the issue of image fusion, but its disadvantages should not be ignored as well.

(a) A lot of parameters need to be set in the traditional PCNN model, and it is not easy for us to adjust them in practical applications. So far, the performance of the parameters largely relies on the experience of experts. What is the worse, a group of parameters producing good effects on certain occasions may not be suitable for other applications.

(b) The mechanism that the dynamic threshold θ_{ij} is exponential decayed and varies periodically is not an effective mode of controlling synchronous pulse bursting. It doesn't consist with the reality of human optical response to intensity variety. Meanwhile, it costs a great deal of computational resources. Moreover, unlike the linear decay, the mode of exponential decay directly causes the treatments of both high intensity pixels and low intensity ones are partial.

(c) The output Y_{ij} in basic PCNN has only two values to choose, namely 0 and 1. Therefore, the impulse sequence resulted from the basic PCNN is binary, which is not beneficial to subsequent image processing.

In order to overcome the defects mentioned above and further enhance the efficiencies of the basic PCNN, several appropriate improved measures are required. In this paper, we tend to modify the basic PCNN structure in terms of parameters reducing and efficiencies increasing. The concrete improved measures of IPCNN are given as follows.

(a) For simplicity, the signal of the feeding channel F_{ij} is simplified to the normalized pixel intensity I_{ij} as follows.

$$F_{ij}[n] = I_{ij} \tag{1}$$

(b) In reference [32], a new PCNN model called unit-linking PCNN was proposed. In the linking channel of the unit-linking PCNN, let $N(ij)$ denote the neighborhood with the neuron N_{ij} as its center. Note that when calculating the linking input L_{ij}, we should exclude N_{ij} from $N(ij)$. If any neuron except N_{ij} in $N(ij)$ fires, the linking input L_{ij} will be set as 1, otherwise L_{ij} is 0. In other words, if N_{ij} fires, any other neuron which has not fired yet in the region $N(ij)$ but has a similar input to N_{ij} may be also encouraged to fire. Obviously, this model reduces the number of undetermined parameters a lot and makes the linking inputs of unit-linking neurons uniform. As a result, the mechanism of unit-linking PCNN can be utilized to complete the task of linking channel in IPCNN, whose expression is given as follows.

$$L_{ij}[n] = \begin{cases} 1, & if \sum_{k \in N(ij)} Y_k(n) > 0 \\ 0, & otherwise \end{cases} \tag{2}$$

(c) How to determine the proper iterative number n is a knotty problem all the time. If n is too small, the neurons can't be activated adequately to make use of the characteristics of the synchronous impulse, so that the performance of image processing is not satisfactory commonly. On the other hand, if n is exceedingly large, it will not only sharply increase computational complexity, but result in an adverse influence on the visual effects of the final image. Consequently, it is very necessary to develop an efficient scheme of setting the iterative number.

In order to settle the above problem, the time matrix model [33], whose size is the same as that of the image, is adopted in IPCNN in this paper. With the help of the time matrix T, the iterative number can be determined adaptively according to the intensity distribution of pixels in images. The mechanism of time matrix T can be described as:

$$T_{ij}[n] = \begin{cases} n, & if\ Y_{ij} = 1 for\ the\ first\ time \\ T_{ij}[n-1], & otherwise \end{cases} \tag{3}$$

With regard to Eq. (3), there are several aspects required to be explained and noted. (i) T_{ij} will keep invariable if N_{ij} does not fire all the time. (ii) If N_{ij} fires for the first time, T_{ij} will be set as the ordinal value of corresponding iteration. (iii) Once N_{ij} has already fired, T_{ij} will not alter again. Its value will be saved as the ordinal value of iteration during which N_{ij} fired for the first time even N_{ij} may fire later. As known to us, the pixels having similar intensity values often share the same or approximate firing times. Accordingly, their corresponding values of the elements in T are also near. Once all pixels have fired, the whole iteration process is over, and the value of the largest element in T is the total iteration times.

(d) In order to eliminate the irrationality of exponential decay in the basic PCNN, the mode of linear decay is utilized in IPCNN as a substitute for the former. Its mathematical expression is listed as follows.

$$\theta_{ij}[n] = \theta_{ij}[n-1] - \Delta + V_{\theta} Y_{ij}[n] \tag{4}$$

Where step Δ is a positive constant. It guarantees that the dynamic threshold θ_{ij} decreases linearly with the iterative number n increasing. V_{θ} is usually set as a relatively large value to ensure that the firing times of each neuron will not exceed one at most.

Moreover, we have to note that the expression of the total internal activity in the basic PCNN is still in use in IPCNN.

In conclusion, IPCNN has overcome the drawbacks of the basic PCNN to a great extent.

(a) There are only four parameters required settings in IPCNN, while the original basic PCNN has ten parameters in all. Obviously, the number of parameters has been reduced a lot in IPCNN.

(b) As shown in Eq. (4), the decayed mechanism of the dynamic threshold θ_{ij} is linear in IPCNN, which not only conforms to the human visual characteristics, but is more helpful to enhance the computational efficiency compared with the basic PCNN.

(c) In IPCNN, the original function of Y has been replaced by the time matrix T, and the element T_{ij} commonly involves too many values. In comparison to the basic PCNN, the time matrix in IPCNN can provide rich temporal and spatial information simultaneously, which is more beneficial to subsequent image processing.

3.2 Parameters Determination of IPCNN

In the basic PCNN, there are ten different parameters in all, namely I_{ij}, β, M_{ijkl}, W_{ijkl}, α_F, α_L, α_{θ}, V_F, V_L and V_{θ}. In comparison, IPCNN proposed in this paper has much fewer parameters to determine. It only has four variables including I_{ij}, β, V_{θ} and Δ. Where I_{ij}

is set as the normalized intensity of neuron N_{ij}; the function of V_θ is to ensure each neuron can fire only once at most, so letting V_θ be a certain comparatively positive and large value can easily satisfy the requirement. As a result, the settings of I_{ij} and V_θ are not difficult at all. With regard to the step Δ, we can set it as 0.01 to guarantee that the decayed speed of the dynamic threshold θ_{ij} is moderate and acceptable. In conclusion, the linking strength β comes to be the only stress of the entire task of parameters determination.

Currently, during too many applications of image processing, the values of β are commonly set as a constant. However, according to the human visual characteristics, the responses to the region with remarkable features are supposed to be stronger than those with inconspicuous features to a certain extent. Thus, the mode of setting β as a constant is not reasonable and required to be modified. In this paper, the model of local directional contrast (LDC) is established and introduced into IPCNN to decide the value of β, which is defined as follows.

$$\beta_X^{K,r}(i,j) = \frac{|X^{K,r}(i,j)|}{\overline{X_K^0(i,j)}} \tag{5}$$

Where X denotes the source images required fusing; $X^{K,r}(i,j)$ is the coefficients located at (i,j) in the r^{th} directional sub-image at the K^{th} NSP decomposition level; "||" is the symbol of absolute value; $\overline{X_K^0(i,j)}$ denotes the local average value of the low-frequency coefficients from image X at the K^{th} level. The expression of $\overline{X_K^0(i,j)}$ is given like this.

$$\overline{X_K^0(i,j)} = \frac{1}{M \times N} \sum_{r=-(M-1)/2}^{(M-1)/2} \sum_{c=-(N-1)/2}^{(N-1)/2} X_K^0(i+r,j+c) \tag{6}$$

Where M is commonly assumed to be equal to N. $M \times N$ is the size of the neighborhood within the center at (i,j).

Note that Eq. (5) is mainly utilized to decide the values of β in high-frequency directional sub-images. When required to determine the values of β in low-frequency sub-images, we can rectify Eq. (5) as following:

$$\beta_X^{low}(i,j) = \frac{|X^{low}(i,j)|}{\overline{X^{low}(i,j)}} \tag{7}$$

As known to us, there is only one low-frequency sub-image left when the course of multi-scale decompositions ends. In Eq. (7), $\overline{X^{low}(i,j)}$ denotes the local average value of the coefficients in the low-frequency sub-image from X, whose expression can be represented like this.

$$\overline{X^{low}(i,j)} = \frac{1}{M \times N} \sum_{r=-(M-1)/2}^{(M-1)/2} \sum_{c=-(N-1)/2}^{(N-1)/2} X^{low}(i+r,j+c) \tag{8}$$

Obviously, the larger $\beta_X^{K,r}(i,j)$ is, the more remarkable the characteristics of the corresponding pixel (i, j) from image X are; furthermore, N_{ij} is prone to be activated much earlier than others.

The basic framework of the proposed technique is as follows. To begin with, decompose the source images into several layers via CNN. Then, PCNN is responsible for coefficients choice in each pairs of layers. The final fused image can be reconstructed by CNN.

4 Experimental Results and Analysis

In order to demonstrate the effectiveness of the proposed technique, two pairs of simulation experiments are conducted in this section. The experimental platform is a PC with Intel Core i7/2.6 GHz/4G and MATLAB 2013a. The whole section can be classified into two main sections. (a) Methods introduction and performance evaluation. (b) Subjective and objective evaluation on the experimental results.

4.1 Methods Introduction and Parameters Setting

The parameters setting of the proposed method is as follows: $W = [0.707\ 1\ 0.707;\ 1\ 0\ 1;\ 0.707\ 1\ 0.707]$, $\Delta = 15$, $h = 500$. It is noteworthy that it is not necessary for us to modify the parameters manually during the following simulation experiments. For simplicity, we term the proposed method M6.

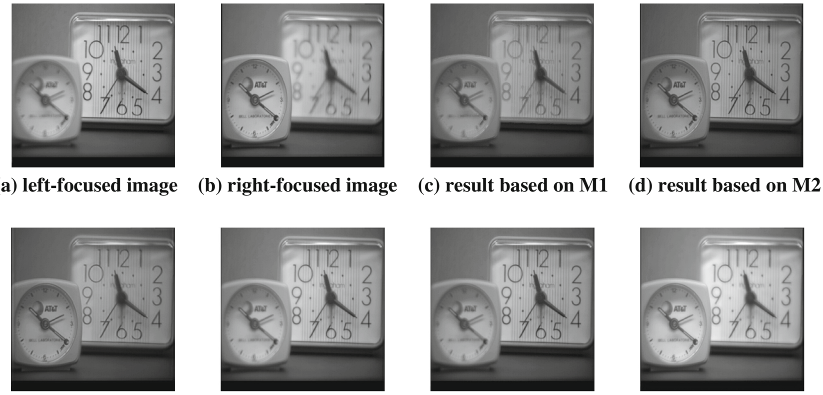
In addition, Five current typical fusion algorithms are adopted to compare with the proposed one in this paper, which are PCNN-based algorithm (M1), NSCT-based algorithm (M2), algorithm in reference [34] (M3), NMF-based algorithm (basic NMF model, M4), WNMF-based algorithm [35] (M5). In M1, the parameters are initialized as follows: $\alpha_F = +\infty$, $\alpha_L = 1.0$, $\alpha_\theta = 0.2$, $V_F = 0.5$, $V_L = 0.2$, $V_\theta = 20$, $W = M = [0.707\ 1\ 0.707;\ 1\ 1\ 1;\ 0.707\ 1\ 0.707]$, β is a constant as 0.2, the number of iterations is initialized as 50, all of the fused coefficients are determined by the firing times of neurons in PCNN. In M2 and M3, the stage number of the multi-scale decomposition in NSCT is set as 3, and the levels of the multi-direction decomposition are 4, 3 and 2 respectively from fine resolution to coarse resolution, besides, the size of the neighborhood is 3 × 3. As for concrete fused schemes, the rule of the selection of maximum coefficients is adopted at the high-frequency level and the average fusion rule is used at the low-frequency level in M2, the initialization of parameters in M3 is the same as that in reference [34]. In M4 and M5, the number of iterations is set as 50, and the group of weight coefficients is given as (0.5, 0.5). In addition, what is necessarily needed to note is that, since the random initialization of parameters W and H commonly has a great influence on the performance of the ultimate fused effect, the fused images, whose information entropy value is the largest in three random simulations, will be chosen to be the final results in M4 and M5 respectively.

In order to testify the superior performance of M6, extensive fusion experiments with two pairs of source images have been performed. The related source images with the size of 512 × 512 have been already accurately registered. The source images used in the following experiments cover 256 gray levels and can be downloaded on the web at http://www.imagefusion.org/. Subjective evaluation system can be adopted to provide direct comparisons. However, it is easily prone to be affected by lots of personal factors, such as eyesight level, mental state, even the mood, and so on. As a result, it is very necessary for us to evaluate the fusion effects based on both subjective vision and objective quality assessment. In this paper, we choose the information entropy (IE), root mean square cross entropy (RCE), standard deviation (SD), and average grads (AG) as evaluation metrics. IE is one of the most important evaluation indexes, whose value directly reflects the amount of average information in the image. The larger the IE value is, the more abundant the information amount of the image is. RCE is used to express the extent of the difference between source images and ultimate fused image. The value of RCE is in inverse proportion to the fusion performance. SD indicates the deviation extent between the gray values of pixels and the average of the fused image. In a sense, the fusion effect is in direct proportion to the value of the SD. AG, which is the last metric, is able to illuminate the clarity extent of the fused image. Being similar to SD, the clarity extent will be better with the AG value increasing.

4.2 Subjective and Objective Evaluation on the Experimental Results

Experimental results of multi-focus image fusion are shown in Fig. 2. Figure 2(a) and (b) are the corresponding source images. The fused images based on M1-M6 are given in Fig. 2(c)–(h).

From the visual angle, we observe that the intensity of the whole fused image based on M1-M3 is not sufficient enough especially the image based on M1 whose indexes of the two clocks are greatly blurred. On the contrary, the overall performances of the final fused image based on M4, M5 and the proposed algorithm are much better. Further compared with M4 and M5, although the definition of the right clock is slightly low, the whole image fusion result is superior to others. The above visual effect is verified in Table 1.



(a) left-focused image **(b) right-focused image** **(c) result based on M1** **(d) result based on M2**

(e) result based on M3 **(f) result based on M4** **(g) result based on M5** **(h) result based on M6**

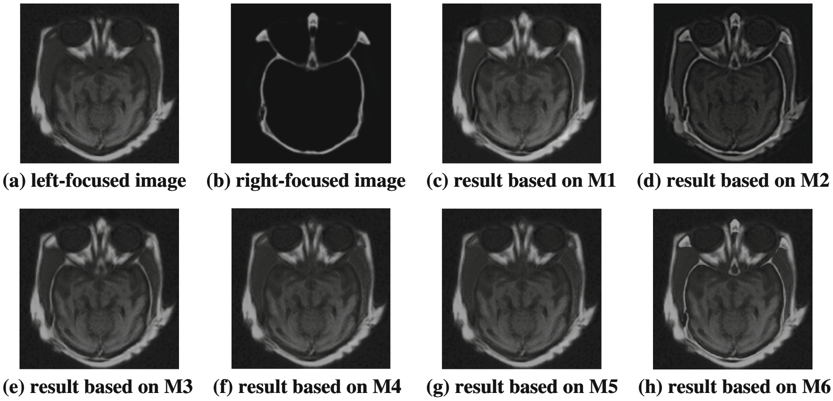
Fig. 2. Multi-focus source images and fused images based on M1–M6

Table 1. Comparison of the fusion methods for multi-focus images

Method	M1	M2	M3	M4	M5	M6
IE	6.967	7.025	7.041	7.331	7.320	**7.598**
MCE	0.035	**0.027**	0.145	0.505	0.495	0.374
SD	26.70	27.62	26.72	34.83	34.84	**39.45**
AG	2.960	3.881	4.013	3.798	3.949	**4.187**

Figure 3(a) and (b) show a medical CT image and a MRI image, whose sizes are 256 × 256, respectively. The fused images based on M1–M6 are given in Fig. 3(c)–(h).

As revealed in Fig. 3, the effects of fused images based on M4 and M5 are not as good as others, whose drawbacks mainly lie in that the information of the source MRI image are not fully described; despite the fact that the performances of the fused images based on M1 and M3 are relatively satisfactory, it is undeniable that the external outlines of these images are not clear enough; A2 overcomes the deficiencies emerged in M1 and M3, but it still has its own problems such as the low intensity of the whole fused image, and the undesirable depiction of the middle part of the image. On the contrary, the proposed method not only has clear external outlines and a rational intensity level, but protects and enhances the details information well. Table 2 reports an objective evaluation of the above mentioned six methods.


(a) left-focused image (b) right-focused image (c) result based on M1 (d) result based on M2

(e) result based on M3 (f) result based on M4 (g) result based on M5 (h) result based on M6

Fig. 3. Medical source images and fused images based on M1–M6

Table 2. Comparison of the fusion methods for medical images

Method	M1	M2	M3	M4	M5	M6
IE	5.801	5.443	5.965	5.770	5.755	**6.243**
MCE	6.547	**2.099**	4.778	6.735	6.701	3.523
SD	28.13	19.24	26.98	25.77	25.88	**30.42**
AG	4.795	4.443	4.310	3.806	3.815	**5.002**

5 Conclusions

In this paper, a new technique for image fusion based on PCNN and CNN is proposed. Experimental results demonstrate that the proposed method has obvious superiorities over current typical ones. The optimization of the proposed method will be the focus in our future work.

Acknowledgements. The authors thank all the reviewers and editors for their valuable comments and works. The work was supported in part by the National Natural Science Foundations of China under Grant 61309008 and 61309022, in part by Natural Science Foundation of Shannxi Province of China under Grant 2014JQ8349, in part by Foundation of Science and Technology on Information Assurance Laboratory under Grant KJ-15-102, and the Natural Science Foundations of the Engineering University of the Armed Police Force of China under Grant WJY-201414.

References

1. Yang, Y., Que, Y., Huang, S., Lin, P.: Multimodal sensor medical image fusion based on type-2 fuzzy logic in NSCT domain. IEEE Sens. J. **16**, 3735–3745 (2016)
2. Ghahremani, M., Ghassemian, H.: Remote sensing image fusion using ripplet transform and compressed sensing. IEEE Geosci. Remote Sens. Lett. **12**, 502–506 (2015)
3. Ali, F.E., El-Dokany, I.M., Saad, A.A., El-Samie, F.E.A.: Curvelet fusion of MR and CT images. Progr. Electromagn. Res. C **3**, 215–224 (2008)
4. Pertuz, S., Puig, D., Garcia, M.A., Fusiello, A.: Generation of all-in-focus images by noise-robust selective fusion of limited depth-of-field images. IEEE Trans. Image Process. **22**, 1242–1251 (2013)
5. Do, M.N., Vetterli, M.: The finite ridgelet transform for image representation. IEEE Trans. Image Process. **12**, 16–28 (2003)
6. Candes, E.J., Donoho, D.L.: Curvelets: A Surprisingly Effective Non-adaptive Representation for Objects with Edges. Stanford University, Stanford (1999)
7. Cao, L., Jin, L., Tao, H., Li, G., Zhang, Z., Zhang, Y.: Multi-focus image fusion based on spatial frequency in discrete cosine transform domain. IEEE Signal Process. Lett. **22**, 220–224 (2015)
8. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multi-resolution image representation. IEEE Trans. Image Process. **14**, 2091–2106 (2005)
9. Miao, Q.G., Shi, C., Xu, P.F., Yang, M., Shi, Y.B.: A novel algorithm of image fusion using shearlets. Opt. Commun. **284**, 1540–1547 (2011)
10. Bhatnagar, G., Wu, Q.M.J., Liu, Z.: Directive contrast based multimodal medical image fusion in NSCT domain. IEEE Trans. Multimedia **15**, 1014–1024 (2013)
11. Easley, G., Labate, D., Lim, W.Q.: Sparse directional image representation using the discrete shearlet transforms. Appl. Comput. Harmon. Anal. **25**, 25–46 (2008)
12. Burt, P.J., Kolcznski, R.J.: Enhanced image capture through fusion. Proc. Conf. Computer Vis. **1**, 173–182 (1993)
13. Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., Benediktsson, J.A.: Model-based fusion of multi- and hyperspectral images using PCA and wavelets. IEEE Trans. Geosci Remot. Sen. **53**, 2652–2663 (2015)
14. Mitianoudis, N., Stathaki, T.: Optimal contrast correction for ICA-based fusion of multimodal images. IEEE Sens. J. **8**, 2016–2026 (2008)

15. Broussard, R.P., Rogers, S.K., Oxley, M.E., Tarr, G.L.: Physiologically motivated image fusion for object detection using a pulse coupled neural network. IEEE Trans. Neur. Net. **10**, 554–563 (1999)
16. Kinser, J.M.: Simplified pulse-coupled neural network. Proc. Conf. Appl. Arti. Neur. Net. **1**, 563–567 (1996)
17. Abdullah, A., Omar, A.J., Inad, A.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web. Eng. **11**, 1–14 (2016)
18. Sathiyamoorthi, V.: A novel cache replacement policy for web proxy caching system using web usage mining. Int. J. Inf. Technol. Web. Eng. **11**, 1–13 (2016)
19. Sylvaine, C., Insaf, K.: Reputation, image, and social media as determinants of e-reputation: the case of digital natives and luxury brands. Int. J. Technol. Hum. Interact. **12**, 48–64 (2016)
20. Wu, Z.M., Lin, T., Tang, N.J.: Explore the use of handwriting information and machine learning techniques in evaluating mental workload. Int. J. Technol. Hum. Interact. **12**, 18–32 (2016)
21. Kong, W.W., Lei, Y., Ren, M.M.: Fusion method for infrared and visible images based on improved quantum theory model. Neurocomputing **212**, 12–21 (2016)
22. Kong, W.W., Wang, B.H., Lei, Y.: Technique for infrared and visible image fusion based on non-subsampled shearlet transform and spiking cortical model. Infrared Phys. Technol. **71**, 87–98 (2015)
23. Kong, W.W., Lei, Y., Zhao, H.X.: Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. Infrared Phys. Technol. **67**, 161–172 (2014)
24. Kong, W.W., Liu, J.P.: Technique for image fusion based on NSST domain improved fast non-classical RF. Infrared Phys. Technol. **61**, 27–36 (2013)
25. Kong, W.W., Lei, Y.J., Lei, Y., Zhang, J.: Technique for image fusion based on non-subsampled contourlet transform domain improved NMF. Sci. China Ser. F-Inf. Sci. **53**, 2429–2440 (2010)
26. Kong, W.W., Lei, Y., Ma, J.: Virtual machine resource scheduling algorithm for cloud computing based on auction mechanism. Optik **127**, 5099–5104 (2016)
27. Kong, W.W., Lei, Y., Zhao, R.: Fusion technique for multi-focus images based on NSCT-ISCM. Optik **126**, 3185–3192 (2015)
28. Kong, W.W.: Technique for image fusion based on NSST domain INMF. Optik **125**, 2716–2722 (2014)
29. Kong, W.W., Lei, Y.: Technique for image fusion between gray-scale visual light and infrared images based on NSST and improved RF. Optik **124**, 6423–6431 (2013)
30. Kong, W.W., Lei, Y.: Multi-focus image fusion using biochemical ion exchange model. Appl. Soft Comput. **51**, 314–327 (2017)
31. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**, 504–507 (2006)
32. Gu, X.D., Zhang, L.M., Yu, D.H.: General design approach to Unit-linking PCNN for image processing. Proc. Conf. Neur. Net. **1**, 1836–1842 (2005)
33. Liu, Q., Ma, Y.D.: A new algorithm for noise reducing of image based on PCNN time matrix. J. Electron. Inf. Technol. **30**, 1869–1873 (2008)
34. Guillamet, D., Vitria, J., Scheile, B.: Introducing a weighted non-negative matrix factorization for image classification. Pattern Recogn. Lett. **24**, 2447–2454 (2003)
35. Li, S.Z., Hou, X.W., Zhang, H.J.: Learning spatially localized, parts-based representation. Proc. Int. Conf. Comput. Vis. Pattern Recogn. **1**, 207–212 (2001)

Fast Iterative Reconstruction Based on Condensed Hierarchy Tree

Wan Fang, Jin HuaZhong(✉), Lei GuangBo, and Ruan Ou

Hubei University of Technology, Wuhan, China
devwaf@qq.com, galaxy0522@163.com

Abstract. Based on the traditional iterative reconstruction workflow, a fast iterative reconstruction algorithm FIRA is proposed. First, using the image feature points extracted by SIFT algorithm, calculation of image similarity based on the minimum hash algorithm in LSH model is performed. Then, the iteration order is specified through hierarchical clustering. In the iterative process, the orientation estimation of images is carried through the clustering result coming from hierarchical tree. The optimization of parameter estimation is performed by bundle adjustment, and finally produce 3d mesh models. The experimental results show that the method could bring high efficiency and eliminate the accumulated error of adjustment calculation.

Keywords: Iterative reconstruction · Minhash · Hierarchy tree

1 Introduction

There were two methods of 3d reconstruction based on multiple images, the first is the whole reconstruction, second is iterative reconstruction. The advantage of iterative reconstruction is that the relationship between multiple images could be used to restore relationship, which could also be applied in some noncontinuous scenes. So it is suitable for 3d reconstruction of random shooting or sequential image, and also for large-scale image processing.

The disadvantage is the calculation of image sequential type, the accumulated error, which could spread to the subsequent image. And with the increase of the number of images, the efficiency of nonlinear optimization in single-step iterative calculation would be reduced very fast.

2 Related Research

In recent years, a lot of related research has emerged in the field of iterative reconstruction. Snavely [1] first computes the uncertainty of each pair of overlapping images, the image position covariance, and links them together to estimate a lower bound. Crandall [2] uses discrete confidence propagation to estimate camera parameters, and defines the constraints between camera and match points based on Markov random field (MRF), and uses LM algorithm for nonlinear optimization. Sattler T. [3] provides a framework for actively searching for additional matches, based on both 2D-to-3D and

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_39

3D-to-2D search. Due to active search, the resulting pipeline is able to close the gap in registration performance. Fiore [4] uses Orthogonal decompositions to isolate the unknown depths of feature points in the camera reference frame and solve it using the singular value decomposition (SVD). Simon [5] uses multi-user image collections from the Internet and examines the distribution of images in the collection to select a set of canonical views to form the scene summary, using clustering techniques on visual features. Snavely N. [6] uses image-based rendering techniques to smoothly transition between imagegraphs, while also enabling full 3D navigation and exploration of the set of images and world geometry, along with auxiliary information such as overhead maps. Bodisszomoru [7] assumes piecewise planarity of man-made scenes and exploit both sparse visibility and a fast over-segmentation of the images. Reconstruction is formulated as an energy-driven, multi-view plane assignment problem, which they solve jointly over super pixels from all views while avoiding expensive image consistency computations. Zhang G. [8] proposes an efficient non-consecutive feature tracking framework to match interrupted tracks distributed in different subsequences or even in different videos. There framework consists of steps of solving the feature "dropout" problem when indistinctive structures, noise or large image distortion exists, and of rapidly recognizing and joining common features located in different subsequences.

Although there are so many frameworks and algorithms came up, the problem of local minimum and accumulative error have not have a good workout, specially when the image numbers increase quickly. And, efficiency and accuracy need to consider simultaneously.

3 FIRA Algorithm

A fast iterative reconstruction algorithm (FIRA) is proposed to solve the problem of the low efficiency of large-scale image matching and the iterative sequence selection in the iterative reconstruction, which can improve the efficiency of reconstruction algorithm and the accuracy of reconstruction results. In Fig. 1, We describe the workflow of the algorithm FIRA, and we will focus to explain each step in detail.

3.1 Image Similarity Calculation

The problem of image similarity matching has been a hot issue. The most image feature in this paper is match point, and we introduce a LSH model which is widely used in text search field into our work. LSH [9] (also called local sensitive hashing), is a probabilistic method of linear time complexity in neighbor searching which is proposed by Datar. We apply the pattern model to image similarity matching, and verify this mode by the algorithm. Now, I will introduce the principle and implementation of the method.

1. **minhash**

Broder [10] define and study the notion of min-wise independent families of permutations. He was motivated by the fact that such a family (under some relaxations) is

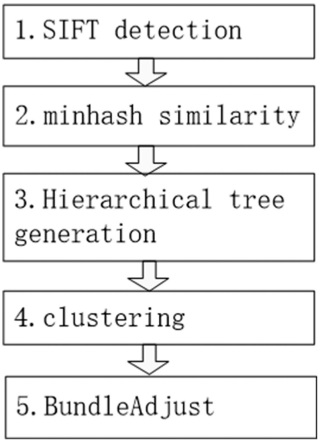


Fig. 1. Workflow of FIRA algorithm

essential to the algorithm used in practice by the AltaVista web index software to detect and filter near-duplicate documents. In the course of our investigation they have discovered interesting and challenging theoretical questions related to this concept and present the solutions to some of them and we list the rest as open problems. In fact it is a hash algorithm, whose another name is minimum priority independent permutation hash, used to quickly estimate the similarity of the two datasets. The difference between minhash and traditional hash algorithm is that the information of minhash record is more concise, which makes the efficiency higher.

Given two data sets: $A = \{a_1, a_2 ... a_n\}, B = \{b_1, b_2 ... b_n\}$ and a hash function set: $H = \{h_1, h_2 ... h_n\}$, $h_i^{min}(A)$ and $h_i^{min}(B)$ represent minima Hash of A and B. Then, the similarity comparison A and B turn into comparison of $h_i^{min}(A)$ and $h_i^{min}(B)$. In the similarity comparison, we use jaccard similarity [11] which is known as jaccard coefficient, is a method of measuring two collection similarity. For two non-empty sets A and B, the jaccard similarity is defined as formula (1), the more similar the two sets are, the greater is the value.

$$sim_{jacc}(A, B) = \frac{A \cap B}{A \cup B} \in [0, 1] \tag{1}$$

When minhash method is used to represent images, for each image there is only a very small amount of information need to be stored for the most features of the image are compressed. Now we give several definitions as the instruction of our algorithm.

Definition 1. For the same hash function, the probability of two sets to generate the same minhash value is the similarity of set A and B, is also the probability when the smallest hash value of A and B is equal to each other, and it equals to jaccard distance, as formula (2).

$$P\{h_i^{min}(A) = h_i^{min}(B)\} = sim_{jacc}(A, B) = \frac{A \cap B}{A \cup B} \in [0, 1] \tag{2}$$

Definition 2. Given k minhash values, the similarity of the two data sets is AS shown in formula (3).

$$sim\,(A, B) = \frac{\left|\{i|1 \le i \le k^\wedge h_i^{min}(A) = h_i^{min}(B)\}\right|}{k} sim\,(A, B) \cdot k \tag{3}$$

In other words, the similar degree of between A and B is estimated with the number of hash functions, the theoretical bounds is the number of hash functions, the distribution of $sim(A, B).k$ is summary of a series of independent random variables $X_1, X_2, \ldots X_k$, expected value of X_1 is:

$$E(X_1) = P\left[h_i^{min}(A) = h_i^{min}(B)\right] \tag{4}$$

Definition 3. The error of $sim(A, B)$ estimation: in order to make the estimation upper bound of $sim(A, B)$ is $\theta(0 \le \theta \le 1)$, the confidence interval is $1 - \delta(0 \le \delta \le 1)$, the number of minhash must meet the constraint:

$$\theta k \ge \frac{2 + \theta}{\theta^2} \ln(\frac{2}{\delta}) \tag{5}$$

Deviation of similarity of $sim(A, B)$ with *Jacarrd* similarity can be presented by $P[|sim(A, B) - p| \ge \theta]$, p is *Jacarrd* similarity, the Chernoff bound of $P[|sim(A, B) - p| \ge \theta]$ is as:

$$P[|sim(A, B) - p|] \le 2e^{-\frac{\theta^2}{2+\theta}k} \tag{6}$$

When k satisfy this definition, $sim(A, B)$ can be used to substitute *Jacarrd* similarity.

For a given set a, with n and k minhash, the space required to be stored is O(k). So the above minhash is generated in O(nk), and the time of comparing two data sets is O(k). Because the comparing time does not depend on the size of the data set, this property is of considerable significance for image comparison that contains a large number of characteristics.

Each image that needs to be processed can be identified as a collection of sift feature points, representing a collection c with each column of the matrix, and row r of the matrix represents all possible elements in the collection. If the collection c contains element r, the element of the c column in the matrix is 1, or 0. This matrix is called a characteristic matrix, often very sparse. In Fig. 3, the number represents the feature points, and the letter represents the camera. We can find it a very sparse matrix, which is very suitable for minhash calculation (Fig. 2).

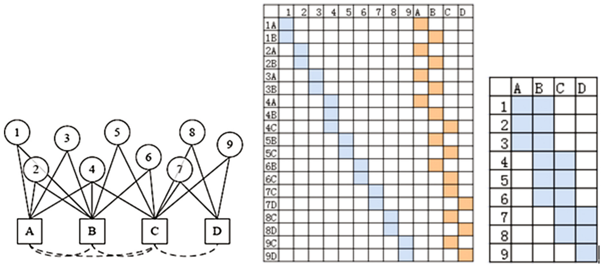


Fig. 2. Left is the Correspondence between cameras and feature points. Middle is jaccob matrix and right is Correspondence with form of matrix representation

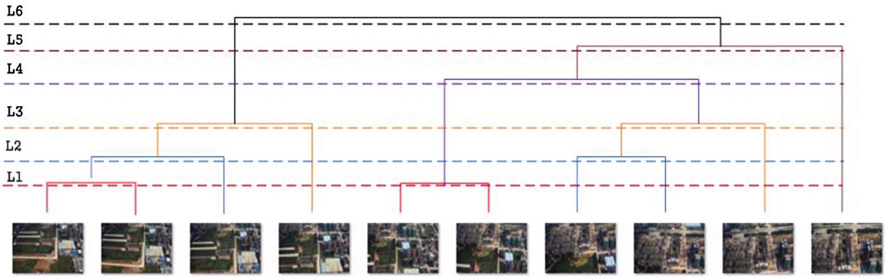


Fig. 3. Hierarchical tree generation

2. Similarity computation algorithm for image pair

Given the image set $p = \{image_i\}$, i = 1, 2...n, we first extract the sift features of all images, forming feature point set $f = \{feature_i\}, i = 1, 2...m$, the algorithm is similarity Compute whose process is as follows:

```
input： image set $p = \{image_i\}$, i = 1,2...n, $f = \{feature_i\}, i = 1,2...m$
output： $Sim[p_i, p_j]$
computeTrack( )： compute feature points for each image for IDF；
Construct k hash function
perform k hashes of all images
foreach $image_i$ and $image_j$
    $Sim[p_i, p_j] = sim(p_i, p_j)$
endfor
```

Computetrack computes the IDF (inverted index) of the sift feature points of each image, the output is the sparse matrix of the distribution of each feature point in each image. Using the hash function builder, specifying the scope of the mapping, we can easily generate the specified number of hash functions. Using the k hash function on each feature point in each image (the information in each feature point contains image sequence, pixel coordinates etc.), named $h(feature_i), i = 1, 2\ldots m$. We can get K $hmin_k(image_i)$, which represent he feature point with the minimum hash value after the hash function transformation. At this point, each image dimension is k, which is significantly lower than the original image feature point dimension n. We use formula (7) to compute degree of similarity among images.

$$sim\left(p_i, p_j\right) = |hmin_k(image_i) \;\cap hmin_k(image_j)|/hmin_k(image_i) \;\cup hmin_k(image_i) \tag{7}$$

The ratio of intersection and set in k minimum hash value is the similarity of two images. In fact, using the jaccard distance to calculate the approximate similarity of two images is enough in our experiments. Because some images have little overlap, some even none. So, if all the SIFT points perform brute force comparison, efficiency will be very low.

3. **Complexity of similarity computation**

We compared with the methods used by Farenzena [12]. The similarity of the two methods is that the feature points are extracted with SIFT, and the SIFT feature points are used for similarity calculation. Farenzena's method needs to calculate the skewer information of the feature points, and also need to calculate the convex hull range of each image and the intersection of each pair of image pairs, the computational complexity is $O(m^2n^2)$.

In our method, only the hash value calculated by k minhash can be directly obtained the similarity of the image, the cost is the calculation of hash value, so the complexity of the algorithm depends on the setting of hash function. The hash function can be selected very flexibly, in the simplest linear hash function, the algorithm complexity is $O(n)$.

Farenzena's method uses all the feature points in the image. For low resolution images, usually hundreds to thousands of feature points are to be used, and high resolution image, such as more than 10 million pixels, the number of feature points are in tens of thousands. At this time, cost of computing will become significantly larger. In contrast, the minhash is based on local hash comparison, the computing dimension is fixed as the number of hash function, this number is far smaller than the number of feature points, algorithm complexity is:

$$T(n) = O(Kn) + O\left(K^2n^2\right) = O\left(K^2n^2\right) \tag{8}$$

In the high resolution image, k is less than m, so the performance of this method will be significantly higher.

3.2 Hierarchical Tree Generation

In iterative reconstruction, the process begins iteratively after the initial image pair is established. Each iteration method needs to calculate the number of projection on the selected image, and select the most number of images to add iterative calculation. Two cases need to be addressed, one is to add a single image to the existing GOI (group of Image), one is to add GOI to another GOI. The key point is the sequence of image to be added into iterative calculation.

In our algorithm, a hierarchical tree is pre-generated which can accurately record the iterative sequence. It can ensure the efficiency of the iterative process, on the other hand, the reliability of the iterative algorithm can be predicted in advance.

The basic idea of hierarchical clustering is to regard all data points as a class at the beginning, and then gradually clustering until the iterative conditions not satisfied. Generally this clustering method uses the euclidean distance or Minkowski distance as the measure of distance. Because the distance measurement does not have transitivity, it often brings difficulties to clustering process, so many related works in order to build a reasonable hierarchical tree of clustering.

In our algorithm, each small data set is the image, because the similarity between image pairs has been calculated by minhash, so the image feature points in clustering does not need to be considered, which brings great convenience. And the jaccard similarity itself is normalized to [0,1], so the aggregation clustering algorithm can be directly applied. For the calculation of similarity between classes and classes, three method can be used as LS method ($single - linkage$, minimal distance among group elements as the distance between clusters), CL method ($complete - linkage$, max distance among group elements as the distance between clusters) and AL method ($average - linkage$, namely clustering, the average distance of the elements in the group as the distance between clusters). $single - linkage$ is used here to achieve better efficiency. When the two object are image and class, we also use $single - linkage$.

When the condensed hierarchy tree is generated, there is one or more clustering merging operations at each level, because the similarity between image pairs has been obtained by minhash. The clustering of similarity, image and image clustering can be completed quickly through single clustering. The goal of clustering is to form a class that contains all images or to form a target range. The process of clustering is the process of iterative calculation.

When the hierarchical tree generation finished, iterative sequence has been decided in advance. This is very useful in algorithm testing and in need of data estimation when the computation not yet start.

4 Experimental Results

4.1 Performance

Using Canon 5D lens on UAV to capture images as experimental data. Using the bundler software [13] and the FIRA algorithm proposed in this paper, these images are reconstructed, and we compared their results.

In Tables 1 and 2, we can get some statistic data from two experiments. Bundler and FIRA use same feature extracting method (SIFT), use different match method which result in significant difference. We can find from series experiments, when matching points became a lot, brute force matching come greatly time-consuming.

Table 1. Performance comparison of bundler algorithm and FIRA algorithm

3000	Pixels (million)	Image nums	SIFT points	Feature match	Tree build	Bundle	Total
Bundler	1500	65	13	356		86	455
FIRA	1500	65	13	141	29	42	225

Table 2. Performance comparison of bundler algorithm and FIRA algorithm

3000	Pixels (million)	Image nums	SIFT points	Feature match	Tree build	Bundle	Total
Bundler	3500	75	25	877		125	1027
FIRA	3500	75	25	188	41	85	339

In fact, images which can be used to reconstruct the scene usually are high resolution, and overlap area is large. We can say some images are for one group that they all have same overlap area. Images in different group should not participate the matching calculation.

In Fig. 4, x-axis represent pixel number, as its unit is ten million, and y-axis represents reconstruction time, in seconds. It is obvious that when the image size become large, bundles become very slow, while FIRA keep a steady speed.

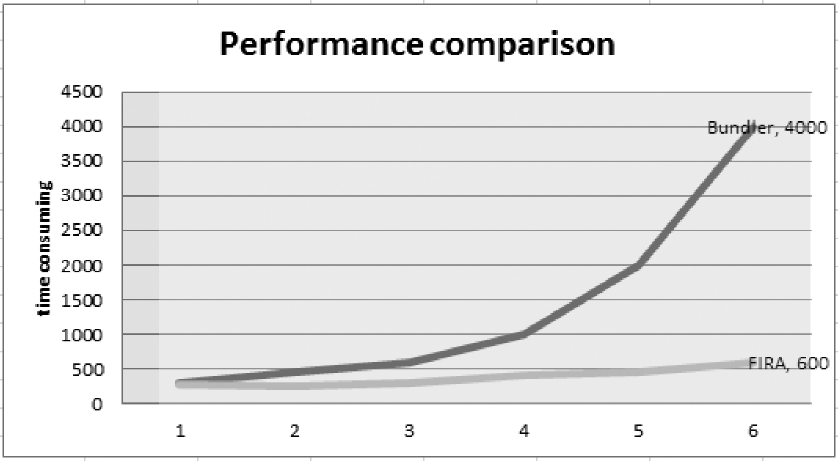


Fig. 4. The x-axis represents pixel number, unit of ten million, The y-axis represents reconstruction time, in seconds

4.2 Accuracy

Because the GPS information of the camera is known, it can be used to verify the results of the iterative reconstruction. As shown in Fig. 5, the red and blue points represent the projection of the camera position calculated by the bundler and FIRA

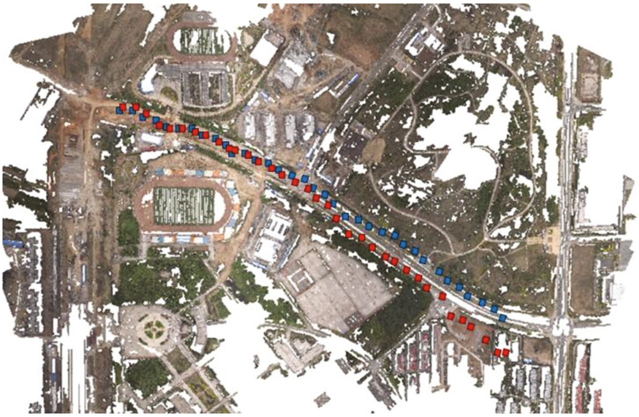


Fig. 5. The FIRA algorithm can overcome the drift problem in traditional iteration

algorithms in the digital reconstructed point cloud. The position of the blue camera is basically correct, which can prove FIRA algorithm can reduce cumulative error and is independent of the initial image selection, so we think it can eliminate the drift problem. The position of the blue camera will gradually shift with the propulsion of the route, which is the effect of cumulative error. For bundler algorithms, the overall calculation will tend to fail because of the improper selection of the initial image.

5 Summary

In this paper, FIRA algorithm is described in detail which can produce fast and stable 3d reconstruction from images. We use minimum hash algorithm in LSH model to calculate image similarity. And then, we construct a hierarchical tree using clustering algorithm to decide participation sequence in the iteration computation for all the images. Experiments prove that FIRA can overcome the drift problem in images iteration computation and bring good performance for 3d reconstruction from images.

Acknowledgement. The work was supported by the Educational Commission of Hubei Province of China (No. D20151401) and the Green Industry Technology Leading Project of Hubei University of Technology (No. ZZTS2017006).

References

1. Snavely, N., Seitz, S.M., Szeliski, R.: Skeletal graphs for efficient structure from motion. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), Anchorage, AK (2008b)
2. Crandall, D., Owens, A., Snavely, N., et al.: Discrete-continuous optimization for large-scale structure from motion. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 3001–3008 (2011)

3. Sattler, T., Leibe, B., Kobbelt, L., et al.: Improving image-based localization by active correspondence search. In: European Conference on Computer Vision, pp. 752–765 (2012)
4. Fiore, P.: Efficient linear solution of exterior orientation. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 140–148 (2001)
5. Simon, I., Snavely, N., Seitz S.M., et al.: Scene summarization for online image collections. In: International Conference on Computer Vision, pp. 1–8 (2007)
6. Snavely, N., Seitz, S.M., Szeliski, R., et al.: Photo tourism: exploring photo collections in 3D. In: International Conference on Computer Graphics and Interactive Techniques, vol. 25 (3), pp. 835–846 (2006)
7. Bodisszomoru, A., Riemenschneider, H., Van Gool, L., et al.: Fast, approximate piecewise-planar modeling based on sparse structure-from-motion and superpixels. In: Computer Vision and Pattern Recognition, pp. 469–476 (2014)
8. Zhang, G., Liu, H., Dong, Z., et al.: Efficient non-consecutive feature tracking for robust structure-from-motion. IEEE Trans. Image Process. **25**(12), 5957–5970 (2016)
9. Datar, M., Immorlica, N., Indyk, P., et al.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 253–262 ACM (2004)
10. Broder, A.Z., Charikar, M., Frieze, A.M., et al.: Min-wise independent permutations. J. Comput. Syst. Sci. **60**(3), 630–659 (2000)
11. Jaccard, J. (ed.): Interaction Effects in Logistic Regression, vol. 135. Sage, Thousand Oaks (2001)
12. Farenzena, M., Fusiello, A., Gherardi, R.: Structure-and-motion pipeline on a hierarchical cluster tree. In: 2009 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1489–1496. IEEE (2009)
13. http://www.cs.cornell.edu/~snavely/bundler/

An Optimal Model of Web Cache Based on Improved K-Means Algorithm

Qiang Wang(✉)

Information Technology Teaching Center,
Tianfu College of Southwestern University of Finance and Economics,
Chengdu, Sichuan Province, China
36522561@qq.com

Abstract. Replacement algorithm optimization is the core of cache model research. On the basis of the cache replacement model RFS, through long-term observation and analysis to the real network logs, find that the fluctuation of the access interval change rate is more valuable in predicting the new objects arrival. Therefore, in this paper, we first get the access heat level through clustering the access interval change rate with the improved K-means clustering algorithm; and then establish HSF optimal web cache model with the access heat level which named H, web object size which named S and web object freshness which named F. The replacement strategy of HSF model's is: First, replace the lowest heat level of the web object; replace the biggest size one, if H is the same; replace The lowest freshness one if H and S are the same. The simulation shows that the HSF model had the better hit rate and the byte hit rate, and the lower the access delay than the RFS.

1 Introduction

In recent years, with the rapid development of the network, more and more users are accessing the internet, so that the network carrying the business is more and more heavy, and the Network traffic is exploding. According to statistics, Internet users and business almost doubled every six months, the application of the network has grown at an unprecedented rate. The user's access to the web server is so large that the system is difficult to bear its pressure. It has become the focus of attention how to ease the growth of business traffic, reduce access latency in the network, improve network performance and so on.

Web cache is an effective technology to ease this contradiction, the core of the Web cache is the replacement strategy, the cache strategy not only affects the client's browsing web page speed, but also relates to the performance of the target server and the overall performance of the middle of the communication network. Based on the web cache model RFS [1], through the long-term observation and analysis of the actual web data access situation, find that the fluctuation of the access interval change rate is more valuable in predicting the new objects arrival, so this paper presents an optimal web cache model HSF based on the improved K-means algorithm based on the RFS optimization model HSF. The simulation experiments show that the HSF model is better than the RFS.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_40

2 The Performance of the Web Cache Replacement Model

The goal of the cache replacement strategy is to achieve good cache performance, Cache performance is usually based two evaluation factors: the hit rate, the byte hit rate [1–3].

The hit rate (abbreviated as HB): the percentage of requests that the received the service from the cache to the total request.

The byte hate rate (abbreviated as BHR): the percentage of the amount of bytes received from the cache to the total requested bytes.

The download total delay abbreviated as TD is the total time downloading the web object from the server to the client.

In fact, there are contradictions between the different indicators of the algorithm. such as the HR focus on reducing the user's response time but the BHR focuses on reducing bandwidth overhead. Because the cache object size is different, it can increase the requested hit rate if the web cache saves the web object as small as possible in the cache but not necessarily get a high byte hit rate (it has a low byte hit rate when the user requests a large file). From another point of view, if the cache saves a larger file, the BHR can improve but the HR can decline. Network users are more likely to reduce the average delay time, at this point to improve the HR is more important; but ISP hopes to reduce the cost of network bandwidth, at this time the BHR should be maximized. So you find a balance as much as possible when you design of the algorithm. Now the web users are increasingly concerned about the speed of web pages, and now more and more experiments [4] also used the download delay as an important evaluation measure of cache performance. So we use the hit rate (HB), the byte hit rate (BHR) and the download delay as a web cache performance evaluation criteria. The concept of the download delay is following.

The download delay (abbreviated as LR): the total time downloading the web object from the server to the client.

3 The Existing Web Replacement Strategy

The classic web cache replacement strategy [5, 6] has the following。

(1) LRU: LUR is least recently used, the web object which is least visited recently is removed from the web cache. Because of simple, good effective and better time complexity (O(1)), the LRU is one of the most commonly used algorithms in Web cache. But the LRU does not take into account other factors such as the web object access frequency, object size, and so on. Some papers [7, 8] think that the hit hate is only 30% to 50% the LRU replacement algorithm for web cache.

(2) FLU: FLU is least frequently used, removed the least visited web object, its algorithm complexity is O(logn). But it may appear the phenomenon of "cache garbage". If the higher access frequency of the object has been kept in the cache for a period of time.

(3) SIZE: Replace the web object according to the size, removed the object with the largest disk capacity, if the object size is the same, the time is the second

determinant. If some small web objects have not been accessed, so these small web objects stay in the cache for a long time and cause “cache pollution”.

(4) GDSF: GDSF is greedy dual-size frequency. It takes into account the object’s localities, size, delay cost, frequency and so on. It calculate the cache value for each web object, and then removed the object of the lowest value. the considerations of the GDSF are more comprehensive, but the algorithm is not only restricted by the access history and time but also may increase the access delay.

With the development of information technology, in recent years, there have been attempts to use data mining methods for WEB cache research and some results.

In literature [9] and literature [10], the current user access is predicted by the markov chain prediction model constructed with web log, and then replaced the object of the unpredictable object set with the GDSF replacement strategy.

In literature [11], The intelligent cache data mining model based on decision tree is constructed to process the replacement of the web cache.

In literature [1], the RFS cache replacement model was constructed with the most famous RFS model theory in data mining (R (recently) is a recent visit time a Web page is accessed, F is the number of times the Web page is accessed for a period of time, S is the size of the Web page), and then subdivided the Web access value with the k-mean clustering, when the web cache needs to replace, it would replace the web object of the small value.

This paper is based on the literature [1], after the long-term observation and study of web cache, found that the change rate of the access interval has a higher accuracy for the hit rate. So in this paper, we can obtain the access heat class of the web object after clustering the access interval change rate with improved K-Means clustering algorithm, and then we construct the HSM model of the optimization model of RFS model with the access heat class, the size and the freshness of the web object.

4 Web Cache Optimization Model Based on Improved K-Means Clustering Algorithm

4.1 Several Definitions

Definition 1 the access interval: The re-reference distance of the object within the cache which is the number of cache visits between this hit the object and the last hit the object.

Definition 2 the access interval change rate: The ratio of the access interval of a web object to the number of times the web object is accessed throughout the study time. the literature [2, 3] considers that the access interval rate of change is the key measure of the web object’s access heat, the greater the change in access interval is, the greater the heat of access is, The formula of the change rate of the access interval is:

$$ad\_value_i^n = \begin{cases} \boldsymbol{INITIAL\_VALUE, if\ obj_i} \\ \qquad \boldsymbol{is\ a\ new\ object} \\ ad\_value_i^{n-1}.\lambda.\frac{avg\_acc\mathrm{int}vl_i}{avg\_acc\mathrm{int}vl_i + now\_acc\mathrm{int}vl_i}, \\ \boldsymbol{if\ avg\_accintvl_i < now\_accintvl_i} \\ ad\_value_i^{n-1}.(1+\lambda.\frac{now\_acc\mathrm{int}vl}{avg\_acc\mathrm{int}vl_i + now\_acc\mathrm{int}vl_i}) \\ \boldsymbol{if\ avg\_accintvli > now\_accintvli} \\ \qquad 0 < \lambda < 1 \end{cases} \quad (1)$$

In formula, $avg\_accintvl_i$ is far below now$_accintvl_i$.

Definition 3 the access heat: The access heat has contact with the average access interval of the web object, the smaller the average access interval of a web object is, the more frequently ac the web object is accessed, so the higher the access heat of the web object is.

Definition 4 the freshness: the freshness refers to the length of the time the web object enters the web cache. If two web objects has not been replaced after entering the web cache, the freshness of the later entrant is higher.

4.2 Establishment of HSF Model Based on Improved K-Means Algorithm

On the basis of the cache replacement model RFS in the literature [4], through long-term observation and analysis to the real network logs, find that the fluctuation of the access interval change rate is more valuable in predicting the new objects arrival. After data mining the access interval change rate with the improved k-means clustering algorithm, then we get the web objects access heat group of the different levels. Then we establish HSF optimal model of web cache with the access heat level which named H, web object size which named S and web object freshness which named F. When the cache container needs to be replaced, first replace the web object of the lowest H, because the web object with an H value is a web group, the web object may be equal to or more than two, if H is the same so replace the web object of the lowest H, if H and S are the same so replace the web object of the lowest F.

4.3 The K-Means Clustering Analysis Algorithm

The cluster analysis [12, 13] is an important branch of data mining, It mainly studies the problem of "things clustering" in statistics. The K-means algorithm is the most commonly used clustering algorithm, the K-means algorithm [14, 15] uses distance to measure the degree of dissimilarity between two objects, and it thinks that the closer the two objects are, the greater their similarity are. After giving a fixed K value, randomly assigns all objects to K non-empty clusters, then calculated the average of each cluster, and use the average to represent the corresponding cluster, each object is reallocated to its nearest cluster according to its distance from the center of each cluster, until no more new distribution occurs. The distance calculation method is mainly

Euclidean distance, Manhattan distance and so on, generally the most used is Euclidean distance. The Euclidean distance calculation formula as shown in Eq. (2).

$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots |x_{ip} - x_{jp}|^2} \tag{2}$$

The K-means algorithm is summarized as following:
Input: the number of clusters k and the database containing n objects.
Output: k clusters that meet the minimum variance standard.

① Select k objects from the n data objects as the initial cluster center
② Repeat:

Each object is reassigned to the most similar cluster based on the average of the objects in the cluster.

Update the average of the clusters, that is, calculate the average of the objects in each cluster.

Until each cluster no longer changes.

But the algorithm has two drawbacks:

① Selection of different initial values may result in different results, that is, the clustering value k is difficult to determine accurately.
② The K-means algorithm is based on the objective function of the algorithm, It usually uses the gradient method to solve extremes, since the search direction of the gradient method is alone in the direction of energy reduction, so it makes the algorithm easy to fall into the local extreme value [16]. In this paper we improve the first shortcoming, and try to optimize the initial clustering number k, and propose an improved k-means algorithm.

4.4 The Improved K-Means Algorithm

We utilize the check clustering validity function "Davies-Bouldin index" proposed by Ray S and Turi RH [17] to calculate the optimal number of clusters k, and then combined with the K-means algorithm. K is defined as the following formula (3).

$$DB(U) = \frac{1}{C} \sum_{i=1, i \neq j}^{c} \max \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\} \tag{3}$$

In formula (3), U↔X: X1 ∪ …Xi ∪ …Xc; Xi is the ith cluster in U, δ(Xi, Xj) is the distance between clusters Xi and cluster Xj. C is the number of clusters in U, The calculation includes calculating the distance in the cluster and the distance between clusters. The distance between clusters and the distance between clusters are calculated using Eqs. (4) and (5), respectively.

$$\delta(S,T) = d(vs, vt) \tag{4}$$

In formula (2), $vs = \frac{1}{|S|}\sum_{x \in S} x, vt = \frac{1}{|T|}\sum_{y \in T} y$

$$\Delta_1(S) = \max\{d(x,y)\}\ \mathrm{x{\in}S, y{\in}S} \tag{5}$$

In formula (5), S is a class of U; D (x, y) is the distance between the sample x and the sample y in S; |S| represents the number of samples contained in cluster S.

We use the clustering validity function Davies-Bouldin index function Davies-Bouldin index to improve the K-means algorithm.

The improved K-means algorithm is summarized as follows:

Input: Maximum number of clusters kmax and database containing n objects.

Output: K clustering the clustering validity function reaches the optimal.

① for K = 2 to Kmax do

Select k objects from the n data objects as the initial cluster center

Repeat;

According to the average of the objects in the cluster, each object is reassigned to the most similar cluster;

Update the average of the clusters, that is, calculate the average of the objects in each cluster.

Until each cluster no longer changes.

Computer DB (U) with the formula (1), turn to formula (1).

② Select K to optimize the clustering efficient function DB (U)
③ Output clustering results
④ Termination

The improved K-means can enter the estimated maximum number of clusters kmax, in the algorithm, the validity function Davies-Bouldin index can be used to make DB (U) optimal and get the best K value. Compared with the K-means algorithm, it is theoretically proved that the improved algorithm is better.

5 Experiment

We do the experiment using about 400,000 the real access Log data of a website, We assume that in the formula (1): λ = 0.6. After doing the data pre-processed, the 3/4 of the pre-processed data was selected as the training set for HSF modeling and clustering, another 1/4 as a test data for testing, all the data is running under Matlab2010.

5.1 Matrix Transformation

Each web object is sampled to study its access interval rate of change, The access interval data for each web sample object can be represented by the following vector:

$$\mathrm{Y} = (\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{i}, \ldots \mathrm{y}_\mathrm{n})$$

The data n can be determined according to the needs of the study.

So for m web object samples we can get the corresponding access interval matrix as follows:

$$l = \left\{ \begin{array}{c} y_{1,1}, y_{1,2} \cdot \cdots \cdot y_{1,n} \\ y_{2,1}, y_{2,2} \cdot \cdots \cdot y_{2,n} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ y_{m,1}, y_{m,2} \cdot \cdots \cdot y_{m,n} \end{array} \right\}$$

According to the definition, the access interval change rate matrix is as follows:

$$F = \left\{ \begin{array}{l} \frac{y_{1,2}-y_{1,1}}{total_{y1}}, \frac{y_{1,3}-y_{1,2}}{total_{y1}} \cdots\cdots \frac{y_{1,n}-y_{1,n-1}}{total_{y1}} \\ \frac{y_{2,2}-y_{2,1}}{total_{y2}}, \frac{y2,3-y2,2}{total_{y2}} \cdots\cdots \frac{y_{2,n}-y_{2,n-1}}{total_{y2}} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \frac{y_{m,2}-y_{m,1}}{total_{ym}}, \frac{y_{m,3}-y_{m,2}}{total_{ym}} \cdots\cdots \frac{y_{m,n}-y_{m,n-1}}{total_{ym}} \end{array} \right\}$$

We use the K-means algorithm and the improved k-means algorithm to cluster the change rate of web object access interval respectively.

5.2 K-Means Algorithm Clustering Effect

After doing some experiment based on the k-means algorithms with k = 3, 4, 5, 6 separately

The distance l_k between clusters corresponding to each k value is as follows.

k = 2, l_k = 0.402

k = 3, l_k = 0.314

k = 4, l_k = 0.199

k = 5, l_k = 0.221

k = 6, l_k = 0.245

We found that l_k is the smallest when k = 4, that is, when k = 4, the k-means algorithm has the best clustering effect on the experimental data. But we have to verify the results of clustering when k = 3, k = 4, k = 5, k = 6 respectively and then determine the best k value.

5.3 The Improved K-Means Algorithm Clustering Result

In the improved k-means algorithm, set the maximum number of clusters k_max = 6, the Davies-Bouldin index function is used to calculate the optimal number of clusters as shown in the Table 1.

From Table 1 we can see when k = 4, DB (U) is the smallest. That is, when k = 4, the clustering effect of the Improved K-means Algorithm is the best.

Table 1. kmax = 6, DB (U) calculation results

Number of clusters(k)	Effective function DB(U)
2	0.398
3	0.319
4	0.197
5	0.218
6	0.241

From the above results we can know that the K-mean and the improved K-means all can achieve the best clustering effect, However, in the k-means algorithm, the optimal K value of clustering effect is difficult to determine precisely, but in the improved k-means algorithm, we can get the best K value by optimizing the effective function DB (U). The experiments confirmed the above theoretical results.

When k = 4, the number of objects in each cluster and the center point of each cluster are shown in the following Table 2.

Table 2. The center point of each cluster when the access change rate is the segmentation variable

Cluster	The cluster center of each cluster					
	The number of web (w)	Change-rate 1	Change-rate 2	Change-rate 3	Change-rate 4	Change-rate 5
C1	6	−0.118	0.157	0.038	−0.026	0.015
C2	9	0.04	−0.049	0.045	−0.035	0.009
C3	8	0.032	0.0.31	0.01	−0.016	0.031
C4	7	0.027	−0.033	−0.032	0.175	−0.157
Cluster	The cluster center of each cluster					
	Change-rate 6	Change-rate 7	Change-rate 8	Change-rate 9	Change-rate 10	Change-rate 11
C1	0.005	−0.016	0.035	−0.039	−0.002	0.026
C2	−0.007	0.036	−0.065	0.147	−0.143	0.045
C3	−0.017	0.013	−0.013	−0.048	0.168	−0.142
C4	0.034	−0.013	0.009	−0.011	−0.009	0.025

When k = 4, the distribution of the center points of each cluster is shown as Fig. 1.

In this paper, after clustering the web object access interval rate with the improved K-means algorithm, the web object access heat is divided into four categories as the Fig. 2.

From the figure we can see, the web object access heat level is divided into C1, C2, C3, C4 four categories. But the vast majority of access to heat concentrated in the

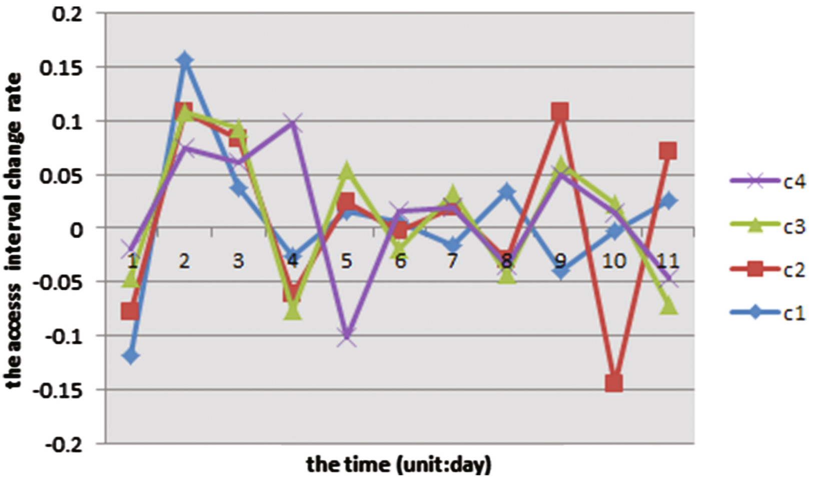


Fig. 1. The distribution of the center points of each cluster when k = 4

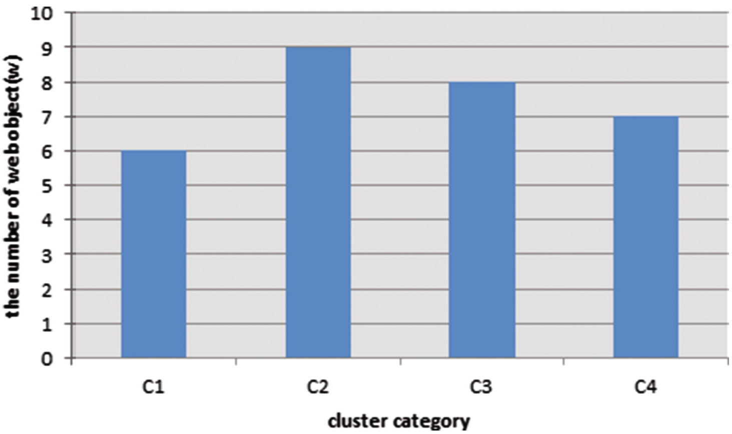


Fig. 2. The web object access heat distribution after clustering

figure C2, C3 level, C1 type for the most popular web objects about 60,000 or so, C4 is the lowest set of access to the object set, it is also the object that was first replaced in the HSF model.

5.4 Comparison of RFS Model and HSF Model

In this paper, we compared the RFS model with the HSF model in terms of performance of the web cache replacement strategy such as hite rate, the byte hit rate, and total delay. The results of the comparison are shown in Figs. 3, 4 and 5.

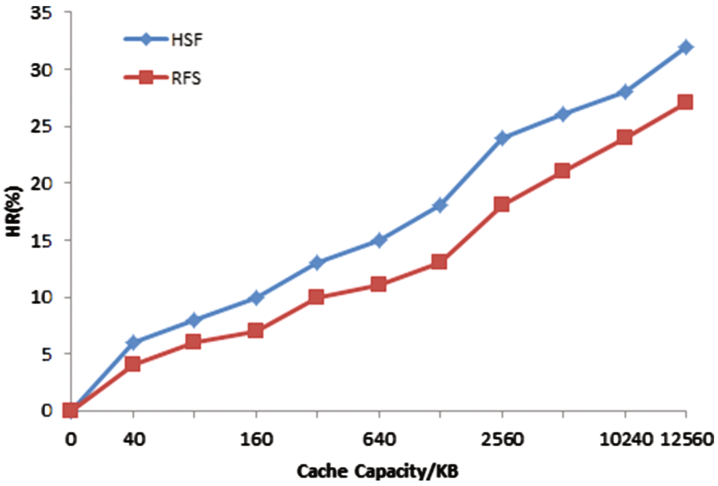


Fig. 3. Hit rate comparison of RFS and HSF

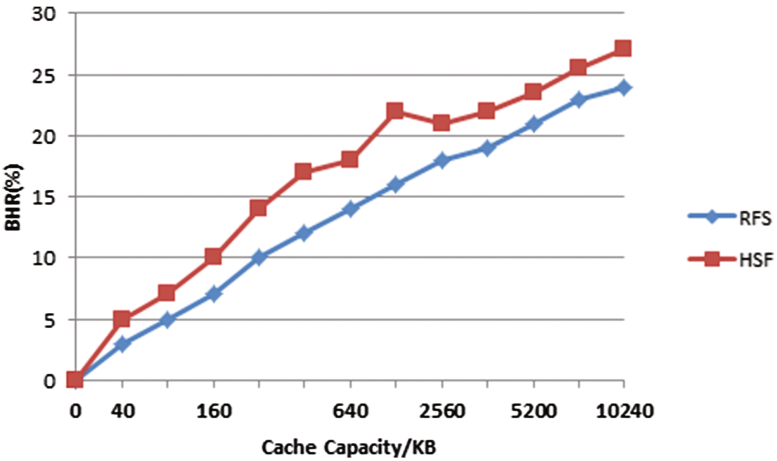


Fig. 4. The byte hit rate comparison of RFS and HSF

From the experimental results we can be seen, when the cache capacity is small, the HSF model has little effect on web cache performance such as the hit rate, byte hit rate than the RSF model, but as the cache capacity increases, the HSF model performance is significantly better than the RSF model.

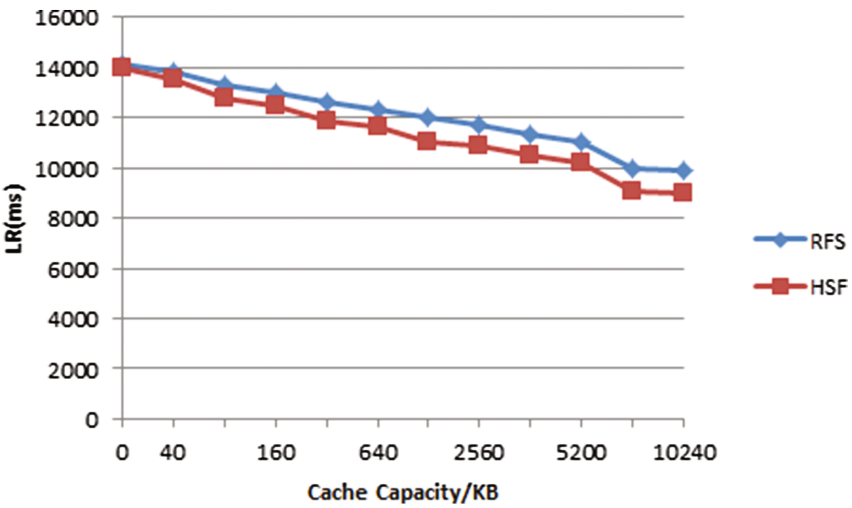


Fig. 5. The total comparison of RFS and HSF

6 Conclusion

In this paper, On the basis of the cache replacement model RFS, through long-term observation and analysis to the real network logs, find that the fluctuation of the access interval change rate is more valuable in predicting the new objects arrival, then proposed the HSF optimization model. The optimization model first obtain the web object access heat level through clustering the access interval change rate with the improved K-means algorithm, and then sort the web objects according to the access heat level, the size, and the freshness, When the buffer is full, Perform the system of the last out according to the order of access heat level, size, freshness. Compared to the RFS model, the HSF model not only achieved a higher hit rate and byte hit rate, but also optimized the total delay.

References

1. Zhang, J.: Replacement strategy of web cache based on data mining. In: 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (2015)
2. Li, Q., He, H., Fang, B.: A web cache replacement policy based on object property. Intell. Comput. Appl. (2014)
3. Li, Q., He, H., Fang, B.: A cache replacement strategy for web content distribution. High - Tech Commun. doi:10.3772/j.issn.1002-0470.2013.07.005
4. Zhang, Y., Shi, L., Wei, L.: Study on optimal model of web cache. Comput. Eng. (2009)
5. Meng, C., Ye, H.: A survey of web cache replacement algorithms. Fujian Comput. (2009)
6. Xiao, L.: Research and application of caching technology. Comput. CD Softw. Appl. (2012)
7. Abrams, M., Standridge, C.R., et al. Caching proxies: limitations and potentials. In: Proceedings of 4th www Conference, Boston, USA, pp. 119–133 (1995)

8. Shi, L., Ye, H., Wei, L., Lian, W.: Relationship between hit ratio and byte hit ratio of web caching. Comput. Eng. **33**, 84–86 (2007)
9. Huang, X., Zhong, Y.: Web cache replacement algorithm based on multi-markov chains prediction model. Microelectron. Comput. (2014)
10. Han, X., Tian, Y.: Web cache replacement algorithm based on prediction. Comput. Eng. Des. (2010)
11. Fan, X.: Intelligent model of Web cache based decision tree classification. Soft. Technol. (2011)
12. Boyinbode, O., Le, H., Takizawa, M.: A survey on clustering algorithms for wireless sensor networks, pp. 137–150. doi:10.1504/IJSSC.2011.040339
13. Richling, S., Hau, S., Kredel, H., Kruse, H.-G.: Operating two InfiniBand grid clusters over 28 km distance, pp. 313–325. doi:http://dx.doi.org/10.1504/IJGUC.2011.042946
14. Wu, J.:Customer segmentation analysis based on k-means algorithm—The case analysis of a company. Lanzhou Commercial College, pp. 12–17, June 2014
15. Zhu, M.: Data Mining. China University of Science and Technology University Press (2002)
16. Huang, G., Wang, X.: An improved artificial ant colony algorithm based on grid partitioning strategy. Microelectron. Comput. **24**(7), 83–86 (2007)
17. Ray, S., Turi, R.H.: Determination of number of clusters in K-means clustering and application in colour image segmentation, ICAPRDT 1999, pp. 27–29 (1999)

Detecting Crowdsourcing Spammers in Community Question Answering Websites

Kaiqing Hao and Lei Wang(✉)

Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province,
School of Software, Dalian University of Technology, Dalian, China
haokqgood@gmail.com, lei.wang@dlut.edu.cn

Abstract. The growth of online crowdsourcing marketplaces has attracted massive normal buyers and micro workers, even campaigners and malicious users who post spamming jobs. Due to the significant role in information seeking and providing, CQA (Community Question Answering) has become a target of crowdsourcing spammers. In this paper, we aim to develop a solution to detect crowdsourcing spammers in CQA websites. Based on the ground-truth data, we conduct a hybrid analysis including both non-semantic and semantic analysis with a set of unique features (e.g., profile features, social network features, content features and linguistic features). With the help of proposed features, we develop a supervised machine learning solution for detecting crowdsourcing spammers in Community QA. Our method achieves a high performance with an AUC (area under the receiver-operating characteristic curve) value of 0.995 and an F_1 score of 0.967, which significantly outperforms existing works.

1 Introduction

Due to the increasing growth of Web 2.0 technologies, Internet users can share their thoughts, personal feelings and experiences in many social websites. One popular direction is the Community Question and Answer (CQA) portals, such as Yahoo! Answers and Baidu Zhidao [19]. Meanwhile, the growth of online crowdsourcing marketplaces, such as ZBJ [17] and Fiverr, makes it easier to organize micro tasks supported by grassroots participants. Compared to regular jobs in a "traditional" organization, crowdsourcing micro jobs are simple (e.g., can be finished with several minutes), and paid with tiny rewards (often between $0.25 and $1). American crowdsourcing service providers made a revenue of 1 billion US dollars in 2016 and showed annual growth of 45.5% from 2011 to 2016 [7].

However, crowdsourcing services attract not only normal users, but also spammers [9,12]. For example, researchers showed that two competitive Chinese companies (360 and Tencent) hired paid posters to influence the opinion of other people towards their products in 2010 [3]; a famous Chinese director admitted publicly that his team employed Internet water army to improve the

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_41

score of his film in 2012 [18]. These crowdsourcing paid workers present a serious threat to the Internet.

In this paper, we study the crowd workers who are hired from the Internet crowdsourcing systems. We track their misbehaviors, utilize unique characteristics to detect crowdsourcing spammers in a CQA website. Unlike most of the existing works which focused on online social network websites (e.g., Twitter [15], Weibo [13]), we extend the research to CQA websites. More specifically, we first collect data from one of the most favourite Chinese crowdsourcing markets, namely ZBJ [17], and the biggest Chinese QA website, namely Baidu Zhidao [19]. Then we propose semantic analysis based on non-semantic analysis with the datasets. Finally, we propose a machine learning solution to detect spammers on CQA websites based on semantic features and non-semantic features.

The main contributions of this paper can be summarized as follows:

- First, we collect ground-truth data by linking a crowdsourcing marketplace and a CQA website. The data sets will serve as the basis of our study.
- Second, with the assistance of real-world data, we conduct a hybrid analysis with both non-semantic analysis and semantic analysis. To our knowledge, a set of non-semantic features (e.g., *survival answer rate*) are the first defined to help analyze CQA. Meanwhile, we are the first to apply semantic features to spammers detection in CQA.
- Finally, we propose a supervised machine learning model to detect crowdsourcing spammers on Zhidao based on above features and analysis. Experiments show that our method is superior to existing works, and achieves an AUC value of 0.995 and an F_1 score of 0.967.

The remainder of this paper is organized as follows. Section 2 introduces background and the strategy of employing crawlers. We analyze the non-semantic and semantic characteristics in Sect. 3. In Sect. 4, we show the classification model. Section 5 presents the experiment result. Section 6 is the related work. We give our conclusions in Sect. 7.

2 Background and Data Collection

We first introduce the relationship between crowdsourcing marketplaces and target QA websites. Then we we present our data collection methodology.

2.1 Link Crowdsourcing Services to CQA Websites

As a typical crowdsourcing marketplace, any user (buyer) can post malicious micro-jobs (e.g., promotion campaigns and spreading rumors) on ZBJ, pay money for qualified crowd workers (sellers) and any user (seller) can present submissions to the buyer.

From a buyer's perspective, he or she first posts a job on the crowdsourcing website such as ZBJ, and waits for bids or submissions. Once the buyer gets a submission, he or she should estimate it as qualified submission or non-qualified submission. Finally, the buyer rewards for qualified workers. From a worker's perspective, he or she browses jobs on ZBJ, and views job requirements. Then the worker visits CQA websites and publishes spam contents. More specifically, the worker need to use one account to ask a question and use another to answer it. Finally, the worker submits URLs of question pages to the crowdsourcing markets for buyers to estimate. If their submissions are estimated as qualified, they can receive the reward. We describe the process with Fig. 1.

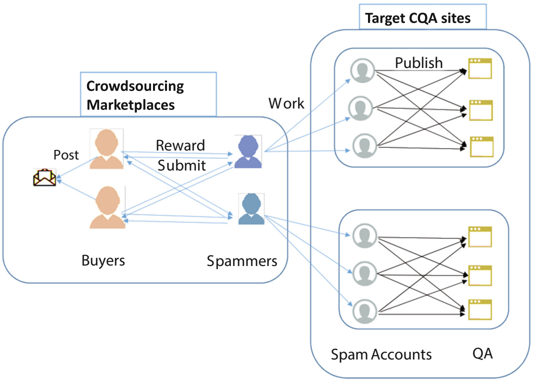


Fig. 1. The process of spammers linking crowdsourcing markets to CQA websites

2.2 Data Collection

We collect ground-truth data from ZBJ and Zhidao.

ZBJ Dataset. By searching "Baidu Zhidao" in Chinese and other similar keywords, we collected all the tasks which are targeting on Baidu Zhidao from ZBJ. For each task, we extracted each submission submitted by workers. The most of the submissions are page URLs from Zhidao. We collected ZBJ dataset from January 10, 2016 to January 13, 2016. As a result, we found 3006 tasks, extracted 14516 Zhidao page URLs, and 5642 of them were still active.

Zhidao Dataset. We design two crawlers: the question crawler for Zhidao question pages and the user crawler for Zhidao profile pages. We first extract start URL links to Zhidao question pages from ZBJ dataset. Then the question crawler visits start links, collects all QA contents (e.g., questioner/answerer links, question/answer contents, time of question/answer posted, etc.). We collect 5642 active questions from 14516 URL links. Other links are no longer active (deleted by administrators or users).

After we have crawled Zhidao question pages, we employ the user crawler to access links of questioner and answerers. We crawl profile information such as users' wealth points, experience points, number of answers, best answer ratio, number of excellent answers, answer link lists. As a result, we crawled 3115 answerers pages and 3536 questioner pages, received 6089 spammers considering duplicate. Next, for each user, we collected the most recent 100 questions that users answered, and got 70673 question pages. We consider these questioners and answerers are spam workers because normal users are less likely to participated in spam QA pairs [14].

Finally, we describe how we collect normal user datasets. We randomly select users from Zhidao in different categories. By manually checking Q&A contents, our volunteers helped us label normal users. Our volunteers should be familiar with spam templates and read no less than 20 Q&A contents of answers to ensure accuracy. We labeled 756 normal users, crawled their profile pages and the most recent 50 questions they answered. At last we got 22165 questions. We crawled Zhidao dataset during a three-month period from March in 2016 to May in 2016. We list the summary of Zhidao dataset in Table 1.

Table 1. Summary of the Zhidao dataset

User pages	ZBJ dataset spammers	6089
	Normal users	756
Question pages	ZBJ dataset spammers	5642
	Spammers answered recently	70763
	Normal users answered recently	22165

3 Analysis of Non-semantic Features and Semantic Features

We first conduct a non-semantic analysis regarding spammers and legitimate users. A set of special features are first defined to analyze CQA. Then, we present semantic analysis with our data sets.

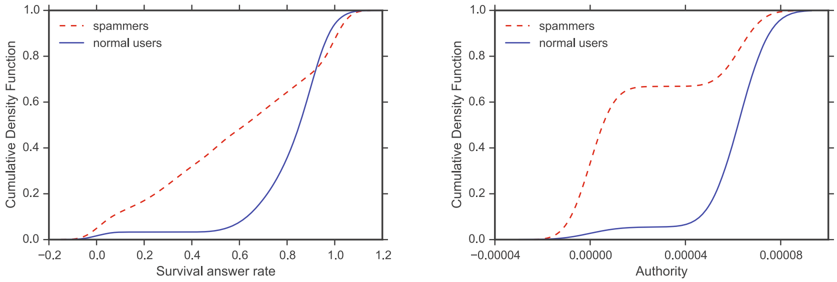
3.1 Non-Semantic Analysis

Non-semantic analysis includes profile analysis and social network analysis.

Profile Analysis. Although the profile pages provide a few basic features (e.g., number of questions, number of excellent answers and the best answer rate), it is hard to tell the differences between spammers and normal users for most of basic profile features. Therefore we extract new attributes (e.g., survival answer rate, the ratio between wealth points and experience points) from basic features.

As far as we know, we first defined survival answer rate as the rate of the number of *survival answers* and the number of answers. We call *survival answers* as answers whose links can be accessed. Our motivation is that questions or answers which contain spam content are harder to survive because of low quality of content and other user's report.

We find it notably different between spammers and normal users from Fig. 2(a). Only 10% legitimate users have a value of survival answer rate less



(a) The CDF of the survival answer rate

(b) The CDF of the authority

Fig. 2. Two non-semantic characteristics of spammers and normal users

than 0.6, whereas almost half of spammers have a value of survival answer rate under 0.6. This shows spam contents are harder to exist than legitimate contents.

Social Network Analysis. In Community QA, users form a social network by asking and answering. We construct a social network $G = (U, E)$ based on relations among users. Each node on the graph denotes a user u_i. If user u_i has a question answered by another user u_j, there will be a edge from u_i to u_j. Finally, we constructed a directed graph G with average degree of 0.933.

We analyze some graph-based features (e.g., authority and centrality). We plot the CDF of one graph-based feature, authority, in Fig. 2(b). While 65% of spammers have a very low authority (< 0.00002), most of normal users have a higher authority (> 0.00004).

3.2 Semantic Analysis

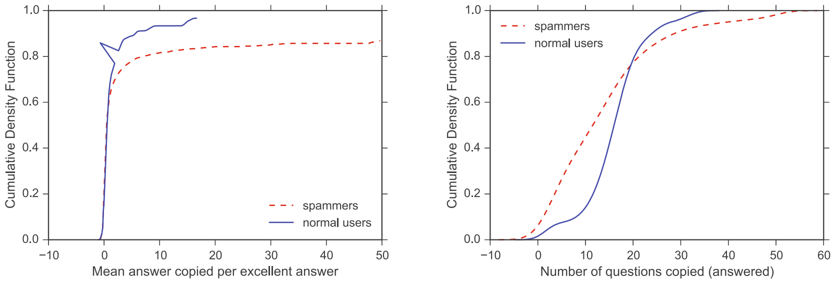
Semantic analysis contains QA content analysis and linguistic analysis.

Content Analysis. We perform content analysis on question similarity and answer similarity. We have two motivations: workers tend to copy others content for profit maximization; workers obtain spam information from same buyer during one task.

We first segment QA content to process Chinese word using *Jieba* [4] which is one of the best Chinese word segmentation modules. Then we calculate term frequency-inverse document frequency (TF-IDF) as the weight of words. Finally, we use latent semantic indexing (LSI) model to compute the similarity of two questions or answers.

We set a threshold value (e.g., 0.85 in later experiment) which is an empirical value. For an answer a_j, we regard it as a duplicate of another answer a_i if the similarity between a_j and a_i is above the threshold value. Figure 3(a, b) show the difference of two similarity features between spammers and normal users.

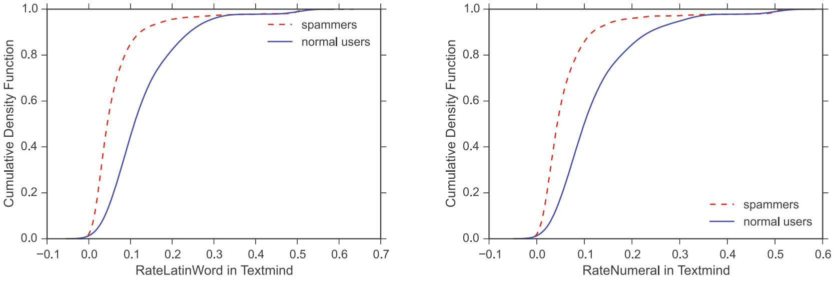
Linguistic Analysis. Lee *et al.* [9] have found that workers are less personal in the messages they post than non-workers on Twitter. Workers on Twitter engage in different language use, especially in *swearing*, *anger*, and *use of 1st*



(a) Mean number of answers copied per excellent answer

(b) Number of questions copied that users answered

Fig. 3. Two content characteristic of spammers and normal users



(a) *RateLatinWord* in Textmind

(b) *RateNumeral* in Textmind

Fig. 4. Two linguistic characteristics of spammers and normal users

person singular. To understand the linguistic characteristics of spammers in CQA websites, we use Textmind [5], a Chinese language psychological analysis system to analyze the preferences and degrees of different categories in text.

First, we count the number of words in QA contents. Next, we count the number of words in each category (102 in Textmind). Then we use Textmind to output a 102 dimension vector for each QA content. Figure 4 shows the top two distinguishing linguistic characteristics: *RateLatinWord* and *RateNumeral* which mean rate of Latin word and rate of numeral, respectively.

4 Detect Crowdsourcing Spammers

We first summarize all features. Then we present top features which have positive power to distinguish crowdsourcing spammers and normal users.

4.1 Features

According to the non-semantic analysis in Sect. 3.1 and semantic analysis in Sect. 3.2, we leverage 131 features which are computed from raw-data and group them into the following 5 categories:

- **Profile features (PF)** extracted from profile pages, including 6 basic features and 6 additional features (e.g., experience points and survival answer rate).
- **QA features (QAF)** extracted from question pages, including 7 features of QA pair (e.g., the mean length of questions).
- **Graph-based features (GF)** extracted from the social network graph G (e.g., authority and between centrality).
- **Content features (CF)** extracted from QA contents, including 7 similarity features (e.g., the number of answers copied).
- **Linguistic features (LF)** extracted from QA contents using Textmind.

We list all 131 features in Table 2.

Table 2. Features

Group	Features
PF (12)	Number of questions, Number of answers, Experience points,
	Number of excellent answers, Wealth points, Best answer ratio,
	Mean experience points per answer, Mean experience points per question plus answer, Mean wealth points per answer,
	Mean wealth points per question plus answer, The ratio between wealth points and experience points, Survival answer rate
QAF (7)	Mean length of answers, Mean length of questions
	Mean time difference between a question be posted and its answers,
	Mean length of questions that users answered,
	Mean length of question titles that users answered,
	Mean number of answers per question that users answered,
	Mean number of answers per question that user asked
GF (3)	Hub, Authority, Betweenness centrality
CF (7)	Number of answers copied, Mean number of answers copied per answer,
	Number of questions copied, Number of questions copied (answered)
	Mean number of questions copied (asked), Mean number of questions
	copied (answered), Mean number of answers copied per excellent answer,
LF (102)	102 Textmind features, which are *Function word, Pronouns, Personal Pronouns, 1st Person Singular, 1st Person Plural, 2nd Person, 3rd Person Plural, 3rd Person Singular, 3rd Person Plural, Non-personal Pronouns, Article, Verb, Auxiliary Verb, Past Tense, Present Tense, Future Tense, Adverb, Preposition, Conjunction, Negate, Quantifier, Number, Swear, Postposition, Special Art, Quantity Unit, Inter Junction, Multi Function, Tense Mark, Past Mark, Present Mark, Future Mark, Progress Mark, Social, Family, Friend, Humans, Affect, Positive Emotion, Negative Emotion, Anxiety, Anger, Sad, Cognitive Mechanic, Insight, Cause, Discrepancy, Tentative, Certain, Inhibition, Exclusive, Precept, See, Hear, Feel, Biology, Body, Health, Sexual, Ingest, Relative, Motion, Space, Time, Work, Achieve, Leisure, Home, Money, Religion, Death, Assent, Filler, Psychology, Love, 11 Punctuation marks(e.g., Period, Comma, Colon, Semi Comma.), Word Count, Word per Sentence, Rate Dict Cover, Rate Numeral, Rate Four Char Word, Rate Six Char Word, Rate Latin Word, Number of Emotion, Number of URLs, Number of Hash Tags.*

4.2 Feature Selection

We try to find top features which have positive power to distinguish crowdsourcing spammers and normal users. We computed the χ_2 value [16] for each feature. The larger χ_2 value of a feature has, the more significant the feature expresses. Table 3 presents top ten most useful features using χ_2 test. *Mean number of answers copied per excellent answer* is the most significant feature to differentiate spammers and normal users. It can be explained by that buyers tend to demand workers' answers to be the best answers and be liked. Additionally, survival answer rate also puts up a high performance. It shows that this concept we defined is meaningful and significant in CQA. Overall, there are 4 non-semantic features and 6 semantic features in the top ten features.

Table 3. Top ten features

Group	Rank	Features
CF	1	Mean number of answers copied per excellent answer
PF	2	Survival answer rate
GF	3	Authority
CF	4	Mean number of questions copied (asked)
CF	5	Number of questions copied (answered)
CF	6	Number of answers copied
LF	7	*RateLatinWord* in Textmind
PF	8	Experience points
LF	9	*RateNumeral* in Textmind
PF	10	Wealth points

5 Results and Evaluation

We show experiment results and evaluation with our data sets.

5.1 Settings and Metrics

We select 3 popular classifiers, including Naive Bayes, SMO and Random Forest, because of their effectiveness and representativeness. Our studies use weka machine learning toolkits [6]. For each experiment, we apply 10-fold cross-validation to avoid over-fitting. Additional, we apply filters to discretize a range of numeric attributes into nominal attributes. Other settings will be default according to the implementations in Weka. We use Area Under the ROC Curve (AUC) as the main metrics to measure the experiment performance because of its high robustness. We also report Precision, Recall, F_1 score as reference measures.

5.2 Classification Results

We train the datasets with basic classifiers. The AUC value varies with 0.5 and 0.926. Naive Bayes and Random Forest show the best performance under a default setting. Experiment results are showed in Table 4.

Table 4. Classification Results using semantic features and non-semantic features

Classifier	Precision	Recall	F_1 score	AUC
Naive Bayes	0.920	0.918	**0.919**	0.924
SMO	0.917	0.921	**0.919**	0.718
Random forest	**0.928**	**0.922**	0.890	**0.926**

Our datasets show imbalanced distribution. Imbalanced data would influence the classification training models, especially for those machine learning algorithms without class balance designs. We use synthetic minority over-sampling technique (SMOTE) [1] to process imbalanced data.

Performance of all the classifiers improves significantly by using SMOTE. Table 5 shows results obtained by using 3 classifiers with SMOTE. Random Forest classifier performs best among 3 classifiers. In reality, Random Forest got 0.967 TP rate and 0.033 FP rate. These two metrics are also the best in all experiments.

Table 5. Classification Results with SMOTE

Classifier	Precision	Recall	F_1 score	AUC
Naive Bayes	0.930	0.927	0.927	0.987
SMO	0.954	0.954	0.954	0.954
Random forest	**0.967**	**0.967**	**0.967**	**0.995**

5.3 Compare with Existing Methods

To evaluate the performance of our approach, we implement some methods in existing works as baselines. We compare our work with the studies of Xu *et al.* [14]. We chose their work as baselines because their work is up-to-date studies in detecting crowdsourcing spammers in CQA. Xu *et al.* computed 20 attributes in their studies, and employed the J48 and PART classifiers. We implement 4 baselines with features proposed by their work.

In contrast, our approach has the same settings with baselines. Table 6 presents the results measured with AUC. Our proposed models perform better than all baselines. It indicates that our approach is more effective and more robust than existing works.

Table 6. The Comparison of AUC

	Original baseline	Our method
Baseline1: J48 classifier	0.638	**0.717**
Baseline2: PART classifier	0.753	**0.825**
Baseline3: J48 classifier + SMOTE	0.942	**0.985**
Baseline4: PART classifier + SMOTE	0.949	**0.986**

6 Related Work

Spammers analysis and detection in CQA. In the literature, research on CQA portals focuses on estimating the quality of QA pairs. Li *et al.* [11] showed the differences between low-quality QA pairs and spam QA pairs. They proposed a propagation algorithm which uses promotion channels to calculate spamming scores of users and channels. However, our datasets show that there are much less channels arising in crowdsourcing tasks, because it is harder for answers with promotion channels to live. Chen *et al.* [2] found three new features for CQA, then they used logistic regression to detect commercial campaigns. Xu *et al.* [14] carried out a detail analysis on the crowdsourcing system, and detected spammers with profile attributes and social network attributes.

Spammer analysis and detection in online social network. Recently, a plethora of methods and systems have been designed to characterize spammers and detect their mis-behaviors in many large-scale social systems. Yang *et al.* analyze spammers' social network on Twitter [15]. Wang *et al.* studied adversarial spammers with Weibo dataset [13]. Compared with these systems, social relationships in community QA are not strong and constant. That enhances the difficulty of detecting spammers in CQA websites.

Crowdsourcing marketplaces. Crowdsourcing services (e.g., Amazon Mechanical Turk and Fivver) have attracted people to do paid micro-tasks. Despite its advantages, these system can also be malicious, called crowdturfing (crowdsourcing + astroturfing). Wang *et al.* [12] investigated two Chinese crowdsourcing websites and tracked spam campaigns in Weibo. Lee *et al.* [8] studied similar problems in three English-based crowdsourcing websites. Their team built crowdturfing task detection classifiers to filter malicious tasks in another crowdsourcing marketplace [10].

7 Conclusions

We propose a supervised machine learning solution for detecting crowdsourcing spammers in Community Question Answering websites. We employ crawlers to collect data from ZBJ and Zhidao. With the assistance of the datasets, we define new unique features in CQA. Then we conduct a hybrid analysis with both non-semantic features and semantic features to detect crowdsourcing spammers. Experiment results show a high performance with an AUC of 0.995 and a F_1 score of 0.967.

References

1. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
2. Chen, C., Wu, K., Srinivasan, V., Bharadwaj, K.: The best answers? think twice: online detection of commercial campaigns in the CQA forums. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE (2013)
3. Chen, C., Wu, K., Srinivasan, V., Zhang, X.: Battling the internet water army: detection of hidden paid posters. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE (2013)
4. Fxsjy: jieba chinese text segmentation (2016). https://github.com/fxsjy/jieba. Cited 4 May 2016
5. Gao, R., Hao, B., Li, H., Gao, Y., Zhu, T.: Developing simplified chinese psychological linguistic analysis dictionary for microblog. In: International Conference on Brain and Health Informatics, pp. 359–368. Springer (2013)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. ACM SIGKDD Explor. Newslett. **11**(1), 10–18 (2009). doi:10.1145/1656274.1656278
7. IBISWorld: Crowdsourcing Service Providers in the US: Market Research Report. Technical report, IBISWorld (2016). http://www.ibisworld.com/industry/crowdsourcing-service-providers.html. Cited 4 May 2016
8. Lee, K., Caverlee, J., Cheng, Z., Sui, D.Z.: Campaign extraction from social media. ACM Trans. Intell. Syst. Technol. **5**(1), 1–28 (2013). doi:10.1145/2542182.2542191
9. Lee, K., Tamilarasan, P., Caverlee, J.: Crowdturfers, campaigns, and social media: tracking and revealing crowdsourced manipulation of social media. In: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM (2013)
10. Lee, K., Webb, S., Ge, H.: The dark side of micro-task marketplaces: characterizing fiverr and automatically detecting crowdturfing. In: Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM (2014)
11. Li, X., Liu, Y., Zhang, M., Ma, S., Zhu, X., Sun, J.: Detecting promotion campaigns in community question answering. In: 24th International Joint Conference on Artificial Intelligence-IJCAI-15 (2014)
12. Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., Zhao, B.Y.: Serf and turf. In: WWW 2012 (2012)
13. Wang, G., Wang, T., Zheng, H., Zhao, B.Y.: Man vs. machine: Lractical adversarial detection of malicious crowdsourcing workers. In: 23rd USENIX Security Symposium, USENIX Association, CA (2014)

14. Xu, A.: Revealing, characterizing, and detecting crowdsourcing spammers: a case study in community Q & A. In: IEEE INFOCOM 2015 (2015)
15. Yang, C., Harkreader, R., Zhang, J., Shin, S., Gu, G.: Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In: WWW 2012 (2012)
16. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning ICML (1997)
17. ZBJ: ZBJ.com. (2016). http://www.zbj.com/. Cited 15 Mar 2016
18. Zeng, J.: Lu Chuan respond to the "Shuijun" event (2012). http://ent.qq.com/a/20121204/000350.htm. Cited 5 May 2016
19. Zhidao: Baidu Zhidao (2016). http://zhidao.baidu.com/. Cited 1 Jan 2016

A Spam Message Detection Model Based on Bayesian Classification

Yitao Yang[1,2](✉), Runqiu Hu[2], Chengyan Qiu[2], Guozi Sun[2], and Huakang Li[2]

[1] Information Technology Department, Nanjing Forest Police College,
No. 28 Wenlan Rd., Nanjing, China
youngyt@gmail.com

[2] School of Computer, Nanjing University of Posts and Telecommunications,
No. 9 Wenyuan Rd., Nanjing, China
{B14040735,B09040502,sun,huakanglee}@njupt.edu.cn

Abstract. In recent years, we have witnessed a dramatic growth in spam mail. Other related forms of spam are also increasingly exposed the seriousness of the problem, especially in the short message service (SMS). Just like spam mail, the problem of spam message can be solved with legal, economic or technical means. Among the technical means, Bayesian classification algorithm, which is simple to design and has the higher accuracy, becomes the most effective filtration methods. In addition, from the perspective of social development, digital evidence will play an important role in legal practice in the future. Therefore, spam message, a kind of digital evidence, will also become the main relevant evidence to the case. This paper presents a spam message detection model based on the Bayesian classification algorithm. And it will be applied to the process of SMS forensics as a means to analyze and identify the digital evidence. Test results show that the system can effectively detect spam messages, so it will play a great role in judging criminal suspects, and it can be used as a workable scheme in SMS forensics.

Keywords: Spam message filtering · Machine learning · Bayesian classification · Digital forensics

1 Introduction

Since the world's first message successfully sent to mobile phone via PC in 1992, on the United Kingdom Walter Fung's GSM, the mobile communication technology has developed very rapidly. Mobile phone has been called the "fifth media", following newspaper, radio, television, and the Internet. Currently, the smart mobile devices are very popular, but accompanied by the worsening security problems, especially the threat of spam message. According to the 360 Internet Security Center's "Chinese spam messages harassing phone intercept Governance Report", in the year 2012, 360 mobile guards cumulatively blocked 71238913342 spam messages for users, intercepted messages more than 195 million daily, and respectively, the amount of interception in the first, second, third and fourth quarters, was 3.8 billion, 6.7 billion, 27.2

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_42

billion and 33.5 billion, it showed high growth in the year. The results of micro blogging sample survey, launched by 360 mobile guards, showed that 64.3% of the respondents often received spam messages and harassing calls [1]. We can see that spam messages have been a very serious problem, and have caused considerable harm to the society. Therefore, not only a spam filtering system on the mobile terminal is needed, but also restraining the recklessly spread of spam messages from the source through legal methods is necessary.

A number of technical approaches have been applied for the problem. We can divide them into two groups: content-based approaches and non-content-based approaches. Social network analysis [2, 3] is a typical non-content-based approach. It aims at predicting whether a spender is a spammer or not. This approach is often used by telecom operators instead of mobile phone users. Content-based approaches, including automatic text classification techniques, are main approaches applied for the SMS spam filtering problem. Support Vector Machines (SVMs) [4], k-nearest neighbor algorithm [5, 6], logistic regression algorithm [7] and Winnow algorithm [8] are included. Among these methods, SVMs are considered to be the most suitable one [9]. In recent research, some evolutionary algorithms, for example, the evolutionary algorithm based on the artificial immune system [10], have also been applied on the problem and researchers have drawn comparison among the performance of them [11]. Hybrid approaches have also been proposed which combine content-based filtering with challenge-response, a technique that sends a reply to the message sender and requires the sender to give a reply. A typical approach of this is CAPTCHA algorithm [12] which sends an image to the message sender and requires a reply to confirm whether the sender is a machine or not. In addition, when we focused on the client-based approach, Bayes algorithm becomes a good choice. It is firstly used to extract key-words in spam SMS filtering system [13] and improved by adding a cost function to false positives [14] later. Researchers then install it on the mobile devices [15] and find that it is effective to filter spam SMS.

On the legal means, the current relevant provisions about electronic evidence in the legal norms are still very immature in our country. In order to ensure that the parties can collect the evidence needed in the proceedings, the laws of western countries have provided some means and procedures of gathering evidence for the parties. But in our country, the traditional system of evidence legislation found in the three major procedural law and relevant judicial interpretations, which formed a relatively perfect litigation evidence collection forensics system, don’t have specific provisions on electronic evidence, only slightly in some administrative regulations and judicial interpretations. The civil procedure law of our country only provides that the parties have the responsibility to provide evidence, but do not give them the appropriate methods and procedures for collecting. Even the “rules of evidence”, highly valued by others, is also basically blank on this issue. We should, therefore, summarize practical experience, and learn from the relevant systems of other countries, to improve the system of electronic evidence in our country [16], in which the standardized forensic process is particularly important to ensure the authenticity, integrity and legitimacy of electronic evidence.

This paper presents a spam message detection model based on Bayesian classification algorithm, and proposes a method for spam messages forensics, which can test and identify the evidence according to the specifications for electronic forensics. Ultimately, we achieved the specifications to collect evidence of spam message, and enhanced the reliability, authenticity and legitimacy of the electronic evidence. In this paper, the following part is organized as follows: Sect. 2 is the theory of Bayesian spam message detection model and the process to establish it; Sect. 3 describes the flow to obtain the spam message evidence; Sect. 4 conducted a number of experiments to prove the correctness and effectiveness of the test model, and analyzed the results father; Sect. 5 is the summary.

2 The Bayesian Spam Detection Model

2.1 Introduction to Bayesian Algorithm

The term Bayesian refers to Thomas Bayes (1702–1761), who proved a special case of what is now called Bayes' theorem in paper titled "An Essay towards solving a Problem in the Doctrine of Chances" [17]. In the 1980s, Bayesian methods are widely accepted and used, such as in the fields of machine learning and talent analytics. Bayesian method is separated from the traditional probability theory, and it is on the basis of probability theory, specifically designed to deal with uncertain problems. In this paper, we will design a model of spam detection based on Bayesian classification algorithm.

Conditional Probabilities and Bayes' Theorem [18]

The probability of a hypothesis B conditional on a given body of data A is the ratio of the unconditional probability of the conjunction of the hypothesis with the data to the unconditional probability of the data alone.

Definition

The probability of B conditional on A is defined as

$$\mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{\mathrm{P}(\mathrm{AB})}{\mathrm{P}(\mathrm{A})} \tag{1}$$

provided that both terms of this ratio exist and P(A) > 0.

Here are some straightforward consequences of (1):

- Probability. P(A) is a probability function.
- Logical Consequence. If A entails B, then P(B|A) = 1.
- Mixing. $\mathrm{P(B)} = \mathrm{P(B|A)P(A)} + \mathrm{P(B|A^C)P(A^C)}$

The most important fact about conditional probabilities is undoubtedly Bayes' Theorem. It relates the "direct" probability of a hypothesis conditional on a given body of data, P(B|A), to the "inverse" probability of the data conditional on the hypothesis, P(A|B).

Bayes' Theorem

Suppose that there are n possible states of the world, labeled $\mathrm{B}_{1,}$ $\mathrm{B}_{2,}$ ..., Bn. The prior probability that Bi is the true state is P(Bi). Let A be some event which has probability

P(A|Bi) of occurring given that Bi is the true state of the world. Then the overall (prior) probability that the event A occurs is

$$\mathrm{P(A)} = \mathrm{P(A}|B_1)\mathrm{P}(B_1) + \mathrm{P(A}|B_2)\mathrm{P}(B_2) + \cdots + \mathrm{P(A}|B_n)\mathrm{P}(B_n)$$

Given that the event A has been observed to have occurred the posterior probability that B is the true state of the world is

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A|B)P(B)}}{\mathrm{P(A)}} \tag{2}$$

When both P(A|B) and P(A|B^{C}) are known, an experimenter need not even know A's probability to determine a value for P(B|A) using Bayes' Theorem.

Bayes' Theorem (2nd form).

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A|B)P(B)}}{P(A|B)P(B) + P(A|B^{\mathrm{C}})P(B^{\mathrm{C}})} \tag{3}$$

In this guise, Bayes' Theorem is particularly useful for inferring caused from their effects, since it is often fairly easy to discern the probability of an effect given the presence or absence of a putative cause.

2.2 The Bayesian Spam Detection Model

From the content, spam detection can be seen as a binary classification problem, the messages are divided into spam and message.

As SMS are in text form, the computer cannot recognize them, it is necessary to preprocess the text message. We can mark the texts by a set of words, such as a text D can be expressed as a set with w1, w2, …, wn, i.e. D = {w1, w2, …, wn}, and the set of words {w1, w2, …, wn} is independent distributed. We stipulate that the probability in a given class C the ith word wi appeared is defined as $\mathrm{P}(w_i|C)$. So for a given class C, if the text D = {w1, w2, …, wn} appeared, the probability is

$$\mathrm{P}(D|C) = \prod\nolimits_i P(w_i|C) \tag{4}$$

The question we have to answer is that what is the probability text D belongs to class C, in other words, what is $\mathrm{P}(C|D)$.

We get formula:

$$\mathrm{P(C|D)} = \frac{P\mathrm{(D|C)P(C)}}{P(D|C)P(C) + P(D|C^{\mathrm{C}})P(C^{\mathrm{C}})} \tag{5}$$

derived from the Bayes' formula 3.

$$\mathrm{P(B|A)} = \frac{\mathrm{P(A|B)P(B)}}{P(A|B)P(B) + P(A|B^{\mathrm{C}})P(B^{\mathrm{C}})}$$

when applied to spam message detection there are only two independent category, S and ¬S (spam and non-spam message), here each element (SMS) is either spam message or not. So the probability text D belong to class S (the class spam) can be written as:

$$\begin{aligned} P(S|D) &= \frac{P(D|S)P(S)}{P(D|S)P(S) + P(D|\neg S)P(\neg S)} = \frac{P(D|S)}{P(D|S) + \frac{P(\neg S)}{P(S)}P(D|\neg S)} \\ \text{order } k &= \frac{P(\neg S)}{P(S)}, then\ P(S|D) = \frac{P(D|S)}{P(D|S) + kP(D|\neg S)} \\ &= \frac{\prod_i P(w_i|S)}{\prod_i P(w_i|S) + k\prod_i P(w_i|\neg S)} \end{aligned} \tag{6}$$

when $P(S|D)$ exceeds a certain threshold, we determine that text D is a spam message.

Since $P(w_i|S) = P(S|w_i)P(w_i)/P(S)$, $P(\neg S|w_i) = 1 - P(S|w_i)$, the formula 6 can be written as:

$$P(S|D) = \frac{\prod_i P(S|w_i)}{\prod_i P(S|w_i) + k^{1-\mathrm{n}}\prod_i(1 - P(S|w_i))} k = \frac{P(\neg S)}{P(S)} \tag{7}$$

Applying a Bayesian classifier for classification is divided into two stages: the first step is the Bayesian classifier learning phase, namely construct classifiers from sample data; the second stage is the application of Bayesian classifier, namely compute the conditional probability of the disaggregated data and classify them. In this section, we will describe how to create a Bayesian classifier, and in Sect. 3, we then analyze in detail how to apply Bayesian classifier model in analyzing evidence. When building the Bayesian classifier, the main purpose is estimating the prior probabilities of class S and $\neg S$ based on the training set, i.e. P(S) and P(¬S), in addition, we also need to obtain P(D|S) and P(D|¬S). We can get the prior probability from training set very easily:

$$\mathrm{P}(S_i) = \frac{\text{the number of text messages belonging to class } S_i}{\text{the total number of messages in the training set}}$$

The conditional probability of message D $\mathrm{P(D}|S_i)$ can be obtained from the conditional probability of the characteristics in the message, i.e.

$$\mathrm{P(D}|S_i) = \prod_i P(w_i|S_i),$$

$$\text{where } \mathrm{P}(w_i|S_i) = \frac{\text{the number of messages containing the word } w_i \text{ in class } S_i}{\text{total number of messages in class } S_i}$$

Constructing a Bayesian classifier steps are as follows:

(1) Collect a large number of spam and non-spam messages, setting up spam message set and non-spam message set.

(2) Preprocess messages, including word segmentation and removing stopwords, then extracting separate string tokens from the message, and calculating each word frequency, frequency the string appeared. Process spam message set and non-spam message set respectively according to this method.

(3) Each message set has a hash table, hashtable_good correspond to the non-spam message set and hashtable_bad correspond to the spam message set. In these hash table, the "Key" column is each token string, and the "Value" column is the number of occurrences of the token, respectively, we indicate them with $good(w_i)$ and $bad(w_i)$.

(4) Considering hashtable_good and hashtable_bad, infer when a token appears in a new text, what is the probability the message is spam. Thereby we create the third hash table hashtable_probability, the "Key" column is still the token strings, the "Value" column is the probability the message is spam when it contains the token, denoted by $\mathrm{P}(\mathrm{S}|w_i)$, calculated as follows:

$$r_{\mathrm{g}} = \mathrm{good}(w_i)/G, r_{\mathrm{b}} = \mathrm{bad}(w_i)/B$$

$$\mathrm{P}(\mathrm{S}|w_i) = r_{\mathrm{b}}/\left(kr_{\mathrm{g}} + r_{\mathrm{b}}\right)$$

w_i in the formula is the token string used to calculate the probability, good and bad represent the hash table created in step (2), G and B respectively represent the number of non-spam messages and spam messages, and k = G/B.

Thus, the learning process of training sets end. According to the established hash table, we can apply formula 7 to estimate the P(S|D), the probability a new message is spam.

$$P(S|D) = \frac{\prod_i P(S|w_i)}{\prod_i P(S|w_i) + k^{1-n}\prod_i (1 - P(S|w_i))} \quad k = \frac{P(\neg S)}{P(S)}$$

3 Spam Message Forensics

3.1 Background

With the arrival of the information age, information technology, digitization has penetrated into people's daily work and life, instant messaging has become one of people's indispensable carry-on items. Served as a communication tool in people's life, it brings convenience and fun, but at the same time, it recorded some dark side of the society. We have reason to believe that messages, the product of daily communication, will be the crucial evidence against the criminal suspect.

So when the party institutes legal proceedings to the court, the spam messages that the senders sent will be the most favorable evidence. How to guarantee the credibility and probative value of the electronic evidence (spam message), not only need a clear legal norm, but also need standardized forensic techniques, and in which the standardized electronic evidence forensics process is particularly important to ensure the authenticity, integrity and legitimacy of the evidence.

3.2 Forensics Process

Conventional process of electronic evidence forensics includes the following steps:

(1) Protect the site: avoid any change to the system settings, damage to the hardware, and data corruption or virus infection.
(2) Obtain the evidence: use disk mirroring tools (such as the safe back, SnapBack DatArret and DIBS RAID, etc.) and forensics tools (e.g. EnCase) backupping the target system image and collecting evidence, then filing them.
(3) Preserve the evidence: due to the volatile of electronic evidence, it is required to carry out the original target system image backup, then encrypt and store them.
(4) Identify the evidence: solve the integrity verification of evidence and determine whether they meet the criteria.
(5) Analyze the evidence: look for the matching key words or key phrases in the two data streams (general data and peripheral data), to discover the link between electronic evidence and the criminal facts.
(6) Track: the purpose of tracking is to find the source of the attack (sources of equipment, software and IP address, etc.), and usually the tracking process for forensics must be logged (including journal of operating system, firewall, IDS and application software, etc.). Also we can use some related equipment or set traps (honeypot) to track and capture the criminal suspect.
(7) Present the evidence: indicate the extracted time, location, machine, extraction and witnesses of the evidence, and then submit it to the judicial organ in the visible form, and provide a complete chain of supervision.

In this article, we assume that the sender of the spam messages is known, and Judiciary requires operators to produce its message records communicating with the outside world in the last period of time for analysis, and working as evidence of it's illegal or not. Given this evidence collection process is different from the conventional process, so the forensic process will be different, mainly in the following steps:

(1) Obtain the evidence: apply for investigation order, and obtain related message records from telecommunication operators. According to the requirements of relevant laws and regulations, the data provided by the operators must include the sender's and the recipient's name, phone numbers, sending and receiving time, and the context of the messages.
(2) Preserve evidence: the data provided by the operators are just a copy of the original data, although the probability that operator loses the data is very small, it can provide the original data unlimited, but in order to simplify the process of identification, it is absolutely necessary to fix the data when we get them, against unintentional or malicious damage during the analysis of data. Fixed methods can be: ① make a copy of the original data using Hash operation, extracting the

fingerprints; ② make several copies with the method "mirrored" and store the original image backup using appropriate storage media. We recommend that store two of the copies in the write-once media, which can prevent extracted evidence accidentally be changed; ③ when making the copies, record the time, place, method of copying, handling personnel, the used software and hardware, and the process of replicating as a third parity, in order to ensure the legality, authenticity, relevance and integrity of the electronic evidence.

(3) Analyze the evidence: classify the large number of data provided by operators with the established Bayesian spam detection model, and calculate the proportion of spam messages in all messages. Steps using the detect model for analysis are as following:

   (a) The program scans the messages to be analyzed one by one, and identifies all the token strings according to the step (2) building the model.

   (b) Query the hash table hashtable_probability, get all the "Values" of the "Keys". Assuming that there n tokens getting from the message, w1, w2… wn.

   (c) Extract features. Find out 15 the most eye-catching identifiers (tokens deviate from the neutral 0.5 furthest) from n tokens, and determine the probability of the entire message is spam message with them. The method to select the features is: mark the value that token wi correspond to in hashtable_probability as ti, make vi = |ti – 0.5|, and order vi (1 ≤ i ≤ n) from large to small, form a set V = {v1, v2, …, vn}, its subset V' = {v1, v2, …, v15} is the selected set of features.

   (d) Express the 15 most eye-catching identifiers with w1, w2…w15, and $\mathrm{P}(\mathrm{S}|w_1, w_2, \cdots, w_{15})$ represents when a message contains w1, w2…w15 at the same time, what is the probability the message is a spam message. It can work out by formula 7:

$$\mathrm{P}(\mathrm{S}|w_1, w_2, \cdots, w_{15}) = \frac{\prod_{i=1}^{15} P(S|w_i)}{\prod_{i=1}^{15} P(S|w_i) + k^{-14} \prod_{i=1}^{15} (1 - P(S|w_i))}$$

When $\mathrm{P}(\mathrm{S}|w_1, w_2, \cdots, w_{15})$ exceeds a threshold, we determine the message as spam.

Present the evidence. At present the laws in our country have not clearly stipulated the legal form of electronic evidence, but usually to some tangible electronic documents such as E-mail, web pages, adopt the practice of notarization forensics, turning them into a notarial deed, recording and displaying these contents.

4 Experiments and Discussion

4.1 Experiment Objectives

The main purpose of this experiment is to check the effect of the Bayesian spam detection model, and determine whether it can provide credible evidence to the court trial and be applied to the judicial appraisal according to the result.

4.2 Experimental Environment

(1) Hardware Configuration:

Model: HP ProBook 4416s
CPU: AMD Athlon(tm) II Dual-Core M300 2.00 GHz
System memory: 1.00 G
System type: 32-bit operating system

(2) Software Configuration:

System version: Windows 7 u
Development environment: Android Developer Tools v21.1.0-569685
JDK 1.6.0_16

4.3 Experimental Process

Test Data. Since short message involves a lot of personal privacy, there is not a corpora publicly available up-to-date. Collecting from network and daily life, we construct a private short message set which contains 2000 normal messages and 2000 spam messages. Messages after processing are given in the form of a document, which does not contain any personal privacy information except the content of message.

Test Method. Divide two types of messages equally. By choosing one set respectively from the two types of message set randomly, we have constructed a training set with 1000 spam messages as well as 1000 normal messages and a testing set with 1000 spam messages as well as 1000 normal messages.

The common evaluation measures include true positive, true negative, false positive, false negative, detection rate, false positive rate and overall accuracy. Their corresponding definitions are as follows [19]: True positives (TP): The number of spam message classified as spam.

True Negatives (TN): The number of normal message classified as normal.
False Positives (FP): The number of normal message falsely classified as spam.
False Negatives (FN): The number of spam message falsely classified as normal message.
Recall: TP/(TP + FN).
Fallout: FP/(TN + FP).

Precision: TP/(TP + FP).
Accuracy: (TP + TN)/(TP + FP + FN + TN).

Test Results

The distribution of the test results is shown in Table 1:

Table 1. The distribution of the test results

TP	TN	FP	FN
968	990	10	32

(Annotations: the total number of messages N = 2000)

4.4 Result Analysis

The Recall shows the spam detection accuracy of a classifier. A higher Recall indicates better spam detection. Fallout indicates the false detection of incoming messages. A classifier with a high Fallout will move normal messages into the spam folder without user notification.

Based on Table 1, we can get the performance evaluation of the model shown in Table 2:

Table 2. Performance evaluation

Recall	Precision	Accuracy	Fallout
96.80%	98.98%	97.90%	1.00%

From the test results, we can see that the performance of this system can completely meet the needs of spam detection. Its high recall rate 96.80% is very strong for spam recognition. When it is used to spam messages forensics, it can effectively identify the spam message from a large number of messages sent by the suspects. In addition, the system's 0.8 ms per message is also affordable, it can provide the analyzing results in time.

5 Conclusion

Under the condition of the serious lack of relevant legal norms about digital evidence in our country, on the one hand, relevant departments should further enrich the special rules that are related to digital evidence, on the other hand, China's three major procedural law should make some necessary regulations aimed at digital evidence. Our country should change the present research situation of the digital evidence. It must get rid of the current phenomenon that law and the technology are "two pieces of skin", strengthen the cooperation between experts of legal and electronic, to jointly study the countermeasures to solve the problem of digital evidence. The spam message forensics process proposed in this paper is combined with the Bayesian spam detection model. So it's making some efforts on this condition.

Bayesian classifier's filtering mechanism is based on the content, it has a high degree of automation, so it's suitable for filtering the ads or fraud messages. Messages sent by the illegal service provider (the subject of spam senders) are mostly advertising and fraud messages, so the Bayesian classifier fully meet the demand of analyzing messages sent by illegal service provider. Therefore, the forensics process proposed in this paper has a high practical value. Of course, this is only the initial stage, the perfection of legal system of digital evidence in China still need the joint efforts of many legal experts and electronic experts.

References

1. Ma, J., Zhang, Y., Liu, J., et al.: Intelligent SMS spam filtering using topic model. In: 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS), pp. 380–383. IEEE (2016)
2. Huang, W.L., Liu, Y., Zhong, Z.Q., et al.: Complex network based SMS filtering algorithm. Acta Automatica Sinica **35**(7), 990–996 (2009)
3. Wang, C., Zhang, Y., Chen, X., et al.: A behavior-based SMS antispam system. IBM J. Res. Dev. **54**(6), 651–666 (2010)
4. Xiang, Y., Chowdhury, M., Ali, S.: Filtering mobile spam by support vector machine. In: Debnath, N. (ed.) Proceedings of the Third International Conference on Computer Sciences, Software Engineering, Information Technology, E-business and Applications, pp. 1–4 (2004)
5. Healy, M., Delany, S., Zamolotskikh, A.: An assessment of case-based reasoning for short text message classification. In: Creaney, N. (ed.) Proceedings of 16th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2005), pp. 257–266 (2005)
6. Duan, L.Z., Li, A., Huang, L.J.: A new spam short message classification. In: Proceedings of the 1st International Workshop on Education Technology and Computer Science, Wuhan, Hubei, China, pp. 168–171 (2009)
7. Zheng, X.X., Liu, C., Zou, Y.: Chinese Short Messages service spam filtering based on logistic regression. J. Heilongjiang Inst. Technol. **24**(4), 36–39 (2010)
8. Cai, J., Tang, Y.Z., Hu, R.L.: Spam filter for short messages using Winnow. In: 7th International Conference on Advanced Language Processing and Web Information Technology, Liaoning, China, pp. 454–459 (2008)
9. Gómez Hidalgo, J.M., Bringas, G.C., Sánz, E.P., García, F.C.: Content based SMS spam filtering. In: Bulterman, D., Brailsford, D.F. (eds.) Proceedings of the 2006 ACM Symposium on Document Engineering, DocEng 2006, pp. 107–114. ACM, New York (2006)
10. Zhang, J., Li, X.M., Xu, W., et al.: Filtering algorithm of spam short messages based on artificial immune system. In: 2011 International Conference on Electrical and Control Engineering, ICECE 2011 Proceedings, Yichang, China, pp. 195–198 (2011)
11. Mahmoud, T.M., Mahfouz, A.M.: SMS spam filtering technique based on artificial immune system. IJCSI Int. J. Comput. Sci. Issues **9**, 589 (2012)

12. Junaid, M.B., Farooq, M.: Using evolutionary learning classifiers to do mobile spam (SMS) filtering. In: Proceedings of Genetic and Evolutionary Computation Conference, GECCO 2011 (2011)
13. Wu, N., Wu, M., Chen, S.: Real-time monitoring and filtering system for mobile SMS. In: Proceedings of 3rd IEEE Conference on Industrial Electronics and Applications, pp. 1319–1324 (2008)
14. Jie, H., Bei, H., Wenjing, P.: A Bayesian approach for text filter on 3G network. In: Proceedings of the 6th International Conference on Wireless Communications Networking and Mobile Computing, pp. 1–5 (2010)
15. Deng, W.W., Peng, H.: Research on a Naive Bayesian based short message filtering system. In: Proceedings of the International Conference on Machine Learning and Cybernetics, pp. 1233–1237. IEEE (2006)
16. Totaro, G., Bernaschi, M., Carbone, G., et al.: ISODAC: a high performance solution for indexing and searching heterogeneous data. J. Syst. Softw. **118**, 115–133 (2016)
17. Bayes, T.: An essay towards solving a problem in the doctrine of chances vol. 1, no. 2, pp. 726–730 (1763)
18. Joyce, J.: Bayes' Theorem (2003). http://www.science.uva.nl/∼seop/entries/bayes-theorem/
19. Shih, D.-H., Jhuan, C.-S., Shih, M.-H.: A study of mobile SpaSMS filtering system. In: The XVIII ACME International Conference on Pacific RIM Management, Canada, 24–26 July 2008 (2008)

Spam Mail Filtering Method Based on Suffix Tree

Runqiu Hu[1] and Yitao Yang[2(✉)]

[1] School of Computer, Nanjing University of Posts and Telecommunications,
No. 9 Wenyuan Road, Nanjing, China
B14040735@njupt.edu.cn

[2] Information Technology Department, Nanjing Forest Police College,
No. 28 Wenlan Road, Nanjing, China
youngyt@gmail.com

Abstract. In recent years, e-mail technology is prospering, bringing efficiency to people from all over the world. It is not limited to time and space, making the transmission of information more convenient. However, the emergence of spam has also brought people a lot of trouble. Thus, spam filtering research is necessary. Traditional spam filtering is mainly based on black and white list technology. Over the past decade, with the development of machine learning, Bayesian classifier has also come into use. However, support for Chinese mail has always been unsatisfactory. This paper proposes a Chinese spam filtering method based on suffix tree, which solves the problem of Chinese character processing and compares it with traditional methods from the aspects of time and space complexity and accuracy.

Keywords: Spam mail filtering · Suffix tree · Natural language processing · Machine learning

1 Introduction

With the popularization of email, many Internet users benefit from its convenience of delivering the message. However, problems of spam mails have also been emerging because of its fast, unlimited transmission anytime, anywhere. According to The Research Report on Chinese enterprise mailbox security in 2016, the number of spam mails received by the enterprise user of email service in China is over 20 million, which consists 69.8% of the total. As for general netizens, in the first half of 2014, each receives 17.5 spam mails every week on average, which takes up 68.6% of all mails received, increasing by 2.2% compared to last season.

The threat under spam mails are as follows [1]:

- Bandwidth occupation
- Internet resource waste
- Network security attack like mail bomb [5].

Therefore, the better method of spam mail filtering is necessary to provide a more efficient and secure network communication.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_43

1.1 Black/White List Filtering

Black/White List Filtering is mainly based on the source address of the email. The white lists contain legitimate or spam senders or relay servers; the filter classifies black list entries as spam. There are many well-developed Anti-Spam Alliance all over the world and the Real time Blackhold List is provided [3]. This filtering method depends on the technique of Hash table. The domains are stored in the Hash table. An email is judged as spam if the address of which can be found in the table. As searching algorithm on Hash technique performs excellent, the method outstands in low requirement of time and space complexity. However, the list needs to be updated constantly, which cannot be done until new spam mails arrive and this brings inflexibility to this method.

1.2 Naïve Bayes Classification

Filtering method based on Bayes Classification focus most on Bayes Theorem [4]. A model on spam and ham mail should be established with known mails and then judge if the mail is spam according to the model. The most important thing for naïve Bayesian classification is attribute extraction, and the proper attributes will bring high accuracy. Bayesian filter often choose words in the message as attributes [7].

This method is good at its relatively better performance on time and space complexity, with a satisfactory accuracy as well. While it has been widely adopted by many email products these years, its great dependence on the training set and lower performance under Chinese mail are still problems to be settled.

1.3 K-NN Algorithm

In pattern recognition, the k-nearest neighbours algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression [9] (Fig. 1).

1.4 Suffix Tree Algorithm

The three methods mentioned above can easily and briefly express the text but they sometimes ignore the context [2], and for the real time updated training set, the model needs to be rebuilt entirely. Thus it's not an online-algorithm, which can process the input in a serializable approach without having to know the whole input. These methods are inadequate for increasingly updated spam libraries. The spam filtering method based on suffix tree is outstanding in the performance of construction, search and updating, and the filtering of spam can be carried out more efficiently in combination with the development of parallel computing in recent years (Fig. 2).

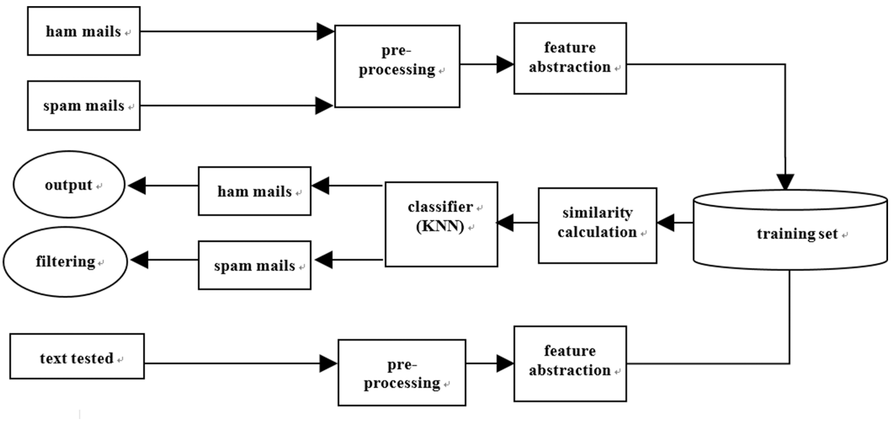


Fig. 1. The procedure of spam mail filtering with k-NN algorithm

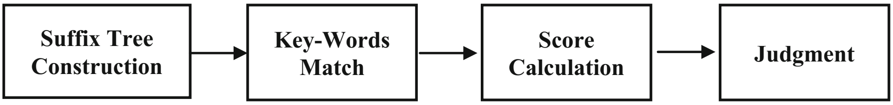


Fig. 2. Four steps taken to filter spam mails using Suffix Tree

For the construction part, Suffix Trie and Suffix Tree are two data structures widely used in spam filtering, as well as other fields regarding Natural Language Processing.

Take the string *T* = *“BANANAS”* as an example, a suffix trie is built as below [6].

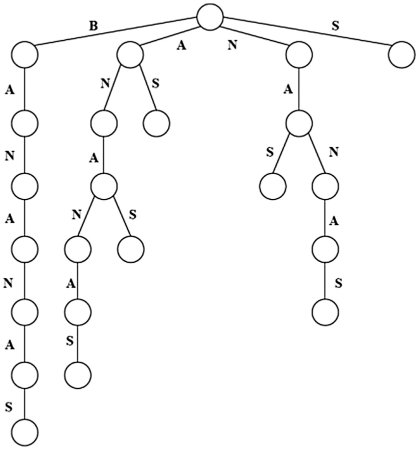


Fig. 3. Example of Suffix Trie with string *T* = *“BANANAS”*. Every empty circle represents a Suffix Node, the lines connected are called edges

Starting from the root node, every suffix of “BANANAS” is inserted into the tree. Therefore, the time complexity of searching is O(n), whereas both the time and space complexity of construction reach up to O(n^2).

Based on Suffix Trie, if nodes with only one child are removed, another tree will be constructed called Suffix Tree (Fig. 4).

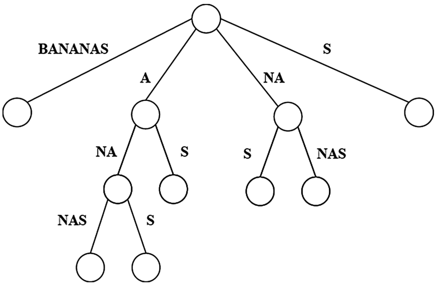


Fig. 4. Suffix Tree constructed based the Suffix Trie in Fig. 3. The nodes with only one child are removed

Therefore, an algorithm of establishing a suffix tree was put forward by McCreight [8], which consists of two steps: constructing a suffix trie and deleting the nodes with only one child.

The algorithm of constructing suffix tree above can only meet the requirement on single string search. However, in practical application, what is more often required is string match on multiple files with multiple length. It is of course a solution to generate more than one suffix tree. However, it is easy to find and count the number of substrings in a tree if all the information can be stored in one tree.

So there are two problems remaining to settle. One is how to deal with a number of substrings, especially for the read and write of files. The other one is how to use a better algorithm than McCreight to have a better performance.

2 Ukkonen’s Algorithm

As is mentioned above, McCreight is an offline algorithm that is not suitable for dealing with dynamically changing data sets.

In computer science, Ukkonen’s algorithm is a linear-time, online algorithm for constructing suffix trees, proposed by Esko Ukkonen in 1995 [10].

Using the Ukkonen algorithm as an online algorithm, the construction can be completed without knowing all the input at the beginning, and its time complexity is O (n), with the spatial O (n^2).

The construction of a suffix tree using Ukkonen’s algorithm starts from root, with one node added each phase.

Definition 1: An **active point** is defined to be the point from which traversal starts for next extension or next phase. Active point always starts from root. Other extension will get active point set up correctly by last extension.

Definition 2: An **active node** is defined to be the node from which active point will start.

Definition 3: An **active edge** is defined to be the node used to choose the edge from active node, which has index of character. The length to go on active edge is called active length.

Definition 4: If an edge is split and a new node is inserted, and if that is not the first node created during the current step, the link connected from the previously inserted node to the new node through a special pointer is called a **suffix link.**

Let $\mathrm{p} > 0, \mathrm{q} > 0, p < \mathrm{q}$ and $\mathrm{S}[\mathrm{p}, \mathrm{p}+1, \ldots, \mathrm{q}]$ is set of nodes beginning from p to q. The following three rules are supposed to be followed.

Rule 1: For phase $\mathrm{q}+1$ if $\mathrm{S}[\mathrm{p}, \mathrm{p}+1, \ldots, \mathrm{q}]$ ends at last character of leaf edge then add $\mathrm{S}[\mathrm{q}+1]$ at the end.

Rule 2: For phase $\mathrm{i}+1$ if $\mathrm{S}[\mathrm{p}, \mathrm{p}+1, \ldots, \mathrm{q}]$ ends somewhere in middle of edge and next character is not $\mathrm{S}[\mathrm{q}+1]$ then a new leaf edge with label $\mathrm{S}[\mathrm{q}+1]$ should be created

Rule 3: For phase i + 1 if $\mathrm{S}[\mathrm{p}, \mathrm{p}+1, \ldots, \mathrm{q}]$ ends somewhere in middle of edge and next character is $\mathrm{S}[\mathrm{q}+1]$ then remain the tree as before.

An example of the construction of the tree with string xyzxb$ is given to illustrate the rules.

Below are optimized techniques used in the algorithm:

1. While traveling down if number of characters on edge is less than number of characters to traverse then skip directly to the end of the edge. If number of characters on label is more than number of characters to traverse, then go directly to that character concerned.
2. Instead of storing actual characters on the path store start and end indices on the path.
3. Stop process as soon as rule 3 is hit. Rule 3 can be considered as a stopper.
4. Keep a global end on leaf to do rule 1 extension (Fig. 5).

Theorem: Using the skip/count trick, any phase of Ukkonen's algorithm takes O(m) time.

Here gives the proof of the theorem:

- There are i + 1 $\leq$ m extensions in phase i + 1;
- In a single extension, the algorithm walks up at most one edge, traverses one suffix link, walks down some number of nodes, applies the extension rules and may add a suffix link;
- The up-walk decreases the current node-depth by at most one.
- Each suffix link traversal decreases the node-depth by at most another one.
- Each down-walk moves to a node of greater depth.

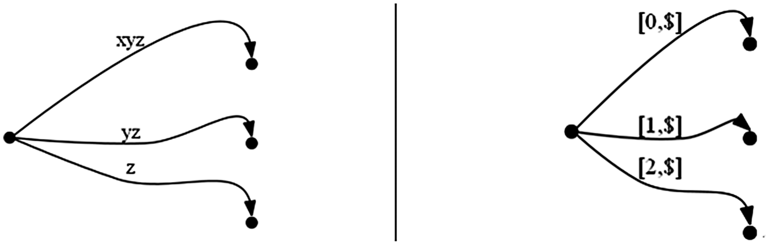


Fig. 5. Comparison of using global end and original characters

- Over the entire phase the node-depth is decremented at most 2 × m times.
- No node can have depth greater than m, so the total increment to current node-depth (down walks) is bounded by 3 × m over the entire phase.

3 Search and Matching Algorithm

The search algorithm has long been researched, with many well-known ones like KMP, Boyer Moore, Sunday, Robin-Carp, etc. These algorithms show great performances under different circumstances. Since a suffix tree is created, the search and match algorithm should base on the tree.

3.1 Break the Limitation of Language

However, the previous research focus largely on English text. For Chinese and other languages, the technique is seldom mentioned. The trouble lies in the encoding of Chinese characters.

Chinese characters uses Unicode for encoding. The first 128 Unicode code points are the same as ASCII. From 11904 starts Chinese, Japanese and Korean characters. Here, we use the encoded numbers instead of real characters to represent the words because an English character and a Chinese character differ from the number of bytes taken in memory, which can bring huge trouble when searching and matching. This takes some preprocessing time, but makes it possible to deal with the spam mail in any other languages all over the world theoretically.

For example, “HI大家好” is stored as “72,73,22823,23478,22909” for processing.

3.2 Search and Matching

Here presents an example using string “abaaba” to demonstrate the search and matching algorithm.

Firstly, a suffix tree should be constructed under Ukkonen’s algorithm (Fig. 6).

Then the matching can start. Suffix tree string matching algorithm is very simple, starting from the root node, to the child node and the search string to be matched, if the

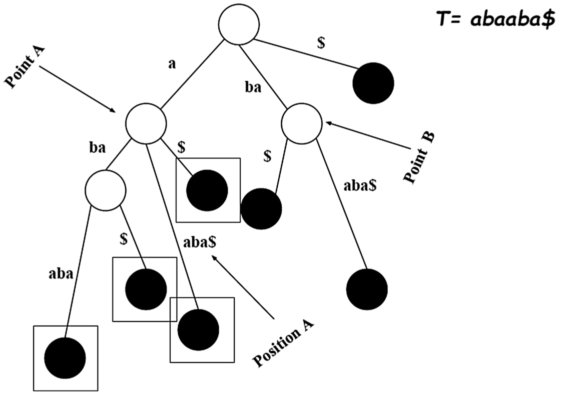


Fig. 6. Suffix Tree built using Ukkonen's algorithm with string *T* = *abaaba$*, black nodes are leaf nodes

search string to be completed before the end of the match failed to meet, then there is no such substring else the substring exits. If the matching node has a child "$", then the string is not the only a substring, but a complete suffix.

For example, the search of substring "baa", starting from the root node, and arrives at Point B. "ba" matches successfully, followed by "a" and the match is completed. "baa" is not a suffix because the next character is not " $ ".

Search of the substring "aaa" first arrives at Point A, and matches successfully. Then come letter "a", where Position A is reached. Now the match is unable to go on, with the string to be searched has not yet get to the end, so the match fails.

To obtain the number of occurrences of a substring in the entire string, first match the substring from the root node, and then count the number of all the leaf nodes below. As mentioned earlier, the addition of "$" makes all the leaf nodes unique, so once there is a leaf node below the node, there must be a unique string containing the string.

For example, Point A in the figure represents the substring "a". Notice the four leaf nodes with the square covered, which represent the four suffixes containing "a". As shown, "a" appears four times in the string.

In programming, counting the number of such nodes, only needs to add the parent variable in the parent Suffix Node class, then use of depth-first traversal algorithm to get the number of strings appearing.

This method of counting the number of times shares the time complexity with the binary tree traversal, which is O(n). Where n is the number of nodes. Among the following traditional statistical methods, suffix tree method outstands due to less time complexity as the table below, where M is the length of the string, N is the sub-string length (Table 1).

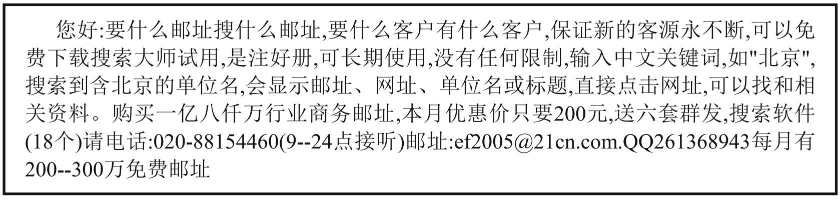
Table 1. Time complexity of some string searching algorithm

Algorithm	Preprocessing	Matching
Brute force	O(0)	O(N × M)
KMP	O(M)	O(N)
Boyer moore	$O(N+M^2)$	O(N)
Sunday	O(M)	O(N × M)
Robin-Carp	O(0)	O(N × M)
Suffix tree matching	O(0)	O(N)

4 Experiment

4.1 Spam Corpus Collection

Chinese corpus in the TREC 2006 Spam Track Public Corpora corpus (referred to as trec06c) is chosen as the spam mail database in this experiment. It contains 64619 Chinese spam with the contents of different types of aspects, the following is one of the samples (Fig. 7):



您好:要什么邮址搜什么邮址,要什么客户有什么客户,保证新的客源永不断,可以免费下载搜索大师试用,是注好册,可长期使用,没有任何限制,输入中文关键词,如"北京",搜索到含北京的单位名,会显示邮址、网址、单位名或标题,直接点击网址,可以找和相关资料。购买一亿八仟万行业商务邮址,本月优惠价只要200元,送六套群发,搜索软件(18个)请电话:020-88154460(9--24点接听)邮址:ef2005@21cn.com.QQ261368943每月有200--300万免费邮址

Fig. 7. Chinese Spam mail sample

4.2 Spam Mail Key-Words Database

The key words database comes from known spam mail wordlist provided by Emarsys, a company of B2C Marketing Automation and the blacklist of spam messages of Kingsoft Mobile Guard.

Considering longer items have relatively lower chance to be matched, items longer than 5 Chinese characters are split into shorter ones. For instance, “代理开具发票” is split into “代理”, “开具”, “发票”.

Some words unlikely to be filtered are also observed, which should be removed to make the match more efficient. Therefore, we choose the first 2000 spam mails and calculate the time of occurrences and deleted items like “依据法律”, “全新呼机”, which never appears in the 2000 mails.

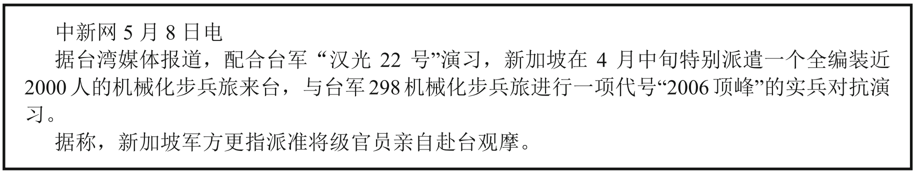
The key words data set containing 155 items is listed below (Table 2):

Table 2. Key words used to filter

券	折	100%	仅	订	汇	票	交易
提现	信用	八卦	保证	免费	只需	收入	表格
为您	朋友	访问	满意	广告	现金	CD	光盘
名人	认证	全新	机会	购买	支票	注意	点击
避免	比较	费用	帮助	合法	这里	隐藏	主页
在家	损失	保险	投资	名牌	密码	优惠	每天
营销	绩效	每周	一次	电话	媒体	申请	商业
学位	会员	价格	百万	奇迹	开放	打印	财富
机遇	订单	问题	价值	生产	体检	收益	不需
向您	承诺	责任	利润	现在	股票	联系	私人
热线	咨询	发票	见谅	打扰	光临	促销	移民
留学	出国	商务	小区	二手	中奖	激情	抽奖
教育	小学	中学	辅导	积分	客户	先生	小姐
短信	离婚	手机	电脑	法院	汇款	上门	回复
宣传	周末	酒店	公寓	教师	加盟	培训	机票
社区	大礼	信托	英语	汽车	代理	中心	航空
隆重	别墅	代开	代办	研修	纯天然	最低价	专卖店
写字楼	最低价格	报名	在线	长途电话	网络营销	行动	打印机
互联网	领域	限制	保密	复制	容易	索取	经验
信用卡	债务	拒绝					

4.3 Ham Text Dataset

This article selects news text library November 27, 2006 provided by Sougou, which contains political, economic, cultural and other aspects. The following is one of the samples (Fig. 8):



中新网 5 月 8 日电

据台湾媒体报道，配合台军“汉光 22 号”演习，新加坡在 4 月中旬特别派遣一个全编装近 2000 人的机械化步兵旅来台，与台军 298 机械化步兵旅进行一项代号“2006 顶峰”的实兵对抗演习。

据称，新加坡军方更指派准将级官员亲自赴台观摩。

Fig. 8. Chinese Ham mail sample

4.4 Evaluation Indicator

This article classify the mail as spam or ham based on score calculated. Let each word be L_i, and the weight of each word is written as p_i, where $i \in [0,154]$. The final score for a message is recorded as Score (mail) [10].

$$\mathrm{Score(mail)} = \sum_{i=1}^{155} \mathit{Exist}(L_i)p_i \tag{1.}$$

where,

$$\mathit{Exist}(L_i) = \begin{cases} 0, & \mathrm{L_i}\ \mathit{exists\ in\ the\ mail} \\ 1, & \mathrm{L_i\ doesn'}\mathit{t\ exist\ in\ the\ mail} \end{cases} \tag{2.}$$

4.5 Weight Choosing·

For $\mathrm{p_i}$,

$$\mathrm{p_i} = \frac{\mathit{length}(L_i) \times \mathit{frequency}(L_i)}{\mathit{length}(L)} \tag{3.}$$

where length($\mathrm{L_i}$) represents the length of the substring, frequency($\mathrm{L_i}$) represents the time of occurrence in the string, length(L) represents total number of words in the key words list, which is a constant equals 316.

To evaluate the correctness of this indicator, select 2000 e-mails (not the same as training set) and calculate the scores. Notice that calculated score range from 0 to more than 100. Some spam mail got the score of 0, and some normal ones got relatively high score.

We choose 3,4,5,6,7,8,9,10 as spam and normal text of the critical value of λ, mails scored greater than λ is considered spam, otherwise it is ham mail.

As the weight of the errors for normal mail being judged to be spam (denoted as SH) and for spam being judged to be normal mail (recorded as HS) is difficult to define, we set the weight of error of SH: HS to be 3: 7,4: 6, 5: 5,6: 4,7: 3.

So the accuracy can be expressed as

$$\mathrm{Accuracy} = \left(1 - \frac{\mathrm{SH}}{2000}\right) \times \mathrm{weight(SH)} + \left(1 - \frac{\mathrm{HS}}{2000}\right) \times \mathrm{weight(HS)} \tag{4.}$$

For example, if 20 mails out of 2000 spam mails are judged as ham mail and 40 mails out of 2000 ham mails are judged as spam mail, given the ratio of weight 3:7 the accuracy should be 96.2% (Fig. 9).

By adding the trend line, we can get the coordinates of the intersection of several curves (4.78, 77.8%). Selecting the threshold(λ) as 4.78 can ignore the problem of accuracy caused by SH and HS weight ratios. Although not optimal, the accuracy of the 77.8% is still superior to the Bayesian classifier method (Table 3).

Note that Suffix Tree needn't pre-processing, for it is an online algorithm, which can deal with stop-list as well as Lemmatizer.

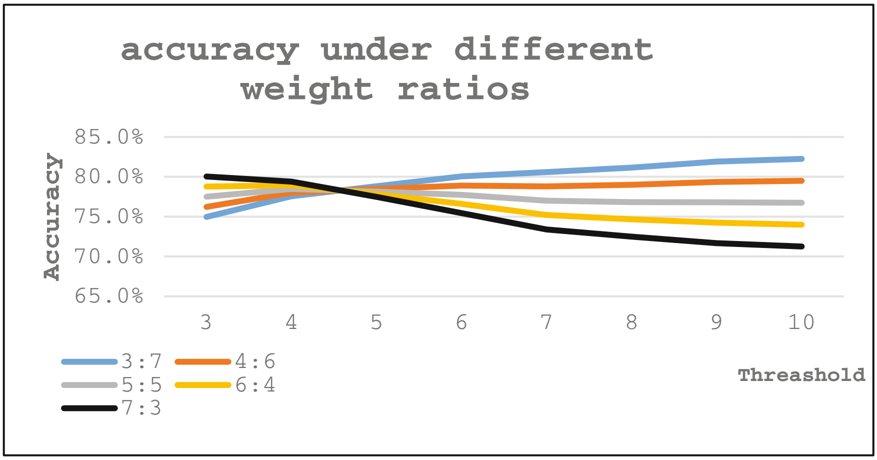


Fig. 9. Results of accuracy under different weight ratios. Threshold ranges from 3 to 10, with different colors indicate the ratio of weight to deal with error of SH and HS

Table 3. Comparison between Suffix Tree and Naive Bayes Classifier

Pre-processing	Word numbers	Accuracy(%)
Non	200	73.82
Stop-list	200	73.40
Lemmatizer	300	63.67
Lemmatizer + stop-list	300	63.05
Suffix Tree	155	77.80

4.6 Complexity Analysis

For time complexity, the comparison with some typical string searching methods is listed below (Table 4):

m is the number of features which is the length of the text in the suffix tree algorithm. m' represents the number of iterations in the Booting algorithm. Suffix tree filtering method has a significant advantage when adding or removing content from a training set, since Ukkonen is an online algorithm that does not need to be rebuilt.

Table 4. Comparison between complexity of different string processing algorithm

Filtering method	Training time	Classifying time	Training set updating time
Naïve Bayes	O(mN)	O(m)	O(mN)
Elastic Bayes	O(mN)	O(mN)	O(mN)
SVM	$O(m^2N^2)$	$O(m^2N)$	$O(m^2N^2)$
Boosting	$O(m'mN^2)$	$O(m')$	$O(m'mN)$
Suffix tree	O(mN)	O(m)	O(m)

For space complexity, the memory issue mentioned above in Suffix Trie is no longer a problem under Ukkonen's algorithm with the use of compression technology. The key to the compression technology is that each operation is just movement of two pointers (the first element and the tail element). However long string as it can be, the storage takes only space of two flags. Therefore, the space complexity of the algorithm is O (m), where m is the length of the text.

5 Conclusion

This article introduces a spam filtering method based on suffix tree. There has been great deal of spam filtering technology with the recent development of machine learning, most of which filter through the calculation of probability and vector model. The development of cloud computing and big data makes the natural language processing out of the limitation of space and time. Through the suffix tree online algorithm, the context can be considered when filtering, making the judgment more precise and reasonable.

For further research, keywords extraction can be completed with the suffix tree from a large number of emails and can be classified with different types, making the weight distribution more valid, and the choice of threshold more scientific.

References

1. Clark, J., Koprinska, I., Poon, J.: A neural network based approach to automated e-mail classification. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence, WI 2003, pp. 702–705. IEEE (2003)
2. Jianlong, T., Ji, Z., Li, G.: Method of spam filtering based on general suffix tree model. Comput. Eng. **33**(9), 100–102 (2007)
3. Kim, J., Chung, K., Choi, K.: Spam filtering with dynamically updated URL statistics. IEEE Secur. Priv. **5**(4), 33–39 (2007)
4. Schneider, K.M.: A comparison of event models for Naive Bayes anti-spam e-mail filtering. In: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics, vol. 1, pp. 307–314. Association for Computational Linguistics (2003)
5. Takemura, T., Ebara, H.: Spam mail reduces economic effects. In: 2008 Second International Conference on the Digital Society, pp. 20–24. IEEE (2008)
6. Pampapathi, R., Mirkin, B., Levene, M.: A suffix tree approach to anti-spam email filtering. Mach. Learn. **65**(1), 309–338 (2006)
7. Firte, L., Lemnaru, C., Potolea, R.: Spam detection filter using KNN algorithm and resampling. In: 2010 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 27–33. IEEE (2010)
8. McCreight, E.M.A.: Space-economical suffix tree construction algorithm. J. ACM (JACM) **23**(2), 262–272 (1976)
9. Ukkonen, E.: On-line construction of suffix trees. Algorithmica **14**(3), 249–260 (1995)
10. http://spamassassin.apache.org/. Apache SpamAssasin

MP3 Audio Watermarking Algorithm Based on Unipolar Quantization

Wu Wenhui[1(✉)], Guo Xuan[1,2], Xiao Zhiting[1], and Wen Renyi[2]

[1] PLA Academy of National Defense Information,
Wuhan 430010, Hubei, China
wjguoxuan@163.com
[2] Officers' College of Chinese Armed Police Force,
Chengdu 610213, Sichuan, China

Abstract. This paper analyzes the current situation of MP3 application and the advantages of wavelet transform in audio watermarking, and effectively uses of the unipolar quantization method to propose an MP3 audio watermarking algorithm based on wavelet transform. The algorithm firstly decodes the MP3 audio, then uses the third-order discrete wavelet transform, embeds the low-frequency coefficients with the method of unipolar quantization, and finally obtains the watermarked MP3 audio file. Simulation results show that the algorithm has good auditory transparency and strong robustness by low-pass filtering, resampling, whitening, and cropping attacks on watermarked audio signals. The watermarking can be fast extracted, and meet the real-time requirements.

1 Introduction

Digital watermark is a branch of information hiding, but also one of the earlier information security technologies for commercial applications. MP3 format audio with good sound quality, [1–3] small size and get a variety of portable player support, many advantages make it the most popular audio format on the Internet. The current infringement of Internet music has occurred, therefore, the MP3 audio watermarking research for the protection of digital audio copyright is important [4–6].

At present, foreign research on MP3 audio watermarks includes: MP3Stego technology proposed by Cambridge researchers in the UK. It is to adjust the quantization error in MP3 encoding to adjust the encoding length of the current audio frame and realize the embedding of data. The shortcoming of the algorithm is lacking robustness in attacking by decompression and then recompression. In the literature [7], an algorithm is proposed to embed watermark information using the linbits of MP3 bit stream. The algorithm is relatively easy to apply in real-time application, but because the linbits of different types of songs are different, the difference between embedding information amount of the files is very big. In the literature [8], an algorithm is proposed to embed the watermark by adjusting the energy difference. By repeating the iteration of an adjustment factor, most of the noise caused by embedding watermark is under the psychological masking curve. The disadvantage is that the algorithm is too complicated. In the literature [9], the watermark is embedded by modifying the first five

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_44

coefficients of the modified discrete cosine transform (MDCT) in each coding unit. The MP3 bit stream are modified in most of these algorithms which are too dependent on the characteristics of MP3 bit stream, watermarks often can't afford to MP3 decompression/compression attacks. Aiming at this problem, we propose an MP3 audio watermarking algorithm based on wavelet transform, which mainly solves the following two problems: first, to satisfy the good auditory transparency, make the watermark robustness for MP3 decompression/compression attack; second, the algorithm is designed to be as simple as possible to meet real-time applications.

In the transform domain algorithm of digital audio watermarking, it is transformed completely from the time domain to the frequency domain in discrete Fourier transform (DFT) and discrete cosine transform (DCT), but the time domain characteristics can't be preserved. The discrete wavelet transform (DWT) has a good time-frequency characteristics, and the ability of a local observation in the time domain and frequency domain. [10] The multi-resolution analysis capability of DWT meets the requirements of the human auditory system and is suitable for audio signal processing, so it is necessary to design an audio digital watermarking system with DWT. It is because of good time-frequency correlation characteristics, as well as multi-resolution analysis of the idea, making discrete wavelet transform into a powerful tool for analyzing and processing audio signals. The discrete wavelet transform is applied to the MP3 audio digital watermarking technology, which makes the MP3 watermark based on wavelet transform has very important application significance.

2 MP3 Audio Watermarking Based on Discrete Wavelet Transform

The watermarking algorithm mainly includes two processes: watermark embedding and watermark extraction, the flow chart is as follows (Fig. 1):

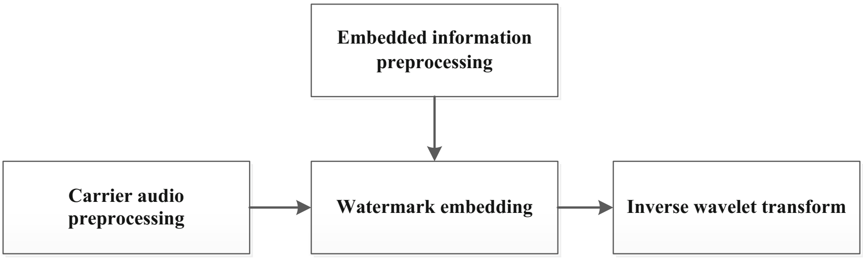

Embedded information preprocessing
Carrier audio preprocessing
Watermark embedding
Inverse wavelet transform

Fig. 1. The flow chart of watermark embedding

2.1 Watermark Embedding

Firstly, the MP3 audio is decoded to get the PCM signal in the algorithm [11], then the PCM signal is segmented, and then the wavelet decomposition is carried on each section to get the low frequency coefficients, finally the watermark embedding to the

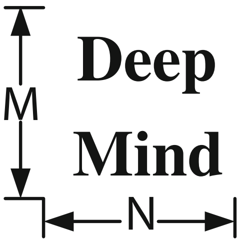


Fig. 2. Embedded information

low frequency coefficient is carried on. Watermark embedding is divided into four steps: embedded information preprocessing, carrier audio preprocessing, watermark embedding and wavelet inverse transformation:

(1) Embedded information preprocessing

Let W be the watermark information, which is a binary image of size $M \times N$, as shown in Fig. 2, the watermark information can be expressed as follow:

$$W = \{w(i,j), 0 \leq i < M, 0 \leq j < N, w(i,j) \in \{0,1\}\}$$

Where $w(i,j) \in (0,1)$ is the pixel value of the binary image, and reducing the dimension of 2D watermarking image W, let V be a one-dimensional sequence:

$$V = \{v(k) = w(i,j), 0 \leq i < M, 0 \leq j < N, k = M \times i + j\} \tag{1}$$

V is encrypted with a small m sequence of length $M \times N$ to generate sequence V':

$$V' = \{v'(k) = v(k) \otimes m(k)\} \tag{2}$$

($\otimes$ represents XOR)

(2) Carrier audio preprocessing

Set L for a piece of MP3 audio, first a mp3 decoding to get PCM code, and the PCM code is divided into N equal segments, let segment n is X_n, and then 3-level wavelet decomposition is done for each small segment X_n [12, 13]. As shown in Fig. 3:

Where L represents the low frequency part, H represents the high frequency part, the end of the serial number represents the number of wavelet decomposition layers. 3-level wavelet decomposition formula is shown as follow:

$$X = H_1 + LH_2 + LLH_3 + LLL_3 \tag{3}$$

(3) Unipolar Quantized Embedded Watermarking

The first M maximum positive values of the high-frequency coefficients LLL_3 which are decomposed by the 3-level wavelet, and the unipolar quantization is carried out to realize watermark embedding [14]. The quantization steps are as follows:

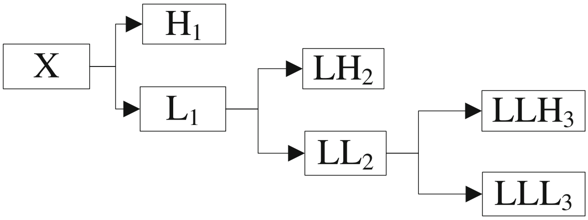


Fig. 3. 3-level wavelet decomposition

First, determine the meaning of the coordinates, the horizontal axis represents the sequence LLL_3; the vertical axis has two meanings that, when used in numerical calculations, are used to measure the magnitude of LLL_3, when used to represent watermark information, the vertical axis is divided by the quantization step α by the A interval set and the B interval set. Regardless of the size of LLL_3, it represents the watermark information “0” when it belongs to the A interval set, it represents the watermark information “1” when it belongs to the B interval set. (We specify the quantization interval end point as follows: it belongs to the A interval set when its value is an even number of quantization steps, it belongs to the B interval set when its value is an odd number of quantization steps). It is shown as Fig. 4, LLL_3 represents the sequence $(0, 1, 1, 1, 0, 1)$.

Secondly, the first operation of the LLL_3 to do $\lfloor \underline{L}LL_3 div\,\underline{\alpha} \rfloor$ operations. Let the business is β and the remainder is γ, i.e., $LLL_3 = \alpha\beta + \gamma(0 \leq \gamma < \alpha)$. For β modulo 2 operation, the modulo value is φ, i.e., $\varphi = \beta \bmod 2$ $(\varphi \in (0, 1))$. According to the value of φ to determine the interval that LLL_3 belongs to, when $\varphi = 0$, LLL_3 belongs to A interval set; when $\varphi = 1$, LLL_3 belongs to B interval set.

When embedded, the values of the watermark information w and φ are used to decide whether to modify LLL_3 or not, if $w = \varphi$, then do not make changes, or else:

$$LLL_3' = \begin{cases} LLL_3 - \frac{\alpha}{2}, if 0 \leq \gamma < \frac{\alpha}{2} \\ LLL_3 + \frac{\alpha}{2}, if \frac{\alpha}{2} \leq \lambda < \alpha \end{cases} \tag{4}$$

Where $\alpha/2$ is the embedding strength, and when the embedding intensity is large, the influence on the carrier audio signal is large, but at this time the watermarking system is more robust; when the embedding intensity is small, the embedded watermark is weak and cannot be perceived by the human ear, but poor resistance to attack.

Suppose that the watermark sequence to be embedded is $w = \{0, 0, 1, 0, 1, 1\}$, as shown in Fig. 4 $LLL_3 = \{0, 1, 1, 1, 0, 1\}$. Then apply the above rules to make some changes to LLL_3, set the modified sequence LLL_3', as shown in Fig. 5, after modification LLL_3' is $\{0, 0, 1, 0, 1, 1\}$. Where • represents the original coefficient LLL_3, ° represents its modified coefficient LLL_3'.

(4) Wavelet inverse transformation

A new wavelet coefficient is combined by the modified low frequency coefficient and the original high frequency coefficients:

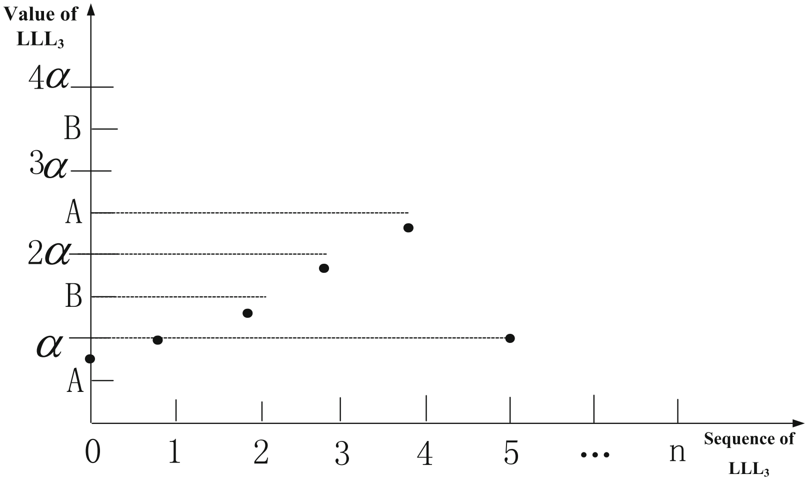


Fig. 4. Single polar quantization

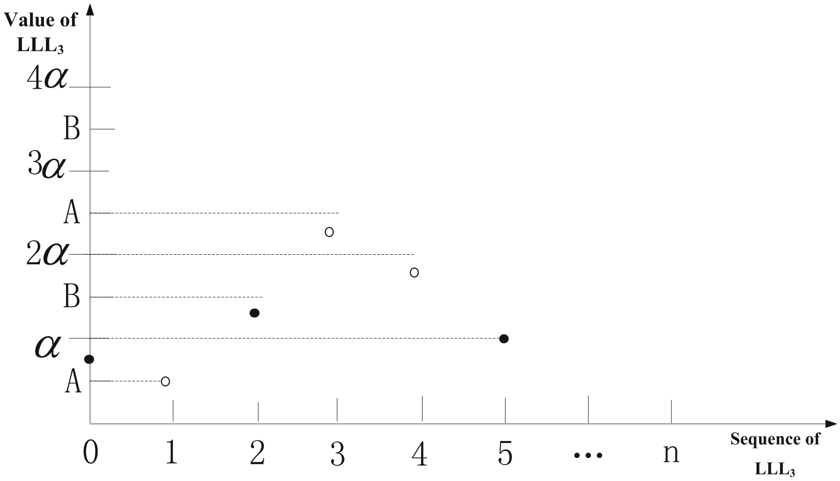


Fig. 5. Modified LLL_3'

$$X' = H_1 + LH_2 + LLH_3 + LLL_3' \tag{5}$$

And then X' is transformed by inverse wavelet transform to get the PCM code containing watermark information, at last MP3 is encoded using the PCM to get the watermarked MP3.

2.2 Watermark Extraction

In the practical, the watermark extraction is more complicated than the watermark embedding in the algorithm complexity and time. This algorithm belongs to the watermark blind extraction, and the extraction process does not need the original audio. Watermark extraction is simple and quick, only three steps, the extraction time is relatively fast, the process is as follows:

(1) The MP3 to be detected is decoded to get PCM code, and then it is divided into n equal segment;

(2) 3-level discrete wavelet transforms are performed for each segment, taking the first M maximum positive values of the low frequency coefficients LLL_3'. According to the extraction formula, the watermark information is extracted and stored in the one-dimensional sequence v'':

$$v'' = \lfloor LLL_3' div \alpha \rfloor \bmod 2 \tag{6}$$

(3) The one-dimensional sequence v'' is decrypted and lifted to a two-dimensional image W':

$$W' = \{w'(i,j) = v''(k) \otimes m(k), k = i \times N + j, 0 \le i < M, 0 \le j < N\} \tag{7}$$

3 Simulation

In order to verify the correctness, feasibility and efficiency of the algorithm, we tested the auditory transparency, robustness and algorithm performance of audio digital watermark. The computer used in the experiment was configured as Corei3-3220 CPU (3.30 GHz), 4 GB RAM, and the software used is Visual Studio 2008 and Matlab2013a. The test subjects were used for mono audio MP3 files with a length of 30 s, a sampling frequency of 44.1 kHz and a bit rate of 128 kbps, and the test object included six different types of music clips, including classical, country, dance, jazz, Pop and rock, using a 64 × 64 binary image as a watermark.

3.1 Auditory Transparency

In order to evaluate the effect of the watermark on the audio effect of the audio file, the perceptual evaluation criteria are adopted, including the subjective quality evaluation method and the objective quality evaluation method.

First use the SDG for subjective quality evaluation, and the original audio and embedded watermarked audio files are played back to 20 auditor tests without telling the listener any watermark embedding information. Experiments show that the vast majority of testers cannot tell the difference between the original audio and the embedded watermarked audio. Let the testers mark the audio files according the SDG standard, and then the average score of the same type is taken as the final SDG score of

Table 1. SDG subjective quality test results

Music type	Classical	Country	Dance	Jazz	Pop	Rock
Value of SDG	–0.32	–1.2	–0.18	–0.68	–0.64	–0.52

the type audio. Subjective test results are listed in Table 1, in the list, the test results are close to 0, indicating that the original audio files and watermarked audio files in the auditory effect is almost the same.

Audio quality evaluation criterion BS.1387 recommended by ITU-R is usually used for audio encoder quality evaluation, but also as a good objective quality evaluation criterion for audio watermarking technology. The variable outputted by model is combined with the neural network to give a magnitude as the objective discrimination of the auditory mass ODG (Objective Difference Grade). The test software Opera TM is used to compare the quality of the file (Table 2).

Table 2. The objective quality test results

Music type	Classical	Country	Dance	Jazz	Pop	Rock
Value of ODG	–1.2	–0.7	–1.1	–0.8	–0.7	–0.7

The test results are close to –1, indicating that the original audio files and watermarked audio files in the auditory effect is basically the same.

Both the subjective quality test results and the objective quality test results show that the watermark embedding algorithm used in this paper has good auditory transparency.

3.2 Robustness

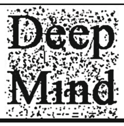
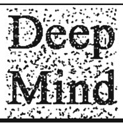
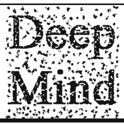
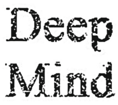
Robustness is a very important indicator of watermarking (exception to fragile watermarking). The embedded watermark must have a certain amount of energy to ensure this feature. And because of the contradiction between the non-perceivability and the robustness of the watermark, the robustness is limited to improve. On the other hand, in the case of a certain perceived quality, the watermark robustness is also related to the amount of embedded information, the embedding strength, the size and characteristics of the digital media. In order to test the robustness of embedded watermarks, several attacks on watermarked audio signals is performed, including low-pass filtering, resampling, white noise, and clipping.

(1) Add noise. A white noise is superimposed on the watermarked audio signal with an average of 0 and a variance of 0.01.

(2) Wave filtering. Butterworth low-pass filter with the order of 6, the cut-off frequency of 10 kHz is selected to filter the audio signal.

(3) Resampling. First up-sampling is performed on the audio signal; the sampling frequency is increased to 88.2 kHz from the origin, and then the extraction technology is used to restore the original sampling frequency to 44.1 kHz.

Table 3. Robustness test results

Attack type	No attack	Noise adding	Filtering	Resampling	Cutting
Watermark image					

(4) Clipping. After removing one-tenth of the sample points in the audio signal, the watermark is then extracted.

As shown in Table 3, the algorithm is effective and has good robustness to noise reduction, filtering and resampling attacks, but ability of anti-cutting attacking is slightly weaker. Since the audio is divided equally and then embed the watermarking in the algorithm, so the watermark information in the part that is cut will be lost, but the embedding will divide the clipped portion into the entire watermark image, thereby reducing the effect of cutting on the watermark to a certain extent.

4 Conclusion

This paper presents an MP3 audio watermarking algorithm based on wavelet unipolar quantization. Based on the characteristics of MP3 in the coding process, the algorithm combines the advantages of wavelet transform and the advantages of unipolar quantization. In the algorithm, the MP3 is decoded, embedded watermark, and then encoded. The watermark can withstand noise, low-pass filter for common signal processing and attack, and has a strong robustness for heavy sampling. Because this algorithm belongs to the watermark blind extraction, the extraction does not need the original audio, it makes the extraction convenient and quick, and so it has a wide range of application prospects. At the end of this paper, the experiment is carried out. The experimental data show that the algorithm achieves the requirement of imperceptibility and robustness, and ensures the practicality of the scheme.

References

1. Aissa, F., Malki, M., Elçi, A.: A similarity measure across ontologies for web services discovery. Int. J. Inf. Technol. Web. Eng. **11**(1), 22–43 (2016)
2. Al-Kabi, M., Wahsheh, H., Alsmadi, I.M.: Polarity classification of arabic sentiments. Int. J. Inf. Technol. Web. Eng. **11**(3), 32–49 (2016)

3. Wang, Y., Ma, J., Lu, X., Lu, D., Zhang, L.: Efficiency optimisation signature scheme for time-critical multicast data origin authentication. Int. J. Grid Util. Comput. **7**(1), 1–11 (2016). http://dx.doi.org/10.1504/IJGUC.2016.073771
4. Ramyachitra, D., Pradeep Kumar, P.: Frog leap algorithm for homology modelling in grid environment. Int. J. Grid Util. Comput. **7**(1), 29–40 (2016). doi:10.1504/IJGUC.2016.073775
5. Al-Jumeily, D., Hussain, A., Fergus, P.: Using adaptive neural networks to provide self-healing autonomic software. Int. J. Space-Based Situated Comput. **5**(3), 129–140 (2015). doi:10.1504/IJSSC.2015.070953
6. Rodas, O., To, M.A.: A study on network security monitoring for the hybrid classification-based intrusion prevention systems. Int. J. Space-Based Situated Comput. **5** (2), 115–125 (2015). doi:10.1504/IJSSC.2015.069240
7. Kim, D.-H., Yang, S.-J., Chung, J.-H.: Additive data inserting into MP3 bit stream using linbits characteristics. In: Proceedings on ICASSP 2004, vol. 4, pp. 181–184 (2004)
8. Qiao, L.T., Nahrstedty, N.: Non-invertible watermarking methods for MPEG encoded audio. In: SPIE Proceedings on Security and Watermarking of Multimedia Contents, vol. 3675, pp. 194–202. University of Illinois at Urbana-Champaign, Singapore (1999)
9. Wang, C.T.: A new audio watermarking based on modified discrete cosine transform of mpeg/audio layer III. In: Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control, vol. 2, pp. 984–989 (2004)
10. Liu, F.: Research of Audio Digital Watermarking Based on Wavelet Transform. Beijing University of Chemical Technology, Beijing (2010)
11. Zhou, J.W.: Research of Audio Digital Watermarking Based on Wavelet Transform. Wuhan University of Technology, Wuhan (2007)
12. Luan, M.M.: Study on Digital Watermarking for MP3 Audio. Southwest Jiaotong University, Chengdu (2010)
13. Niu, X.X., Yang, Y.X.: A New algorithm for digital watermarking based on the wavelet transform. Chin. J. Comput. **23**(1), 21–27 (2000)
14. Yuan, Z.L., Wen, Q.Y., Niu, X.X., et al.: Voice hiding algorithm based on quantization coding. J. China Inst. Commun. **23**(5), 108–112 (2002)

Multi-documents Summarization Based on the TextRank and Its Application in Argumentation System

Caiquan Xiong(✉), Yuan Li, and Ke Lv

School of Computer Science, Hubei University of Technology,
Wuhan 430068, China
x_cquan@163.com

Abstract. In the group argumentation environment, a large amount of text information will be produced. How to find the specific speeches of experts from many similar speeches and extract their common summary is of great significance to improve the efficiency of experts' argumentation and promote consensus. In this paper, the heuristic method is first used to cluster the speech texts and find the similar speech sets. Then, we use TextRank algorithm to extract multiple document summary, and feedback the summary to the experts. The experimental results show that the efficiency of the experts' argumentation is improved and the decision-making is promoted.

1 Introduction

Argumentation is a process of activating group thinking, resolving conflicts and seeking common understanding [1]. In the course of the argumentation, how to quickly and efficiently make the final decision is extremely important.

In terms of argumentation text analysis, Bai B etc. [2] proposed a hot-spot extraction method based on topic clustering in the argumentation system. In their paper, they proposed a hot-spot extraction formula to extract hot topics and popular viewpoints to assist experts, which can promote the rapid generation of the final decision. Xiong CQ and Li DH [3] proposed a heuristic clustering algorithm for the consistency analysis of expert opinions, and used the visualization technology to display the clustering results, which makes the expert's thinking converge quickly. Wang A and Li YD [4] proposed a method for generating the summary in the argumentation, by using the probability mixture model to extract the experts' topic set and generating a summary according to the topic evolution, and thus the interaction between experts can be promoted and the method is useful to assist and summarize the decisions made in the conference. In the process of open-ended team argumentation, there will be a large number of electronic argumentation information, and Li XM and Li J [5] proposed a method which can quickly dig out and identify large-scale argumentation topics and then visualize them to members, thus stimulating the innovative thinking of members. Zhang PZ [6] proposed a method of topic automatic clustering analysis based on SOM (Self-organizing map, SOM) in the argumentation system, which mainly to solve the problem of information overload. However, the method proposed in Ref. [6] is only the clustering analysis of the

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_45

experts' speech information, which does not continue to make summarization, such as abstract extraction, from these speeches based on clustering analysis.

The other works focused on automatic summarization technology. Lin CY etc. [7] developed a multi-document summarization system called NeATS, which attempts to extract relevant or interesting portions from a set of documents about some topic and present them in coherent order. Hearst MA [8] proposed an technique for subdividing texts into multi-paragraph units that represent passages or subtopics. The algorithm is fully implemented and is shown to produce segmentation that corresponds well to human judgments of the subtopic boundaries. Multi-paragraph subtopic segmentation should be useful for many text analysis tasks, including information retrieval and summarization. Yan [9] proposed an SRRank algorithm and two extensions to make better use of the semantic role information. In a summary based on Abstractive Summarization, Alexander M. Rush [10] presented a neural attention-based model for abstractive summarization through recent developments in neural machine translation. However, the most existing ranking algorithms have effectiveness, but cannot be used in the argumentation system.

This paper proposes a new text analysis method for argumentation support system. The method first uses clustering algorithm to cluster the experts' speeches and get different text sets, then extracts the summaries from the text sets. Finally, a tool based on d3.js is used to visualize the results of text analysis. The experimental results show that this method can improve the efficiency of argumentation and promote the decision-making.

2 Multi-document Summarization Method Based on TextRank Algorithm

In the process of argumentation, experts will produce a lot of speech text information that have diverse and complex structure. But the computer cannot directly utilize the data. During the text preprocessing, the two essential parts are the operation of text segmentation and the removal of stop words. It refers to cutting a sentence into individual words, while filtering out some words of the high frequency with no practical meaning. The process of multi-document Summarization are shown in Fig. 1.

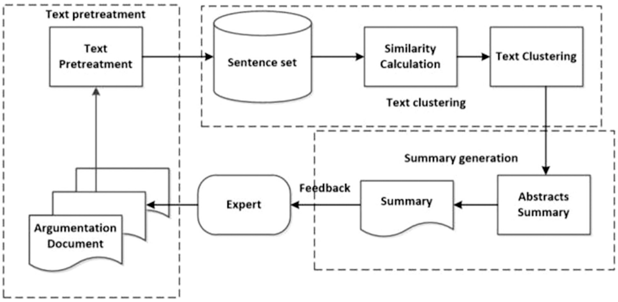


Fig. 1. The flow chart of multi-document summarization

2.1 Text Preprocessing

For the program-proposed stage that a large number of texts are generated, and the information in these tests is of structural diversity and complexity. We can make a preprocessing of the text, which is the operation of word segmentation, removal of words and so on, and the text information can be transformed into information which can be recognized and processed by the computer. At the same time, we can extract the characteristic words which are keywords, which can achieve the function of reducing dimension and improving the efficiency of calculation (Fig. 2).

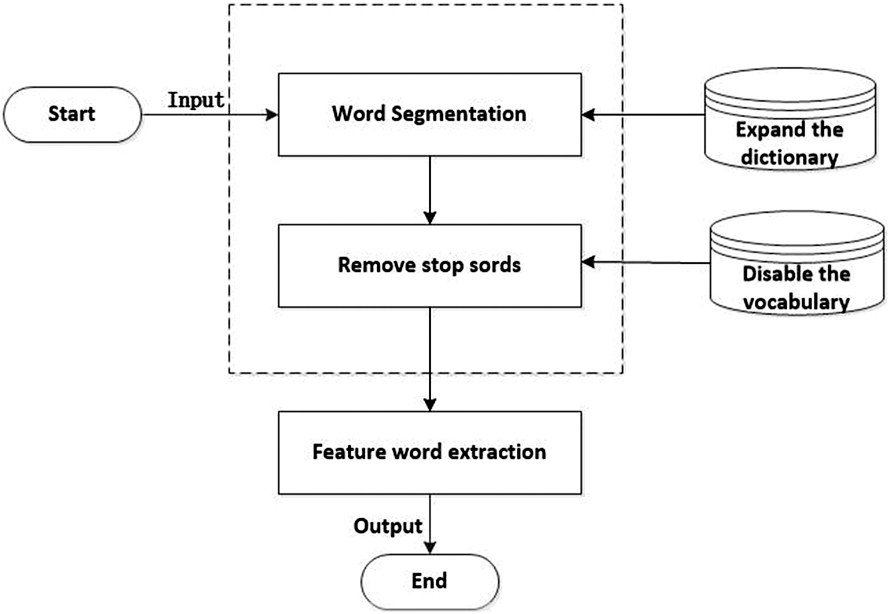


Fig. 2. Text preprocessing

2.2 Text Preprocessing

2.2.1 Text Feature Weighting

TF-IDF [11] is a classical statistic-based feature extraction algorithm. This algorithm takes the ratio of the number of occurrences of a feature word in a document to the number of documents containing that word as the weight of the word. The two main concepts are as follows.

(1) Term Frequency. Term frequency refers to the ratio of the number of occurrences of a word to the total number of words in a document. The calculation of term frequency feature need to remove the words whose frequency are smaller than a certain value, so as to reduce the dimension of feature space.

(2) Inverse Document Frequency. The more times a word appears in documents, the lower importance the word to a specific document should be.

In this paper, we select the characteristic words in the text as features. For a word w_j in a certain argumentation text d_i, its value *tf* can be expressed as: $tf_{ij} = \frac{w_{i,j}}{\sum_k w_{k,j}}$, $w_{i,j}$ is the word frequency in the text, $\sum_k w_{k,j}$ is the sum of all the words in the text, *idf* is computed as $idf_i = \log \frac{|D|}{\left|\left\{j: w_i \in d_j\right\} + 1\right|}$, where $|D|$ represents the number of texts, $\left|\left\{j : w_i \in d_j\right\} + 1\right|$ refers to the number of texts containing word w_j, and in order to ensure that the denominator is not zero, 1 is added in the denominator, then the *TF* −*IDF* value of the document can be obtained as: $TF - IDF = tf_{ij} \times idf_i$, which is used to calculate the similarity between the calculated heuristic clustering text.

2.2.2 Vector Space Model (VSM)

Vector space model (VSM) [12] is used to transform the content of the text to a vector in the vector space and to represent the semantic similarity as spatial similarity. The basic idea of vector space is to represent the text information in a vector, and to represent the similarity between words in the text as the similarity between vectors, and to calculate the similarity between texts by the distance of vectors when the document has been vectorized.

Using D to represent the document, T to represent the feature, the feature set can be expressed as $D(t_1, t_2, \ldots, t_n)$, and also can be expressed as $D=((t_1, w_1), (t_2, w_2), \ldots\ldots, (t_n, w_n))$, where t_i represents the i-th feature, w_i represents the weight of the i-th feature, it can be shown in Table 1:

Table 1. Textual representation in vector space

	t_1	t_2	…	t_n
D_1	w_{11}	w_{12}		w_{1n}
D_2	w_{21}	w_{22}		w_{2n}
…	…	…	…	…
D_n	w_{n1}	w_{n2}		w_{nn}

2.2.3 Text Similarity Calculation

Text similarity calculation is mainly used to compare the degree of similarity between the text, and in the calculation of text similarity, cosine computation is most commonly used. And vector space model is used to map each object to a feature vector, i.e., $D(w_1, w_2, \ldots, w_n)$, then a space vector angle will form between any two space vectors, the cosine of the angle is the similarity measure of these two texts. Therefore, the similarity calculation model of text i and text j is expressed as:

$$sim\left(d_i, d_j\right) = \frac{\sum_{k=1}^{n} w_k(d_i) \times w_k\left(d_j\right)}{\sqrt{\left(\sum_{k=1}^{n} w_k^2(d_i)\right)\left(\sum_{k=1}^{n} w_k^2\left(d_j\right)\right)}}. \tag{1}$$

where n is the length of text vector and k satisfies $1 \leq k \leq n$, and $sim\left(d_i, d_j\right)$ is the similarity between text d_i and d_j, denoted by, where $0 \leq sim^{i,j} \leq 1$. $\sum_{k=1}^{n} w_k(d_i) \times$

$w_k(d_j)$ is the vector inner product of d_i and d_j, denominator $\sqrt{\left(\sum_{k=1}^{n} w_k^2(d_i)\right)}$ and $\sqrt{\left(\sum_{k=1}^{n} w_k^2(d_j)\right)}$ are the length of the vector d_i and d_j respectively.

2.2.4 Expert Opinion in Text Clustering

In this paper, the clustering algorithm in Ref. [3] is used to perform text clustering. First, the expert set $E=\{e_1, e_2, \ldots, e_m\}(m \geq 2)$ is introduced, and a similarity threshold f is defined at the same time. And then the expert set is traversed over each expert. The expert e_1 is put in the first cluster C_1 and removed in the expert set E, and the first element e_i from E is continued to be compared with the remaining clusters. If there is a cluster C_t, each element of its similarity with e_i are greater than or equal to δ, and this similarity is the biggest, then e_i will be incorporated into C_t and removed from E at the same time, otherwise e_i will be put in a new cluster and removed from E until E is empty, the generated clusters will be output when the clustering ends.

2.3 Multiple Document Summary

2.3.1 Text Summary Algorithm

With the development of the Internet, the automatic summarization technology, first proposed in 1985 [13], has made great progress recent years. The commonly used text summarization techniques can be mainly divided into two categories: automatic summary based on extraction as well as based on abstraction. The abstraction-based automatic document summarization requires that the computer can first understand the meaning of the expressions in the document, and then express concisely via human language. Such as Dichotomy [14], Hidden Markov Model [15, 16] and Bayesian Analysis [17]. Due to the limitation of current natural language processing technology, this method is not effective, so this paper adopts the method of extraction-based summarization.

Graph-based sorting algorithm, which is used in our work, is mainly inspired by Google's PageRank [18] and Kleiberg's HITS algorithm [19]. Base on that, the LexRank [20] algorithm proposed later use sentences to construct the nodes of the graph, and take the similarity between sentences as the weight of graph edge. When the similarity is greater than a threshold value, the two nodes are connected otherwise they are not connected, thereby the original weighted graph is converted to an unweighted graph. On this basis, Mihalcea proposed TextRank algorithm [21], where the edges are weighted by the cosine similarity, and the weighted graphs are generated to create text summaries according to the order of the sentences.

In a cluster, texts are around a single topic, and the relationships between the texts are very close. The more times some sentences about the topic appear, the more likely they are the summary sentences of the text cluster. On the basis of the above idea, we use graph-based sorting algorithm to select summary sentences to constitute the summary of the text cluster. The basic principle of the graph-based sorting is voting. In the graph, if there is a node A pointing to another node B, then B is given a "vote", that is, the importance of a node is mainly dependent on all nodes pointing to it, a node pointing to another node is equivalent to that the pointed node obtains a "vote", the greater the number of votes are, the greater the importance of this node is. In other words, the importance of a node depends not only on the number of votes it get but also on the importance of the pointing node.

To extract the summary in the text we use the TextRank algorithm. The general model of TextRank can be expressed as a directed and weighted graph $G = (V, E)$, which consists of node set V and edge set E, and E is a subset of $V \times V$. The weight of the edge between any two nodes v_i, v_j in the graph is w_{ij}. For a given node v_i, $\ln(v_i)$ is a set of nodes pointing to this node, and $out(v_j)$ is the set of nodes to which v_i points. The score for node v_i is defined as follows:

$$WS(v_i) = (1-d) + d * \sum_{v_j \in \ln(v_i)} \frac{w_{ij}}{\sum\limits_{v_k \in out(v_j)} w_{jk}} WS(v_j). \tag{2}$$

where d is the damping coefficient, generally defined as 0.85, w_{ij} is the similarity between sentences, $WS(v_i)$ is the weight of j for the last iteration, $\ln(v_i)$ denotes a set of sentences pointing to sentence v_i, $out(v_j)$ denotes a set of sentences pointed by sentence v_j. Using TextRank to extract the text summary, we need to pay attention to two important aspects. The first is that the initial setting of the sentence weight is generally defined as 1. The second is that the essence of this algorithm is an iterative algorithm, thus we need to set the threshold of iteration, and the general convergence threshold is set to 0.001, if the weight of vertex reaches the threshold we stop the iteration.

2.3.2 Similarity Calculation Between Sentences

To extract the summary of a document, we first need to calculate the similarity between the sentences in the document after text preprocessing. Here we use the BM25 algorithm, BM25 (Best Match25) is usually used for searching relevance score:

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)}. \tag{3}$$

k denotes a modulating factor, k is generally fixed at 1.5, b denotes another factor with a value of 0.75, $|D|$ denotes the length of the sentence, $f(q_i, D)$ denotes the frequency of the word q_i in the sentence $D, avgdl$ denotes the average length of the sentence in the document. The formula for $IDF(S_i)$ is:

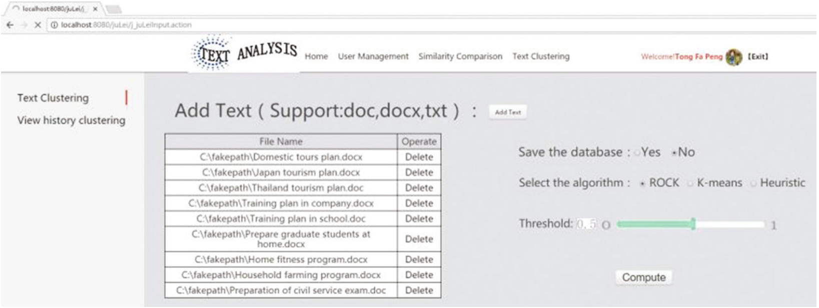
$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}. \tag{4}$$

where N is the total number of sentences in the cluster and $n(q_i)$ is the total number of sentences containing all the words q_i in the cluster. 0.5 represents the smoothing factor. According to the definition of IDF, we can know that for a given set of documents, the more documents that contain q_i, the lower the weight of q_i will be, otherwise the less, the higher. Specifically, in a document set, if a number of documents contains q_i, the distinguishing ability of q_i will be very low, and the importance of using q_i will be very low, too.

3 Application Effect Analysis

In the operation of the system, we randomly selected nine undergraduate students and took “Summer Planning of 2017” as the discussion topic to start a discussion. The nine students were selected from different grades for the purpose of non-interference. Each student wrote a plan about the summer, we collected these plans, arranged them as nine text files and named this files according to their content. And then these plans are divided into four main categories, namely, tourism, staying at home, practicing in the company and preparing civil servant examinations. Different clusters were obtained by adjusting the thresholds, and the summaries of different clusters were extracted, and then compared with the manually made summaries.

The nine documents in Fig. 3 are selected for our clustering experiment. Through text analysis and processing by the system, we can get the following results. Figure 4 shows that when the threshold f is 0.01, all the experts’ speeches are clustered together to form only one text set while extracting a text summary for it. Figure 5 shows that when the threshold f is 0.05, the effect of experts’ speeches clustering is consistent with the expected results while extracting a text summary for the “home plan text” set. It can



Add Text (Support:doc,docx,txt) :

File Name	Operate
C:\fakepath\Domestic tours plan.docx	Delete
C:\fakepath\Japan tourism plan.docx	Delete
C:\fakepath\Thailand tourism plan.doc	Delete
C:\fakepath\Training plan in company.docx	Delete
C:\fakepath\Training plan in school.doc	Delete
C:\fakepath\Prepare graduate students at home.docx	Delete
C:\fakepath\Home fitness program.docx	Delete
C:\fakepath\Household farming program.docx	Delete
C:\fakepath\Preparation of civil service exam.doc	Delete

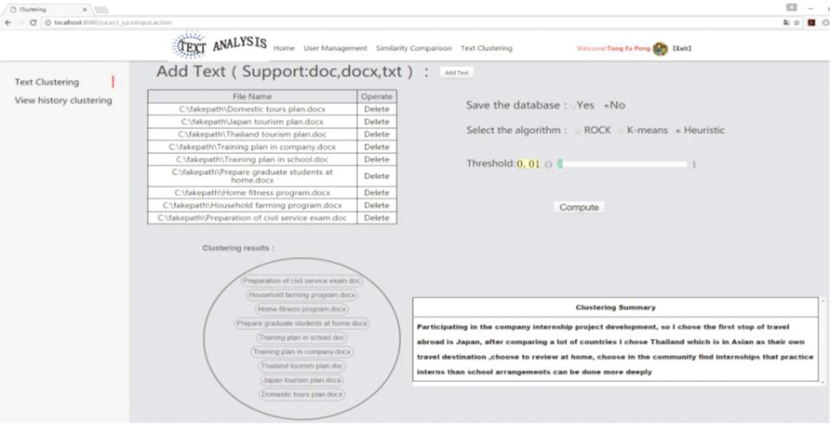
Save the database : ○Yes ●No

Select the algorithm : ● ROCK ○ K-means ○ Heuristic

Threshold: 0. 5 0 1

Compute

Fig. 3. Upload text



Add Text (Support:doc,docx,txt) :

File Name	Operate
C:\fakepath\Domestic tours plan.docx	Delete
C:\fakepath\Japan tourism plan.docx	Delete
C:\fakepath\Thailand tourism plan.doc	Delete
C:\fakepath\Training plan in company.docx	Delete
C:\fakepath\Training plan in school.doc	Delete
C:\fakepath\Prepare graduate students at home.docx	Delete
C:\fakepath\Home fitness program.docx	Delete
C:\fakepath\Household farming program.docx	Delete
C:\fakepath\Preparation of civil service exam.doc	Delete

Save the database : ○Yes ●No

Select the algorithm : ○ ROCK ○ K-means ● Heuristic

Threshold: 0. 01 0 1

Compute

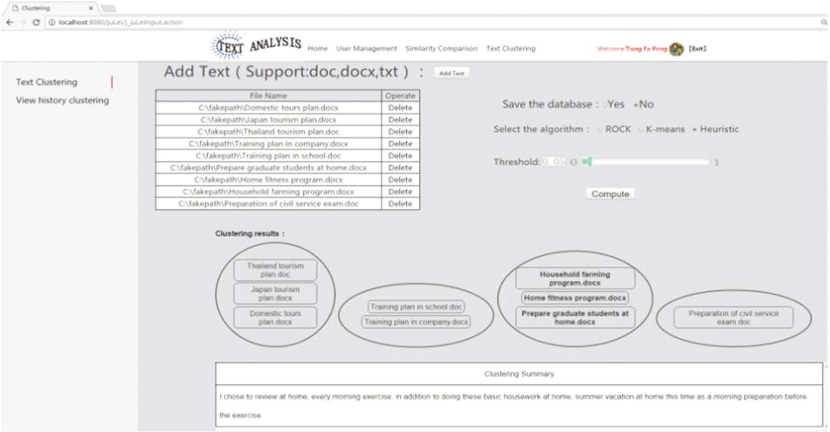
Clustering results :

Preparation of civil service exam.doc
Household farming program.docx
Home fitness program.docx
Prepare graduate students at home.docx
Training plan in school.doc
Training plan in company.docx
Thailand tourism plan.doc
Japan tourism plan.docx
Domestic tours plan.docx

Clustering Summary

Participating in the company internship project development, so I chose the first stop of travel abroad is Japan, after comparing a lot of countries I chose Thailand which is in Asian as their own travel destination ,choose to review at home, choose in the community find internships that practice interns than school arrangements can be done more deeply

Fig. 4. Clustering results at $f = 0.01$ and Clustering summary


Clustering
localhost:8080/juLei/j_juLeiInput.action
TEXT ANALYSIS
Home
User Management
Similarity Comparison
Text Clustering
Welcome!Tong Fa Peng
[Exit]
Text Clustering
View history clustering
Add Text
Compute
Clustering results :
Thailand tourism plan.doc
Japan tourism plan.docx
Domestic tours plan.docx
Training plan in school.doc
Training plan in company.docx
Household farming program.docx
Home fitness program.docx
Prepare graduate students at home.docx
Preparation of civil service exam.doc

Add Text (Support:doc,docx,txt) :

File Name	Operate
C:\fakepath\Domestic tours plan.docx	Delete
C:\fakepath\Japan tourism plan.docx	Delete
C:\fakepath\Thailand tourism plan.doc	Delete
C:\fakepath\Training plan in company.docx	Delete
C:\fakepath\Training plan in school.doc	Delete
C:\fakepath\Prepare graduate students at home.docx	Delete
C:\fakepath\Home fitness program.docx	Delete
C:\fakepath\Household farming program.docx	Delete
C:\fakepath\Preparation of civil service exam.doc	Delete

Save the database : ○ Yes ● No

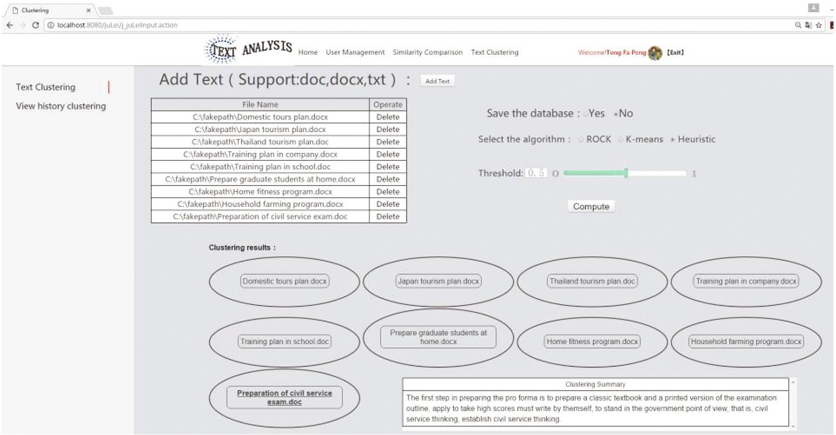
Select the algorithm : ○ ROCK ○ K-means ● Heuristic

Threshold: 0.05 0 1

Clustering Summary

I chose to review at home, every morning exercise, in addition to doing these basic housework at home, summer vacation at home this time as a morning preparation before the exercise

Fig. 5. Clustering results at $f = 0.05$ and Summary of the home plain text


Clustering
localhost:8080/juLei/j_juLeiInput.action
TEXT ANALYSIS
Home
User Management
Similarity Comparison
Text Clustering
Welcome!Tong Fa Peng
[Exit]
Text Clustering
View history clustering
Add Text
Compute
Clustering results :
Domestic tours plan.docx
Japan tourism plan.docx
Thailand tourism plan.doc
Training plan in company.docx
Training plan in school.doc
Prepare graduate students at home.docx
Home fitness program.docx
Household farming program.docx
Preparation of civil service exam.doc

Add Text (Support:doc,docx,txt) :

File Name	Operate
C:\fakepath\Domestic tours plan.docx	Delete
C:\fakepath\Japan tourism plan.docx	Delete
C:\fakepath\Thailand tourism plan.doc	Delete
C:\fakepath\Training plan in company.docx	Delete
C:\fakepath\Training plan in school.doc	Delete
C:\fakepath\Prepare graduate students at home.docx	Delete
C:\fakepath\Home fitness program.docx	Delete
C:\fakepath\Household farming program.docx	Delete
C:\fakepath\Preparation of civil service exam.doc	Delete

Save the database : ○ Yes ● No

Select the algorithm : ○ ROCK ○ K-means ● Heuristic

Threshold: 0.5 0 1

Clustering Summary

The first step in preparing the pro forma is to prepare a classic textbook and a printed version of the examination outline, apply to take high scores must write by themself, to stand in the government point of view, that is, civil service thinking, establish civil service thinking

Fig. 6. Clustering results at $f = 0.5$ and Summary of civil service

be found that if the actual clustering results are consistent with the expected results, then the resulting summaries are more accurate. Figure 6 shows that when the threshold f is 0.5, since the similarity between texts can not reach the range of the threshold, thus each can only be clustered into a cluster respectively. Just click the “civil service text” set to generate the text set summary.

The above analysis shows that the text analysis method proposed in this paper can be used to extract summaries of experts’ speeches, which can improve the efficiency of experts’ argumentation, and achieve optimal argumentation effect.

4 Conclusion

In this paper, a text analysis method for the argumentation environment is proposed. By clustering a large number of texts, the similar opinions from different experts are clustered together. Then the TextRank algorithm is used to process text summaries, and the results will feedback to the experts. The results show that the efficiency of experts' argumentation is improved, and the decision-making is promoted. However, our text analysis method is not enough in the accuracy of abstract extraction. Because of the graph-based summary generation method, the generated summary can only depend on the sentences in the text set and in the future further research will focus on the generation of summary according to the semantics of the document content.

Acknowledgements. This research is supported by National Natural Science Foundation of China under grant number 61075059, 61300127.

References

1. Xiong, C.Q., Li, D.H., Zhang, Y.: Clustering analysis of expert's opinion and its visualization in hall for workshop of meta-synthetic engineering. Pattern Recogn. Artif. Intell. 282–287 (2009)
2. Bai, B., Li, D.H., Xiong, C.Q.: Hot extraction based on topic cluster in discussion support system. Comput. Digit. Eng. **38**, 81–85 (2010)
3. Xiong, C.Q., Li, D.H.: Model of argumentation. J. Softw. **20**, 2181–2190 (2009)
4. Wang, A., Li, Y.D.: Probabilistic mixture model based summarization approach for CWME discussions. Comput. Sci. 191–194 (2011)
5. Li, X.M., Li, J., Zhang, P.Z.: Topic identification and visualization for open team innovation argumentation. J. Syst. Manag. 1–7 (2015)
6. Jiang, Y.Z., Zhang, P.Z., Zhang, X.X.: Research on intelligence visualization in group argument support system. J. Syst. Manag. **12**, 1–11 (2009)
7. Lin, C.Y., Hovy, E.H.: From single to multi-document summarization. In: Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002, pp. 457–464 (2002)
8. Hearst, M.A.: TextTiling: segmenting text into multi-paragraph subtopic passages. Comput. Linguist. **23**, 33–64 (1997)
9. Yan, S., Wan, X.: SRRank: leveraging semantic roles for extractive multi-document summarization. IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 2048–2058 (2014)
10. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. Comput. Sci. (2015)
11. Salton, G., Yu, C.T.: On the Construction of Effective Vocabularies for Information Retrieval. Birkhauser (1973)
12. Salton, G.: A vector space model for automatic indexing. Commun. ACM **18**, 613–620 (1975)
13. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**, 159–165 (1958)
14. Kupiec, J.: A trainable document summarizer. In: International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68–73 (1995)

15. Schlesinger, J.D., Okurowski, M.E., Conroy, J.M., O’Leary, D.P., Taylor, A., Hobbs, J., Wilson, H.T.: Understanding machine performance in the context of human performance for multi-document summarization (2002)
16. Conroy, J.M., O’leary, D.P.: Text summarization via hidden markov models. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 406–407. ACM (2001)
17. Aone, C., Okurowski, M.E., Gorlinsky, J.: Trainable, scalable summarization using robust NLP and machine learning. In: Proceedings of the 17th International Conference on Computational Linguistics, vol. 1, pp. 62–66. Association for Computational Linguistics (1998)
18. Page, L.: The PageRank citation ranking: bringing order to the web, vol. 9, pp. 1–14 (1998). http://www-db.stanford.edu/∽backrub/pageranksub.ps
19. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. J. ACM **46**, 604–632 (1999)
20. Erkan, R., Dragomir, R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Qiqihar Junior Teachers Coll. 22(2004) (2011)
21. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. Unit Scholarly Works, pp. 404–411 (2004)

An Unconstrained Face Detection Algorithm Based on Visual Saliency

Ying Tong[1], Rui Chen[1](✉), Liangbao Jiao[1], and Yu Yan[2]

[1] Department of Communication Engineering, Nanjing Institute of Technology,
Nanjing 211167, China
{tongying, chenrui, jiaoliangbao}@njit.edu.cn
[2] Jiangsu Province Traditional Chinese Medicine Hospital, Nanjing 210000,
China
yanyucan@126.com

Abstract. This paper a novel face detection method based on visual saliency mechanism to improve the accuracy of unconstrained face recognition. Log Gabor transformation is used to extract visual features, and obtain facial saliency map by using stable balance measurement method based on Graph-Based Visual Saliency. Then binary image is obtained by segmenting facial saliency map with maximum entropy threshold and the rectangle area is marked by setting the centroid of object region as the center. Face region is detected from the original image according to the rectangle area. Experimental results on LFW database show that our algorithm can effectively remove the background interference without losing any face information and quickly precisely detect the face region which is more conducive to the unconstrained face recognition.

1 Introduction

Recently, more researchers begin to study the problem of face detection in the unconstrained environment. For example, Zhu proposed a hybrid model based on tree structure [1]. All the face labels are modeled, and the facial topological changes caused by the viewpoint can be captured by a global mixture model. Hakan proposed a face and facial label detection method based on the hierarchical classification [2]. In his method, face candidate regions are determined by the root detector, then the label facial position can be found by the sub-detector. Yan [3] proposed a face detection method based on structure model with partial subtype. In this method, partial subtype can control the local facial deformation, and the part deformation can control global facial deformation. To eliminate the background interference and identify which person a face image belongs to precisely, researchers normally use AAM model [4] to get the key points of the face region, and then the facial features are extracted in the center of key points [5].

In the field of computer vision, visual attention mechanism is utilized to identify which areas of image are remarkable, and get the visual saliency map by gray value scale. In [6], Itti proposed a visual saliency detection algorithm based on feature fusion, which has landmark significance. The basic idea is to establish the color, direction and brightness channels which model the biological visual characteristics, and strengthen

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_46

the corresponding features in the image, and then combine the characteristics of each channel to form a visual saliency map. On the basis of Itti algorithm, Harle and etc. proposed a GBVS (Graph-Based visual saliency) detection algorithm [7], which adopted Markov transfer matrix to measure the saliency map with Itti's visual feature extraction method. Subsequently, Hou [8] proposed a saliency detection algorithm based on frequency domain analysis, and obtained the image residual spectrum as a significant feature by the Fourier statistical properties from the scene image. Goferman proposed a CASD (Context Aware Saliency Detection) model [9]. Considering the global contrast and spatial correlation, Cheng proposed a target detection method based on region contrasting [10]. With the rapid development of the machine learning in various research fields, some saliency detection algorithms based on neural networks or Deep Learning are also proposed [11].

Although these methods have done, how to improve the object detection algorithms based on visual saliency are mainly taken into account two factors: the choice of visual features and the partition of object area. In this paper, we propose a novel unconstrained face detection algorithm based on visual saliency which both considers the two factors, the schematic diagram is shown in Fig. 1. The rest of this paper is organized as follows: Sect. 2 presents a novel saliency detection algorithm based on Log Gabor features and GBVS. Section 3 introduces the segmentation algorithm based on maximum entropy. Section 4 introduces the face detection algorithm based on the object centroid. Section 5 reports the experimental simulation results and Sect. 6 is the conclusions.

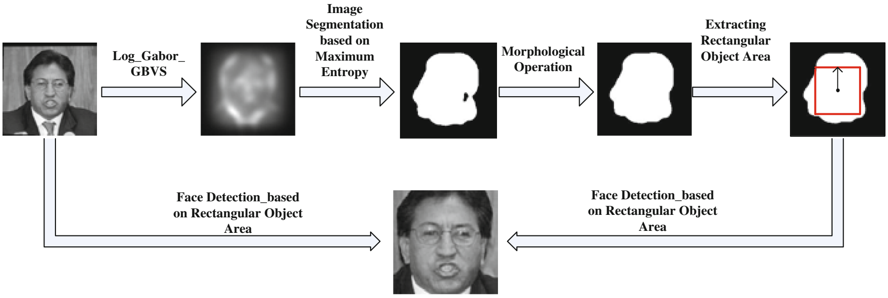


Fig. 1. The schematic diagram of the proposed algorithm

2 Saliency Detection Algorithm Based on Log_Gabor_GBVS

Due to adopting the stable balance state as the metric of channel saliency and globally automatically managing the saliency relationship between each pair of image pixels, the GBVS model has a better adaptability for the complex background, and the advantage of Log Gabor transformation, we combine Log Gabor transformation with GBVS algorithm to get more precise visual saliency map. The key merit of the proposed method is that Log Gabor transformation could extract more abundant low-level visual features and GBVS could more accurately obtain visual saliency map. The detailed steps are as follows.

(1) Extract multiscale and multi-direction visual feature maps with Log Gabor transformation, its multiscale maps are get by 1/2, 1/4, 1/6, 1/8 down-sampling original images, and its directional parameters are 0°, 45°, 90° and 135° These visual feature maps are the multi-channel low-level feature maps which size is $n \times n$.

(2) Calculate the weighted matrix of multi-channel feature maps. The directional weight $W((i,j),(p,q))$ from pixel (i,j) to pixel (p,q) is defined as:

$$W((i,j),\,(p,\,q)) = d((i,j)||(p,\,q)) \cdot F(i-p,j-q) \tag{1}$$

where $d((i,j)||(p,\,q))$ represents the similarity between the gray value $M(i,\,j)$ of pixel $(i,\,j)$ and the gray value $M(p,\,q)$ of pixel $(p,\,q)$, which can be calculated by formula (3); $F(i-p,j-q)$ is the distance weights between two pixel points, $F(a,\,b) = \exp\left(-\frac{a^2+b^2}{2\sigma^2}\right)$, When the distance between two points is lower, the distance weights $F(\cdot)$ is higher, the distance weight is lower on the contrary.

$$d((i,j)||(p,\,q)) = \left|\log\frac{M(i,\,j)}{M(p,\,q)}\right| \tag{2}$$

(3) The size of the distance weight matrix is $n^2 \times n^2$. Normalize it and obtain the Markov state transfer matrix.

(4) Markov transfer matrix is iterated until Markov chain reaches a stationary distribution. The stationary distribution of Markov chain represents the time-consuming from one pixel to the other pixel. The visual feature between two pixels is more similar, the distance weight is higher, and the transfer probability is greater. So the time-consuming between two pixels is shorter.

(5) Obtain the principle eigenvectors of stable Markov matrix. Rearrange these principle eigenvectors into two dimensional maps $(n \times n)$, which is the active maps, and normalize them.

(6) Add up the normalized active maps of each channel together, and get the final visual saliency map.

3 Segmentation Method Based on Maximum Entropy Criterion

The entropy represents the information of images and the entropy is proportional to the information. So we automatically get threshold by maximum entropy criterion and segment the visual saliency map. The straightforward steps are as follows:

(1) $f(x,y)$ is the visual saliency map which size is $M \times N$, and its gray-value level is $G = \{0,1,\ldots,L-1\}$;

(2) n_i represents the number of pixels with gray-value i in $f(x,y)$, then p_i indicates the probability of gray-value i, $p_i = \frac{n_i}{M \times N}$, $i = 0, 1, \ldots, L-1$;

(3) According to the theory of maximum entropy, $f(x,y)$ is divided into two parts: the object area and the background area, and the probability of each area can be respectively represented as $P_O(t) = \sum_{i=0}^{t} p_i$ and $P_B(t) = \sum_{i=t+1}^{L-1} p_i$, where t is the segmentation threshold, and $P_O(t) + P_B(t) = 1$;

(4) Then the entropy of the object area and the background area is respectively represented by:

$$H_O(t) = \ln \sum_{i=0}^{t} \frac{p_i}{P_O(t)} \tag{3}$$

$$H_B(t) = \ln \sum_{i=t+1}^{L-1} \frac{p_i}{P_B(t)} = \ln \sum_{i=t+1}^{L-1} \frac{p_i}{1 - P_O(t)} \tag{4}$$

(5) The total entropy of $f(x,y)$ is $H(t) = H_O(t) + H_B(t)$. When the maximum of $H(t)$ is reached, the gray-value t is the optimal segmentation threshold.

4 Face Detection Algorithm Based on Object Region Centroid

The flow chart of the proposed face detection algorithm based on object region centroid is shown in Fig. 2.

The detailed steps are as follows.

(1) Input the original image $I(x, y)$, using the proposed Log_Gabor_GBVS algorithm to extract a significant figure $S(x, y)$ in the target area;

(2) Obtain the optimal threshold based on the maximum entropy criterion, and get the two value image $M_1(x, y)$ after segmenting the saliency map $S(x, y)$;

(3) Determine whether there is redundancy in the two value image $M_1(x, y)$. If there is, go to step (5); if there isn't, go to step (6);

(4) Determine whether the two values image $M_1(x, y)$ is completely covered by the target area. If it isn't, go to step (5); if it is, go to step (6);

(5) Implement morphological operations on the two value image $M_1(x, y)$ to get the connected two valued image $M_2(x, y)$. If template $M_1(x, y)$ remaining outside the target face region, delete the redundant area in $M_1(x, y)$, only retain the useful region. If $M_1(x, y)$ does not completely cover the face region, implement closing operation on $M_1(x, y)$; if $M_1(x, y)$ lost too much information, implement hole filling on $M_1(x, y)$;

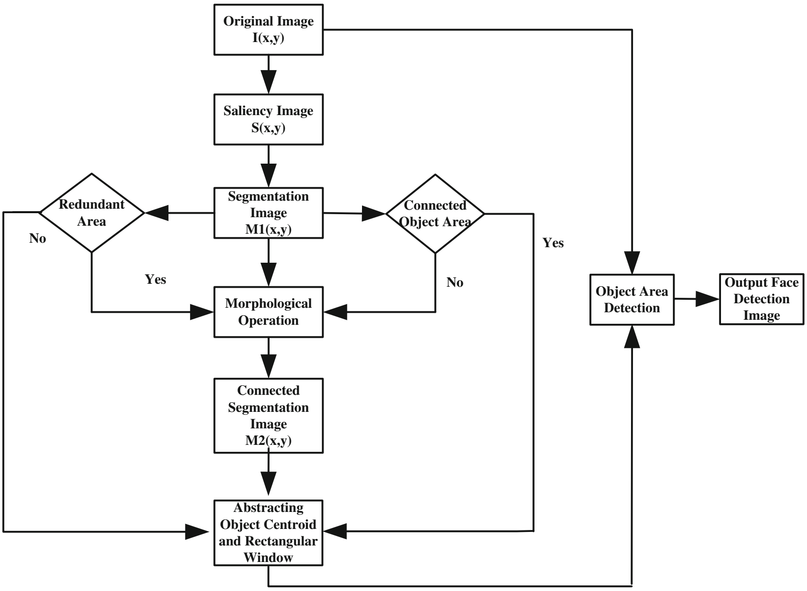

Original Image I(x,y)
Saliency Image S(x,y)
Segmentation Image M1(x,y)
Redundant Area
Connected Object Area
No
Yes
Yes
No
Morphological Operation
Connected Segmentation Image M2(x,y)
Abstracting Object Centroid and Rectangular Window
Object Area Detection
Output Face Detection Image

Fig. 2. Flow chart of the proposed algorithm

(6) Determine the mass center of the target area. Let the initial variable *sum_x* record the sum of the horizontal coordinates, *sum_y* record the sum of the vertical coordinates in the target area of the connected two value image $M_2(x, y)$, and *area* record the number of pixels in the target area. Traversal $M_2(x, y)$, if the pixel value is 1, then update the above three variable by

$$area = area + 1$$

$$sum\_x = sum\_x + x$$

$$sum\_y = sum\_y + y$$

Otherwise, no update. After traversal $M_2(x, y)$, the centroid coordinates p of the target area $\left(p_x = \frac{sum\_x}{area}, p_y = \frac{sum\_y}{area}\right)$ are calculated by the three values *sum_x*, *sum_y* and *area*. Taking the centroid as the center, the minimum distance from the centroid to the side as the edge, you can get the rectangular window of the face target.

(7) According to the location of the rectangular window in the human face, the final image is detected in the original image $I(x, y)$.

5 Experiment Results

We used LFW (Faces in Labeled the Wild) face database [12] for the simulation experiment. This database contains 13233 face images from 5749 colleagues, image size is 250 × 250, including 1680 people who have more than two pictures, remaining 4069 people only have one picture, some images contain not only a face, but the target face is located in the center of the picture, all other faces except the target face are regarded as the background, each picture are assigned an unified format identification name.

When the unconstrained face recognition is carried out on the LFW database, the samples should be selected firstly in the database. From the LFW database, this article select the people who has more than 20 pictures as the experimental object, get 62 kinds of people, a total of 3023 pictures. Randomly selected 10 images of each kind of people as the training sample, the remaining picture as the test sample, apply the unconstrained face recognition method. Using the algorithm in this paper and the classic Viola-Jones face detection algorithm, apply the unconstrained face detection algorithm on the LFW database after sample selection, get two new databases, denoted as LFW_Log_Gabor_GBVS and LFW_Viola_Jones. HOG operator is used to extract the features from these two databases, and the SVM classifier is used for classification. The experimental results are shown in Table 1.

Table 1. Comparison of unconstrained face recognition results from three databases

Image database	Recognition rate (%)	Recognition time(s) (recognition + classification)
LFW	31.79	211.7
LFW_Viola_Jones	66.63	69.9
LFW_Log_Gabor_GBVS	**72.53**	**57.8**

From Table 1 we can see that, without the preprocessing of face detection the original LFW database, the unconstrained face recognition rate is only 31.79%. This is because the original samples from LFW database contain a large amount of real and complex background, which can serious influence the accuracy of face recognition. Conversely, if we preprocess on the original database for face detection, both the Viola-Jones face detection algorithm, and detection algorithm in this paper, can effectively improve the accuracy of unconstrained face recognition. The recognition rate of LFW_Viola_Jones database is increased to 66.63%, and the recognition rate of LFW_Log_Gabor_GBVS database is increased to 72.53%, which is much higher than the recognition rate of the original LFW database. Therefore, face detection is a very effective and necessary preprocessing step before the unconstrained face recognition.

At the same time, we can also see that the recognition rate of LFW_Viola_Jones database is increased by 5.9% compared to the LFW_Log_Gabor_GBVS database, this is because the proposed detection algorithm can more effectively detect irregular face than the Viola-Jones algorithm, it can remove the background interference with no loss of any face information, and reach a good balance. This has further verified the conclusion in the Sect. 6.

6 Conclusions

In this paper, a unconstrained face detection algorithm based on visual attention mechanism is proposed. Firstly, extract visual feature using Log_Gabor transform, to obtain the saliency map of target face in complex environment, the GBVS measurement method based on steady-state equilibrium is adopted. Then, the threshold segmentation and morphological optimization are used to get two value face image using the maximum entropy criterion. Finally, set the target centroid as the center, set the minimum distance from the region side to the centroid as the edge, cut out the rectangular face region in the original image, achieve the face target detection in complex environment. Compared with the classical Viola-Jones algorithm, this algorithm results are better in detection of the irregular faces. The method can remove the background interference with no loss of the facial information, and time is not significant increased, it is a practical and effective face detection algorithm. At the same time, the application on the unconstrained face recognition is further proved this conclusion.

Acknowledgement. This work is financially supported by NSFC Projects No. 61471162, Program of International science and technology cooperation (0S2015ZR1083), NSF of Jiangsu Province (BK20141389, BK20160781), NSF of Hubei Province (2014CFB589), Technology Research Program of Hubei Provincial Department of Education (D20141406), Key Laboratory of meteorological detection and information processing in Jiangsu Province (KDXS1503), and NIT fund for Young Scholar (CKJB201602).

References

1. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)
2. Hua-Chun, Y., Wang, X.A.: A study on components and features in face detection. Int. J. Inf. Technol. Web. Eng. **10**(3), 33–45 (2015)
3. Yan, J., Zhang, X., Lei, Z., et al.: Face detection by structural models. Image Vis. Comput. **32**(10), 790–799 (2014)
4. De la Torre, F., Collet, A., Quero, M., et al.: Filtered component analysis to increase robustness to local minima in appearance models. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
5. Ding, C., Choi, J., Tao, D., et al.: Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(3), 518–531 (2016)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
7. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Advances in Neural Information Processing Systems, pp. 545–552 (2006)

8. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
9. Alamareen, A., Aljarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web. Eng. **11**(3), 1–14 (2016)
10. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. IEEE Trans. Pattern Anal. Mach. Intell. **34**(10), 1915–1926 (2012)
11. Cheng, M.M., Mitra, N.J., Huang, X., et al.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
12. Kaneko, K., Okada, Y.: Facial expression system using Japanese emotional linked data built from knowledge on the web. Int. J. Space-Based Situated Comput. **4**(3/4), 165 (2014)
13. Huang, G.B., Ramesh, M., Berg, T., et al.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07-49, University of Massachusetts, Amherst (2007)

Pavement Crack Detection Fused HOG and Watershed Algorithm of Range Image

Huazhong Jin(✉), Fang Wan, and Ou Ruan

School of Computer Science, Hubei University of Technology,
Wuhan, China
galaxy0522@163.com

Abstract. Pavement crack detection plays an important role in pavement maintaining and management. In recent years, pavement crack detection technique based on range image is a recent trend due to its ability of discriminating oil spills and shadows. Existing pavement crack detection methods cannot effectively detect transverse and network cracks, because these methods generally represent the crack geometry feature using single laser scan line, which cannot take the effects of spatial variability, anisotropy and integrity into account. Aiming at the deficiency of existing algorithms, the pavement crack detection method fused histogram of oriented gradient and watershed algorithm is proposed. Firstly, crack edge strength and orientation are detected by histogram of oriented gradient in pavement range image. Then, the traditional watershed algorithm is improved by using the crack edge orientation in order to better extract the crack object. Experiment results show that the proposed method can accurately detect different types of crack objects and identify the severity of crack damage simultaneously.

1 Introduction

Crack is one of the most common diseases of pavement; automatic detection of cracks is of great significance for highway inspection and maintenance management. At present, the crack detection algorithm mainly takes the gray and morphological characteristics of the two-dimensional pavement image as a criterion for identifying cracks; Due to the limitation of the hardware condition of the image acquisition system and the influence of external illumination, with the interference of the road surface oil pollution, shadow, tire traces, random noise and other factors, the pavement sampling data is not reliable, and the misjudgment rate is large. Therefore, the traditional pavement crack detection method based on image gray information cannot achieve satisfactory results [1–3]. 3-D laser visual detection technology generates pavement range image by collecting three-dimensional data of pavement. Compared with the traditional method, range image is not sensitive to the pavement oil pollution, shadow, dark shading, and contains rich pavement geometric characteristics, and has the advantages of high accuracy and high resolution. Then, the detection method of range image has become one of the new research directions of pavement crack detection [4–6].

Since 1966, the international organization for standardization firstly put forward the concept of three-dimensional pavement detection [7], Canada, the United States, Japan

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_47

and other developed countries have carried out the research and development of corresponding detection technology and equipment. The 3D detection technology of pavement crack mainly includes laser holographic imaging technology, laser radar, stereo vision and line structured light 3-D detection technology [8]. In the 1990s, some research results have been published. In 1997, Laurent and Hebert of Canadian INO Company proposed a crack detection algorithm based on laser 3D pavement profile [9]. The algorithm takes 3 cm as sampling interval on the profile, and detects the crack signal by setting the crack width threshold and depth threshold. However, when the crack is just in the sampling interval, the algorithm causes the error detection of the crack signal. In 1999, Bursanescu et al. adopted adaptive width of moving window function to filter the profile, so as to extract the location and width of the crack [10]. The effect of crack detection is very sensitive to window width, which causes miss detection of the crack signal. At present, there are few studies on the use of three-dimensional laser data to extract the crack feature in China. Sun Xiaoming proposed a crack extraction method based on sparse decomposition [11]. This method uses the combination of trapezoidal function and Gaussian function to describe the geometric characteristics of the crack signal, and applies the matching tracking algorithm of sparse representation to achieve the matching and separation of cracks in pavement contour. Because there are a lot of spaces between aggregate particles in asphalt mixture, the pavement macro texture is considered as pseudo crack. Therefore, the method cannot distinguish between true and false cracks. Tang Lei et al. [12] used space detection operator based on differential geometry to extract the crack information of three-dimensional terrain surface. This method maps 2-D image to 3-D terrain surface. Because the actual pavement image contains a lot of variety of disease and noise information will be brought into three-dimensional data, the effect of crack detection is not ideal.

At present, some research achievements have been made on the use of pavement range image to extract cracks, but most of them only focus on the pavement laser scan line, which lack the description of the whole pavement crack feature. Due to the influence of light stripe direction and the crack geometry on the laser scan line, most of the existing researches are limited to the one dimensional problem. Although the pavement longitudinal cracks can be detected well, it is difficult to accurately extract the cracks in other directions. The geometry of pavement cracks is complex and varied, and the main types of pavement cracks include lateral and network cracks, which are often characterized by variability, anisotropy and global characteristics. If only considering the geometric information of the scan line, the spatial distribution feature of pavement cracks cannot be accurately described.

Aiming at the geometric shape and feature of the pavement cracks in range image, the pavement crack detection fused histogram of oriented gradient and watershed algorithm is proposed to solve the limitation of crack extraction method based on laser scan line. The proposed method can not only detect the different types of cracks, but also identify the severity of the crack damage.

2 3-D Measurement of Line Structured Light and Crack Feature of Pavement Range Image

The 3-D vision detection technology based on line structured light is a kind of non-contact measurement technology of laser vision detection, which can quickly obtain high-precision information of 3-D object surface, and has been widely used in accurate measurement of 3-D objects [13, 14].

The basic principle of measurement is shown in Fig. 1. The line structured light source projects a light plane in the space. When the light plane intersects with the object surface, a light stripe is generated on the object surface. Then the light stripe is collected by the 3-D camera. With the change of the object surface fluctuation, the corresponding deformation of the light stripe will occur. According to the triangulation principle, the 3-D profile information of the measured object is obtained from the deformed light stripe [15].

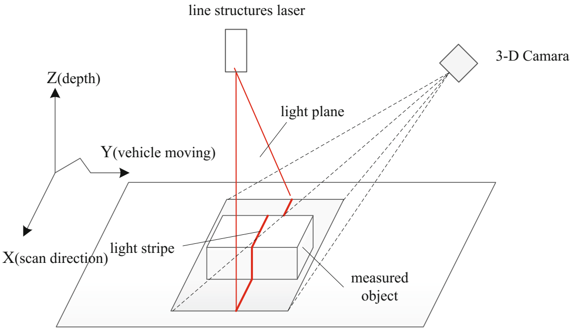
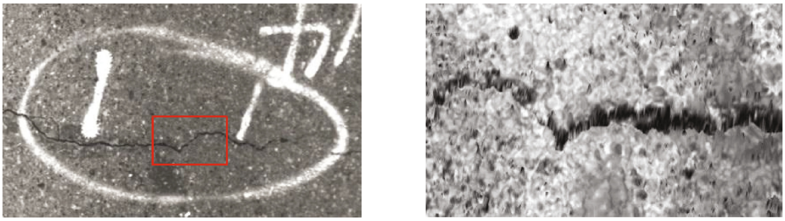


Fig. 1. Illustration of 3-D laser detection system

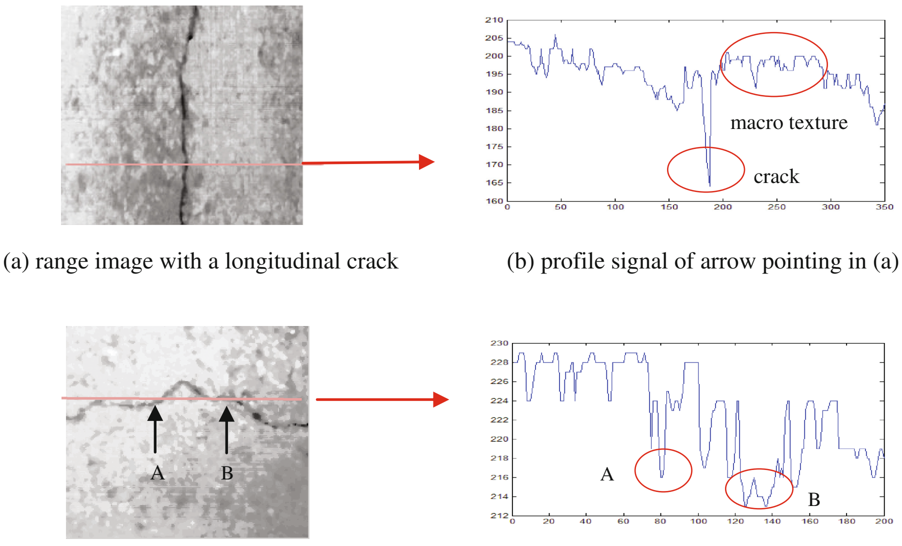
The surface point cloud data of line structured light vehicle detection technology has rich geometric, high accuracy and high density features. Therefore, the high-precision point cloud data can accurately express the spatial location and distribution information of the pavement crack, as shown in Fig. 2. Figure 2a is a pavement gray image, and Fig. 2b is a range image of the red rectangular frame enlarged in Fig. 2a.



(a) pavement gray image

(b) range image of the red rectangular frame enlarged in (a)

Fig. 2. Pavement gray image and range image



(a) range image with a longitudinal crack

(b) profile signal of arrow pointing in (a)

(c) range image with a transverse crack

(d) profile signal of arrow pointing in (c)

Fig. 3. Road crack image and the corresponding profile signal

The pavement image acquired by the line structured light detection technology is extracted from the light strip center, and the pavement three-dimensional profile is obtained, as shown in Fig. 3. Figure 3a and c are pavement range normalized images. Figure 3a shows that the laser beam is projected onto the pavement with a longitudinal crack, and the Fig. 3c indicates that the laser beam is projected onto the pavement with a transverse crack. Figure 3b and d represents the pavement profile signals at the projected light stripe in Fig. 3a and c, respectively.

According to the analysis of Fig. 3b, the crack geometry has the following features: (1) the crack edge is composed of a set of irregular points, which has a certain direction and a large edge gradient amplitude; (2) linear aggregation in the crack direction; (3) For the spatial distribution, the cracks have continuity and proximity on the adjacent scan lines; At the same time, the crack depth is larger than that of macro texture and noise, and the crack width is different. Because the actual situation of the pavement is more complex, especially the existence of strong absorption and strong reflection objects, and the debris shielding, the laser beam cannot completely projected onto the 3-D camera. It makes the crack geometric information loss. At the same time, the crack geometric forms such as direction, depth, width, continuous and discontinuous are randomly changed, which bring laser beam abnormal deformation, resulting in crack geometry degradation in range image. As shown in Fig. 3c, the laser beam is projected onto the transverse crack. Because there is a certain angle between the crack in A and the light stripe direction, crack geometric features are directly obtained on the scanning line of 3-D camera, as shown in Fig. 3d. However, the crack in B coincides with the light stripe; its geometric features are not obvious on the corresponding scan line, as shown in Fig. 3d. The pavement condition complexity and crack geometry diversity can also cause that range image appears invalid zero value and abnormal value phenomena, and produce a lot of noise. From the above analysis, according to the principle of pavement range information extracted by line structure light, combined with the pavement complexity and the crack diversity, the crack object is small discontinuous and low signal-to-noise ratio in pavement range image. However, its geometric features show that the gradient amplitude near crack edge dramatically changes, appear the linear distribution of the point set in the spatial, and emerge the approximate continuity of gradient amplitude in crack direction. Therefore, this paper use the high accuracy and high resolution pavement range image to detect the crack object by the gradient amplitude and orientation change, considering the crack variability and anisotropy in the overall spatial distribution.

3 Crack Edge Detection by HOG

Histogram of Oriented Gradient (HOG) by Dalal and Triggs in 2005 is a feature descriptor for object detection in computer vision and image processing [16]. This method divides the whole detection window into a small grid or two semicircle areas in one direction. The gray histogram of image is computed in these grid or semicircle regions, and the histogram gradient is calculated in this direction. The gradient feature of the whole image is represented by linking the histograms of oriented gradient of all regions. HOG is essentially based on statistical pattern recognition method to describe edge gradient distribution and orientation. Because of the sharp change of crack depth value and edge linear aggregation, the method can be used to describe the gradient amplitude and orientation change of it. In this paper, the circular HOG method for detection area is used to statistics and describes the crack geometric feature.

As a nonparametric estimation of the feature distribution, histogram can express the statistical features of the local region in image. Histogram gradient method is to estimate the regional probability density function, which can compute the difference between regions by histogram distance function. It measures the local area difference, and describes the similarity between different regions. As a nonparametric estimation method of feature distribution, and histogram has the advantages of translation and rotation invariance, it has been widely used.

Histogram can express the statistical characteristics of image local area, but the similarity between regional characteristics is usually measured by distance function. χ^2 is intuitive, and has the advantages of simple and fast calculation of the difference between different image regions. In this paper, the traditional χ^2 statistic method is transformed into χ_θ^2 distance function with different radius and direction. Its formula is

$$\chi^2(g_\theta, h_\theta) = \frac{1}{2}\sum_{k=1}^{K} \frac{(g_\theta(k) - h_\theta(k))^2}{g_\theta(k) + h_\theta(k)} \tag{1}$$

In the above formula, g_θ and h_θ are the histogram of the two half circular disc which neighborhood radius is r. k is the number of histogram bins. θ represents the

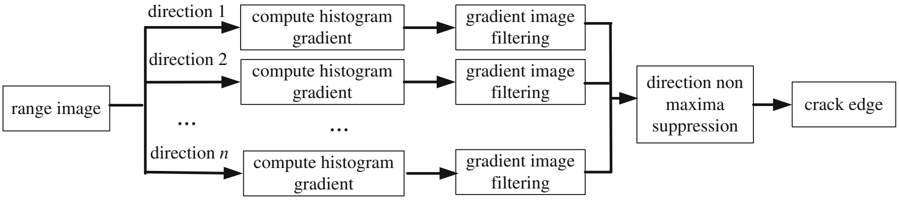


Fig. 4. Crack edge detection based on HOG

radians in a particular direction, $\theta = m \times \frac{\pi}{n}$, $m \in [0, n)$; m represents a specific slice, n denotes the number of partitions equal to π.

The crack edge extraction algorithm based on HOG is shown in Fig. 4.

In the pavement range image, the neighborhood information of each pixel is computed by the disk, and the crack edge feature descriptor is generated. Specific crack edge extraction algorithm includes the following steps:

(1) Compute histogram gradient

The histogram gradient of all scan points in the fixed disk is computed in eight directions at any position of the pavement range image. First, the histogram of half disk region of each position in the direction θ is calculated, and the depth value statistics in the corresponding neighborhood are obtained. The Gauss function is used to compute the one-dimensional filter convolution for histogram bins to obtain a smooth function curve. The parameters of the Gaussian function are determined by the number of histogram bins and width factor σ. Then, the normalized histogram bins is needed. The normalization is performed in each disk, and the general normalized function has the following four:

(a) L1-norm: $v \rightarrow v/\left(\|v\|_1 + \varepsilon\right)$;
(b) L2-sqrt: $v \rightarrow v/\sqrt{\|v\|_1 + \varepsilon}$;
(c) L2-norm: $v \rightarrow v/\sqrt{\|v\|_2^2 + \varepsilon^2}$;
(d) L2-Hys: the method is the same as above, the maximum limit to 0.2, and renormalized processing.

The normalized histogram is used to estimate the depth information, so that the depth feature is robust to noise and edge orientation change. L1-norm method is used in this paper. Finally, the histogram gradient of each half disk region of each position in the direction θ is calculated by formula (1).

(2) Gradient image filtering

Because the acquisition of pavement depth value is affected by environmental lighting conditions, the complexity of ground conditions and the geometric morphological diversity of crack, range image contain a large number of random noises. These noises will bring in the gradient image and generate multiple detection peaks. These peaks make the crack edge curve not smooth, and need to be detected and removed. In common curve smoothing method, the mean filter smoothing method and the weighted average method do not consider the change trend of the curve itself. However, the least square smoothing method assumes that the curve has some mathematical characteristics, and uses polynomial to fit the curve. In the curve smoothing method, Savitzky-Golay filtering algorithm is a classical least square smoothing method. It smooths the curve by using the simplified least square fitting convolution method and calculates the order derivatives of the smooth curve [17].

Usually, the curve is the p polynomial, that is,

$$Y_i = a_0 + a_1 i + a_2 i^2 + \cdots + a_p i^p. \tag{2}$$

In the formula, Y_i represents the smoothed value of the point i. The error of the fitting curve is computed by the polynomial above.

$$S = \sum_{j=-m}^{m} (Y_j - y_j)^2 \tag{3}$$

In formula (3), y_j represents the value before smoothing, and the smooth window size is $k = 2\ m + 1$. In order to minimize the error S, S is used as partial differential, and the partial differential of S is equal to zero.

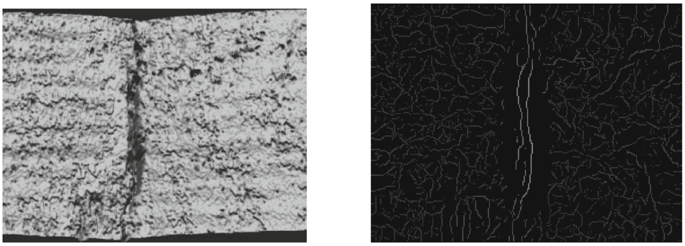
In this paper, two order Savitzky-Golay filters are used to fit the edge points of each crack in the eight directions. For each scanning point, a parabolic surface is used to fit the ellipse. The long axis of the elliptical region is the disk radius, and the short axis is 1/4 of the long axis. The fitting direction is $\pi/2$, $3\pi/8$, $\pi/4$, $\pi/8$, 0, $7\pi/8$, $3\pi/4$, $5\pi/8$. The experimental results show that the two order Savitzky-Golay filter is used to enhance the local extremum of edge curve, remove the noise, smooth out the peaks, and preserve the details of the edge.

(3) Direction non maxima suppression

In order to locate the crack composite edge point, direction non maxima suppression needs to be computed for the gradient values of eight directions in the previous step. The angle corresponding to the maximum gradient in eight directions is used as the edge orientation of the pixel. The gradient value of two adjacent pixels is viewed along the vertical direction of the pixel dominant direction. If the gradient value of the pixel is greater than or equal to the value of the adjacent point, the gradient value of the pixel is preserved. Otherwise, it is set to 0.

The above algorithm extracts the crack edge of deep image; the experimental results are shown in Fig. 5.

From the analysis of Fig. 5b, the HOG method can well detect the details of the crack edge and pavement macro texture. The crack edge is clear, the edge strength is bigger, the edge direction is consistent with the crack direction; the edge of macro texture is too broken, the edge strength is weak, and its shape is random distribution. Therefore, the HOG method can detect the crack objects by setting the edge strength threshold and orientation. However, it is also seen from Fig. 5b that the crack edge is not closed, which indicates that the HOG method cannot extract the crack surface target well. Although it can describe the strength and orientation of the edge, it is unable to identify and deal with the edge discontinuity, which causes the edge of the area crack to



(a) original range image （b）crack edge

Fig. 5. Crack edge detection by HOG

not be closed. Therefore, this paper will introduce the watershed transform algorithm to further improve the area cracks detection.

4 Crack Detection by Direction Watershed Algorithm

Watershed transform is a kind of image segmentation method based on mathematical morphology, which is essentially the process of simulating the flooding surface [18]. The traditional watershed transform algorithm has many advantages, such as simple, fast. It can detect the weak edge object and get the complete boundary of the object. Because it is usually transformed in the gradient image, the transformation result is affected by the noise and other factors. So it has a large number of pseudo local regions, which appears too much phenomenon. Because of the pavement complex condition, its range image is prone to noise in the image processing. In particular, the uncertainty and complexity of the spatial distribution of cracks, the traditional watershed method is prone to a large number of broken areas. At the same time, the cross is easily emerged near the strong boundary. These facts make the segmentation effect worse. In order to improve the segmentation effect, this paper proposes an improved watershed algorithm by the edge direction. The main idea is to extract regions from the previous edge image by the traditional watershed transform. Then the edge directions are fitted to modify the edge strength. Lastly, region detection results are obtained by edge strength.

The direction watershed transform algorithm to extract the crack region includes the following steps:

(1) The original pavement image is processed to obtain its corresponding gradient image by HOG algorithm. The global gradient maximums in gradient image pixels are extracted from the eight directions. The pixel of the gradient maximum value is used as the edge in order to obtain the crack edge image, as shown in Fig. 5b.
(2) The minimum depth value of the edge image is taken as the seed, and the watershed transform algorithm is used to get the segmentation regions.
(3) Along the crack edge, the constrained Delaunay triangulation is computed to modify the edge of the previous step. The two endpoints of the edge arc are used as the point set in the two-dimensional real domain V. The corresponding edge is embedded as a constraint edge, and the two ends of the constraint edge must be in

the result of the partition. The Delaunay triangulation is computed to obtain DT = (V, E), whose point set V is the corresponding edge vertex. Then, the point set v is used to modify the original edge. The constrained edges are embedded without changing the set of boundary points, which can easily deal with the discontinuous, steep line and so on. On the other hand, due to the Delaunay triangulation net with properties of empty circumcircle and minimum angle maximization, it can maximally satisfy the approximate equilateral triangle (angle) of a triangle. The purpose is to avoid too narrow and sharp triangle on the crack edges.

(4) Extraction of adjacent area label. The width of adjacent regions is determined by the scale factor of the corresponding edge length. The label and size of the region is determined by the different edge length, and the vertex-edge map is recorded.

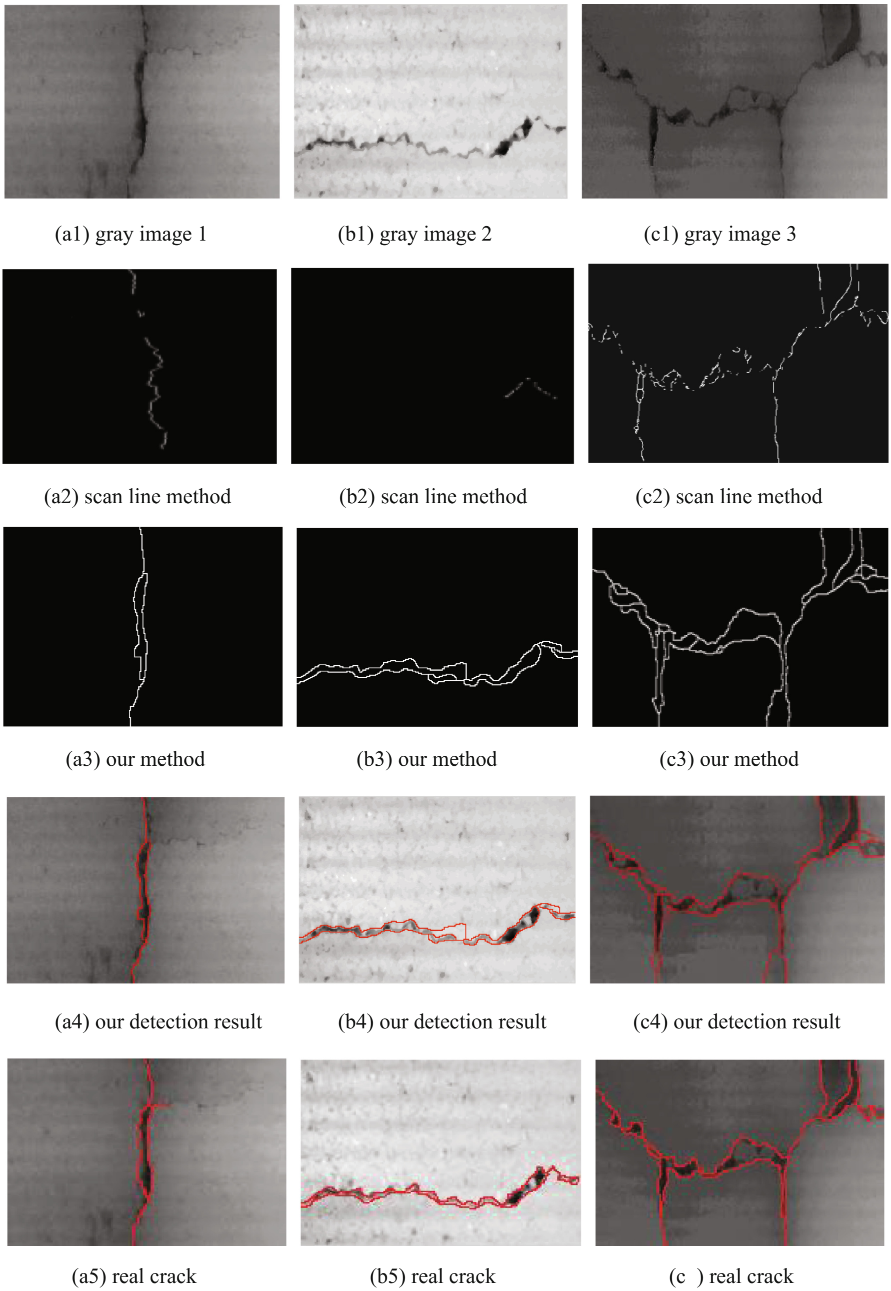
(5) Boundary fitting. According to the local geometry and length of each arc, the arc direction of each point is estimated. If the distance between the two ends of the arc and any point on the arc is larger than a given threshold, it is necessary to use the segment to fit the arc. By iterative, the fitting process is stopped when the distance is not greater than the set threshold. Hence, the arc is represented as a scale invariant piecewise segment by the approximation method. The point (x, y) direction on the corresponding segment is expressed by $o(x, y) \in [0, \pi)$.

(6) Extract the label of the closed edge.

5 Experiment and Analysis

In this paper, the experimental data are collected by the vehicle range camera, the vehicle speed is 80 km/h, the sampling interval of vehicle moving direction is 1 mm, and the sampling interval of vertical vehicle moving direction is 1.8 mm. In order to verify the effectiveness of the proposed method, the range images with longitudinal, transverse, massive and network cracks are selected. The experimental images are cut from the original range images, and the size is 200 × 200 scan points. These images are preprocessed by median filtering and normalization. Because HOG method computes histogram gradient and the gradient image filters from eight directions, it consumes a large amount of computing resources. Because the computation of histogram gradient and filtering gradient image consume a large amount of computing resources, the sampling radius of the detection area is selected as 3 point spacing to improve the real-time performance.

The experiment is to test the effectiveness of the proposed method for different types of cracks, compared with the laser scanning line and the artificial method. Pavement range images comprising longitudinal, transverse and block cracks are respectively shown in Fig. 6a, b and c. Figure 6a1, b1 and c1 respectively represents the corresponding pavement gray image. Crack detection results by the scan line method are shown in Fig. 6a2, b2 and c2. Because the method can better describe the feature of the crack signal perpendicular to the vehicle moving direction, the results show that the longitudinal crack detection effect is more ideal from Fig. 6a2; Crack detection effect is poor in Fig. 6b2 and c2. It can be seen that the scan line method is a



(a1) gray image 1

(b1) gray image 2

(c1) gray image 3

(a2) scan line method

(b2) scan line method

(c2) scan line method

(a3) our method

(b3) our method

(c3) our method

(a4) our detection result

(b4) our detection result

(c4) our detection result

(a5) real crack

(b5) real crack

(c) real crack

Fig. 6. Experiment 1 of crack extraction

kind of treatment method for the simple cracks, especially the longitudinal crack. Therefore, the scan line method cannot effectively extract the more complex types of cracks.

The detection results by the proposed algorithm are shown in Fig. 6a3, b3 and c3. Our detection results and original image are superimposed on the Fig. 6a4, b4 and c4. From overlay analysis, it can be seen that our algorithm can reduce the fragment and discontinuity of the crack, and can also close the crack edge better.

The crack objects are displayed in a closed boundary. Because the proposed algorithm reduces the broken edge and enhances the integrity of the crack edge, it improves the extraction effect of the crack edge. Real cracks are obtained by field measurements and manual editing in Fig. 6a5, b5 and c5. Compared with the results of manual segmentation, the crack detection results of this method are very close to the real crack. The detection results of the more complex network cracks can also be observed in Fig. 7.

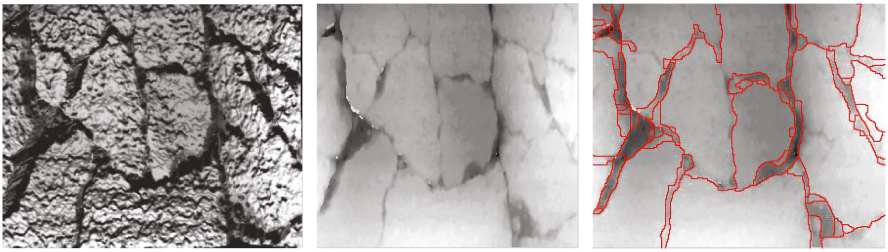

(a) range image (b) gray image (c) detection result

Fig. 7. Experiment 2 of crack extraction

In order to analyze and evaluate the effect of crack detection, this paper introduces the Performance Evaluation of Tracking and Surveillance (PETS) method. Based on the measurement method of PETS, the degree of mismatch between pairs of scanning points is measured by comparing the true sample segmentation (manual segmentation) and the detection of samples (algorithm segmentation), which is called the Negative Rate Metric (NR) [19, 20]. That is

$$NR = \frac{1}{2}(NR_{fn} + NR_{fp}). \tag{4}$$

The formula $NR_{fn} = \frac{N_{fn}}{N_{tp} + N_{fn}}$ indicates that the number of real samples that are not detected accounts for the proportion of all "real samples"; the amount of data mistaken for the real sample in the observation accounts for the proportion of all real samples. The ratio is called by error rate.

The formula $NR_{fp} = \frac{N_{fp}}{N_{fp} + N_{tn}}$ indicates that the number of non-real samples is proportional to all the non-real samples, that is false negative rate.

A data set containing 150 pavement range images is used to verify the effectiveness of the proposed method. The data set is divided into four data sets, which contain four kinds of typical cracks, namely, longitudinal, transverse, block and network cracks. Among them, the data set number is 1, including 56 longitudinal crack images; the data set of 2 contains 30 transverse crack images; the data set of 3 contains 34 block crack images; the data set of 4 contains 30 network crack images. According to the extracted edge intensity k = 0.12, cracks are extracted on the above four typical pavement images.

According to the formula (4), the error rate, false negative rate and mismatch degree of scan line method and our method are computed respectively. Their values correspond to the mean of different data subsets, as shown in Table 1.

Table 1. Evaluation of crack extraction results based on NR method

Method	Data set	Error rate	False negative rate	Mismatch
Scan line method	1	0.001345	0.3541	0.1777
	2	0.001463	0.8622	0.4318
	3	0.002321	0.4710	0.2367
	4	0.003432	0.5230	0.2632
Our method	1	0.002236	0.1655	0.0839
	2	0.005216	0.1236	0.0644
	3	0.004911	0.1121	0.0585
	4	0.005160	0.1210	0.0631

The experimental results show that misidentification rate, false negative rate and mismatch of our method is far lower than those of scanning line method. The main reason is that the crack features on the scan line are incomplete. Because the longitudinal feature is obvious on the scan line, this method is good for the longitudinal crack detection. However, the scan line method cannot fully describe the spatial distribution characteristics of the transverse, block and network cracks. Therefore, the scan line method is not ideal for the detection of non-longitudinal cracks. Because our proposed method takes into account the variability, anisotropy and global feature of the crack spatial distribution, it has strong identification ability to the edge and damage degree, and has high stability and reliability.

6 Conclusion

In this paper, aiming at the shortcomings of the existing pavement crack detection algorithm, the pavement crack detection combined HOG and watershed algorithm is proposed. Based on the gradient magnitude and orientation of the pavement range image, the crack edge is detected by the HOG. The watershed algorithm is used to get the connected domain in crack edge image, and the crack closure boundary is obtained. The proposed method can accurately detect crack edge and identify the severity of crack damage simultaneously. In practical engineering applications, the detection method can be used to identify linear and area objects, and get the true situation of the pavement cracks. The experimental results show that the proposed method can better overcome the macro texture and noise, and has high reliability and stability.

Acknowledgments. This work is funded by Hubei Provincial Department of Education, No. 2014277.

References

1. Liu, W., Xie, K., Pu, Z.: Review of pavement automatic detection system. J. China Foreign Highw. **27**(2), 30–33 (2007)
2. Mcghee, K.H.: NCHRP synthesis 334: automated pavement distress collection techniques. TRB, Washington, D.C. (2004)
3. Wang, K.C.P.: Designs and implementations of automated systems for pavement surface distress survey. J. Infrastruct. Syst. **6**(1), 24–32 (2000)
4. Cheng, H.D., Miyojim, M.: Automatic pavement distress detection system. Inf. Sci. **108**(1), 219–240 (1998)
5. Di, M.P., Piccolo, I., Cera, L.: Automated distress evaluation. In: Proceedings of 4th International SIIV Congress, pp. 12–14. International SIIV Congress, Palermo (2007)
6. Tsai, Y.C.J., Li, F.: Critical assessment of detecting asphalt pavement cracks under different lighting and low intensity contrast conditions using emerging 3D laser technology. J. Transp. Eng. **138**(5), 649–656 (2012)
7. Jianfeng, W.: Research on Vehicle Technology on Road Three-Dimension Measurement, pp. 23–45. Chang'an University, Xi'an (2012)
8. Gavilan, M., Balcones, D., Marcos, O., et al.: Adaptive road crack detection system by pavement classification. Sensors **11**(10), 9628–9657 (2011)
9. John, L., Jean-Francois, H.: High performance 3D sensors for the characterization of road surface defects. In: Machine Vision Applications 2002, pp. 11–13. Nara-ken New Public Hall, Nara (2002)
10. Liviu, B., Maher, H.: Three-dimensional laser ranging image reconstruction using three-line laser sensors and fuzzy methods. In: Proceedings of SPIE the International Society for Optical Engineering, vol. 3835, pp. 106–117. SPIE, Boston (1999)
11. Sun, X., Huang, J., Liu, W., et al.: Pavement crack characteristic detection based on sparse representation. J. Adv. Sig. Process. **1**, 1–11 (2012)
12. Lei, T., Chunchun, Z., Wang, H., et al.: Automated pavement crack detection based on image 3D terrain model. Comput. Eng. **34**(5), 20–21 (2008)
13. Wong, A.K.C., Niu, P., He, X.: Fast acquisition of dense depth data by a new structured light scheme. Comput. Vis. Image Underst. **98**(3), 398–422 (2005)
14. Peter, L., Andrew, B.: Real-time tracking of surfaces with structured light. Image Vis. Comput. **13**(7), 585–591 (1995)
15. Valkenburg, R.J., McIvor, A.M.: Accurate 3D measurement using a structured light system. Image Vis. Comput. **16**(2), 99–110 (1998)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, pp. 886–893. IEEE, San Diego (2005)
17. Savitzky, A., Golay, M.J.E.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **36**(8), 1627–1639 (1964)

18. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. **13**(6), 583–598 (1991)
19. Young, D.P., Ferryman, J.M.: PETS metrics: on-line performance evaluation service. In: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 317–324. IEEE, Beijing (2005)
20. Ellis, T.: Performance metrics and methods for tracking in surveillance. In: Third IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, pp. 26–31. IEEE, Copenhagen (2002)

Compressed Video Sensing with Multi-hypothesis Prediction

Rui Chen[1], Ying Tong[1], Jie Yang[1], and Minghu Wu[2](✉)

[1] School of Communications Engineering, Nanjing Institute of Technology,
No. 1 Hongjing Avenue Jiangning Science Park, Nanjing, China
{chenrui, tongying, yangjie}@njit.edu.cn

[2] Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy
and Operation Control of Energy Storage System,
Hubei University of Technology, Wuhan, China
wuxx1005@163.com

Abstract. This paper proposes a novel framework of multi-hypothesis compressed video sensing. Multi-hypothesis prediction and bi-directional motion estimation are applied to generate side information candidates. The correlation coefficients between the non-key frame and the three candidates are calculated respectively for selecting the most similar side information. The simulation results show that the proposed framework can provide better recovery performance than the framework using original MH-BCS-SPL algorithm.

1 Introduction

Recently, CS (Compressed Sensing) has revolutionized the signal sampling and processing systems by integrating the compression and sensing. The application of CS for video, known as Compressed Video Sensing (CVS), has been studied from different aspects. At the encoder, the input video frames are grouped into group of pictures consisting of a key frame and a number of non-key frames. In [1], key-frames are encoded using traditional MPEG/H.264 encoding [2], while non-key frames are sensed using a CS measurement matrix and reconstructed using side information generated from the neighboring reconstructed key frames. The disadvantage of this framework is that the complex MPEG/H.264 encoding is still required. In [3], the CS measurement is applied to both key frames and non-key frames. The key-frames are reconstructed using Gradient Projection for Sparse Reconstruction (GPSR) [4], and the non-key frames will be reconstructed with side information generated from the decoded key frames.

To alleviate the huge computation and memory burden, L. Gan presents a block-based CS (BCS) for 2D images with the assumption of the independence among blocks in [5]. Do and etc. uses the adjacent blocks in the precious decoded frame to represent the block in the current frame to improve the accuracy of side information, and developed a residual reconstruction method [6]. S. Mun extends Gan's BCS and cast the reconstruction in the domain of recent transforms that feature a highly directional decomposition in [7]. These methods are known as Single-Hypothesis Motion Compensation (SH-MC) schemes, which has some disadvantages. At the encoder, it

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_48

imposes a transmission overhead to send the block motion vectors besides the increase in the computational complexity at the encoder-side due to the motion estimation search. Moreover, the SH-MC implicitly assumes that the motions occurring in the video frames are of uniform block translational model. As this assumption does not always hold, the blocking artifacts appear in the recovered frame [8].

To address these issues, E. Tramel and etc. propose a strategy for incorporating multi-hypothesis motion compensation (MH-MC) into BCS with smooth projected Landweber (MH-BCS-SPL) reconstruction of video to get more accurate prediction by finding a linear combination of all the blocks/hypotheses in the search window [8]. The MH-MC techniques improve the recovery performance at the expense of more complexity at the decoder. A combination of the MH and SH reconstruction schemes is used in [9], and an elastic net-based MH-MC is suggested in [10] which achieved acceptable performance at the expense of more complexity compared to Tikhonov-regularization reconstruction. In [11], the authors propose hypothesis set updating and dynamic reference frames selection algorithms. An approach to deploy the MH prediction in measurement domain and pixel domain successively is presented in [12] to develop a two-stage MH reconstruction scheme, and R Li and etc. presented the space-time quantization and motion-aligned reconstruction to improve the performance of CVS system [13].

However, there are still remain some problems to be solved. The generation rule of side information (SI) is usually simple due to releasing the computation burden of the coders. The non-key frame reconstruction process also cannot perform effectively with the rough prediction. Thereby, we focus on a novel framework where three side information candidates are calculated for choosing to improve the traditional MH prediction algorithm. The remaining part of the paper is organized as follows. Section 2 reviews the compressed sensing and the existing MH-BCS-SPL scheme. Section 3 describes our proposed compressed video sensing framework and the implementation details of the framework. The simulation results are presented in Sect. 4. Finally, Sect. 5 concludes our work.

2 Compressed Sensing Overview

Compressed sensing combines signal acquisition and dimensionality reduction by measuring a projection of the signal x of dimensionality N using some basis, Φ, of dimensionality $M \times N$, where $M << N$. The measurement vector y is obtained as:

$$y = \Phi x \tag{1}$$

Where $x \in R^N$, $y \in R^M$. If x is sufficiently sparse in some transform basis Ψ, then x is reconstructed from y by the optimization as:

$$\hat{x} = \arg\min_{x \in R^N} \|\Psi x\|_1, \quad s.t. \quad y = \Phi x \tag{2}$$

Where Ψ and Φ are sufficiently incoherent and M is sufficiently large. The compressed sampling rate is defined as $R = M/N$. Usually, Φ is a random matrix such that it is incoherent with any chosen Ψ. In practical applications, most natural signals are not truly sparse in any transform basis Ψ. Then a common variant of the reconstructed problem of (2) is to relax the equality for a bound as follows,

$$\hat{x} = \arg\min_{x \in R^N} \|\Psi x\|_1, \quad s.t. \quad \|y - \Phi x\|_2 \leq \varepsilon \tag{3}$$

To solve the relaxed reconstructed problem of (3), L. Gan suggests a blocked-based CS (BCS) [11] that removes the global sampling of x by a dense Φ and replaces it with a block-diagonal measurement matrix. When the same Φ_B is used for every block, Φ takes on a block-diagonal form as

$$\Phi = \begin{bmatrix} \Phi_B & 0 & \ldots & 0 \\ 0 & \Phi_B & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & 0 & \Phi_B \end{bmatrix} \tag{4}$$

Such that (1) can be rewritten in block-by-block fashion as $y_i = \Phi_B x_i$, Where x_i is block i of the image. The size of Φ_B is $M_B \times B^2$, and the subrate of BCS is $R_B = M_B/B^2$. In this paper, we use BCS-SPL in [5] for recovery.

3 The Proposed CVS Scheme Based on MH

3.1 The Block Diagram of the Proposed CVS Scheme

The block diagram of our proposed CVS system is presented in Fig. 1. The CVS encoder is on the left side of the dotted line and the decoder is on the right. At the encoder, video frames are divided into key frames and non-key frames.

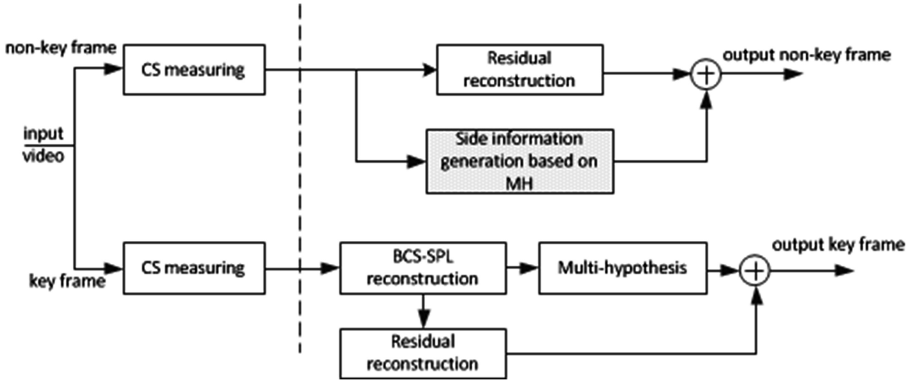


Fig. 1. Block diagram of the proposed CVS codec.

Let x_1 denote the key frame. At the decoder, the reconstruction process for the key frames is: First, the key frame x_1 will be initial reconstructed by using BCS-SPL algorithm. Then the prediction frame $\tilde{x}_1$ can be obtained by MH prediction implemented to the measurements of the x_1 (i.e. y_1) and its initial reconstruction. The residual between x_1 and $\tilde{x}_1$ is $R_1 = x_1 - \tilde{x}_1$. Because y_1 is acquired simply by taking the inner products of x_1 with the rows of Φ_1, the projection of R_1 into the measurement basis is

$$D_1 = \Phi_1 R_1 = \Phi_1(x_1 - \tilde{x}_1) = y_1 - \Phi_1 \tilde{x}_1 \tag{5}$$

where Φ_1 and y_1 are the measuring matrix and the measurements of key frame x_1 respectively, and D_1 is the measurements of residual R_1. Then D_1 is reconstructed by BCS-SPL algorithm to get the initial residual $\tilde{R}$, so the reconstructed key frame x_1 can be obtained by $\tilde{x}_1$ and the initial residual $\tilde{R}$ as $x_1 \approx \tilde{x}_1 + \tilde{R}$.

For non-key frames, they should be joint decoded by using side information. Let x_2 denote the non-key frame, it will be decoded by using side information generated from the key frame x_1. As shown in Fig. 1, side information (denoted as *SI*) is generated by using MH prediction. The residual R_2 between x_2 and *SI* is

$$D_2 = \Phi_2 R_2 = \Phi_2(x_2 - SI) = y_2 - \Phi_2 \cdot SI \tag{6}$$

where Φ_2 is the measuring matrix for non-key frame x_2 and y_2 is the measurements of x_2. Similarly, we can get the side information residual $\widetilde{SI}$ after reconstructing D_2, and the non-key frame x_2 can be reconstructed by $x_2 \approx SI + \widetilde{SI}$.

3.2 Side Information Estimation Based on MH in Measurement Domain

In the framework as shown in Fig. 1, the reconstruction quality of non-key frame depends hugely on the quality of the generated side information. To make full use of the similarity between the two consecutive key frames and the non-key frame respectively, the proposed side information generation algorithm based on MH in measurement domain is shown in Fig. 2.

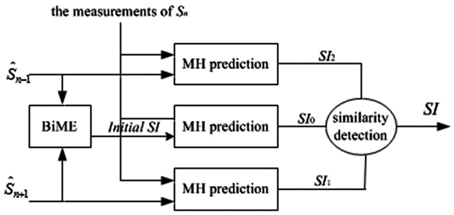


Fig. 2. The block diagram of side information generation based on MH.

Let $\hat{S}_{n-1}$ and $\hat{S}_{n+1}$ be temporally adjacent two reconstructed key frames and S_n be the non-key frame. As shown in Fig. 2, first, the initial side information *SI* is generated by the bi-directional motion estimation (BiME). The SI_0 is the multi-hypothesis prediction result of the initial side information and the measurements of the non-key frame S_n. Similarly, we can obtain the multi-hypothesis prediction SI_1 and SI_2. Then, we obtain three side information candidates. The similarity between the measurements of S_n and SI_i (i = 0, 1, 2) is calculated respectively, and the most similar SI_i is selected to reconstruct the non-key frames. The correlation coefficient function $r(y_1, y_2)$ that we adopt to measure the correlation between two frames is defined as

$$r(y_1, y_2) = \frac{\sum_{i=1}^{N} [y_1(i) - \bar{y}_1][y_2(i) - \bar{y}_2]}{\sqrt{\sum_{i=1}^{N} [y_1(i) - \bar{y}_1]^2}\sqrt{\sum_{i=1}^{N} [y_2(i) - \bar{y}_2]^2}} \tag{7}$$

where y_1 and y_2 are the different measurement vectors of block, N is the length of a measurement vector. Then, the *SI* generation procedure is described as follows.

3.3 Multi-hypothesis Prediction for Non-key Frame Reconstruction

In this section, we will describe the non-key frame reconstruction based on multi-hypothesis prediction. Let x be the original image and $\tilde{x}$ be its prediction image, the residual R between x and $\tilde{x}$ is $R = x - \tilde{x}$. In measurement domain, the residual R can be calculated by $D = \Phi(x - \tilde{x}) = y - \Phi\tilde{x}$. Then the approximate reconstructed $\hat{x}$ is obtained by $\hat{x} = \tilde{x} + R(D, \Phi)$, where $R(D, \Phi)$ denotes as one image compressed sensing reconstruction algorithm. So, the recovery quality is heavily dependent on the accuracy of the prediction image $\tilde{x}$. The problem of predicting the most similar image to the original image can be expressed as

$$\tilde{x} = \arg \min_{p \in p(X_{ref})} \|x - p\|_2^2 \tag{8}$$

where $p(X_{ref})$ can be the neighboring key frame or the side information generated by motion estimation. Due to the original image is unknown at the decoder, we replace x with its approximation $\hat{x}$ and (8) can be rewritten as

$$\tilde{x} = \arg \min_{p \in p(X_{ref})} \|\hat{x} - p\|_2^2 \tag{9}$$

Then, $\hat{x}$ can be transformed to the measurement domain and calculated as

$$\hat{x} = \arg \min_{p \in p(X_{ref})} \|y - \mathbf{\Phi} p\|_2^2 \tag{10}$$

Because of the measured value y is available at the decoder, so we can improve the accuracy of the prediction. Equation (10) can be solved by multi-hypothesis prediction. Each block that needs to be predicted is considered as the optimal linear combination of edge information or multiple blocks in the key frame as $\widetilde{x_i} = \mathbf{H}_i^s \omega$, where ω is the optimal linear combination coefficient, H_i^s is a $\mathrm{B}^2 \times MB$ matrix consisting of all candidate blocks, M is the total number of the hypothesis prediction blocks, each column of H_i^s is the column representation of each hypothesis prediction block. We substitute (10) into (9) and get

$$\omega = \arg\min_{\omega} \|\mathrm{y}_i - \mathbf{\Phi}\mathbf{H}_i^s \omega\|_2^2 + \lambda \|\Gamma\omega\|_2^2 \tag{11}$$

where $\lambda\|\mathit{\Gamma}\omega\|_2^2$ is the penalty term, λ is the Lagrange parameters. $\mathbf{\Gamma}$ is a diagonal matrix as

$$\mathbf{\Gamma} = \begin{bmatrix} \|y_i - \mathbf{\Phi}h_1\|_2^2 & \cdots & 0 \\ \cdots & \ddots & \cdots \\ 0 & \cdots & \|y_i - \mathbf{\Phi}h_k\|_2^2 \end{bmatrix} \tag{12}$$

where $h_1,\ldots,\ h_k$ are the columns of $\mathbf{H}_\mathrm{i}^\mathrm{s}$. For each block, then ω can be calculated directly by the usual Tikhonov solution as

$$\omega = \left((\Phi H_i^s)^T(\Phi H_i^s) + \lambda^2 \mathbf{\Gamma}^T \mathbf{\Gamma}\right)^{-1} (\Phi H_i^s)^T y_i \tag{13}$$

By taking (13) into (11), the prediction block $\tilde{x}_i$ can be achieved. Finally, all the predicted blocks are put together providing the *SI* frame.

4 Experimental Results

To evaluate our proposed framework, we take the standard test video sequences with QCIF format available at http://trace.eas.asu.edu/yuv/ in our experiments. The sampling rate of key frames is 0.7 and the sampling rate of non-key frames varies from 0.1 to 0.5. The block size B = 16, and the spatial window size w is set to be ±15 pixels.

The average PSNR performances with different sampling subrates for the four sequences (i.e. *Akiyo*, *Coastguard*, *Foreman* and *Stefan*) using the proposed algorithm and the original MH-BCS-SPL algorithm are shown in Table 1 and the graphic descriptions are shown in Fig. 3. As can be seen in Fig. 3, our MH-BCS-SPL framework gives better reconstruction quality across the range of tested subrates. It also can be seen that for sequences with fast or complex motion, such as the *Coastguard* and the *Foreman* sequences, the proposed method shows significant performance gains. However, for the *Akiyo* sequence with low motion, the performance are fluctuant and the performance gains are not substantial due to the background color in great changes.

Table 1. Average PSNR(dB) of non-key frames with different subrates.

Sequences	Algorithm	Sampling rate				
		0.1	0.2	0.3	0.4	0.5
Akiyo	Original MH-BCS-SPL	37.43	38.65	40.35	41.32	42.89
	Proposed MH-BCS-SPL	38.87	39.74	41.64	42.02	43.31
Coastguard	Original MH-BCS-SPL	26.64	27.06	30.12	31.17	32.85
	Proposed MH-BCS-SPL	27.71	28.15	30.45	32.34	34.72
Foreman	Original MH-BCS-SPL	31.08	31.85	33.84	34.39	37.73
	Proposed MH-BCS-SPL	31.32	32.81	34.53	35.65	38.25
Stefan	Original MH-BCS-SPL	23.45	24.87	26.08	27.77	29.34
	Proposed MH-BCS-SPL	23.59	24.93	26.41	27.84	29.55

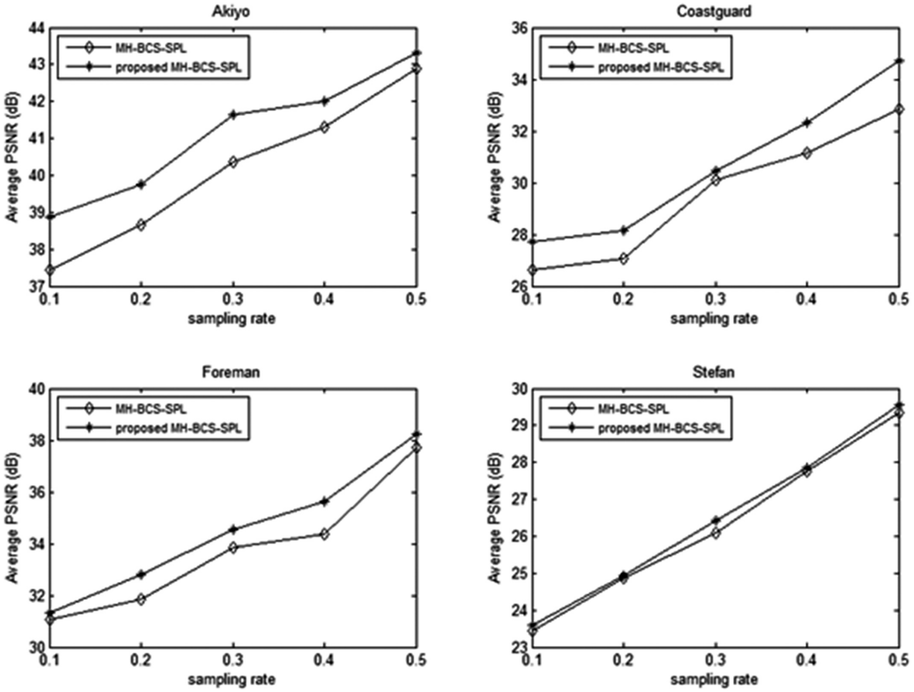


Fig. 3. Reconstruction quality of non-key frames with different subrates.

5 Conclusions

In this paper, a new distributed compressed video sensing framework based on MH prediction is proposed to capture and compress videos at low complexity encoder and efficiently reconstruct videos at the decoder. The proposed framework can estimate the initial side information by MH prediction and BiME. The side information is selected according to the correlation coefficients and used to recover the non-key frames. Our simulation results demonstrate that the proposed framework can provide better reconstruction quality than the original MH-BCS-SPL algorithm.

Acknowledgments. This research was supported by NSFC Projects No. 61471162, Program of International science and technology cooperation (0S2015ZR1083), NSF of Jiangsu Province (BK20141389, BK20160781), NSF of Hubei Province (2014CFB589), and Technology Research Program of Hubei Provincial Department of Education (D20141406).

References

1. Prades-Nebot, J., Ma, Y., Huang, T.: Distributed video coding using compressive sampling. In: Picture Coding Symposium, pp. 1–4. IEEE Xplore (2009)
2. Prадidtong Ngam, C., Natwichai, J.: Content-based video search on peer-to-peer networks. Int. J. Grid Util. Comput. **2**(3), 234–242 (2011)
3. Kang, L.W., Lu, C.-S.: Distributed compressive video sensing. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1169–1172. IEEE (2009)
4. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE J. Sel. Top. Sign. Proces. **1**(4), 586–597 (2007)
5. Gan, L.: Block compressed sensing of natural images. In: International Conference on Digital Signal Processing, Cardiff, UK, pp. 403–406, July 2007
6. Do, T., Yi, C., Nguyen, D., et al.: Distributed compressed video sensing. In: IEEE International conference on Image Processing (ICIP), Cario, Egypt, pp. 1393–1396. IEEE (2009)
7. Mun, S., Fowler, J.E.: Block compressed sensing of images using directional transforms. In: Data Compression Conference, p. 547. IEEE Computer Society (2010)
8. Fujimoto, T., Endo, R., Shigeno, H.: P2P video-on-demand streaming using caching and reservation scheme based on video popularity. Int. J. Grid Util. Comput. **3**(2/3), 188–199 (2012)
9. Tramel, E.W., Fowler, J.E.: Video compressed sensing with multihypothesis. In: Data Compression Conference, DBLP, pp. 193–202 (2011)
10. Zhu, J., Cao, N., Meng, Y.: Adaptive multihypothesis prediction algorithm for distributed compressive video sensing. Int. J. Distrib. Sens. Netw. **2013**(8), 718–720 (2013)
11. Chen, J., Wang, N., Xue, F., et al.: Distributed compressed video sensing based on the optimization of hypothesis set update technique. Multimedia Tools Appl. 1–20 (2016)
12. Ou, W.F., Yang, C.L., Li, W.H., et al.: A two-stage multi-hypothesis reconstruction scheme in compressed video sensing. In: IEEE International Conference on Image Processing, pp. 2494–2498. IEEE (2016)
13. Li, R., Liu, H., He, W., et al.: Space-time quantization and motion-aligned reconstruction for block-based compressive video sensing. KSII Trans. Internet Inf. Syst. **10**(1), 321–340 (2016)

Security Analysis and Improvements of Three-Party Password-Based Authenticated Key Exchange Protocol

Qingping Wang, Ou Ruan(✉), and Zihao Wang

School of Computer Science, Hubei University of Technology,
Wuhan, China
ruanou@163.com

Abstract. Three-party password-based authenticated key exchange (3PAKE) protocol allows two clients, each sharing a password with a trusted server, to establish a secret session key with the help of the server. It is a practical mechanism for establishing secure channels in the communication networks. Recently, Xu et al. proposed a 3PAKE protocol without the server's public key. They claimed that their protocol could withstand various attacks. In this paper, we show Xu et al.'s protocol is insecure against the stolen-verifier attack. Furthermore, we propose an improved 3PAKE protocol to overcome the weakness of Xu et al.'s protocol. Security and performance analysis shows that our protocol not only overcomes the security weakness, but also is more efficient. Therefore, our protocol is more suitable for the practical applications.

1 Introduction

Password-based authenticated key exchange (PAKE) protocols allow two or more specified parties to authenticate each other and establish a high-entropy secret session key by using only the weak, low-entropy and easily memorable passwords. This authenticated key exchange scheme is the most widely used in practice because no additional devices such as smart cards or hardware tokens is needed, but just a human-memorable password for authenticating the parties.

Bellovin and Merritt [1] first proposed a two-party PAKE protocol in 1992. The protocol allowed two parties to authenticate each other via a public, insecure network and establish a secure session key which is to be used for protecting their subsequent communication. Then, many efficient and practical PAKE protocols [2–6] have been proposed. The above two-party protocols were not scalable in a large-scale peer-to-peer system, since every pair of communication parties needs to share a password, so that each party in an n-party system has to maintain $n-1$ passwords [7]. To solve this problem, Three-party password-based authenticated key exchange (3PAKE) protocols were introduced [8–15]. However, these 3PAKE protocols still existed some security problems such as on-line undetectable password guessing attack [16] and off-line password guessing attack [10].

In order to increase the efficiency and preventing various attacks, in 2005 Lee et al. [17] proposed an efficient verifier-based key agreement protocol for three parties

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_49

without server's public key. Lee et al. claimed the proposed protocol could resist various attacks and provide the perfect forward secrecy. Wang et al. [18] pointed out that it would be more dangerous when suffers from the impersonation attack in 2006. After the defects of Lee-3PAKE protocol are discovered, there are a lot of improved protocols, which are based on the three-party authenticated key exchange protocol. Kwon J O et al. [19] designed a secure three-party password authentication key agreement protocol, but the communication cost and computation cost of the protocol were larger than [17]. Li et al. [20] proposed an efficient three-party password-based authenticated key exchange protocol based on bilinear pairings. Recently, Xu et al. [21] proposed an efficient 3PAKE according to the Lee-3PAKA protocol, combined with symmetric encryption.

In this paper, we show that Xu et al.'s scheme is vulnerable to the stolen-verifier attack. In addition, we propose an improved scheme to solve this problem. The protocol also enjoys low computational complexity and is suitable for resource-constrained devices.

The rest of this paper is organized as follows. In Sect. 2, we review Xu et al.'s scheme. In Sect. 3, a stolen-verifier attack against their scheme is described in details. In Sect. 4, we propose an improved scheme and the security and performance analyses are discussed in Sect. 5. The paper is concluded in Sect. 6.

2 Review of Xu et al.'s Protocol

This section revisits the 3PAKE protocol proposed by Xu et al. [21].

2.1 Notations

The notations used throughout this paper are summarized in Table 1.

Table 1. Notations for the proposed protocols

Notation	Description
A	Alice's public identity
B	Bob's public identity
S	Authentication Server's public identity
M	The attacker
pw	A weak password
E	Symmetric encryption
D	Symmetric decryption
a, b, c, d	Session-independent random numbers
p	A large prime
g	A generator g in the cyclic group Z_p^*
$\mathrm{H}(\cdot)$	A collision-resistant one-way hash function
$\oplus$	Bit-wise exclusive-OR (XOR) operation
K	A session key
V	Verifier computed from a password
c^{-1}	Inverse of c on Z_p^*

2.2 Protocol Description

For a detailed analysis, we review Xu et al.'s 3PAKE protocol [21]. The details of this protocol, shown in Fig. 1, are as follows:

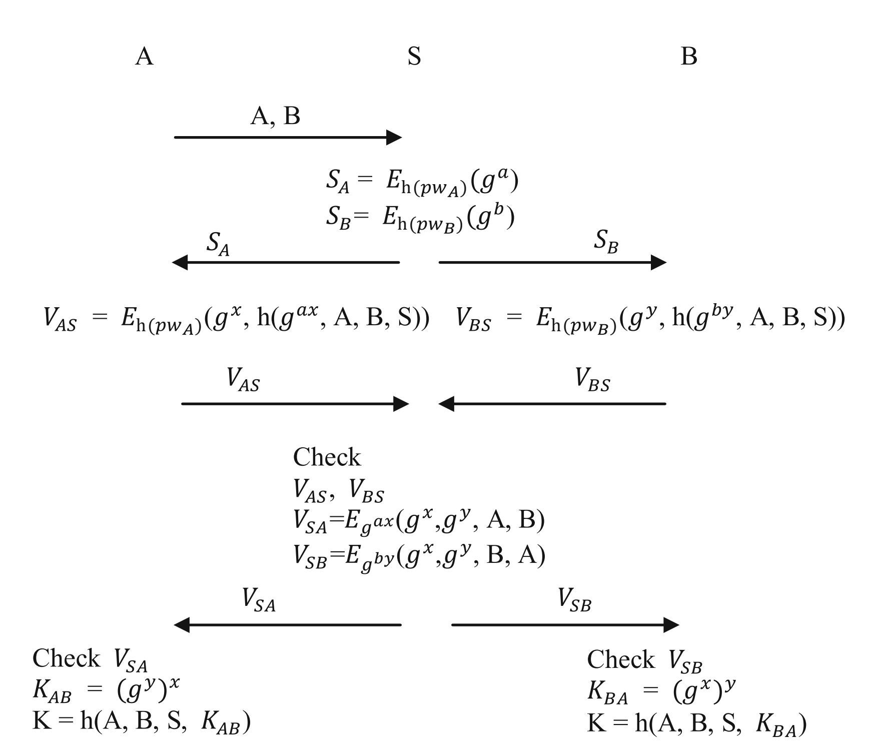


Fig. 1. Authentication and key exchange phase of Xu et al.'s protocol

Before the running of the protocol, Alice and Bob sends their verifiers $h(pw_A)$ and $h(pw_B)$ to S through a secure channel. S stores $h(pw_A)$ and $h(pw_B)$ in a password table.

Round 1: User A sends A and B to S.

$$A \rightarrow S : A, B.$$

Round 2: After receiving the messages sent by A, S randomly chooses a and b, computes $S_A = E_{h(pw_A)}(g^a)$, $S_B = E_{h(pw_B)}(g^b)$ and then sends S_A and S_B to A and B, respectively.

$$S : a, b \in_R Z_p^*$$

$$S : S_A = E_{h(pw_A)}(g^a).$$

$$S : S_B = E_{h(pw_B)}(g^b).$$

$$S \rightarrow A : S_A.$$

$$S \rightarrow B : S_B.$$

Round 3: After receiving the message sent by S, A computes $g^a = D_{\mathrm{h}(pw_A)}(S_A)$ and $V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ by choosing x $\in_R Z_p^*$, and sends V_{AS} to S. Similarly, after receiving the message from S, B computes $g^b = D_{\mathrm{h}(pw_B)}(S_B)$ and $V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{by}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ by choosing y $\in_R Z_p^*$, and sends V_{BS} to S.

$$\begin{aligned}
&\mathrm{A} : g^a = D_{\mathrm{h}(pw_A)}(S_A).\\
&\mathrm{A} :\ \mathrm{x} \in_R Z_p^*.\\
&\mathrm{A} : V_{AS} = E_{\mathrm{h}(pw_A)}\ (g^x,\ \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\mathrm{A} \rightarrow \mathrm{S} : V_{AS}\\
&\mathrm{B} : g^b = D_{\mathrm{h}(pw_B)}(S_B).\\
&\mathrm{B} : \mathrm{y} \in_R Z_p^*.\\
&\mathrm{B} : V_{BS} = E_{\mathrm{h}(pw_B)}\ (g^y,\ \mathrm{h}(g^{by}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\mathrm{B} \rightarrow \mathrm{S} : V_{BS}
\end{aligned}$$

Round 4: After receiving the messages sent by A and B, S checks whether $V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ and $V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{by}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ hold or not. If it holds, S computes $V_{SA} = E_{g^{ax}}(g^x, g^y, \mathrm{A}, \mathrm{B})$ and $V_{SB} = E_{g^{by}}(g^x, g^y, \mathrm{B}, \mathrm{A})$ and sends V_{SA} and V_{SB} to A and B, respectively. Otherwise S aborts the protocol.

$$\begin{aligned}
&\mathrm{S} : \mathrm{Checks}\\
&V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{ay}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\mathrm{S} : V_{SA} = E_{g^{ax}}(g^x, g^y, \mathrm{A}, \mathrm{B}).\\
&\mathrm{S} : V_{SB} = E_{g^{by}}(g^x, g^y, \mathrm{B}, \mathrm{A}).\\
&\mathrm{S} \rightarrow \mathrm{A} : V_{SA}.\\
&\mathrm{S} \rightarrow \mathrm{B} : V_{SB}.
\end{aligned}$$

Finally: After receiving the message sent by S, A checks whether $g^x \in (g^x, g^y, \mathrm{A}, \mathrm{B})$ hold or not, If it holds, A computes $K_{AB} = (g^y)^x$. Otherwise A aborts the protocol. Similarly, after receiving the message sent by S, B checks whether $g^y \in (g^x, g^y, \mathrm{B}, \mathrm{A})$ hold or not, If it holds, B computes $K_{BA} = (g^x)^y$. Otherwise B aborts the protocol. Finally, A and B compute a common session key K = h(A, B, S, K_{AB}) = h(A, B, S, K_{BA}) = h(A, B, S, g^{xy}), respectively.

$$\begin{aligned}
&\mathrm{A} : \text{Checks} \\
&\qquad g^x \in (g^x, g^y, \mathrm{A}, \mathrm{B}). \\
&\mathrm{A} : K_{AB} = (g^y)^x. \\
&\mathrm{A} : \mathrm{K} = \mathrm{h}(\mathrm{A}, \mathrm{B}, \mathrm{S}, K_{AB}). \\
&\mathrm{B} : \text{Checks} \\
&\qquad g^y \in (g^x, g^y, \mathrm{B}, \mathrm{A}). \\
&\mathrm{B} : K_{BA} = (g^x)^y \\
&\mathrm{B} : \mathrm{K} = \mathrm{h}(\mathrm{A}, \mathrm{B}, \mathrm{S}, K_{BA})
\end{aligned}$$

3 Attacks on Xu et al.'s 3PAKE Protocol

In this section, we show that Xu et al.'s 3PAKE is vulnerable to stolen-verifier attack.

Through the security analysis of the Xu-3PAKE protocol, the author points out that the protocol provides forward security and resist man-in-the-middle attack, Denning-Sacco attack, password guessing attack, stolen-verifier attack and replay attack. Among them, the author claims that the protocol cannot be directly impersonate the user when the adversary obtains the authentication value of a user's password on the server, but in fact it still cannot resist the attack of the stolen-verifier.

According to the security model proposed by Dolev and Yao [23], an active attacker can control the communication channels through intercepting the communication and inserting data into the channels. Below are the details of our attacks.

The attack of the stolen-verifier:

We assume that *M* is an attacker who has got *A*'s verifier V_A. *M* can impersonate *A* to communicate with *B* by performing the following steps.

Round 1: Like the normal interaction, M sends S the message(A, B).

$$\mathrm{M} \rightarrow \mathrm{S} : \mathrm{A}, \mathrm{B}.$$

Round 2: After receiving the messages sent by M, S randomly chooses a and b, computes $S_A = E_{\mathrm{h}(pw_A)}(g^a)$, $S_B = E_{\mathrm{h}(pw_B)}(g^b)$ and then sends S_A and S_B to A and B, respectively. But S_A is intercepted by M.

$$\begin{aligned}
&\mathrm{S} : \mathrm{a}, \mathrm{b} \in_R Z_p^*. \\
&\mathrm{S} : S_A = E_{\mathrm{h}(pw_A)}(g^a). \\
&\mathrm{S} : S_B = E_{\mathrm{h}(pw_B)}(g^b). \\
&\mathrm{S} \rightarrow \mathrm{M} : S_A. \\
&\mathrm{S} \rightarrow \mathrm{B} : S_B.
\end{aligned}$$

Round 3: After intercepting the message in Round 2, M computes $g^a = D_{\mathrm{h}(pw_A)}(S_A)$ and $V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ by choosing x $\in_R Z_p^*$, and sends V_{AS} to S. After receiving the message from S, B computes $g^b = D_{\mathrm{h}(pw_B)}(S_B)$ and $V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{by}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ by choosing y $\in_R Z_p^*$, and sends V_{BS} to S.

$$\begin{aligned}
&\mathrm{M}: g^{a} = D_{\mathrm{h}(pw_A)}(S_A).\\
&\mathrm{M}: \mathrm{x} \in_R \ Z_p^*.\\
&\mathrm{M}: V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\mathrm{M} \to \mathrm{S}: V_{AS}.\\
&\mathrm{B}: g^{b} = D_{\mathrm{h}(pw_B)}(S_B).\\
&\mathrm{B}: \mathrm{y} \ \in_R \ Z_p^*.\\
&\mathrm{B}: V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{by}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\mathrm{B} \to \ \ \mathrm{S}: V_{BS}.
\end{aligned}$$

Round 4: After receiving the messages sent by M and B, S checks whether $V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ and $V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{by}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ hold or not. If it holds, S computes $V_{SA} = E_{g^{ax}}(g^x, g^y, \mathrm{A}, \mathrm{B})$ and $V_{SB} = E_{g^{by}}(g^x, g^y, \mathrm{B}, \mathrm{A})$ and sends V_{SA} and V_{SB} to A and B, respectively, But V_{SA} is intercepted by M. Otherwise S aborts the protocol.

$$\begin{aligned}
&\mathrm{S}: \text{Checks}\\
&\qquad V_{AS} = E_{\mathrm{h}(pw_A)}(g^x, \mathrm{h}(g^{ax}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\qquad V_{BS} = E_{\mathrm{h}(pw_B)}(g^y, \mathrm{h}(g^{ay}, \mathrm{A}, \mathrm{B}, \mathrm{S})).\\
&\mathrm{S}: V_{SA} = E_{g^{ax}}(g^x, g^y, \mathrm{A}, \mathrm{B}).\\
&\mathrm{S}: V_{SB} = E_{g^{by}}(g^x, g^y, \mathrm{B}, \mathrm{A}).\\
&\mathrm{S} \to \mathrm{M}: V_{SA}.\\
&\mathrm{S} \to \mathrm{B}: V_{SB}.
\end{aligned}$$

Round 5: After intercepting the message in Round 4, M checks whether $g^x \in$ (g^x, g^y, A, B) hold or not, If it holds, M computes $K_{AB} = (g^y)^x$. Otherwise M aborts the protocol. After receiving the message sent by S, B checks whether $g^y \in$ (g^x, g^y, B, A) hold or not, If it holds, B computes $K_{BA} = (g^x)^y$. Otherwise B aborts the protocol. Then, M and B compute a common session key K = h(A, B, S, K_{AB}) = h(A, B, S, K_{BA}) = h(A, B, S, g^{xy}), respectively. Finally, B believes the common session key K = h (A, B, S, K_{AB}) is true. B also believes that he communicate with A. In fact, M gets the session key K = h(A, B, S, K_{AB}) and impersonates A to communicate with B.

$$\begin{aligned}
&\mathrm{M}: \text{Checks}\\
&\qquad g^x \in (g^x, g^y, \mathrm{A}, \mathrm{B}).\\
&\mathrm{M}: K_{AB} = (g^y)^x.\\
&\mathrm{M}: \mathrm{K} = \mathrm{h}(\mathrm{A}, \mathrm{B}, \mathrm{S}, K_{AB})\\
&\mathrm{B}: \text{Checks}\\
&\qquad g^y \in (g^x, g^y, \mathrm{B}, \mathrm{A}).\\
&\mathrm{B}: K_{BA} = (g^x)^y\\
&\mathrm{B}: \mathrm{K} = \mathrm{h}(\mathrm{A}, \mathrm{B}, \mathrm{S}, K_{BA})
\end{aligned}$$

4 Improved Scheme

In this section, we present an enhanced protocol to remedy the security loopholes existing in Xu et al.'s protocol. The protocol depicted in Fig. 2 works as follows:

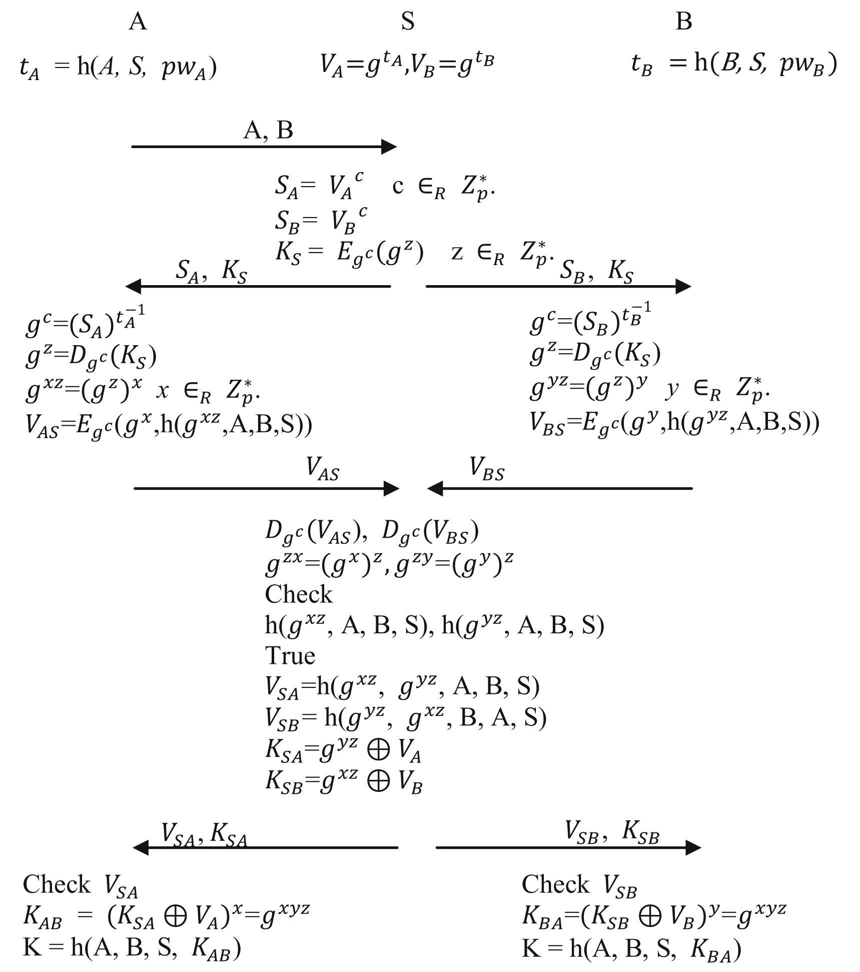

A
S
B
$t_A = h(A, S, pw_A)$
$V_A = g^{t_A}, V_B = g^{t_B}$
$t_B = h(B, S, pw_B)$
A, B
$S_A = V_A{}^c \quad c \in_R Z_p^*.$
$S_B = V_B{}^c$
$K_S = E_{g^c}(g^z) \quad z \in_R Z_p^*.$
S_A, K_S
S_B, K_S
$g^c = (S_A)^{t_A^{-1}}$
$g^z = D_{g^c}(K_S)$
$g^{xz} = (g^z)^x \quad x \in_R Z_p^*.$
$V_{AS} = E_{g^c}(g^x, h(g^{xz}, A, B, S))$
$g^c = (S_B)^{t_B^{-1}}$
$g^z = D_{g^c}(K_S)$
$g^{yz} = (g^z)^y \quad y \in_R Z_p^*.$
$V_{BS} = E_{g^c}(g^y, h(g^{yz}, A, B, S))$
V_{AS}
V_{BS}
$D_{g^c}(V_{AS}), \ D_{g^c}(V_{BS})$
$g^{zx} = (g^x)^z, g^{zy} = (g^y)^z$
Check
$h(g^{xz}, A, B, S), h(g^{yz}, A, B, S)$
True
$V_{SA} = h(g^{xz}, g^{yz}, A, B, S)$
$V_{SB} = h(g^{yz}, g^{xz}, B, A, S)$
$K_{SA} = g^{yz} \oplus V_A$
$K_{SB} = g^{xz} \oplus V_B$
V_{SA}, K_{SA}
V_{SB}, K_{SB}
Check V_{SA}
$K_{AB} = (K_{SA} \oplus V_A)^x = g^{xyz}$
$K = h(A, B, S, K_{AB})$
Check V_{SB}
$K_{BA} = (K_{SB} \oplus V_B)^y = g^{xyz}$
$K = h(A, B, S, K_{BA})$

Fig. 2. The proposed protocol

Before the running of the protocol, Alice and Bob sends their verifiers V_A and V_B to S through a secure channel. S stores V_A and V_B in a password table.

Step 1: User A sends A and B to S.

$$\mathrm{A} \rightarrow \mathrm{S} : \mathrm{A}, \mathrm{B}.$$

Step 2: After receiving the messages sent by A, S randomly chooses z and c, computes $S_A = V_A^c$, $S_B = V_B^c$, $K_S = E_{g^c}(g^z)$ and then sends (S_A, K_S) and (S_B, K_S) to A and B, respectively.

$$\begin{aligned}
&\mathrm{S} : \mathrm{z},\ \mathrm{c} \in_R Z_p^*.\\
&\mathrm{S} : S_A = V_A^c.\\
&\mathrm{S} : S_B = V_B^c.\\
&\mathrm{S} : K_S = E_{g^c}(g^z).\\
&\mathrm{S} \rightarrow \mathrm{A} : S_A, K_S.\\
&\mathrm{S} \rightarrow \mathrm{B} : S_B, K_S.
\end{aligned}$$

Step 3: After receiving the message sent by S, A computes $g^c = (S_A)^{t_A^{-1}}$, $g^z = D_{g^c}(K_S)$, $g^{xz} = (g^z)^x$, and $V_{AS} = E_{g^c}(g^x, \mathrm{h}(g^{xz}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ by choosing x $\in_R Z_p^*$, and sends V_{AS} to S. Similarly, after receiving the message from S, B computes $g^c = (S_B)^{t_B^{-1}}$, $g^z = D_{g^c}(K_S)$, $g^{yz} = (g^z)^y$ and $V_{BS} = E_{g^c}(g^y, \mathrm{h}(g^{yz}, \mathrm{A}, \mathrm{B}, \mathrm{S}))$ by choosing y $\in_R Z_p^*$, and sends V_{BS} to S. Note that t_A = h(A, S, pw_A) and t_B = h(B, S, pw_B).

$$\begin{aligned}
&\mathrm{A} : g^c = (S_A)^{t_A^{-1}}.\\
&\mathrm{A} : g^z = D_{g^c}(K_S).\\
&\mathrm{A} : g^{xz} = (g^z)^x\ x \in_R Z_p^*.\\
&\mathrm{A} : V_{AS} = E_{g^c}(g^x, \mathrm{h}(g^{xz}, \mathrm{A}, \mathrm{B}, \mathrm{S}))\\
&\mathrm{A} \rightarrow \mathrm{S} : V_{AS}.\\
&\mathrm{B} : g^c = (S_B)^{t_B^{-1}}.\\
&\mathrm{B} : g^z = D_{g^c}(K_S).\\
&\mathrm{B} : g^{yz} = (g^z)^y\ y \in_R Z_p^*.\\
&\mathrm{B} : V_{BS} = E_{g^c}(g^y, \mathrm{h}(g^{yz}, \mathrm{A}, \mathrm{B}, \mathrm{S}))\\
&\mathrm{B} \rightarrow \mathrm{S} : V_{BS}.
\end{aligned}$$

Step 4: After receiving the messages sent by A and B, S computes $g^{xz} = (g^x)^z$, $g^{yz} = (g^y)^z$ by $D_{g^c}(V_{AS})$ and $D_{g^c}(V_{BS})$ and verifies whether h(g^{zx}, A, B, S) = h(g^{xz}, A, B, S), h(g^{zy}, A, B, S) = h(g^{yz}, A, B, S) or not. If they hold, S computes and sends V_{SA} = $\mathrm{h}(g^{xz}, g^{yz}, \mathrm{A}, \mathrm{B}, \mathrm{S})$, $K_{SA} = g^{yz} \oplus V_A$. and V_{SB} = $\mathrm{h}(g^{yz}, g^{xz}, \mathrm{B}, \mathrm{A}, \mathrm{S})$, $K_{SB} = g^{xz} \oplus V_B$. to A and B, respectively. Otherwise, S terminates the protocol.

$$\begin{aligned}
&\mathrm{S}: D_{g^c}(V_{AS}),\ D_{g^c}(V_{BS}).\\
&\mathrm{S}: g^{xz} = (g^x)^z\ g^{yz} = (g^y)^z\\
&\mathrm{S}: \text{Check}\\
&\qquad \mathrm{h}(g^{zx}, \mathrm{A}, \mathrm{B}, \mathrm{S}) = \mathrm{h}(g^{xz}, \mathrm{A}, \mathrm{B}, \mathrm{S})\\
&\qquad \mathrm{h}(g^{zy}, \mathrm{A}, \mathrm{B}, \mathrm{S}) = \mathrm{h}(g^{yz}, \mathrm{A}, \mathrm{B}, \mathrm{S})\\
&\text{True.}\\
&\mathrm{S}: V_{SA} = \mathrm{h}(g^{xz}, g^{yz}, \mathrm{A}, \mathrm{B}, \mathrm{S}),\ K_{SA} = g^{yz} \oplus V_A.\\
&\mathrm{S}: V_{SB} = \mathrm{h}(g^{yz}, g^{xz}, \mathrm{B}, \mathrm{A}, \mathrm{S}),\ K_{SB} = g^{xz} \oplus V_B.\\
&\mathrm{S} \rightarrow \mathrm{A}: V_{SA}, K_{SA}.\\
&\mathrm{S} \rightarrow \mathrm{B}: V_{SB}, K_{SB}.
\end{aligned}$$

Finally: After receiving the message sent by S, A computes $g^{yz} = K_{SA} \oplus V_A$, than verifies whether h(g^{xz}, g^{yz}, A, B, S) = V_{SA} or not. If it holds A computes $K_{AB} = (K_{SA} \oplus V_A) = g^{xyz}$ and K = h(A, B, S, K_{AB}). Otherwise, A terminates the protocol. Similarly, After receiving the message sent by S, B computes $g^{xz} = K_{SB} \oplus V_B$, than verifies whether h(g^{yz}, g^{xz}, B, A, S) = V_{SB} or not. If it holds B computes $K_{BA} = (K_{SB} \oplus V_B)^y = g^{xyz}$ and K = h(A, B, S, K_{BA}). Otherwise, B terminates the protocol. Finally, Alice and Bob negotiate a common session key K = h(A, B, S, K_{AB}) = h(A, B, S, K_{BA}).

5 Security Analysis and Performance Comparison

5.1 Security Analysis

In this section, we prove the security of 3PAKE using those definitions in [22].

Theorem 1. The proposed protocol provides the property of the perfect forward secrecy.

Proof. Perfect forward secrecy is provided in the situation that even though a password is compromised M cannot derive previous session keys. To analyze this, suppose that M knows the password pw Then M tries to find previous session keys from the information collected by passive attack in past communication sessions, i.e., $K_S = E_{g^c}(g^z)$, g^x, g^y, g^{yz}, g^{xz}. However, she cannot do these using them without solving DLP and DHP. Therefore, the proposed protocol provides the property of perfect forward secrecy.

Theorem 2. The proposed protocol is secure against the Denning-Sacco attack.

Proof. To be secure against the Denning-Sacco attack, the protocol should be designed such that even though a session key is compromised, M cannot compute the password and confirm the correctness of the guessed password. To analyze this, suppose that M knows a session key K = h(A, B, S, K_{AB}). Then M tries to compute the password or confirm the correctness of the guessed password from it and the information collected by passive attack in past communication sessions, i.e., g^x, g^y, g^{yz}, g^{xz}, h(A, B, S, K_{AB}).

However, M cannot do these using them without solving DLP and DHP. Therefore, PAKE is secure against the Denning Sacco attack.

Theorem 3. The proposed protocol is secure against stolen-verifier attack.

Proof. The protocol being secure against stolen-verifier attack means an attacker not being able to pose as a client after compromising the server. In the proposed protocol, if M gains password file, M may know two client's verifiers $V_A = g^{\mathrm{h}(A,S,pw_A)}$ and $V_B = g^{\mathrm{h}(A,S,pw_B)}$. However, M cannot pose as the clients because of not knowing t_A = h (A, S, pw_A) and t_B = h(A, S, pw_B) used in step 3. Therefore, the proposed protocol is secure against server compromise.

Theorem 4. The proposed protocol is secure against man-in-the-middle attack.

Proof. We analyze if a malicious insider M can succeed in launching man-in-the-middle attack. Suppose that M tries to masquerade A or B. However, S can detect this attack when verifying $V_{AS} = E_{g^c}(g^x$, h(g^{xz}, A, B, S)) and $V_{BS} = E_{g^c}(g^y$, h(g^{yz}, A, B, S)). M cannot compute the valid g^{xz} or g^{yz} due to not knowing their correct passwords. Therefore, the improved scheme can resist man-in the-middle attack.

5.2 Efficiency Analysis

Performance of key agreement protocols can be approximated in terms of communication and computation loads. We compare our improved 3PAKE with the protocol of Xu et al. Table 2 shows the comparison regarding with several efficiency factors such as the number of rounds, random numbers, exponentiations, symmetric encryption/decryption, hash functions.

Table 2. The performance comparison

	Xu et al.			Our scheme		
	A	B	S	A	B	S
Random number	1	1	2	1	1	1
Exponentiation	3	3	4	4	4	6
Sym. enc./dec.	3	3	6	2	2	3
Hash function	2	2	2	3	3	4
Round	4			4		

As shown in Table 2, for user A and B, our scheme has one more exponentiation operation and one more hash operation than Xu et al.'s scheme, but our scheme has one less symmetric encryption/decryption computations than Xu et al.'s scheme. For server S our scheme has two more exponentiation operations and two more hash operation than Xu et al.'s scheme, but our scheme has three less symmetric encryption/decryption computations than Xu et al.'s scheme. Usually the cost of symmetric encryption/decryption is much larger than the cost of exponentiation operation (160bit) and hash operation. Thus, our protocol has better performance than Xu et al.'s protocol. Moreover, Xu et al.'s protocol is vulnerable to the stolen-verifier attack and our protocol could overcome such weakness. Therefore, our protocol is more suitable for the practical applications.

6 Conclusion

In this paper, we show that Xu et al.'s 3PAKE protocol is vulnerable to the stolen-verifier attack and propose a new 3PAKE protocol to solve this problem. Security and performance analysis show our protocol overcome the weakness in Xu et al.'s protocol and has better performance. One of our future works is to extend our new scheme to multi-server architecture for the distributed systems.

Acknowledgments. The work was supported by the Educational Commission of Hubei Province of China (No. D20151401) and the Green Industry Technology Leading Project of Hubei University of Technology (No. ZZTS2017006).

References

1. Bellovin, S.M., Merritt, M.: Encrypted key exchange: password based protocols secure against dictionary attacks. In: Proceedings of IEEE Symposium on Research in Security and Privacy, pp. 72–84 (1992)
2. Ruan, O., Kumar, N., He, D.B., Lee, J.H.: Efficient provably secure password-based explicit authenticated key agreement. Pervasive Mob. Comput. **24**(12), 50–60 (2015)
3. Yi, X., Rao, F.Y., Tari, Z., Hao, F.: ID2S password-authenticated key exchange protocols. IEEE Trans. Comput. **65**, 1–14 (2016)
4. Lu, Y., Zhang, Q., Li, J., Shen, J.: Comment on a certificateless one-pass and two-party authenticated key agreement protocol. Inf. Sci. **369**, 184–187 (2016)
5. Zhang, L.: Certificateless one-pass and two-party authenticated key agreement protocol and its extensions. Inf. Sci. **293**(1), 182–195 (2015)
6. Farash, M.S., Islam, S.H., Obaidat, M.S.: A provably secure and efficient two-party password-based explicit authenticated key exchange protocol resistance to password guessing attacks. Concurrency Comput. Prac. Experience **27**(17), 4897–4913 (2015)
7. Xie, Q., Dong, N., Tan, X., et al.: Improvement of a three-party password-based key exchange protocol with formal verification. Inf. Technol. Control **42**(3), 231–237 (2013)
8. Chang, C.-C., Cheng, Y.-F.: A novel three-party encrypted key exchange protocol. Comput. Stan. Interfaces **26**(5), 471–476 (2004)
9. Lee, T.-F., Hwang, T., Lin, C.-L.: Enhanced three-party encrypted key exchange without server public keys. Comput. Secur. **23**, 571–577 (2004)
10. Lin, C.-L., Sun, H.-M., Hwang, T.: Three-party encrypted key exchange: attacks and a solution. ACM Operating Syst. Rev. **34**(4), 12–20 (2000)
11. Sun, H.-M., Chen, B.-C., Hwang, T.: Secure key agreement protocols for three-party against guessing attacks. J. Syst. Softw. **75**(1–2), 63–68 (2005)
12. Islam, S.H.: Design and analysis of a three party password-based authenticated key exchange protocol using extended chaotic maps. Inf. Sci. **312**(C), 104–130 (2015)

13. Amin, R., Biswas, G.P.: Cryptanalysis and design of a three-party authenticated key exchange protocol using smart card. Arab. J. Forence Eng. **40**(11), 1–15 (2015)
14. Lu, C.F.: Multi-party password-authenticated key exchange scheme with privacy preservation for mobile environment. Ksii Trans. Internet Inf. Syst. **9**(12), 5135–5149 (2015)
15. Nam, J., Paik, J., Kim, J., Lee, Y., Won, D.: Server-aided password-authenticated key exchange: from 3-party to group. In: International Conference on Human Interface & The Management of Information, vol. 6771, pp. 339–348 (2011)
16. Ding, Y., Horster, P.: Undetectable on-line password guessing attack. ACM SIGOPS Operating Syst. Rev. **29**(4), 77–86 (1995)
17. Lee, S.W., Kim, H.S., Yoo, K.Y.: Efficient verifier-based key agreement protocol for three parties without server's public key. Appl. Math. Comput. **167**(2), 996–1003 (2005)
18. Wang, R.C., Mo, K.R.: Security enhancement on efficient verifier-based key agreement protocol for three parties without server's public key. Int. Math. Forum **1**(17–20), 965–972 (2006)
19. Kwon, J.O., Jeong, I.R., Sakurai, K., et al.: Efficient verifier-based password-authenticated key exchange in the three-party setting. Comput. Stand. Interfaces **29**(5), 513–520 (2007)
20. Li, W., Wen, Q., Zhang, H.: Verifier-based password-authenticated key exchange protocol for three-party. J. Commun. **29**(10), 149–152 (2008)
21. Xu, et al.: Efficient three-party password-based authenticated key exchange protocol. J. Univ. Electron. Sci. Technol. China **41**(4), 596–598 (2012)
22. Lee, S.W., Kim, W.H., Kim, H.S., et al.: Efficient password-based authenticated key agreement protocol. Lecture Notes in Computer Science, pp. 617–626 (2004)
23. Dolev, D., Yao, A.C.: On the security of public key protocols. IEEE Trans. Inf. Theory **29**, 198–208 (1983)

A Combined Security Scheme for Network Coding

Chao Xu, Yonghui Chen(✉), Hui Yu, Yan Jiang, and Guomin Sun

Department of Computer Science, Hubei University of Technology,
Wuhan, Hubei, China
Hg_cyh@mail.hbut.edu.cn

Abstract. Network coding is theoretically the most efficient coding scheme for decentralized networks with better throughput and better robustness. However, if malicious intermediate nodes launch pollution attacks to the data by triumphantly forging network code, the sink node would suffer from failed decoding with bandwidth wasting, longer delay and more overheads. The classic bit by bit digital signature schemes are elegant, but the computation complexity is high, for each bit have to execute a hash computation. The pollution detection schemes based on null key cannot against colluding attacks. The schemes based on homomorphic MAC ensure the sink nodes verify the data, but those intermediate nodes cannot detect the pollution attacks. Above schemes are not enough efficient. In this paper, we propose a new combined security network coding scheme based on homomorphic MAC and null key that overcome the shortage of each other.

1 Introduction

In 2000, Network coding firstly proposed by Ahlswede et al. [1]. That can theoretically achieve the maximum multicast throughput by performing coding operations on the contents of packets rather than the traditional technique of replication and forwarding. And Li et al. [2] has proved that linear network coding could achieve the optimal throughput in multicast networks. Network coding has many killer applications, such as content distribution [3], wireless networks [4] and Distributed Storage [5], etc.

However, these salient advantages of network coding are reduced by those malicious pollution attacks. Consider a scenario in which a data center is distributing a file to a set of clients via a P2P network. If any intermediate node is malicious and capable, generates corrupted packets and contributes them back to the network. A single corrupted packet can result in tens or even hundreds of polluted ones. This may cause legitimate client unable to recover the file properly.

Recently, great research efforts have been devoted to securing network coding against pollution attacks. The existing secure schemes are categorized into two kinds: information-theoretic schemes [6, 7] and cryptography-based schemes [8–19]. The existing information-theoretic schemes have the advantage of not relying on any computational assumptions, but are limited to offer pollution detection/correction at receivers in an end-to-end fashion. So that it cannot effectively helps the forwarders to prevent the transmission of useless polluted packets. The cryptography-based scheme

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_50

has two categories: public key encryption [8–12], symmetric key encryption [13–16, 20]. The first category is provably secure under the hardness assumptions of well-known cryptographic problems, but the computation overhead is high. The second category involve symmetric key encryption that the computationally is efficient, but the disadvantage is that the secret must be managed carefully and the bandwidth overhead is large.

Existing cryptographic schemes are trying to provide a way for honest nodes to verify authenticity of individual packets. A good scheme should has the additional advantage that intermediate nodes in the network can verify correctness of individual packets, so that corrupted packets can be discarded before polluting those good ones.

In this paper we only consider the cryptography-based schemes. Inspired by [14, 16, 21], we present a combined scheme to prevent the network coding from being polluted in typical multicast networks. The source node computes a signature of MAC for each package. Then calculate a null space that is orthogonal to the augmented and signed source packets. Intermediate nodes check the received packets by verifying null key. The sink nodes verify the MAC signature by using the secret key. Therefore both schemes could overcome the shortage of each other in our scheme.

The rest of this paper is organized as follows. Section 2 introduces the system model. Section 3 shows our schemes. Section 4 is the security analysis. Section 5 is comparison. Section 6 is conclusion.

2 Network Model and Operation

2.1 Network Model

We take a typical multicast scenario, which is easy to be extended to many applications, where a source node S divides the original data into several n-length packets $\bar{\mathbf{v}}_1, \bar{\mathbf{v}}_2, \ldots, \bar{\mathbf{v}}_m$ and then delivers those packets to receivers $\{\mathbf{R}_i\}$, if without encryption. Each packet $\bar{\mathbf{v}}_i$ denotes a vector $(\bar{\mathbf{v}}_{i,1}, \bar{\mathbf{v}}_{i,2}, \ldots, \bar{\mathbf{v}}_{i,n})$ in finite field F_q^n, q is a prime. To encrypt with network coding, each $\bar{\mathbf{v}}_i$ should be augmented as packet $\mathbf{v}_i$ by ith unit vector of dimension m as in [2, 3, 6–8, 12, 14–19, 22–25]:

$$\mathbf{v}_i = \left(\bar{\mathbf{v}}_1, \overbrace{\underbrace{0,\ldots,0,1}_{i},0,\ldots 0}^{m}\right) \in F_q^{m+n} \tag{1}$$

Let V denote the subspace spanned by $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m\}$, thus $\mathbf{v}_i$ is the ith vector of V. While S sends vectors in V to receivers $\{\mathbf{R}_i\}$, the intermediate nodes, in network coding networks, are responsible for replicating the vectors.

For random network coding the source sends linear combinations of packets using randomly selected coefficients. For example, the source packets $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_l\}$, may be linear combined by using coefficient $\alpha_1, \alpha_2, \ldots, \alpha_l \in F_q$.

$$w_{\mathrm{j}} = \sum\nolimits_{\mathrm{i}=1}^{l} \alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}} \tag{2}$$

The coding coefficients can derived from last m symbols of w_{j}. Intermediate nodes linearly combine their received packets for output in a similar way. Then a receiver $\mathbf{R}_{\mathrm{i}}$ can recover V exactly after receiving more than m linearly independent packets.

2.2 Adversary Model

We assume that the source is always secured, but the forwarders are not trustable. The adversaries can fully control the compromised forwarders and launch the pollution attacks from them. In such attacks, the adversaries may intentionally pollute the output packets of the compromised nodes, or directly inject the forged packets into the systems. It is serious that a small number of polluted packets can quickly propagate in the systems and infect a large proportion of nodes. When a forwarder receives a polluted packet, all of its output packets will be polluted. Then, these polluted packets are further used by downstream forwarders for encoding, thus, more and more packets will be polluted. So, it is necessary to filter the polluted packets as early as possible.

In this paper, we try to design an efficient scheme for resisting pollution attacks in network coding system. The scheme should ensure the forwarders to detect and resist the collusion attacks.

3 Our Scheme

3.1 Basic Idea

Homomorphic MAC Scheme. In the homomorphic MAC scheme, the source calculates a tag for each of the original message vectors. Intermediate nodes compute valid tags for random linear combinations using the homomorphic property. Then the sink nodes verify the tags of received vectors and drop all vectors with an invalid tag. Because using symmetrical keys, only the source node and sink nodes have secret key, if intermediate nodes is able to verify tags before forwarding them on to other nodes, they must have a shared secret key with each network node. This is unacceptable. Hence, intermediate nodes cannot check the integrity of each packet using this scheme.

Null Keys Scheme. The null space of the data V, denoted as $\Pi_{\mathrm{V}}^{\perp}$, satisfy $\mathrm{V} \cdot \Pi_{\mathrm{V}}^{\perp} = 0$. With the rank-nullity theorem:

$$\mathrm{rank}(\mathrm{V}) + \mathrm{nullity}(\mathrm{V}) = \mathrm{n} \tag{3}$$

The dimension of the null space of V, $\Pi_{\mathrm{V}}^{\perp}$, is nullity(V). For a $\mathrm{m} \times (\mathrm{m}+\mathrm{n})$ data packet matrix V, the number of linearly independent data packets the source distributing at least is m, from the system model, while the number of linearly independent null keys needed to construct complete null space is n. The source provides the participating nodes with keys to perform the verification process. It sends a random linear combination of the vectors $\Pi_{V}^{\perp}$ on each of its following nodes destined to all the participants including the malicious nodes. Each intermediate node verifies the received blocks using the orthogonality principle. When an intermediate node receives at least one key from its incoming nodes, he may form the null keys K_i to check the received packets $\boldsymbol{\omega}$ whether they satisfy $K_i \cdot \boldsymbol{\omega}^T = 0$.

Because the malicious nodes have access to null keys, they also can forge a data block $\boldsymbol{\omega}$ that satisfies $K_i \cdot \boldsymbol{\omega}^T = 0$ but do not belong to original data. So this scheme cannot resist pollution completely.

Our Idea. Based on those schemes, we present a combined scheme for network coding as shown in Fig. 1. Similar to [21], we pad every augmented source packet with a MAC signature and then find out the null space of the augmented and signed source packets. To verify a packet a relay node just checks whether the packet maps the null space to zero. The MAC signature will be used in the sink nodes to verify the integrity of received packets.

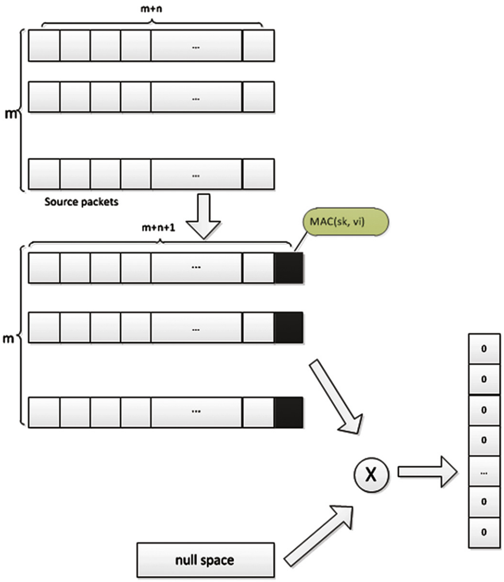


Fig. 1. A combined scheme

3.2 Define Our Scheme

Our scheme is defined as a tuple of five probabilistic polynomial-time (PPT) algorithms: Setup, Sign, Generate, Combine and Verify.

Setup. Input the k-length security parameter 1^k, the system parameters m and n for encoding and authentication, as in system model; Output a prime number q and a secret key sk.

Sign. Input a secret key sk, a vector space identifier id, an augmented vector $\mathbf{v}_i$, $\mathbf{v}_i \in F_q^{n+m}(i = 1, 2, \ldots, m)$; Output $\mathbf{v}_i^* = (\mathbf{v}_i, t_i) \in F_q^{m+n+b}$, where $t_i \in F_q$ is termed as the signature of $\mathbf{v}_i$, b is the length of t_i.

Generate. Input $V^* = \left(\mathbf{v}_1^{*T}, \mathbf{v}_2^{*T}, \ldots, \mathbf{v}_m^{*T}\right)^T$, $V^{*\perp}$, the null space of V^*.

Combine. Input l vectors $\mathbf{v}_1^*, \ldots, \mathbf{v}_l^*$ and l coefficients $\alpha_1, \alpha_2, \ldots, \alpha_l$, where $\alpha_i \in F_q$; Output a vector $\mathbf{y} = (\sum_{i=1}^{l} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{l} t_i)$.

Verify. The intermediate nodes and the sink nodes have different verification. The detailed operation is as follows:

(1) For the intermediate nodes: Input null key K_i, a vector $\mathbf{y} = (\sum_{i=1}^{l} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{l} t_i)$; Output either 0 (accept), or 1 (reject).

(2) For the sink nodes: Input a secret key sk, an identifier id, a vector $\mathbf{y} = (\sum_{i=1}^{m} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{m} t_i)$, null key K_i; Output: either 0 (accept), or 1 (reject).

Our scheme is correct if the following conditions are satisfied.

(1) $\mathbf{Verify}\left(K_i \cdot (\sum_{i=1}^{l} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{l} t_i)\right) = 0$, for the intermediate nodes;

(2) $\mathbf{Verify}(\mathrm{MAC}(sk, id, \sum_{i=1}^{m} \alpha_i \mathbf{v}_i^*), \sum_{i=1}^{m} t_i) = 0$ and $\mathrm{Verify}(K_i \cdot \sum_{i=1}^{m} \alpha_i \mathbf{v}_i^*) = 0$, for the sink nodes.

Attack Game: Our scheme is secure if for any probability, polynomial-time (PPT) adversary A, the probability that A wins the security game defined below is negligible.

Setup. The challenger C generates a random key $k \leftarrow F_q^{m+n}$.

Query. A adaptively submits vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ to challenger C, who runs **Sign** for these vectors and sends the corresponding $\mathbf{v}_1'^*, \mathbf{v}_2'^*, \ldots, \mathbf{v}_m'^*$ to A, where $\mathbf{v}_i'^* = \left(\mathbf{v}_i', t_i'\right)(i = 1, 2, \ldots m)$.

Output. A generates a vector $\boldsymbol{y}^* = (\sum_{i=1}^{l} \alpha_i \mathbf{v}_i', \sum_{i=1}^{l} t_i')$ with $\boldsymbol{y}^* \notin$ span $\left(\mathbf{v}_i^*, \mathbf{v}_2^*, \ldots, \mathbf{v}_m^*\right)$.

(1) For intermediate nodes, if $\mathbf{Verify}(K_i \cdot \boldsymbol{y}^{*T}) = 0$ then A wins, otherwise A loses.

(2) For the sink nodes, if $\mathbf{Verify}\left(\mathrm{MAC}\left(sk, id, \sum_{i=1}^{m} \alpha_i \mathbf{v}_i'\right), \sum_{i=1}^{m} t_i\right) = 0$ and $\mathbf{Verify}(K_i \cdot \mathbf{y}^{*T}) = 0$ then A wins, otherwise A loses.

3.3 Construction Our Scheme

Based on the above definition, we present our construction scheme.

Setup. Given 1^k, m, n; Output secret key sk, a prime number q.

Sign. Given $\mathbf{v}_i \in F_q^{m+n}(i = 1, 2, \ldots, m)$, id, sk; Output $\mathbf{v}_i^* = (\mathbf{v}_i, t_i) \in F_q^{m+n+b}$, $t_i = \mathrm{MAC}(sk, id, \mathbf{v}_i)$, $t \in F_q$, b is the length of the t_i.

Generate. Given $V^* = \left(\mathbf{v}_1^{*T}, \mathbf{v}_2^{*T}, \ldots, \mathbf{v}_m^{*T}\right)^T$; Output $V^{*\perp}$, the null space of V^*, where $\dim(V^{*\perp}) = n + b$, $V^{*\perp} = (Z_1, Z_2, \ldots, Z_{n+b})$

Combine. Given l vectors $\mathbf{v}_1^*, \ldots, \mathbf{v}_l^*$ and l random coefficients $\alpha_1, \alpha_2, \ldots, \alpha_1$, where $\alpha_i \in F_q$; Output $\mathbf{y} = (\sum_{i=1}^{l} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{l} t_i)$.

Verify. The intermediate nodes and the sink nodes have different verification. The calculation is as follows:

(1) For the intermediate nodes: Given K_i (any combination of $(Z_1, Z_2, \ldots, Z_{n+b})$), $\mathbf{y} = (\sum_{i=1}^{l} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{l} t_i)$, $\mathbf{Verify}(K_i \cdot \mathbf{y}^T) = 0$; Output either 0 (accept), or 1 (reject).

(2) For the sink nodes: Given sk, id, $\mathbf{y} = (\sum_{i=1}^{m} \alpha_i \mathbf{v}_i^*, \sum_{i=1}^{m} t_i)$, K_i (any combination of $(Z_1, Z_2, \ldots, Z_{n+b})$), $\mathbf{Verify}(K_i.\mathbf{y}^T) = 0$ and $\mathbf{Verify}(\mathrm{MAC}(\mathrm{sk}, \mathrm{id}, \sum_{i=1}^{m} \alpha_i \mathbf{v}_i^*), \sum_{i=1}^{m} t_i) = 0$; Output either 0 (accept), or 1 (reject).

4 Security Analysis

Theorem 1. Let A be a m × n matrix consisting of m independent blocks of dimension n, in the finite field F_q. The probability that a random n-dimensional vector maps A to zero is $\frac{1}{q^m}$.

Proof: Any n-dimensional vector w that maps A to zero belongs to $\Pi_A^\perp$ the null space of A. Following Eq. (3), $\dim(\Pi_A^\perp)$ is equal to $n - m$. Hence the probability of choosing a random vector that maps A to zero is

$$\Pr\left(Aw^T = 0\right) = \frac{q^{n-m}}{q^n} = \frac{1}{q^m} \tag{4}$$

We prove security assuming G is a secure Pseudo-Random Generator (PRG). For a PRG adversary B we let $\mathrm{PRG} - \mathrm{Adv}[B, G]$ be B's advantage in winning the PRG security game with respect to G.

Theorem 2. For any q, n, m, the MAC scheme HomMac is a secure (q, n, m) homomorphic MAC assuming the PRG G is a secure PRG.

In particular, for all homomorphic MAC adversaries A, there is a PRG adversary B such that:

$$\mathrm{NC} - \mathrm{Adv}[A, \mathrm{HomMac}] \leq \mathrm{PRG} - \mathrm{Adv}[B, G] + \left(\frac{1}{q}\right) \tag{5}$$

Proof: We prove the theorem using a sequence of two games denoted Game 0, 1. For i = 0, 1 let W_i be the event that A wins the homomorphic MAC security game in Game i.

Game 0 is identical to Attack Game applied to the scheme HomMac. Therefore:

$$\mathrm{Pr}[W_0] = \mathrm{NC} - \mathrm{Adv}[A, \mathrm{HomMac}] \tag{6}$$

In Game 1, the output of the PRG used in HomMac is replaced with a truly random string. That is to say, Game 1 is identical to Game 0 except that the challenger C computes $k \leftarrow F_q^{m+n}$ instead of $k \leftarrow G(K)$. Everything else remains the same. Then there is a PRG adversary B such that:

$$|\mathrm{Pr}[W_0] - \mathrm{Pr}[W_1]| = \mathrm{PRG} - \mathrm{Adv}[B, G] \tag{7}$$

The complete challenger in the attack game works as follows:

The adversaries A submit MAC queries $\{\mathbf{v}_i\}_{d=1}^{m}$. The challenger C responses to those queries and eventually the adversary outputs $\mathbf{v}_i'^* = \left(\mathbf{v}_i', t_i'\right)$. The adversary wins only if $\mathbf{v}_i'^* \notin \mathrm{span}\left(\mathbf{v}_1^*, \mathbf{v}_2^*, \ldots, \mathbf{v}_m^*\right)$.

We now show that $\mathrm{Pr}[W_1] = \frac{1}{q}$ in Game 1.

For intermediate nodes, the adversary wins if $\mathbf{Verify}(K_i \cdot \left(\sum_{i=1}^{l} \alpha_i \mathbf{v}_i'^*, \sum_{i=1}^{l} t_i'\right)) = 0$. The probability of the adversary win is:

$$\mathrm{Pr}_{intermediate} = \frac{1}{q^m} \tag{8}$$

For sink nodes, the adversary wins if $\mathbf{Verify}\left(\mathrm{id}, \mathrm{MAC}\left(\mathrm{sk}, \sum_{i=1}^{m} \alpha_i \mathbf{v}_i'\right), \sum_{i=1}^{m} t_i\right) = 0$ and $\mathbf{Verify}(K_i \cdot \left(\sum_{i=1}^{m} \alpha_i \mathbf{v}_i', \sum_{i=1}^{m} t_i'\right)) = 0$. The probability of the adversary win is:

$$\mathrm{Pr}_{sink} = \frac{1}{q^m} \cdot \frac{1}{q} = \frac{1}{q^{m+1}} \tag{9}$$

5 Comparison with Existing Schemes

In this section, we compare the performance of previous schemes and our scheme from the key length and the security parameters, some parameters come from [23]. The security parameter is the probability of a successful forgery. And the smaller is most definitely better (Table 1).

Table 1. Comparison of the performance of known schemes

Schemes	Key length	Security
IP MAC [15]	n + m	1/q
HomMac [14]	n + 2m	1/q
SpaceMac [26]	n + 2m	1/q
HSM [21]	(n + m + 1)r	$1/q^r$
Our scheme	n + 2m	$1/q^{m+1}$

6 Conclusion

We firstly analyze Null Keys and Homomorphic MAC schemes, and then show the limitations of these schemes. Inspired by these schemes, we combine the advantage of these schemes and provide a new scheme to improve the security of network coding. In our scheme, we use the null space to check the receive packets. To prevent collusion nodes through the verification, the source pads each packet with an extra MAC signature, so that even the malicious packets have passed the verification can be found at the sink node.

Acknowledgements. This work is supported by the HBUT PHD Fund BSQD12027 and the Hubei Provincial Educational Research Plan D20131404.

References

1. Ahlswede, R., et al.: Network information flow. IEEE Trans. Inf. Theory **46**(4), 1204–1216 (2000)
2. Li, S.-Y., Yeung, R.W., Cai, N.: Linear network coding. IEEE Trans. Inf. Theory **49**(2), 371–381 (2003)
3. Gkantsidis, C., Rodriguez, P.R.: Network coding for large scale content distribution. In: Proceedings of the IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2005. IEEE (2005)
4. Katti, S., et al.: XORs in the air: practical wireless network coding. In: ACM SIGCOMM Computer Communication Review. ACM (2006)
5. Dimakis, A.G., et al.: A survey on network codes for distributed storage. Proc. IEEE **99**(3), 476–489 (2011)
6. Ho, T., et al.: Byzantine modification detection in multicast networks using randomized network coding. In: Proceedings of International Symposium on Information Theory, ISIT 2004. IEEE (2004)
7. Jaggi, S., et al.: Resilient network coding in the presence of byzantine adversaries. In: 26th IEEE International Conference on Computer Communications, INFOCOM 2007. IEEE (2007)
8. Krohn, M.N., Freedman, M.J., Mazieres, D.: On-the-fly verification of rateless erasure codes for efficient content distribution. In: Proceedings of the 2004 IEEE Symposium on Security and Privacy. IEEE (2004)
9. Gkantsidis, C., Rodriguez. P.: Cooperative security for network coding file distribution. In: INFOCOM (2006)
10. Yu, Z., et al.: An efficient signature-based scheme for securing network coding against pollution attacks. In: The 27th IEEE Conference on Computer Communications, INFOCOM 2008. IEEE (2008)
11. Zhao, F., et al.: Signatures for content distribution with network coding. In: IEEE International Symposium on Information Theory, ISIT 2007 (2007)
12. Boneh, D., et al.: Signing a linear subspace: signature schemes for network coding. In: International Workshop on Public Key Cryptography. Springer (2009)
13. Yu, Z., et al.: An efficient scheme for securing XOR network coding against pollution attacks. In: INFOCOM 2009. IEEE (2009)

14. Agrawal, S., Boneh, D.: Homomorphic MACs: MAC-based integrity for network coding. In: International Conference on Applied Cryptography and Network Security. Springer (2009)
15. Li, Y., et al.: RIPPLE authentication for network coding. In: Proceedings of the INFOCOM 2010. IEEE (2010)
16. Kehdi, E., Li, B.: Null keys: limiting malicious attacks via null space properties of network coding. In: INFOCOM 2009. IEEE (2009)
17. Cheng, C., et al.: Security analysis and improvements on two homomorphic authentication schemes for network coding. IEEE Trans. Inf. Forensics Secur. **11**(5), 993–1002 (2016)
18. Esfahani, A., Mantas, G., Rodriguez, J.: An efficient null space-based homomorphic MAC scheme against tag pollution attacks in RLNC. IEEE Commun. Lett. **20**(5), 918–921 (2016)
19. Liu, G.: Security analysis and improvement of a tag encoding authentication scheme for network coding. Wuhan Univ. J Nat. Sci. **21**(5), 394–398 (2016)
20. Wang, J., et al.: An efficient short null keys based scheme for securing network coding against pollution attacks. In: Internet Conference of China. Springer (2014)
21. Zhang, P., et al.: Padding for orthogonality: efficient subspace authentication for network coding. In: Proceedings of the INFOCOM 2011. IEEE (2011)
22. Charles, D., Jain, K., Lauter, K.: Signatures for network coding. In: 2006 40th Annual Conference on Information Sciences and Systems. IEEE (2006)
23. Li, X., et al.: Two improved homomorphic MAC schemes in network coding. In: 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD). IEEE (2015)
24. Liu, G., Wang, X.: Homomorphic subspace MAC scheme for secure network coding. ETRI J. **35**(1), 173–176 (2013)
25. Wang, Q., et al.: MIS: malicious nodes identification scheme in network-coding-based peer-to-peer streaming. In: Proceedings of the INFOCOM 2010. IEEE (2010)
26. Le, A., Markopoulou, A.: Cooperative defense against pollution attacks in network coding using SpaceMac. IEEE J. Sel. Areas Commun. **30**(2), 442–449 (2012)

Gaussian Scale Patch Group Sparse Representation for Image Restoration

Yaqi Lu[1,2], Minghu Wu[1,2](✉), Nan Zhao[1,2], Min Liu[1,2], and Cong Liu[1,2]

[1] Hubei Key Laboratory for High-Efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan, People's Republic of China
luyaqieee@foxmail.com,
{wuxx1005,nzhao}@mail.hbut.edu.cn

[2] Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar Energy, Hubei University of Technology,
Wuhan 430068, People's Republic of China

Abstract. This passage puts forward a new sparse representation method, to solve the shortage problem of image restoration. First of all, extract the patch groups by utilize the non-local similar patches, and then using the simultaneous sparse coding to develop a non-local extension of Gaussian scale mixture model. Finally integrate the patch group model and Gaussian scale mixture model into encoding framework. Experimental results show that the proposed method achieves leading performance in terms of both quantitative measures and visual quality. In addition, our algorithm generates high-resolution images that are competitive or even superior in quality to images produced by other similar methods.

1 Introduction

Along with the booming information age, digital image has become one of the most significant carriers of information exchange thanks to its performance. The information impact on people's life and work becomes more important [1, 2]. Many outstanding methodologies of image processing by researchers these years [3]. Under the frame of algorithm, there are some good algorithms. Such as, multi-class classification algorithm [4], colony algorithm [5], GMM [6], clustering algorithm [7], correction algorithm [8], and so on. Those methods has obvious improvements rather than traditional algorithms. In the research of image processing, the sparse representation has been widely followed with interests and researches, and the image reconstruction via sparse encoding has been more and more concerned.

After combining with image denoising and sparse encoding reconstruction theories, this passage proposes an image reconstruction algorithm based on Gaussian mixture structure patch group sparse representation, which utilizes encoding methods of sparse non-local regularization and patch group weighting, and to implement patch grouping by grouping the similar patches to a local patch in a large enough neighborhood. This method utilizes a nonlocal extension of Gaussian scale mixture (GSM) model, And

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_51

patch group is formed by grouping the similar patches to a local patch in a large enough neighborhood. Therefore, the sparse representation of image patch can be applied with dictionary to generate a high resolution image. Simultaneously, there is an experimental simulation of comparing it with current PGPD [9] and NCSR [10] methods, and the method proposed in this passage has brought a more outstanding improvement.

2 Sparse Representation Model

The method of non-local self-similarity prior learning based patch group is capable of enhancing denoising performance, so that to learn explicit NSS models from natural images for high performance denoising. And putting nonlocal similar patches into groups. At last, the patch groups are extracted from training images for training non-local self-similarity prior models.

In particular for a noisy observation image that was captured, we obey the patch grouping method to prior model that millions of patch groups are extracted from a set of clean natural images during training phase, each similarity patch can be mapped from corresponding ones in the affiliated group and the grouping similarity patches are gathered into local patches, for each local patch we search for its similar patches in a window centered on it to form a patch group, Then the group mean is calculated from each patch. And sparse encoding can be utilized through dictionary for every patch group and because of that the sparse encoding coefficient follows Laplasse distribution, to weight the sparse coefficient vector, we can engage weight vector way in the case. In the phase of denoising, training Gaussian mixture model is available of providing dictionary and regularization parameter, and weighted sparse encoding model can be used for image denoising.

In this work, we propose a Gaussian scale patch group sparse representation. This passage constructs sparse coefficients vector model with a Gaussian scale mixture model. By characterizing a series of sparse coefficients of similarity patch via same prior distribution, we can effectively exploit both local and nonlocal dependencies among the sparse coefficients, and it considers each of sparse coefficients as a Gaussian distribution model with proportion variable, and the proportion variable can be fetched by sparse prior distribution. Studies have shown that the sparse coefficient and proportional variable predicted by the maximum a posteriori can be efficiently calculated via the method of alternating minimization. The crux of the model is that for a collection of similar patches, that corresponding sparse coefficients should be characterized by the probability density function, and to make the solving more easily.

3 Algorithm Implementation

The passage combines with structural sparse encoding frame, and the Non-local Self-Similarity prior model of an image indicates that it is able to enhance the performance of image denoising and reconstruction when there are many non-local similarity image patches for a single local image patch. And the passage prefers Principal Component Analysis method, by training dictionary for each image patch groups based on PCA based dictionary, and using orthogonal dictionary to simplify the Bias reasoning of sparse model.

The method introduced in the passage is using a natural clean NSS image, and to learn NSS model with high performance from natural image with patched grouping based prior training solution, then to construct Patch Groups Gaussian Mixture Model based learning algorithm by grouping non-local similarity patches, and finally use Simultaneous Sparse Coding to get a non-local extended GSM model. Specific algorithm procedure is shown as Algorithm 1 below.

```
Algorithm 1: The Gaussian Mixture Patch Group Sparse Representation Algorithm
Input: dictionary $\boldsymbol{D}$, noise image $y$, GMM components
Initialization: configure parameter, PCA of initial image patch;
Outer loop:
        1. Evaluate noise standard deviation;
        2. Calculate iteration regularization;
           Inner loop: for each patch group
           a) Calculate average value for each group;
           b) Select Gaussian component, weight code;
           c) Reconstruct each image patch in patch group;
           End for
        3. Update PCA with dictionary for image patch;
        4. Classify image patch and re-construct image $\tilde{x}$;
  End for
Output: the complete denoised image $\hat{x}$.
```

4 Simulation Analysis

In the passage, we selected some black and white images with size 512 × 512. We practically utilize the method PG-GSM introduced in the passage to comparison with current fine methodologies. The noise parameter to original natural image is configured as 4 types defining 5, 10, 30 and 50, and input into the program for simulation. We compares the denoising method revealed in the passage with PGPD and NCSR methods. For each single experiment, different Gaussian noise parameter has been provisioned correspondingly to different images, and the simulation denoising algorithm has been PSRN value tested to each image. The testing result is shown in below.

As Figs. 1 and 2 shows, the Gaussian noise level is lower, such configured as 5 and 10. It can be intuitively seen that the PSNR from our method is higher than one from PGPD method, while higher than one from NCSR method.

As to Figs. 3 and 4, assuming the Gaussian noise level is 30 and 50. It can be seen that in some images the PSNR from our method is higher than one from PGPD method, while higher than one from NCSR method.

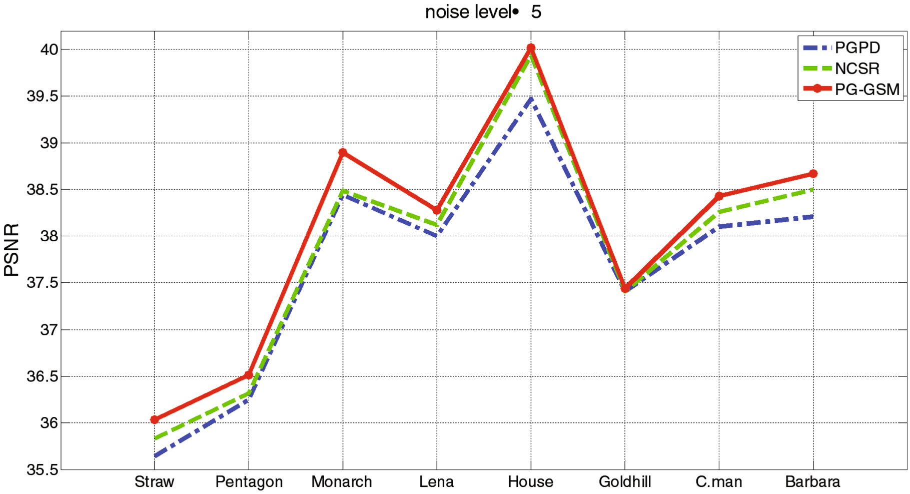


Fig. 1. PSNR(dB)results of different algorithms on natural images. (noise level is 5)

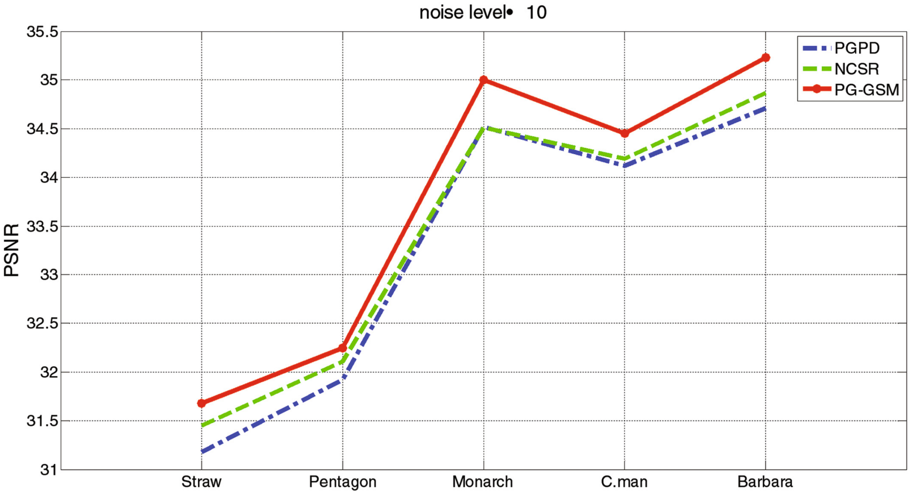


Fig. 2. PSNR(dB) results of different algorithms on natural images. (noise level is 10)

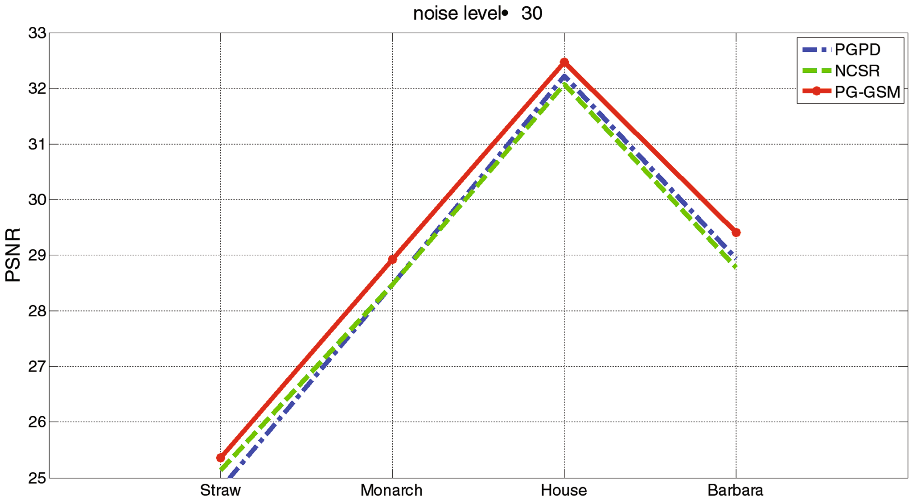


Fig. 3. PSNR(dB) results of different algorithms on natural images. (noise level is 30)

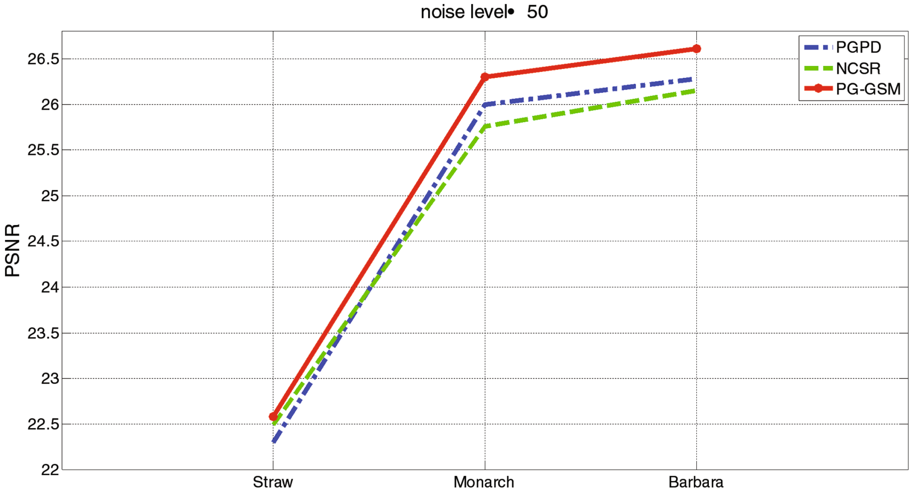


Fig. 4. PSNR(dB) results of different algorithms on natural images. (noise level is 50)

5 Conclusion

The passage proposes a kind of effective method of image recovery, denoising and reconstruction, which is to implement unified modeling combing with weighted grouping patch encoding and Gaussian scale mixture encoding. This passage utilizes patch groups weighted coding model and sparse coding to generate a non-local extended GSM model. this work effect the union between the two models, using alternating optimization evaluates the sparse coefficients, and then the reconstructed image patch is united and restructured.

Acknowledgments. This research was supported by National Natural Science Foundation of China (61471162, 61501178, 61601177); Program of International science and technology cooperation (2015DFA10940); Science and technology support program (R & D) project of Hubei Province (2015BAA115); PhD Research Startup Foundation of Hubei University of Technology (BSQD14028); Open Foundation of Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar Energy (HBSKFZD2015005, HBSKFTD2016002).

References

1. Zabukovec, A., Jaklič, J.: The impact of information visualisation on the quality of information in business decision-making. Int. J. Technol. Hum. Interact. **11**(2), 61–79 (2015)
2. Lin, J.H., Peng, W.: The contributions of perceived graphic and enactive realism to enjoyment and engagement in active video games. Int. J. Technol. Hum. Interact. **11**(3), 1–16 (2015)
3. Alamareen, A., Aljarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web. Eng. **11**(3), 1–14 (2016)
4. Wu, K., Kang, J., Chi, K.: Research on fault diagnosis method using improved multi-class classification algorithm and relevance vector machine. Int. J. Inf. Technol. Web. Eng. **10**(3), 1–16 (2015)
5. Mathiyalagan, P., Suriya, S., Sivanandam, S.N.: Hybrid enhanced ant colony algorithm and enhanced bee colony algorithm for grid scheduling. Int. J. Grid Util. Comput. **2**(1), 45–58 (2011)
6. Vinod, D.S., Mahesha, P.: Support vector machine-based stuttering dysfluency classification using GMM supervectors. Int. J. Grid Util. Comput. **6**(3/4), 143–149 (2015)
7. Boyinbode, O., Le, H., Takizawa, M.: A survey on clustering algorithms for wireless sensor networks. Int. J. Space-Based Situated Comput. **1**(2/3), 130–136 (2011)
8. Sun, N., Murakami, S., Nagaoka, H., et al.: A correction algorithm for stereo matching with general digital cameras and web cameras. Int. J. Space-Based Situated Comput. **3**(3), 169–184 (2013)
9. Xu, J., Zhang, L., Zuo, W.: Patch group based nonlocal self-similarity prior learning for image denoising. In: Proceedings of the 15th IEEE International Conference on Computer Vision, pp. 244–252. Institute of Electrical and Electronics Engineers Inc., Santiago, Chile (2016)
10. Dong, W., Zhang, L., Shi, G., Li, X.: Nonlocally centralized sparse representation for image restoration. IEEE Trans. Image Process. **22**(4), 1620–1630 (2013)

An Efficient Identity-Based Homomorphic Signature Scheme for Network Coding

Yudi Zhang, Yan Jiang, Bingbing Li, and Mingwu Zhang(✉)

School of Computer Sciences, Hubei University of Technology,
Wuhan, Hubei, People's Republic of China
zhangyudi007@gmail.com, bleachsigh@gmail.com, batmangshock@hotmail.com,
csmwzhang@gmail.com

Abstract. Network coding is now widely used to improve the network throughput capacity in lots of applications, such as distributed storage, wireless mesh networks, etc. Unlike the traditional routing scheme in which the network nodes simply relay the received packets, network coding technique requires the intermediate node to combine the received packets together and then re-transmit it repeatedly. However, there is a fatal threat that the malicious intermediate nodes can tamper the data before combining the packets, and thus the standard signature scheme cannot satisfy the security requirement for this application. In this paper, we propose an identity-based homomorphic scheme for network coding which can prevent malicious nodes to produce the pollution attacks. The public key of our scheme is a constant size which is only the hash output of user's identity. We present the detailed construction and analyze the security of the scheme in the random oracle model.

1 Introduction

Network coding was first introduced by Ahlswede [1] as an alternative to traditional "store-and-forward" routing network, it showed that random linear coding can achieve the optimal throughput for multicast [7,16,23] and even unicast transmissions [17,18]. Network coding allows each intermediate node to encode packets en-route. Intermediate nodes can linearly combine the messages received from the uplinks.

Advantages of network coding are huge, such as improving throughput and minimizing the transmission delay of a network. Another benefit is its robustness and adaptability, the destination can recover the original data once when it has received sufficiently many correct packets, even if some packets are lost. Due to these advantages, network coding has been approached for many practical network applications, such as wireless sensor network [14,19,20], video broadcast [9], peer-to-peer content distribution networks [13,15], and distributed storage systems [5,8].

However, network coding are facing some potential security threats. A major concern in network coding is that packets may be modified by malicious nodes

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_52

i.e. pollution attack [11,12]. The pollution attack is originated from any malicious behaviors of untrusted forwarders or adversaries, such as injecting polluted information, modifying and replaying the disseminated messages, which can propagate and pollute multiple packets to the whole networks. If a junk message is mixed by forwarders, the output messages of the forwarder will be contaminated. Since these polluted messages could spread to all downstream nodes by combining junk messages, so such polluted messages should be detected and filtered as early as possible.

Traditional hash function based signature schemes are inapplicable and unsuitable for net work coding, because the original source signatures can be destroyed in the encoding (combining) process, which is performed by each nodes. Krohn et al. [15] proposed homomorphic hashing for preventing pollution attacks. In this scheme, the authentication information and public key are large. Yu et al. [21] suggest a homomorphic signature [10,24] scheme based on the RSA signature scheme, that the forwarders can achieve efficient verification at the expense of increased transmission overhead. The drawback of this approach is that RSA signature is typically very large. Zhao et al. [22] presented a scheme for network coding. In this scheme, they designed a signature that can be used to easily check the membership of a received vector in the given subspace. However, the scheme can only be used for distributing a single file. Boneh et al. [3] also proposed two signature schemes that can be used in conjunction with network coding, the schemes can be viewed as signing linear subspaces.

In this paper, we present an identity-based signature scheme for network coding under the idea of Boneh [3]. Our scheme can be used to provide cryptographic protection against pollution attacks. Same as [3], the destination must receive a minimum number of uncorrupted packets to recover the file. Because our scheme is an identity-base signature scheme, so the intermediate nodes can verify the signature by hash values of user's ID, and discard corrupted packets as well.

2 Preliminaries

In this section, we first review the practical network coding and identity-based scheme briefly, then introduce the model of network coding signature scheme and complexity assumptions.

2.1 Linear Network Coding

In a linear network coding scheme [16], the information source outputs a continuous stream of packets, which can be grouped into blocks with n source packets per block. In the proposed scheme, a file to be transmitted is viewed as an ordered sequence of n-dimensional vectors $\bar{\mathbf{v}}_1, \dots, \bar{\mathbf{v}}_m \in \mathbb{Z}_q^n$, where q is a prime. The vector can be referred as block or packet. The source node creates m augmented vectors $\mathbf{v}_1, \cdots, \mathbf{v}_m$ that:

$$\mathbf{v}_i = (-\bar{\mathbf{v}}_i-, 0, \dots, 0, 1, 0, \dots, 0) \in \mathbb{Z}_q^{n+m} \tag{1}$$

each $\bar{\mathbf{v}}_i$ is appended with the vector of length m containing a single 1 in the ith position. When a node in the network receiving packets $\mathbf{w}_1, \ldots, \mathbf{w}_l \in \mathbb{Z}_q^{n+m}$ on its l incoming communication edges, node i computes the block $\mathbf{w} = \sum_{j=1}^{l} \alpha_{i,j}\mathbf{w}_j$, where $\alpha_{i,j} \in \mathbb{Z}_q$, and the weight $\alpha_{i,j}$ can be chosen randomly and independently by each node. The resulting vector $\mathbf{w}$ is then transmitted on the node's outgoing edges.

When a destination node receives m linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$, it can recover the original file. For a received vector $\mathbf{w}_i$, let $\mathbf{w}_i^L$ denote the left-most n positions of the vector, and let $\mathbf{w}_i^R$ denote the right-most m positions. Then the node computes an $m \times m$ matrix G:

$$G = \begin{pmatrix} -\mathbf{w}_1^R- \\ \vdots \\ -\mathbf{w}_m^R- \end{pmatrix} \tag{2}$$

Then original file $\bar{\mathbf{v}}_1, \ldots, \bar{\mathbf{v}}_m$ is given by:

$$\begin{pmatrix} -\bar{\mathbf{v}}_1- \\ \vdots \\ -\bar{\mathbf{v}}_m- \end{pmatrix} = G \cdot \begin{pmatrix} -\mathbf{w}_1^L- \\ \vdots \\ -\mathbf{w}_m^L- \end{pmatrix} \tag{3}$$

2.2 Identity-Based Signature

In this sub-section, we introduce the definition of identity-based signature scheme (IBS) and it's security model [2,6].

Definition 1. An identity-based signature scheme is consist of four algorithms, **IB.Setup**, **IB.Extract**, **IB.Sign** and **IB.Verify**.

IB.Setup: It takes as the input security parameter, the PKG outputs master secret key **msk** and system parameter **mpk**.

IB.Extract: It takes as input parameters **mpk**, **msk** and an identity **id**, then outputs a private key sk_{id} corresponding to the user with this identity.

IB.Sign: It takes as input **mpk**, a private key sk_{id}, and a message m, then outputs a signature σ.

IB.Verify: It takes as input **mpk**, a signature σ, a message m, and an identity **id**, then outputs 1 if σ is a valid signature, otherwise, outputs 0.

Definition 2. Security model for identity-based signature schemes. [6] An identity-based signature scheme is secure against existential forgery on adaptively chosen message and ID attacks if no polynomial time algorithm $\mathcal{A}$ has a non-negligible advantage against a challenger $\mathcal{C}$ in the following game:

1. $\mathcal{C}$ runs **IB.Setup** of the scheme. The master public parameters are given to $\mathcal{A}$.

2. $\mathcal{A}$ issues the following queries:
   a. Hash function query. $\mathcal{C}$ computes the value of hash function for requested input and sends the value to $\mathcal{A}$.
   b. **IB.Extract** query. $\mathcal{A}$ sends an identity ID to $\mathcal{C}$, then $\mathcal{C}$ runs **IB.Extract** and sends the private key to $\mathcal{A}$.
   c. **IB.Sign** query. $\mathcal{A}$ sends an identity ID and a message m to $\mathcal{C}$, then $\mathcal{C}$ runs **IB.Sign** and returns the signature to $\mathcal{A}$.
3. $\mathcal{A}$ outputs (ID, m, σ) and (ID, m) have never been queried to **IB.Extract** and **IB.Sign** respectively. $\mathcal{A}$ wins the game if σ is a valid signature of m for ID.

2.3 Network Coding Signature Scheme

Charles et al. presented network coding signature scheme [4], this scheme can combine the valid signatures on each vectors without knowledge of the signer's secret key, and generate a valid signature on any linear combination. Boneh et al. proposed two signature schemes [3] for network coding which has constant public-key size.

Definition 3. A homomorphic network coding signature scheme [3] is a tuple of probabilistic, polynomial-time algorithm with the following functionalities:

Setup: On input a security parameter 1^k and an integer N, this algorithm outputs a prime q, a public key PK, and a secret key SK.

Sign: On input a secret key SK, a file identifier $\gamma \in \{0,1\}_k$, and a vector $\mathbf{v} \in \mathbb{Z}_q^N$, this algorithm outputs a signature σ.

Combine: On input a public key PK, a file identifier γ, and a set of tuples $\{(\beta_i, \sigma_i)\}_{i=1}^{l}$ with $\beta_i \in \mathbb{Z}_q$, this algorithm output a signature σ. In this algorithm, if each σ_i is a valid signature on the vector $\mathbf{v}_i$, then σ is a signature on $\sum_{i=1}^{l} \beta_i \mathbf{v}_i$.

Verify: On input a public key PK, an identifier $\gamma \in \{0,1\}_k$, a vector $\mathbf{a} \in \mathbb{Z}_q^N$, and a signature σ, this algorithm outputs 1 if σ is a valid signature, otherwise outputs 0.

2.4 Bilinear Groups and Complexity Assumptions

Definition 4. Let $\mathbb{G}_1$, $\mathbb{G}_2$ and $\mathbb{G}_T$ be three cyclic groups of large prime order q, let e: $\mathbb{G}_1 \times \mathbb{G}_2 \rightarrow \mathbb{G}_T$ be an admissible pairing, it satisfies the following properties:

1. Bilinearity: $e(g^a, h^b) = e(g, h)^{ab}$, for all $g \in \mathbb{G}_1$, $h \in \mathbb{G}_2$ and $a, b \in \mathbb{Z}_q$.
2. Non-degeneration: $e(g, h)$ generates $\mathbb{G}_T$ for any generators g in $\mathbb{G}_1$ and h in $\mathbb{G}_2$.
3. $\mathbb{G}_2 \rightarrow \mathbb{G}_1$ is an efficiently computable isomorphism.

Definition 5. co-CDH Problem. Given $g_1, g_1^a \in \mathbb{G}_1$ and $h \in \mathbb{G}_2$, output $h_2 \in \mathbb{G}_2$.

The advantage of an algorithm $\mathcal{A}$ in solving co-CDH in groups $\mathbb{G}_1$ and $\mathbb{G}_2$ is:

$$\text{Adv co-CDH}_{\mathcal{A}} = Pr[\mathcal{A}(g_1, g_1^a, h) = h^a : a \overset{r}{\leftarrow} \mathbb{Z}_q, h \overset{r}{\leftarrow} \mathbb{G}_2] \tag{4}$$

co-CDH assumption: For every PPT algorithm $\mathcal{A}$, Adv co-CDH$_{\mathcal{A}}$ is negligible.

3 Our Construction

In this section we construct an identity-based homomorphic network coding signature scheme. In our scheme, each node can rapidly drop the packet which has been polluted or forged.

Setup$(1^k, N)$. Given security parameter 1^k and a positive integer N, the PKG chooses the system parameters that include two groups $\mathbb{G}_1$, $\mathbb{G}_2$ of prime order $q \geqslant 2^k$, a generator $h \in \mathbb{G}_2$ and a random $z \in \mathbb{Z}_q$ as a master secret key **msk**. It sets $h^z \in \mathbb{G}_2$. The PKG chooses secure hash functions $H_0 : \{0,1\}^* \to \mathbb{Z}_q$, $H_1 : \{0,1\}^* \times \{0,1\}^* \to \mathbb{G}_1$. The system parameters $mpk = (q, \mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_T, h^z, e, H_0, H_1)$.

Extract(mpk, msk, id). On input global parameters mpk, a master secret key msk and a identity id, this algorithm at random select $r \overset{r}{\leftarrow} \mathbb{Z}_q$ and sets $y = r + z \cdot H_0(h^r, id)$. Then it outputs the secret key $sk_{id} = (y, h^r)$. Note that h^r is a public tuple even it is part of the secret key.

Sign$(mpk, sk_{id}, id, \gamma, \mathbf{v})$. On input parameters mpk, a secret key sk_{id}, a identity id, a file name γ and a vector $\mathbf{v} = (v_1, \cdots, v_N) \in \mathbb{Z}_q^N$, this algorithm outputs the signature

$$\sigma = (\prod_{i=1}^{N} H_1(\gamma, i)^{v_i})^y \tag{5}$$

Combine$(mpk, id, \gamma, \{(\beta_i, \sigma_i)\}_{i=1}^{l})$. On input parameters mpk, a identity id, a file name γ, and $\{(\beta_i, \sigma_i)\}_{i=1}^{l}$ with $\beta_i \in \mathbb{Z}_q$, it outputs

$$\sigma = \prod_{i=1}^{l} \sigma_i^{\beta_i} \tag{6}$$

Verify$(mpk, id, \gamma, \mathbf{a}, \sigma)$. On input parameters mpk, a identity id, a file name γ, and a vector $\mathbf{a} \in \mathbb{Z}_q^N$, if

$$e(\sigma, h) = e(\prod_{i=1}^{N} H_1(\gamma, i)^{a_i}, h^r \cdot (h^z)^{H_0(h^r, id)}) \tag{7}$$

this algorithm output 1; otherwise it outputs 0.

To make the sign algorithm more efficient, we present a variant sign scheme, which will only sign the properly augmented vectors like [3], this is the case for application to network coding. The dimension m of the resulting vector space is known at the time any vector is signed and verified. The signer at random choose $g_1, g_2, \cdots, g_N \overset{r}{\leftarrow} \mathbb{G}_1$ and publish these as part of public key. Sign a vector $\mathbf{v} = (v_1, \cdots, v_N) \in \mathbb{Z}_q^N$ using the identity id, a file name γ, and $n = N - m$, then the signer computes:

$$\sigma = (\prod_{i=1}^{m} H_1(\gamma, i)^{v_{n+i}} \prod_{j=1}^{n} g_j^{v_j})^y \tag{8}$$

4 Security Analysis

In this we prove the security of our identity-based scheme for network coding. We consider the variant scheme (call it IBSNC_1),and the prove the security of the identity-based scheme IBSNC_2 constructed from IBSNC_1.

Lemma 1. *Let $S_2 = (Setup_2, Extract_2, Sign_2, Combine_2, Verify_2)$ be a homomorphic network coding identity-based signature scheme. Then $S_1 = (Setup_1, Extract_1, Sign_1, Verify_1)$ defined as follows is a network coding identity-based signature scheme.*

- $\mathbf{Setup}_1(1^k, N)$ runs $\mathbf{Setup_2}(1^k, N)$ and outputs the results.
- $\mathbf{Extract}_1(mpk, msk, id)$ runs $\mathbf{Extract}_2(mpk, msk, id)$ and output the results.
- $\mathbf{Sign}_1(mpk, sk_{id}, id, \gamma, \mathbf{V})$ runs $\mathbf{Sign}_2(mpk, sk_{id}, id, \gamma, \mathbf{v}_1), \ldots, \mathbf{Sign}_2(mpk, sk_{id}, id, \gamma, \mathbf{v}_m)$, where $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is any basis of $\mathbf{V}$. It then outputs $\sigma = ((\mathbf{v}_1, \sigma_1), \ldots, (\mathbf{v}_m, \sigma_m))$.
- $\mathbf{Verify}_1(mpk, id, \gamma, \mathbf{a}, \sigma)$ parses σ as $((\mathbf{v}_1, \sigma_1), \ldots, (\mathbf{v}_m, \sigma_m))$, and computes coefficients $\{\beta_i\}$ such that $\mathbf{a} = \sum_i \beta_i \mathbf{a}_i$ (Output 0, if on solution exists). Finally, it outputs $\mathbf{Verify}_2(mpk, id, \gamma, \mathbf{a}, \mathbf{Combine}_2(mpk, id, \gamma, \{\beta_i, \sigma_i\}_{i=1}^m))$.

Same as [3], we have $\mathbf{a} \in V \leftrightarrow \mathbf{a} = \sum_{i=1}^m a_{n+i}\mathbf{v}_i$, and we say that a homomorphic network coding identity-based signature scheme S_2 is secure if the network coding identity-based signature scheme S_1 constructed from S_2 as in Lemma 1 is secure.

Definition 6. Like the Definition 2, we now give the security model of network coding identity-based signature scheme. A network coding identity-based signature scheme is secure if no polynomial time algorithm $\mathcal{A}$ has a non-negligible advantage against a challenger $\mathcal{C}$ in the following game:

1. $\mathcal{C}$ runs **Setup** of the scheme. The public parameters are given to $\mathcal{A}$.
2. $\mathcal{A}$ issues the following queries:
   a. Hash function query. $\mathcal{C}$ computes the value of hash function for requested input and sends the value to $\mathcal{A}$.
   b. **Extract** query. $\mathcal{A}$ sends an identity id to $\mathcal{C}$, then $\mathcal{C}$ runs **Extract** and sends the private key to $\mathcal{A}$.
   c. **Sign** query. $\mathcal{A}$ sends an identity id, a sequence of vector subspaces $V_i \subset \mathbb{Z}_q^N$, and γ_i to $\mathcal{C}$, then $\mathcal{C}$ runs **Sign** and returns the signature to $\mathcal{A}$.
3. $\mathcal{A}$ outputs $(id, \gamma^*, \mathbf{a}^*, \sigma^*)$, is has not been queried and either $\gamma^* = \gamma_i$ for all i or $\gamma^* = \gamma_i$ for some i but $\mathbf{a}^* \notin V_i$. $\mathcal{A}$ wins the game if σ is a valid signature of m for ID.

Theorem 1. *Let $IBSNC_2$ be the network coding identity-based scheme constructed from $IBSNC_1$ via the method of Lemma 1. Then $IBSNC_2$ is secure in the random oracle model assuming that the co-CDH problem in $(\mathbb{G}_1, \mathbb{G}_2)$ is infeasible.*

5 Conclusion

We designed an identity-based homomorphic signature scheme for network coding to against pollution attacks. The intermediate nodes can verify each packets using user's ID, and discard the corrupted packets, so the corrupted packets can not be transmit to downstream nodes. This identity-based scheme is very practical and then can be implemented in many applications. Since the key size is constant, it's efficient to verify and transmit during the performance.

Acknowledgements. This work is supported by the National Natural Science Foundation of China under grant 61672010.

References

1. Ahlswede, R., Cai, N., Li, S.Y., Yeung, R.W.: Network information flow. IEEE Trans. Inf. Theory **46**(4), 1204–1216 (2000)
2. Bellare, M., Namprempre, C., Neven, G.: Security proofs for identity-based identification and signature schemes. J. Cryptology **22**(1), 1–61 (2009)
3. Boneh, D., Freeman, D., Katz, J., Waters, B.: Signing a linear subspace: signature schemes for network coding. In: International Workshop on Public Key Cryptography, pp. 68–87. Springer (2009)
4. Charles, D., Jain, K., Lauter, K.: Signatures for network coding. In: 2006 40th Annual Conference on Information Sciences and Systems, pp. 857–863. IEEE (2006)
5. Chen, F., Xiang, T., Yang, Y., Chow, S.S.M.: Secure cloud storage meets with secure network coding. IEEE Trans. Comput. **65**(6), 1936–1948 (2016). doi:10.1109/TC.2015.2456027
6. Choon, J.C., Cheon, J.H.: An identity-based signature from gap Diffie-Hellman groups. In: International Workshop on Public Key Cryptography, pp. 18–30. Springer (2003)
7. Dai, M., Kwan, H.Y., Sung, C.W.: Linear network coding strategies for the multiple access relay channel with packet erasures. IEEE Trans. Wireless Commun. **12**(1), 218–227 (2013)
8. Dimakis, A.G., Godfrey, P.B., Wu, Y., Wainwright, M.J., Ramchandran, K.: Network coding for distributed storage systems. IEEE Trans. Inf. Theory **56**(9), 4539–4551 (2010). doi:10.1109/TIT.2010.2054295
9. Esmaeilzadeh, M., Sadeghi, P., Aboutorab, N.: Random linear network coding for wireless layered video broadcast: general design methods for adaptive feedback-free transmission. IEEE Trans. Commun. **65**(2), 790–805 (2017)
10. Gorbunov, S., Vaikuntanathan, V., Wichs, D.: Leveled fully homomorphic signatures from standard lattices. In: Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, pp. 469–477. ACM (2015)
11. Han, K., Ho, T., Koetter, R., Medard, M., Zhao, F.: On network coding for security. In: Military Communications Conference 2007, MILCOM 2007, pp. 1–6. IEEE (2007)
12. Jaggi, S., Langberg, M., Katti, S., Ho, T., Katabi, D., Médard, M.: Resilient network coding in the presence of byzantine adversaries. In: INFOCOM 2007, 26th IEEE International Conference on Computer Communications, pp. 616–624. IEEE (2007)

13. Jain, K., Lovasz, L., Chou, P.A.: Building scalable and robust peer-to-peer overlay networks for broadcasting using network coding. Distrib. Comput. **19**(4), 301–311 (2007)
14. Katti, S., Rahul, H., Hu, W., Katabi, D., Médard, M., Crowcroft, J.: Xors in the air: practical wireless network coding. In: ACM SIGCOMM Computer Communication Review, vol. 36, pp. 243–254. ACM (2006)
15. Krohn, M.N., Freedman, M.J., Mazieres, D.: On-the-fly verification of rateless erasure codes for efficient content distribution. In: Proceedings of 2004 IEEE Symposium on Security and Privacy 2004, pp. 226–240. IEEE (2004)
16. Li, S.Y., Yeung, R.W., Cai, N.: Linear network coding. IEEE Trans. Inf. Theory **49**(2), 371–381 (2003)
17. Li, Z., Li, B.: Network coding: the case of multiple unicast sessions. In: Allerton Conference on Communications, vol. 16, p. 8 (2004)
18. Lun, D.S., Médard, M., Koetter, R.: Network coding for efficient wireless unicast. In: 2006 International Zurich Seminar on Communications, pp. 74–77. IEEE (2006)
19. Petrovic, D., Ramchandran, K., Rabaey, J.: Overcoming untuned radios in wireless networks with network coding. IEEE Trans. Inf. Theory **52**(6), 2649–2657 (2006)
20. Swapna, B., Eryilmaz, A., Shroff, N.B.: Throughput-delay analysis of random linear network coding for wireless broadcasting. IEEE Trans. Inf. Theory **59**(10), 6328–6341 (2013)
21. Yu, Z., Wei, Y., Ramkumar, B., Guan, Y.: An efficient signature-based scheme for securing network coding against pollution attacks. In: The 27th Conference on Computer Communications, INFOCOM 2008, pp. 1409–1417. IEEE (2008)
22. Zhao, F., Kalker, T., Médard, M., Han, K.J.: Signatures for content distribution with network coding. In: IEEE International Symposium on Information Theory 2007, ISIT 2007, pp. 556–560. IEEE (2007)
23. Zhu, Y., Li, B., Guo, J.: Multicast with network coding in application-layer overlay networks. IEEE J. Sel. Areas Commun. **22**(1), 107–120 (2004)
24. Zkik, K., Tebaa, M., El Hajji, S.: A new secure framework in mcc using homomorphic signature: application in banking data. In: Transactions on Engineering Technologies, pp. 413–427. Springer (2016)

A Three-Dimensional Digital Watermarking Technique Based on Integral Image Cryptosystem and Discrete Fresnel Diffraction

Yiqun Liu[1,2(✉)], Jianqi Zhang[1], Zhen Zhang[2], Haining Luo[2], and Xiaorui Wang[1]

[1] School of Physics and Optoelectronic Engineering,
Xidian University, Xi'an 710071, Shaanxi, China
wjliuyiqun@126.com

[2] Key Laboratory of CAPF for Cryptology and Information Security,
Department of Electronic Technology Engineering,
University of Chinese Armed Police Force, Xi'an 710086, Shaanxi, China

Abstract. This paper presents a three-dimensional (3D) digital watermarking technique based on integral image cryptosystem and discrete Fresnel diffraction (DFD). 3D digital watermarking is generated by computational integral imaging. The 3D digital watermarking is encrypted and embedded by integral imaging cryptosystem that is designed with DFD transform algorithm. Finally, the extracted 3D digital watermarking is decrypted and displayed by integral imaging cryptosystem. The feasibility and effectiveness of the proposed scheme is demonstrated by numerical simulation experiment. The majority of system will improve the security and robustness of 3D digital watermarking. The proposed method can provide a new, real-time, and effective strategy in the security data management of cloud computing and big data.

Keywords: 3D digital watermarking · Integral imaging cryptosystem · DFD · Optical image

1 Introduction

The security of information systems is increasingly crucial in our lives, as everything is going to be connected to the Internet. The main motivation for using optical technology of optics and photonics for information security is that optical waveform possess many complex degrees of freedom such as amplitude, phase, polarization, nonlinear transformations, quantum properties of photons, and multiplexing that can be combined in many ways to make information encryption more secure and more difficult to attack [1–3]. Digital watermarking provides secure means for copyright protection, data management and authentication, encrypted communication, collective computational intelligence in managing disaster situations [4]. Refregier and B. Javidi [5] proposed double random phase encoding technology based on 4*f* systems, which opened a new field of optical information security research, and since then, different transform domain algorithms have been proposed to improve the security level of the optical information hiding with double random phase encoding. Xiang Peng et al. introduced a

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_53

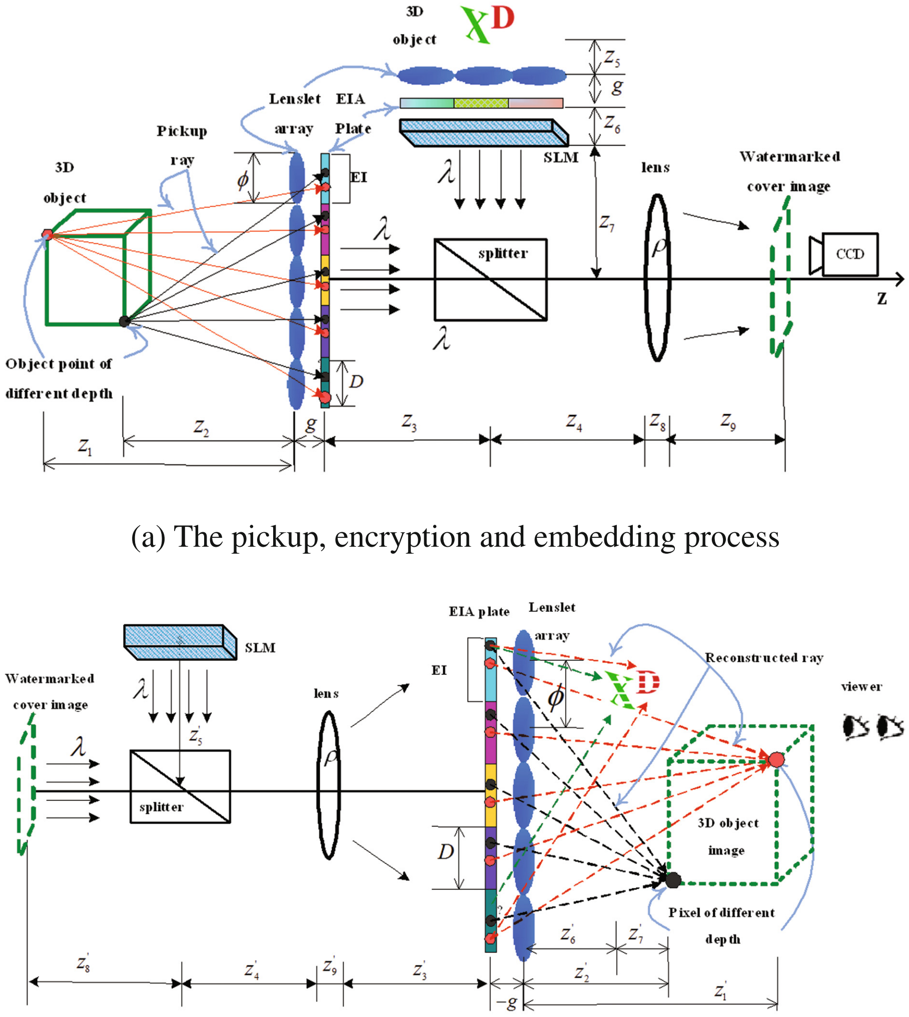
3D digital watermarking algorithm based on virtual optics [6, 7]. The algorithm realizes embedding and blind extraction of 3D digital watermarks utilizing the morphological variation of virtual Fresnel diffraction in 3D space.

Integral imaging has been one of the promising auto-stereoscopic 3D imaging and display techniques since it was proposed by Lippmann in 1908 [8]. It provides viewer full parallax, continuous views which works with incoherent light [1, 9–13]. In 2007, Dong-Choon, *et al.* found that integral imaging could be applied in the research of digital watermarking [14]. The quality of reconstructed images would be significantly degraded by the interference between adjacent pixels because the reconstruction method to calculate integral imaging was actually cascading pixel reconstruction. Both of security and algorithm efficiency are weak. Elemental images (EIs) were recorded, captured, and generated by a lenslet array. The literature [15–18] presents a new method for 3D scene acquisition via reconstruction with multi-spectral information and its Fourier-based encryption using computational integral imaging, by which the field of view, resolution, and information security are increased, respectively. Double random phase encryption (DRPE) in the Fourier domain is employed on Bayer formatted elemental image to encrypt the captured 3D scene. But the security and quality of digital watermarking images was degraded with these algorithms which were reported [12, 15, 19].

This paper proposes a 3D digital watermarking based on integral imaging cryptosystem in DFD domain. 3D digital watermarking is generated, encrypted, embedded, extracted, decrypted and displayed by the integral imaging cryptosystem that is implemented with computational reconstruction algorithm based on smart pseudoscopic-to-orthoscopic conversion (SPOC) model. The feasibility and effectiveness of the proposed system and technique are demonstrated by simulation experiments. The new method is able to meet the requirements of robustness and security. The quality of image can also meet these criterions of the human visual model. The new method has some advantages of optical imaging systems such as multi-dimension, high design freedom, and high robustness. A primary implication of encrypted processing is that the majority of integral imaging cryptosystem will be encryption-in-the-loop applications, and the majority of system will improve the security and robustness of 3D digital watermarking. New scheme and solution are introduced to security data management of big data and cloud computing, copyright protection of 3D multimedia digital products [17, 20–33]. The work also promotes a promising approach to information hiding based on optical technologies.

2 The Principle of the System

According to the theories of Fourier optics, digital watermarking and integral imaging, the general expression of DFD transform [34] is as the formula (1). The schematic of 3D digital watermarking with integral image cryptosystem is designed and shown in Fig. 1. The whole procedure of the scheme consists of two stages: Fig. 1(a) the pickup, encryption and embedding process, Fig. 1(b) the extraction, decrypted and display process.



(a) The pickup, encryption and embedding process

(b) The extraction, decryption and display process

Fig. 1. The schematically of integral imaging cryptosystem

Therein, A and B represent the two planes which separate spatially in the direction of propagation. s, t, z_{AB} and λ denote the sampling number of two adjacent orthogonal pixels, the spacing between the planes, the wavelength of incident light, respectively. We define correlation sampling lengths of the input plane along the x and y axes as Δx and Δy, and the Fourier plane along the ξ and η axes in Fresnel transform domain (FTD) as $\Delta\xi$ and $\Delta\eta$, respectively. C is a complex constant whose value may be calculated by the formula (2).

$$\begin{aligned} DFD[A, B, s, t; z_{AB}, \lambda] &= \tfrac{\exp[j2\pi z_{AB}/\lambda]}{j\lambda z_{AB}} \times \exp[j\tfrac{\pi}{\lambda z_{AB}}(s^2\Delta\xi^2 + t^2\Delta\eta^2)] \\ &\times \sum_{q=0}^{N-1}\sum_{l=0}^{N-1} U_A(q, l)\exp[j\tfrac{\pi}{\lambda z_{AB}}(q^2\Delta x_0^2 + l^2\Delta y_0^2)] \times \exp[-j2\pi(\tfrac{qs}{N} + \tfrac{lt}{N})] \end{aligned} \tag{1}$$

Where

$$C = \frac{\exp[j2\pi z_{AB}/\lambda]}{j\lambda z_{AB}} \tag{2}$$

As we all know, since $DFD[A, B, s, t; z_{AB}, \lambda]$ is complex value in Formula (1), it pickups both the amplitude and phase information of the result signal in the optical implementation described in Fig. 1.

$$\begin{aligned} G_{3D}(\omega, \gamma) = \{&\alpha_1 DFD[C(x, y), L(x, y), s, t; z_C, \lambda] \\ &+ \alpha_2 DFD[W(x, y), L(x, y), s, t; z_W, \lambda] \\ &+ \alpha_3 DFD[R(x, y), L(x, y), s, t; z_R, \lambda]\} \times T(s, t; f) \end{aligned} \tag{3}$$

Assume that $C(x, y)$, $W(x, y)$, $R(x,y)$ represent different plane of two EIA (Elemental Image Array) planes and random phase mask plate (RPMP), respectively. $L(x, y)$ is the front plane of imaging lens in propagation direction. *s, t* are the number of pixel samples. Z_j represents the distance between the different planes, where $j = 1, 2, \cdots 9, C, W, R.$ These symbols $z_C = z_3 + z_4,\ z_W = z_4 + z_6 + z_7,\ z_R = z_4 + z_7$ and g represents the distance between the lenslet array and elemental image plane, D represents the size of elemental images, ϕ represents lenslet spacing, the focal length of the imaging lens ρ is f, $\alpha_1, \alpha_2, \alpha_3$ are encryption weighting factors, $\alpha_1 + \alpha_2 + \alpha_3 = 1$, for example $\alpha_1 = \alpha_2 = 0.4$, $\alpha_3 = 0.2$. They are used to adjust the energy ratio among the DFD transforms of the cover image, 3D digital watermarking and RPMP. The optical transfer function is $T(s, t; f)$. Thus, the transformation process of 3D digital watermarking with integral image cryptosystem in the DFD domain can be described by the following mathematical models, for example, the formula (3). While the optical field distribution for the back plane of the imaging lens can be determined by $G_{3D}(\omega, \gamma)$, and the relevant watermarked and encrypted cover image will be recorded by the CCD (Charge Coupled Device) camera that is accomplished by computer.

Shown in Fig. 1, the stage consists of 3D optical image extraction, decryption and display. The authorized users receive the correct marked encrypted image at first, and then subtract the contribution of the RPMP in the above process to obtain DWM' as shown in the formula (4), therefore, according to the theory of optical Fourier transform, the cipher text images can be decrypted, respectively. Finally, 3D digital watermarking $TDWM'$ can be reconstructed and displayed by integral imaging cryptosystem. Mathematical model of the decryption process can be represented by the formula (4) to (6) as follows.

$$DWM' = G_{3D}(\omega,\gamma) - DFD\{\alpha_3 DFD[R(x,y),L(x,y),s,t;z_{R1},\lambda] \times T(s,t;f)\} \tag{4}$$

$$TDWM' = IDFD[DWM']|_{z=z'_W} \tag{5}$$

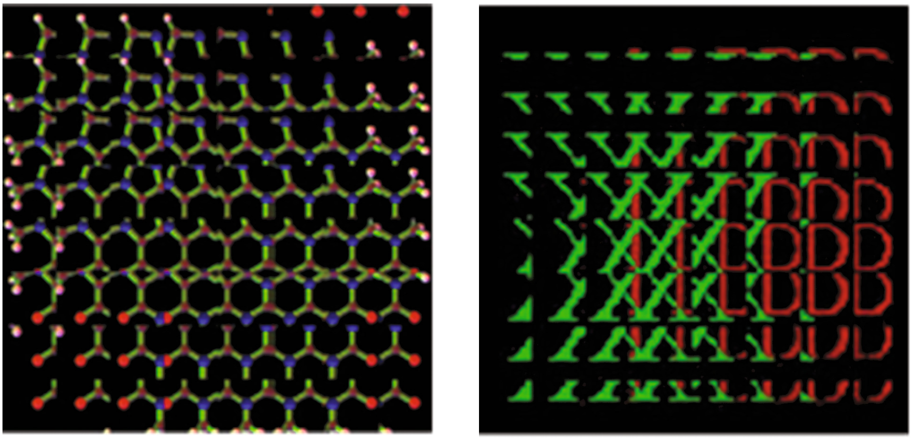
Among them, the diffraction distance z'_W can be calculated by the following formula:

$$\frac{1}{z_W} + \frac{1}{z'_W} = \frac{1}{f} \tag{6}$$

Then 3D digital watermarking is identified and displayed by the integral imaging cryptosystem.

3 Experimental Results and Discussions

As shown in Fig. 2, respectively are the EIAs which are pickuped by integral imaging cryptosystem. Ball-stick EIAs are chosen as the cover images and XD EIAs as digital watermarking images, with sizes from 64 × 64 pixels, to 1024 × 1024 pixels. Joint Photographic Experts Group (JPEG) is the image format.



(a) cover image (b) watermark image

Fig. 2. EIAs of ballstick and XD models

Corresponding 3D digital watermark images of EIAs can be recovered and displayed in the reconstruction algorithm of the integral imaging cryptosystem. The most optimal position and the best pixel are selected as the position of the image point to determine 3D images with the best image point. As shown in Fig. 3, 3D digital watermarks are showed the reconstructed image taken by a camera located 50 inch away from the integral imaging cryptosystem from different viewpoints. Low display quality and resolution of 3D images can be affected by the characteristic parameter and resolution of projectors, technological level of the lenslet array, and so on.

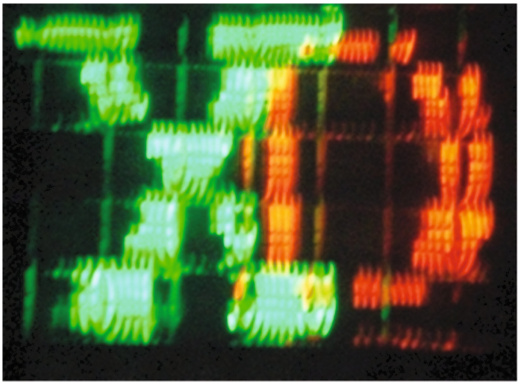


Fig. 3. Display the extracted 3D digital watermarking

Compared with traditional computer cryptography based on mathematical theorems and electronic information security techniques, the comparison analysis is depicted in Fig. 4. As different curves indicate, the 3D digital watermarking images were attacked by active attack, such as uniform noise, Gaussian noise and cropping noise. They are higher anti-noise attacking performance with DFD transform than DWT. XD EIAs are recorded and selected as test images of 3D digital watermarking. It can be known from Fig. 4 that the new proposed system has stronger anti-noise attacking performance after being attacked in three kinds of noises. Anti-attack performance should be continuously reinforced when improving the safety of 3D digital watermarking.

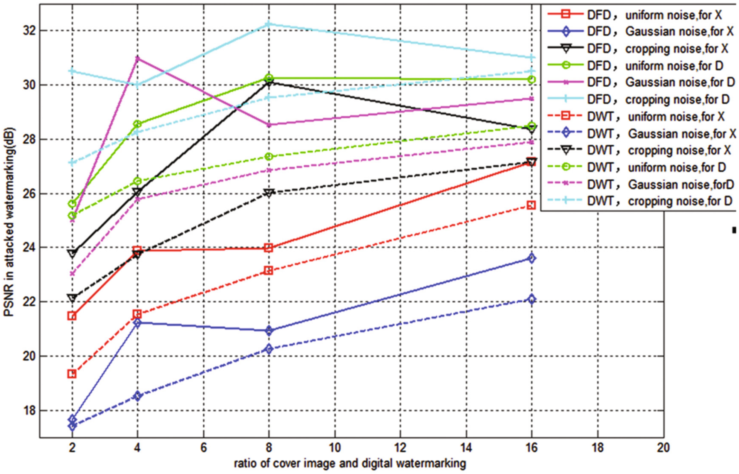


Fig. 4. PSNR values in attacked 3D digital watermarking

With application of RPMP technology, information hiding and digital watermarking system can be enhanced on the safety, reliability and robustness. Noise attack, geometric attack and other illegal attack can be effectively resisted. Since each pixel on output image includes all information of input image, when extracting digital watermarking, 3D watermark image can be restored and displayed in combination with holographic-like properties [14, 19] of EIAs. Therefore 3D digital watermarking can be successfully recovered and displayed with computational reconstruction algorithm of integral imaging.

For 3D digital watermarking achieved in DWT method, the size of digital watermarking image is limited by the size of cover image, and specific proportional relation should be conformed to the relationship. Expected result can be reached only if size ratio of cover image and watermark image meets requirements. However, new proposed technique in the paper overcomes proportion limitation of cover image and watermark image, size ratio of cover image and watermark image could be 1:1, and embedding capacity of digital watermarking can be obviously enlarged.

4 Conclusion

The paper presents a scheme of 3D digital watermarking embedding and extraction algorithm that is designed with 3D integral imaging cryptosystem based on DFD transform. The performance of proposed 3D digital watermarking with integral imaging cryptosystem is characterized by the experimental results. The new method is not only able to meet the requirements of robustness and security, but image quality and display quality achieve the criterion of the human visual model. The proposed method opens up a new research perspective for the copyright protect of 3D digital multimedia products. Although integral imaging is a major technique in the next generation auto-stereoscopic display, most of the basic ideas and 3D digital watermarking algorithms were proposed more than tens of years ago or even 100 years ago, none of them are without critical issues that are obstacles to catching a mass market of the integral imaging. In the future work, we will optimize the algorithm. The property of the real-time and security performance will be also enhanced.

References

1. Markman, A., Carnicer, A., Javidi, B.: Security authentication with a three-dimensional optical phase code using random forest classifier. J. Opt. Soc. Am. A **6**, 1160–1165 (2016)
2. Javidi, B., et al.: Roadmap on optical security. J. Opt. **8**, 1–39 (2016)
3. Pereira, R., Pereira, E.G.: Future internet: trends and challenges. Int. J. Space-Based Situated Comput. (IJSSC) **3**, 159–167 (2015)
4. Bessis, N., Asimakopoulou, E., Xhafa, F.: A next generation emerging technologies roadmap for enabling collective computational intelligence in disaster management. Int. J. Space-Based Situated Comput. **1**, 76–85 (2011)
5. Refregier, P., Javidi, B.: Optical image encryption based on input plane and Fourier plane random encoding. Opt. Lett. **7**, 3 (1995)

6. Peng, X., Zhang, P.: Security of virtual-optics-based cryptosystem. Optik Int. J. Light Electron Opt. **11**, 525–531 (2006)
7. Peng, X., et al.: Three-dimensional vision with dual acousto-optic deflection encoding. Opt. Lett. **15**, 1965–1967 (2005)
8. Lippmann, G.: Épreuves réversibles donnant la sensation du relief. J. Phys. Théor. Appl. **1**, 821–825 (1908)
9. Yontem, A.O., Onural, L.: Integral imaging using phase-only LCoS spatial light modulators as Fresnel lenslet arrays. J. Opt. Soc. Am. A Opt. Image Sci. Vis. **11**, 2359–2375 (2011)
10. Xiao, X., et al.: Advances in three-dimensional integral imaging: sensing, display, and applications. Appl. Opt. **4**, 546–560 (2013)
11. Markman, A., Wang, J., Javidi, B.: Three-dimensional integral imaging displays using a quick-response encoded elementa. Optica **5**, 332–335 (2014)
12. Li, X.W., Cho, S.J., Kim, S.T.: Combined use BP neural network and computational integral imaging reconstruction for optical multiple-image security. Opt. Commun. **1**, 147–158 (2014)
13. Liu, Y., et al.: An improved security 3D watermarking method using computational integral imaging cryptosystem. Int. J. Technol. Hum. Interact. **2**, 1–21 (2016)
14. Hwang, D.-C., Shin, D.-H., Kim, E.-S.: A novel three-dimensional digital watermarking scheme basing on integral imaging. Opt. Commun. **1**, 40–49 (2007)
15. Muniraj, I., Kim, B., Lee, B.-G.: Encryption and volumetric 3D object reconstruction using multispectral computational integral imaging. Appl. Opt. **27**, 25–32 (2014)
16. Lin, J., Nishino, H.: A construction-from-parts-type 3D modeller for digital fabrication. Int. J. Space-Based Situated Comput. **4**, 230–241 (2015)
17. Akase, R., Okada, Y.: WebGL-based 3D furniture layout system using interactive evolutionary computation and its user evaluations. Int. J. Space-Based Situated Comput. **3**, 143–164 (2014)
18. Sun, N., et al.: A correction algorithm for stereo matching with general digital cameras and web cameras. Int. J. Space-Based Situated Comput. **3**, 169–184 (2013)
19. Liu, Y., Wang, X., Zhang, J., Zhang, M., Luo, P., Wang, X.A.: An improved security 3D watermarking method using computational integral imaging cryptosystem. Int. J. Technol. Hum. Interact. (IJTHI) **2**, 1–12 (2016)
20. Steinbauer, M., Anderst-Kotsis, G.: DynamoGraph: extending the Pregel paradigm for large-scale temporal graph processing. Int. J. Grid Comput. **2**, 141–151 (2016)
21. Kohana, M., Okamoto, S.: Access control for a confirming attendance system. Int. J. Space-Based Situated Comput. **2**, 121–128 (2016)
22. Pereira, R., Pereira, E.G.: Future internet: trends and challenges. Int. J. Space-Based Situated Comput. **3**, 159–167 (2015)
23. Mokadem, R., Hameurlain, A.: Data replication strategies with performance objective in data grid systems: a survey. Int. J. Grid Comput. **1**, 30–46 (2015)
24. Chasaki, D., Mansour, C.: Security challenges in the internet of things. Int. J. Space-Based Situated Comput. **3**, 141–149 (2015)
25. Bashar, A.: Graphical modelling approach for monitoring and management of telecommunication networks. Int. J. Space-Based Situated Comput. **2**, 65–75 (2015)
26. Thabet, M., Boufaida, M., Kordon, F.: An approach for developing an interoperability mechanism between cloud providers. Int. J. Space-Based Situated Comput. **2**, 88–99 (2014)
27. Nishimura, M., Nishino, H., Kagawa, T.: A digital contents management system using a real booklet interface with augmented reality. Int. J. Space-Based Situated Comput. **3**, 194–202 (2014)

28. Moore, P., et al.: Detection of the onset of agitation in patients with dementia: real-time monitoring and the application of big-data solutions. Int. J. Space-Based Situated Comput. (IJSSC) **3**, 136–154 (2013)
29. Akase, R., Okada, Y.: WebGL-based 3D furniture layout system using interactive evolutionary computation and its user evaluations. Int. J. Space-Based Situated Comput. (IJSSC) **4**(3–4), 143–164 (2014)
30. Balusamy, B., Krishna, P.V.: Collective advancements on access control scheme for multi-authority cloud storage system. Int. J. Grid Util. Comput. Spec. Issue Intell. Grid Cloud Comput. **6**(3–4), 133–142 (2015)
31. Alamareen, A., Al-Jarrah, O., Aljarrah, I.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web Eng. (IJITWE) **3**, 1–14 (2016)
32. Honarvar, A.R., Sami, A.: Extracting usage patterns from power usage data of homes' appliances in smart home using big data platform. Int. J. Inf. Technol. Web Eng. (IJITWE) **2**, 39–50 (2016)
33. Wang, Y., et al.: Degradation and encryption for outsourced PNG images in cloud storage. Int. J. Grid Util. Comput. **1**, 22–28 (2016)
34. Voelz, D.: Computational Fourier Optics a MATLAB Tutorial, pp. 63–168. SPIE Press, Bellingham (2010)

Building Real-Time Travel Itineraries Using 'off-the-shelf' Data from the Web

Ayushi Gupta(✉), Sharmistha Rai(✉), Himali Singal(✉), Monika Chaudhary(✉), and Rishabh Kaushal(✉)

Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi, India
ayushigupta.noida@gmail.com, raimistha911@gmail.com, himalisingal3@gmail.com, monika.ch13@gmail.com, rishabh.kaushal@gmail.com

Abstract. Existing travel related systems and commonly used websites have some major limitations which cause efforts to be made by the traveler before going out on vacation. Some of these sites allow users to write their personal experiences about visited places but don't produce a proper itinerary, and those which do, focus only on minimizing the travel time between POIs ignoring other important factors like POI ratings, traffic conditions, etc.

Our work focuses on Building Real-Time Travel Itineraries using 'off-the-shelf' data from the Web. The proposed solution solves the existing limitations by using an optimization algorithm, which produces a real-time itinerary after optimizing various important factors like travel time between POIs, traffic conditions, ratings of POIs, to enhance the traveler's experience in a city.

Out of the several optimization approaches available, an algorithm was finalized after comparison of performance and accuracy between the approaches. Best results were obtained in case of a dynamic programming based approach, which optimized both accuracy and performance.

Keywords: Traveling salesman problem · Data collection · POI ratings

1 Introduction

There are many travel websites floating in the market today, whose main motive is to provide any travel enthusiast with all the information about various locations that people may want to visit. Some examples of the most commonly used sites are - TripAdvisor Inc.[1], TripHobo[2], and Tripoto.[3]

[1] https://www.tripadvisor.in/. It is an American travel website company providing reviews of travel-related content along with interactive travel forum.

[2] https://www.triphobo.com/. It is a Pune, India based website developed to formulate basic travel itineraries of famous travel spots.

[3] https://www.tripoto.com/. It is a travel community which is based on data provided by users. It looks like a social networking website where people share blogs about the places they have traveled.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_54

While these might seem like ideal websites for users to help them plan perfect travel itinerary, the major focus of these sites is generally on package bookings and user reviews, or providing platforms for sharing travel experiences. Thus, there is no proper itinerary generation in these cases. Apart from this, though there do exist sites that produce itineraries, but they use static plans and don't formulate one on the spot. So, even if we get a travel itinerary, the generation is not done on-the-fly.

All these travel applications have some or the other shortcoming, which form the motivation of our work. Our work aims to ensure improved user experience and satisfaction. While the existing applications just give a basic review of the POI and provide a commercialized view of traveling, we provide a solution that optimizes the user's time and experience of travel.

We start off by obtaining the list of most popular POIs of the city that user wants to visit. This collection is done by obtaining data from the web, using sites like Foursquare and TripAdvisor and their APIs. This list provides the name of POIs along with their ratings. To ensure an unbiased view, the POI list of both the above mentioned sites are merged to produce another list.

Google Maps API is then used to obtain a matrix, that considers the time to travel between each POI in the list, and the current traffic conditions and provides shortest travel time between each available pair of POI in the list.

After the data collection phase, a dynamic programming based algorithm (modified Held-Karp Algorithm) is used to optimize the obtained factors, and produce a resultant itinerary that allows the user to travel those POIs in an efficient manner. This algorithm was properly validated by comparing the results of other available options like Brute Force approach and Greedy approach. It is then modified to include a profit factor instead of conventional travel time considered in Traveling Salesmen Problems. For a better user experience, the itinerary route is also shown pictorially on Google Maps.

2 Related Work

A lot of work that has been done in the travel sector till now, with a lot of research still going on. Study of these works reveal that there exist some related work offering solutions to the problem, however each has its own drawbacks.

Damianos et al. [7] emphasized on considering POI ratings and travel time while planning a trip, but ignored other relevant factors like traffic conditions, user preferences etc. Tseng et al. [14] proposed a WEB-Based Tour Planning Support System Using Genetic and Ant Colony Algorithms. The paper just focused on possible solutions of how to visit the nodes, but no way on how to decide those nodes. Yahi et al. [15] proposed an algorithm which aimed to create an itinerary by modifying the Traveling Salesman algorithm. The paper associated a profit factor of happiness along with TSP and consequently applied TSP on this modified version of a node to create a path such that the profit i.e. the happiness level is maximized. This paper failed to cater to the aspect of time in the proposed itinerary. Smirnov et al. [13] proposed an approach for tourism application in which a trip was designed on the basis of the time schedules, ease-of-access and the handling abilities of the local transport services, places of interest etc. Although there was a shortcoming that no itinerary was generated. Cacho et al. [2] presented a platform that used social media as a data source to support the decisions of policymakers in the context of smart tourism destinations initiatives. The work didn’t address the common visitor, aiming to travel to a destination, who find it difficult to plan their trip, and focused only on the policy makers. Casillo et al. [3] introduced an Adaptive Context Aware Application for tourism. The app didn’t provide a proper trip schedule for the POIs of the current place, and since the app was context dependent, the visitor could not get the trip plan before actually reaching the place. Dhiratara et al. [5] talked about the analysis of social media responses to determine the tourist visiting pattern for various tourist attractions. Social media provides a live real-time feedback system for these patterns. This helps to create a hierarchy of the various POIs on the basis of their popularity. The shortcoming of this was, even though the travel time constraint was addressed, no proper itinerary was made. Alcoba et al. [1] described the use of sentiment analysis technology to analyze visitor’s reviews in order to design experiences in the tourism business sector. The work allowed visitors to choose whether to visit a POI or not on the basis of sentiment analysis done but didn’t provide a well planned trip to suite visitor’s time of stay or time of travel for a POI. Rossetti et al. [12] considered text based reviews as feedback and used them for analysis and future recommendations. It is simply a recommendation system, not producing any itinerary. Nakatoh et al. [11] did a statistical analysis from tourist reactions. In this paper, the POIs were ranked on the basis of reviews only with no attention towards time constraints. The paper failed to create an itinerary and could only create a rank list for the POIs. Jossé et al. [9] did a static search in a database to create an efficient travel path considering travel time and popularity on the basis of quantitative reviews of earlier visitors. Zacarias et al. [16] proposed a tool which helped to visit the POIs and places like hotels, places to eat, etc. It asked the user to select the POI (which are selected automatically in our work) and aimed at minimizing the travel time only without considering the POI ratings as a profit factor. Drosatos et al. [6] formulated Pythia which was a suggestion system for tourists, with focus on privacy. The app only recommended suitable POIs to a user based on her current location, without any itinerary generation. Liu et al. [10] talked about travel paths on the basis of online feedback by focusing on current failures. A new system was developed which took into account the data from both users and POIs but interest value of the attractions was neglected as the system focused on better commute time. Zhou et al. [17] used geo-tagged images on Flickr to rank and order the various POIs for further recommendations. It failed to create a well-defined itinerary and relied heavily on one source. Cenamor et al. [4] Presented PlanTour, a system that created personalized tourist plans using the human-generated information gathered from social network. No real time data was being used by the system to evaluate the making of the plan. Gu et al. [8] aimed to find all the places of attraction in Shenzhen, China through performing data mining on the check-ins that had occurred over a period of two years. While the paper found all the places to visit in the city, it provided no idea on what order to follow while visiting them.

3 Methodology

The methodology involves two major phases. The first step is Data Collection phase - it involves gathering the list and ranking of POIs of the preferred city. It is followed by the second phase which is the Algorithm Construction for Itinerary - wherein the optimization algorithm is executed on the obtained list of POIs, that includes factors like travel time, traffic conditions, ratings etc. to produce the required itinerary.

3.1 System Design

Figure 1 represents all the components of the system.

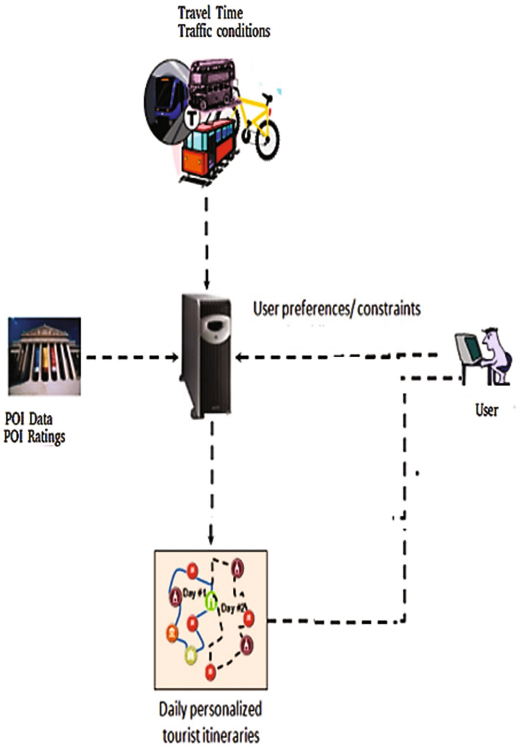


Fig. 1. System flow

3.1.1 User

The user can be any individual traveler or an organization which wants to obtain a travel itinerary as per the information entered by him/her like City of travel, Duration of stay.

3.1.2 POI Data and POI Ratings

POI data refers to the list of POIs along with their rankings for a particular city of interest of the traveler. Rather than using the list provided from a single source the aim is to make the system more reliable and unbiased by integrating resultant list obtained by the two sources namely Foursqaure and TripAdvisor.

3.1.3 Travel Time and Traffic Conditions

To make the user experience better and formulate an efficient itinerary at the end, factors like Travel Time for reaching POIs have to be optimized. Thus, to obtain this information, the Google Maps API is used which provides the final results after taking into consideration the traffic conditions at that time.

3.1.4 Daily Personalized Tourist Itinerary

The desired output of the system, wherein the user of the system is provided with a proper, day-wise plan/tourist itinerary as per the specifications in the form of input provided by them in the beginning. The itinerary will be distributed across the total number of days for which the tourist is traveling.

3.2 Phase-1: Data Collection

The data collection phase involves collecting data from the web, which basically is ‘off-the-shelf-data’. The data here refers to the list of all the major attractions or POIs of a particular city of interest. Along with this, useful information is also obtained about each POI like, rating given to the POI, travel time to a POI from another POI, traffic conditions encountered while reaching a POI, etc. This data is then processed and it serves as input for the optimization algorithm.

3.2.1 List of POIs

The Foursquare API provides methods for accessing a resource such as a venue, tip, or user. Foursquare API lets the user access various methods, of which we use the Explore method. To use this method, an authorization key was obtained with which the API returns the required list of all the POIs near the given input place with the following attributes: ‘name’, ‘rating’, ‘category’, ‘address’, etc. The input is organized in a JSON format, so that the individual keys can be identified and used as input for next steps.

To obtain the list of POIs for a given place as provided by TripAdvisor, TripAdvisor API is used, which also requires an authentication key for its access. A list of POIs in order of their ranking as given on the Trip Advisor site is thereby obtained, such that the first POI is the one with the maximum rating and the last one is the one having the lowest rating. For obtaining an authentic list of POIs for a given city, the list produced by the two data sources is merged, to produce a final resultant list.

3.2.2 Distance Between POIs

The Google Maps API is a service that provides distance and travel time between a source and destination. We send a list of POIs to the API and receive a matrix of travel time for each possible combination of POIs. The latitude and longitude values obtained from the above two sources is sent as origins and destinations to the Google API which returns a JSON response. The response can then be used to generate the matrix where each row and each column correspond to one POI. Each cell represents the optimized travel time between each pair of POIs, after considering traffic conditions, and distance between those POIs.

For the purpose of considering ratings, along with factors like travel time, traffic conditions, and distance between the POIs, a formula is applied to each entry of the Google API matrix. The algorithm then minimizes this value to produce the best possible itinerary.

Formula: each entry m(i,j) of the matrix is replaced by a value m(i,j)/sum of values of all the other factors.

3.3 Phase-2: Algorithms for Itinerary Construction

The problem regarding existing itinerary planning can be solved by providing the traveler with a list of all the major tourist attractions of a city along with the order of traversal.

Travel itinerary planning is a difficult optimization issue, similar to the Traveling Salesman Problem (TSP). Any tour plan, apart from the scenic spots, requires a great consideration of the requirements of tourists. Therefore, such a planning task can be regarded as solving a multi-constrained tour itinerary planning problem.

Various approaches exist to solve this problem which is theoretically NP-hard.

3.3.1 Brute Force

The most direct solution would be to try all permutations (ordered combinations) and see which one is cheapest (using brute force search). The running time for this approach lies within a polynomial factor of O(n!), i.e. the factorial of the number of POIs. Such a high time delay makes this algorithm infeasible for us.

To implement this approach, we have to restrict the number of POIs to 10 so that an output can be obtained in reasonable time.

Pseudocode

```
function algorithm Brute Force ()
T = initial_tour
best_tour = T
min_traveltime = traveltimeT
while there are more permutations of T do
    generate a new permutation of T
    if traveltime(T) <min_traveltime then
        best_tour = T
        min_traveltime = traveltime(T)
    end
end
print best_tour and min_traveltime
```

Algorithm 1: Function algorithm Brute Force ()

3.3.2 Dynamic Programming: Held-Karp Approach

HeldKarp algorithm solves the problem in time $O(n^2 2^n)$. This algorithm gives an optimal solution of the TSP. There is an optimization property for TSP: *Every subpath of a path of minimum distance is itself of minimum distance.*

Pseudocode

```
function algorithm TSP (G, n)
for $k := 2$ to $n$ do
    C({1, k}, k) := $d_{1,k}$
end
for $s := 3$ to $n$ do
    for all $S \subseteq 1, 2, ..., n$, $|S| = s$ do
        for all $k \in S$ do
            C(S, k) = $\min_{m \neq 1, m \neq k, m \in S}$ [C(S - {k}, m) + $d_{m,k}$ ]
        end
    end
end
opt := $\min_{k \neq 1}$ [C({2, 3, . . . , n}, k) + $d_{k,1}$]
return (opt)
```

Algorithm 2: Function algorithm TSP (G, n)

3.3.3 Greedy Approach: Nearest Neighbor Algorithm

The nearest neighbor (NN) algorithm (a greedy algorithm) lets the tourist choose the nearest unvisited POI as his next move. For ‘n’ POIs randomly distributed in a city, the algorithm manages to fetch a path one-fourth of a fraction longer than the shortest possible path.

This algorithm has a complexity of O(n) however it generates a sub-optimal solution.

Pseudocode

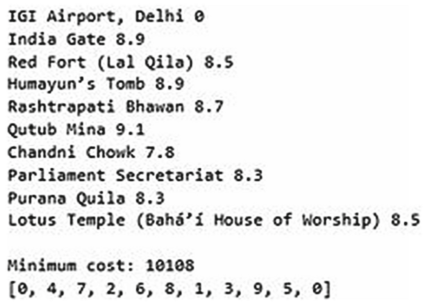
```
function algorithm Greedy (c[n][n], n, time)
maxTime = time
i = 0
current = 0
visited[n] = 0
j = 0
while len(visited) < n +1 do
    minEdge = None
    next = None
    if $c[current][i] \neq 0$ and (minEdge == 0 or c[current][i] < minEdge)
    and $i \notin visited$ then
        minEdge = c[current][i]
        next = i
    end
    cost+=minEdge
    current = next
    visited[++j] = i
end
print cost and visited
```

Algorithm 3: Function algorithm Greedy (c[n][n], n, time)

4 Results and Conclusion

The output obtained for the city of Delhi, India by Brute-Force algorithm is as shown in Fig. 2.

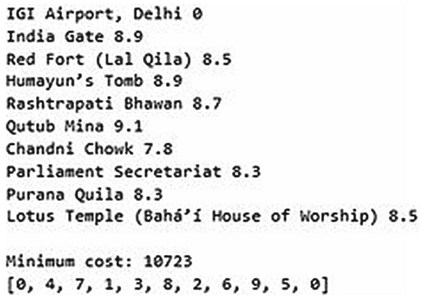
Another possible optimization approach will be dynamic-programming based, called the Held-Karp algorithm. The output obtained after using the Held-Karp algorithm for the list of 10 POIs in Delhi, India is as shown in Fig. 2.



```
IGI Airport, Delhi 0
India Gate 8.9
Red Fort (Lal Qila) 8.5
Humayun’s Tomb 8.9
Rashtrapati Bhawan 8.7
Qutub Mina 9.1
Chandni Chowk 7.8
Parliament Secretariat 8.3
Purana Quila 8.3
Lotus Temple (Bahá’í House of Worship) 8.5

Minimum cost: 10108
[0, 4, 7, 2, 6, 8, 1, 3, 9, 5, 0]
```

Fig. 2. Output obtained after applying the algorithm on the list of POIs. First the list of POIs is given along with their ratings where the numbering starts from 0. Minimum cost is minimum amount of time in seconds that will be required by the traveler to cover all these points. Lastly, the set shows the order in which these POIs are to be covered.



```
IGI Airport, Delhi 0
India Gate 8.9
Red Fort (Lal Qila) 8.5
Humayun’s Tomb 8.9
Rashtrapati Bhawan 8.7
Qutub Mina 9.1
Chandni Chowk 7.8
Parliament Secretariat 8.3
Purana Quila 8.3
Lotus Temple (Bahá’í House of Worship) 8.5

Minimum cost: 10723
[0, 4, 7, 1, 3, 8, 2, 6, 9, 5, 0]
```

Fig. 3. Output obtained after applying Greedy Algorithm on the list of POIs. First the list of POIs is given along with their ratings where the numbering starts from 0. Minimum cost is minimum amount of time in seconds that will be required by the traveler to cover all these points. Lastly, the set shows the order in which these POIs are to be covered.

The results generated by Held-Karp algorithm and Brute Force algorithm are the same.

The output produced by Greedy algorithm for Delhi, India is as shown in Fig. 3. The format of output is same as for Brute Force and Held-Karp, with differences in the values obtained.

4.1 Results

Various possible optimization algorithms were applied to the cities of Delhi, Mumbai, in India for analyzing different parameters and thus reaching upon a proper conclusion. The two parameters analyzed help us conclude that only the brute force approach and the Held-Karp algorithm provide accurate solutions. However, since the brute force approach has a high complexity, it restricts the number of POIs that can be visited (Table 1).

Table 1. Comparison of the idle time obtained, after applying each algorithm on same data. All algorithms cover the same number of POIs in each scenario. It can be seen that the idle time in case of Brute Force and Held-Karp is more than Greedy as they generate the more optimized path for final itinerary generation, while the latter fails to so.

Travel time available	Brute force algorithm	Greedy algorithm	Held-Karp algorithm
20 h	2.6 h	3.2 h	2.6 h
15 h	1.4 h	1 h	1.4 h
10 h	1.2 h	30 min	1.2 h
7.5 h	23 min	11 min	23 min

Furthermore the time complexity of Held-Karp algorithm comes out to be the best one amongst all three. Brute force doesn't work for POIs more than 10, and although Greedy has a complexity of O(n) but gives us a sub-optimal solution only. Therefore, we conclude that the modified Held-Karp algorithm, along with a profit factor is to be used for implementation of the system.

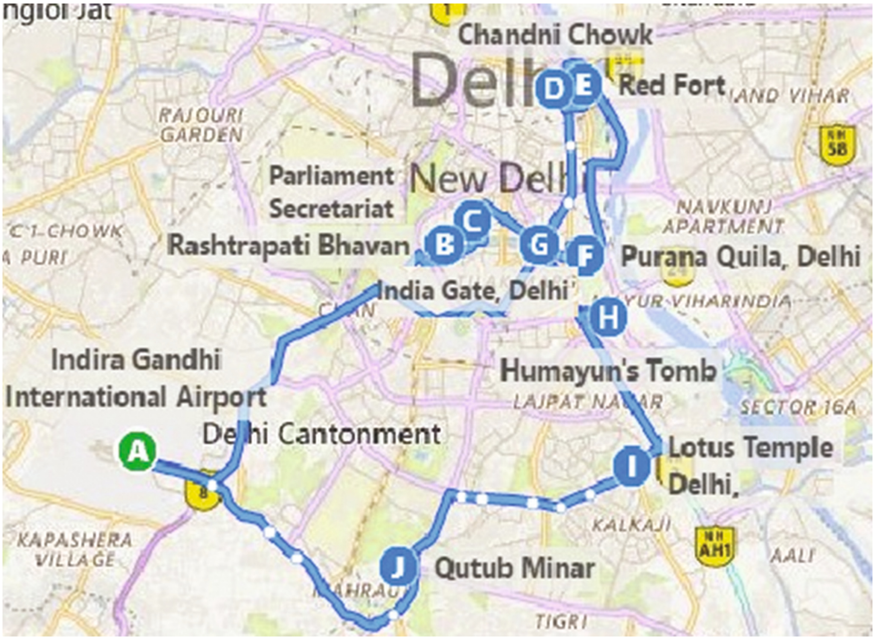


Fig. 4. Final output of the optimized path to be followed by the traveler for Delhi, India. The user's current location is the Airport, labeled A. User then runs our algorithm along with the time available with them for travel which gives this map as a result showing all the POIs that they must cover and the path that they must follow.

The final result of an optimized path for Delhi, India is presented to the user in the form of a map using Google Map API as shown in the Fig. 4. The same analysis is conducted on another city - Mumbai, India. While implementing the algorithm on the new city we see a change in the number of POIs covered for each algorithm, given the same amount of travel time by each traveler as shown in Table 2.

So here we see that we can obtain different conclusive data using same parameter as a base for reference.

4.2 Conclusion

The absence of any itinerary maker which uses real time data for evaluation and generation of an optimized path to travel a city is still an untouched concept,

Table 2. Comparison of the total number of POIs covered for Mumbai, India, after applying each algorithm on same data. All the algorithms work over the same amount of travel time in each scenario. It can be seen that Brute Force and Held-Karp manage to cover more POIs than Greedy in all the cases, for the amount of travel time entered by user.

Travel time available	Brute force algorithm	Greedy algorithm	Held-Karp algorithm
20 h	9	8	9
15 h	6	5	6
10 h	4	3	4

which will play a great role in increasing the user's travel experience. While the existing applications usually just give a basic review of the POIs and provide a commercialized view of traveling, we aim to provide a solution for the user to optimize their time and experience of travel. After application of various algorithms on different cities and consequent analysis of the results, it is safe to conclude that Held-Karp Algorithm will be the best algorithm to implement our Smart Travel Itinerary Generator. The algorithm gives the same result as Brute Force with lesser time complexity and manages to cover all POIs giving a result with higher average rating.

This will play a definite role in enhancing smart tourism in a city and enriching a traveler's experience, subsequently helping in the establishment of smart cities.

4.3 Limitations

Apart from those already considered, many factors can be weighed/included into the Google API Matrix for creating a more useful application such as:

- Including time of stay at each POI, closed days.
- Using traveller's total duration of stay at the vacation city, as in the number of days.
- considering tourists preferences monuments, parks, museums.

Acknowledgement. We have taken a lot of efforts for this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to Dr. Seja K.R. and the review committee for their invaluable guidance and constant supervision in this project.

References

1. Alcoba, J., Mostajo, S., Paras, R., Mejia, G.C., Ebron, R.A.: Framing meaningful experiences toward a service science-based tourism experience design. In: International Conference on Exploring Services Science, pp. 129–140. Springer, May 2016

2. Cacho, A., Figueredo, M., Cassio, A., Araujo, M.V., Mendes, L., Lucas, J., Farias, H., Coelho, J., Cacho, N., Prolo, C., Prolo, C.: Social smart destination: a platform to analyze user generated content in smart tourism destinations. In: New Advances in Information Systems and Technologies, pp. 817–826. Springer (2016)
3. Casillo, M., Cerullo, L., Colace, F., Lemma, S., Lombardi, M., Pietrosanto, A.: An adaptive context aware app for the tourism. In: Proceedings of the the 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016, Data Science 2016, p. 26. ACM, August 2016
4. Cenamor, I., de la Rosa, T., Núñez, S., Borrajo, D.: Planning for tourism routes using social networks. Expert Syst. Appl. **69**, 1–9 (2017)
5. Dhiratara, A., Yang, J., Bozzon, A., Houben, G.: Social media data analytics for tourism - a preliminary study. In: KDWeb (2016)
6. Drosatos, G., Efraimidis, P.S., Arampatzis, A., Stamatelatos, G., Athanasiadis, I.N.: Pythia: a privacy-enhanced personalized contextual suggestion system for tourism. In: 2015 IEEE 39th Annual on Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 822–827. IEEE, July 2015
7. Gavalas, D., Konstantopoulos, C., Mastakas, K., Pantziou, G.: A survey on algorithmic approaches for solving tourist trip design problems. J. Heuristics **20**(3), 291–328 (2014)
8. Gu, Z., Zhang, Y., Chen, Y., Chang, X.: Analysis of attraction features of tourism destinations in a mega-city based on check-in data mining: a case study of Shenzhen, China. ISPRS Int. J. Geo-Inf. **5**(11), 210 (2016)
9. Jossé, G., Schmid, K.A., Züfle, A., Skoumas, G., Schubert, M., Pfoser, D.: Tourismo: a user-preference tourist trip search engine. In: International Symposium on Spatial and Temporal Databases, pp. 514–519. Springer, August 2015
10. Liu, H.L., Li, J.H., Peng, J.: A novel recommendation system for the personalized smart tourism route: design and implementation. In: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp. 291–296. IEEE, July 2015
11. Nakatoh, T., Hirokawa, S.: Extraction of tourism objects from blogs. In: Tourism Informatics, pp. 43–58. Springer, Heidelberg (2015)
12. Rossetti, M., Stella, F., Zanker, M.: Analyzing user reviews in tourism with topic models. Inf. Technol. Tourism **16**(1), 5–21 (2016)
13. Smirnov, A., Shilov, N., Kashevnik, A., Ponomarev, A.: Cyber-physical infomobility for tourism application. Int. J. Inf. Technol. Manage. **16**(1), 31–52 (2017)
14. Tseng, S.Y., Ding, J.W., Chen, R.C.: WEB-based tour planning support system using genetic and ant colony algorithms. J. Internet Technol. **11**(7), 901–908 (2010)
15. Yahi, A., Chassang, A., Raynaud, L., Duthil, H., Chau, D.H.P.: Aurigo: an interactive tour planner for personalized itineraries. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 275–285. ACM, March 2015
16. Zacarias, F., Cuapa, R., De Ita, G., Torres, D.: Smart tourism in 1-click. Procedia Comput. Sci. **56**, 447–452 (2015)
17. Zhou, X., Xu, C., Kimmons, B.: Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform. Comput. Environ. Urban Syst. **54**, 144–153 (2015)

Energy Efficient Integration of Renewable Energy Sources in Smart Grid

Ghulam Hafeez, Nadeem Javaid(✉), Saman Zahoor, Itrat Fatima,
Zahoor Ali Khan, and Safeerullah

COMSATS Institute of Information Technology, Islamabad 44000, Pakistan
nadeemjavaidqau@gmail.com
http://www.njavaid.com

Abstract. With the emergence of smart grid (SG), the residents have the opportunity to integrate renewable energy sources (RESs) and take part in demand side management (DSM). In this regard, we design energy management control unit (EMCU) based on genetic algorithm (GA), binary particle swarm optimization (BPSO), and wind driven optimization (WDO) to schedule appliances in presence of objective function, constraints, control parameters, and comparatively evaluate the performance. For energy pricing, real time pricing (RTP) plus inclined block rate (IBR) is used. RESs integration to SG is a challenge due stochastic nature of RE. In this paper, two techniques are addressed to handle the stochastic nature of RE. First one is energy storage system (ESS) which smooths out variation in RE generation. Second one is the trading/cooperation of excess generation to neighboring consumers. The simulation results show that WDO perform more efficiently than unscheduled in terms of reduction in: electricity cost, the tradeoff between electricity cost and waiting time, and peak to average ratio (PAR). Moreover, incorporation of RESs into SG design increase the revenue and reduce carbon emission.

1 Introduction

Traditional electric grids are designed to carry power from generation system to a large number of consumers. Moreover, traditional electric grids are inefficient to meet the modern challenges like renewable energy (RE) integration, distributed generation (DG), and demand side management (DSM). In this regard, SG has emerged as smart solution. SG is a system that includes a traditional power system and information and communication technologies (ICTs) to form a platform in which customer and utility interact via two-way communication [1]. SG also incorporates RE sources (RESs), energy storage system (ESS), smart meters (SMs), distributed storage (DS), and sensors. SG encourages user participation in energy savings, cooperation through demand response (DR) mechanism, and integration of RESs [2].

RESs are greener alternative to fossil fuel and key contributor to SG so therefore, in recent years, penetration of RESs has increased. It was reported in

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_55

2014 that wind, solar, and biomass power plants provided 60% electricity generation in Denmark; about 30% of electricity demand in Portugal was supplied by nonhydropower renewable; Spain had 29% RE generation. However, integration of RESs to SG poses significant challenges such as stochastic and intermittent nature of RE causes voltage and frequency fluctuations. Energy storage and load scheduling by DR are cost effective to mitigate stochastic and intermittent nature of RE generation [3–5]. DSM and integration of RESs are the key features of SG to cope the gap between demand and supply. DSM is the planning, monitoring strategy of utility to align stochastic demand with supply. It has six agendas: (i) DR program is price-based program in which consumers modify the behavior of their load in response to fluctuations in electricity price (EP) over time. (ii) Peak clipping is one of the DSM utility strategy to reduce the system peak load, operating cost, and load dependence on peak power plants by direct load control. (iii) Valley filling is one of the classical form of the DSM to build up the load during off peak hours (i.e. when EP is less than average price). (iv) Strategic conservation is the utility strategy to change the load shape in order to reduce both net demand and aggregated energy consumption. (v) Strategic load growth is the load shape to increase the sale beyond the valley filling. (vi) Flexible load shape is related to reliability and planning constraints. It estimates the load of demand side and forecast over the planning side. The consumers allowed to make change their load according to various incentives [6]. However, the DSM and cooperation between the distributed RESs are neglected. They formulate the optimal scheduling of appliances as genetic algorithm (GA) in [10], to achieve the desired reduction in electricity cost and peak to average ratio (PAR) in a setting with real time prices (RTP) plus inclined block rate (IBR). Other relevant work [11] utilizes heuristic based energy management controller (EMC) to optimally schedule appliances in the presence of time of use (TOU) plus IBR tariff in order to reduce electricity cost, PAR, and waiting time. However, despite its importance, proper classification of residential appliances, consumers in order to motivate consumers to cooperate/trade energy with the other users and to schedule appliances to achieve related benefits has not been well investigated.

Hence, in this paper we focus on, appliances proper scheduling for DSM and RE integration by ESS and trading/cooperation between the consumers with excess generation in order to handle time varying nature of RESs and align the stochastic demand with supply. We design generic DSM architecture to incorporate heuristic based EMC unit (EMCU) for appliances scheduling and integration of RESs with ESS and cooperation among consumers.

2 Related Work

In order to optimally schedule house hold appliances and integrate RESs numerous techniques have been presented by authors. Their work is summarized in Table 1.

Table 1. Summary of related work

References	Techniques	Objectives	Limitations
Mitigate time varying nature of RES [7]	Game theoretic approach	Minimization of time average cost of energy exchange within the grid	Reduction the number of peak power plants and frequency of interruption are not addressed
Power quality integration of RESs [8]	FACTS+VSM	Mitigation of power quality issues such as voltage fluctuation, frequency fluctuation, and harmonics	Appliances scheduling and cooperation between the DRESs are not considered
Real time residential RE integration [9]	lyapunov optimization and real time algorithm	Minimize overall system cost within finite time horizon	User comfort and PAR are ignored
Optimal power scheduling in HEMS [10]	GA	Electricity cost and PAR reduction	The complexity of system increased due division of scheduling time horizon
An efficient heuristic approach for RESs [11]	Heuristic algorithm+MKP	Electricity bill and PAR minimization and user comfort maximization	Frequency of interruption and demand curve smoothing are not addressed
An incentive based optimal energy consumption scheduling [12]	BPSO+DR+TOU	To increase the electricity bill saving	PAR, user comfort, and RESs integration are ignored
IREMS for dynamic DR in smart building [13]	GA	Mitigating intermittent nature of RERs and reducing electricity cost	Energy trading between REG users and main grid are ignored
An intelligent multi agent control system (MACS) [14]	Heuristic algorithm	Reducing energy consumption without compromising user comfort	Minimizing PAR and demand curve smoothing are not described
Agent based control for decentralized DSM (DDSM) in SG [15]	EGTTs	Peak demand reduction and efficient integration of RESs	Trading and cooperating between REG user are not mentioned
An efficient model of DSM [16]	PAR and electricity bill reduction for residential, industrial and commercial sectors	Heuristic evolutionary approach	Reduction in power consumption and pressure on the environment are ignored

3 System Model

A smart power system composed of one service provider and demand side having three sector residential sector, industrial sector, and commercial sector containing a large number of consumers, however, we specifically focus on residential sector. In this section, appliances classification and RE integration model described.

3.1 Appliances Classification

In this section, appliances are classified into two categories on the basis of performance parameters and interaction to EMCU. Detail explanation of classification is given as:

SA: This refers to the class of appliances which have wireless transceiver and data processor to use the wireless technology (i.e. ZigBee, Z-Wave, and Wi-Fi) to receive the real-time data from SM via EMCU to control their operation. These appliances make the function faster, cheaper, and in more energy efficient way. These appliances are further classified into three categories. The operation timeslots (OTS) and power rating of SA are shown in Table 2.

1. Power elastic appliances: This type of appliances have elasticity in their rated operating power such as air conditioner, water cooler, and refrigerator. These appliances have rated power interval mentioned by the manufacturer on their name plate. These appliances operate at minimum power during on peak hours and operate at rated maximum power during the off-peak hours in order to reduce electricity cost and PAR. We represent such type of appliances by A^s_{pe} and its energy consumption is denoted by E^{pe}_T. The p^{pe}_r is power rating. The energy consumption and electricity cost of power elastic appliances is given as:

$$\begin{aligned} E^{pe}_T &= \sum_{pe \in \{AC,WC,RF\}} (\sum_{t=1}^{120} (p^{pe}_r \times \ S^{pe}(t))) \\ C^{pe}_T &= \sum_{pe \in \{AC,WC,RF\}} (\sum_{t=1}^{120} (p^{pe}_r \times S^{pe}(t) \times \varphi(t))) \end{aligned} \tag{1}$$

   where C^{pe}_T denote electricity cost of power elastic appliances, status of appliances is represented S^{pe}, and $\varphi(t)$ is utility EP.
2. Time elastic appliances: Unlike power elastic appliances, these appliances have fixed rated power and have elasticity in their operation time. Washing machine, clothes dryer, and water motor are encircled under this category. These appliances are allowed to run at any time with in the user defined timeslots in order to reduce electricity cost and PAR. In addition, these appliances can be interrupted, shifted, and shutdown any time if required. This type appliances are represented by A^s_{te} and their energy consumption is denoted by E^{te}_T. The energy consumption and electricity cost of time elastic appliances is calculated as:

$$\begin{aligned} E^{te}_T &= \sum_{te \in \{WM,CD,Wm\}} (\sum_{t=1}^{120} (p^{te}_r \times S^{te}(t))) \\ C^{te}_T &= \sum_{te \in \{WM,CD,Wm\}} (\sum_{t=1}^{120} (p^{te}_r \times S^{te}(t) \times \varphi(t))) \end{aligned} \tag{2}$$

3. Essential appliances: This type of appliances is also called critical appliances. The essential appliances are electric kettle, electric iron, and oven having fixed rated power. These appliances cannot be interrupted, shifted, and shutdown during the operation until to completion. These appliances have prespecified scheduling time horizon in which appliances will operate in order enhance the comfort of the users. Its representation is given by A^s_{ea} and has power rating p^e_r. Daily energy consumption and electricity cost calculation is defined as

follows:

$$\begin{aligned} E_T^{ea} &= \sum_{ea \in \{EK,EI,OV\}} (\sum_{t=1}^{120} (p_r^{ea} \times S^{ea}(t))) \\ C_T^{ea} &= \sum_{ea \in \{EK,EI,OV\}} (\sum_{t=1}^{120} (p_r^{ea} \times S^{ea}(t) \times \varphi(t))) \end{aligned} \tag{3}$$

TA: Unlike SA, this refers to the type of appliances which can be operate and control manually without any interaction to EMCU. These appliances are used by the consumers manually if needed, such as electric bulb, fan, television, and computer. The TA cannot be scheduled because they do not communicate and interact with EMCU whereas the appliances which can be scheduled by EMCU is the only SA.

Table 2. Description of appliances

Category	SA	OTS	Power rating (KW)
Power elastic	Air conditioner	75	[0.8 1.5]
	Water cooler	70	[0.5 1]
	Refrigerator	60	[0.18 0.5]
Time elastic	Washing machine	40	0.7
	Clothes dryer	40	2
	Water motor	36	0.8
Essential	Electric kettle	20	1.5
	Electric iron	30	1.8
	Oven	25	2

3.2 RE Integration Model

Greening the power system aims to modernize the power system so that it can accommodate large-scale integration of RESs such as solar and wind. RESs integration is the practice of developing efficient ways to deliver RE to the grid and neighboring consumers at the time of needs and in order to enhance revenue and reliability of the power system. One is ESS that is capable of storing RE to smooths out fluctuation in the RE. The second way is generation combined with trading/cooperating by exchanging energy among the neighboring consumers. In this case the consumers becomes prosumers because they generate energy for their own use and selling the surplus energy. For this purpose residential consumers on the basis of energy consumption, RE generation, ESS, and energy trading/cooperating are divided into three categories: GEC, SEC, TEC, where GEC consume the energy taken from electricity grid station and neither cooperate nor generate their own energy as shown in Fig. The SEC generate their own energy and as well as take energy from the electricity grid station and from neighboring RE generating consumer to fulfill their electricity demand. Unlike the SEC, the TEC generate, store, and trade/cooperate their energy with the other consumers. The detailed description is as follows.

1. GEC: The GEC using the RTP, price incentive information and take part in the DR to reduce the electricity cost and peak power consumption by shifting their load from on peak power hours to off peak hours because these consumers only depends on electricity grid station. The GEC energy consumption calculation is performed by the following formula:

$$E_{g \in GEC}(t) = \sum_{t=1}^{120} E_g(t) \quad \forall\, t \in T_h \tag{4}$$

2. SEC: The SEC fulfill energy demand by generating their own RE, borrowing energy from neighboring REG consumer, and electricity grid station. In case the energy generated by SEC is more than the demand then it will stop taking energy from the electricity grid station, neighboring REG consumer and will store the surplus energy in the ESS. In case of energy deficiency, the SEC buy energy from the neighboring REG consumer and electricity grid station. However, the SEC buy energy at first priority from the neighbor microgrid in order to reduce the electricity cost. The energy consumed by SEC is given as under:

$$E_{s \in SEC}(t) = \sum_{t=1}^{120} E_g(t) \;+\; \sum_{t=1}^{120} N_b(t) \;-\; R(t) \pm\; S(t) \tag{5}$$

   where $E_g(t)$ is the energy taken at each timeslot from the electricity grid station, $R(t)$ is the electricity bought at each timeslot from neighboring REG consumer, is the SEC RE generation at each timeslots, $S(t)$ is the energy stored are discharged at each timeslot from the ESS.
3. TEC: Unlike, the SEC the TEC fulfill their energy demand by generating their own renewable energy, trading and cooperating with neighboring microgrid. In case when energy is sufficient or no load condition the energy is stored in the ESS and trade energy with other consumer to increase the revenue. The energy consumption can be calculated by the following equation:

$$E_{T \in SEC}(t) = \sum_{t=1}^{120} E_g(t) \mp \sum_{t=1}^{120} N_T(t) \;-\; R(t) \pm\; S(t) \tag{6}$$

   where $N_T(t)$ is the trading or cooperating energy with the other consumer and $S(t)$ storing or discharging energy. Negative sign $-N_T(t)$ shows the selling energy and positive sign means $+N_T(t)$ borrowing energy. Positive sign means $+S(t)$ battery is charging otherwise discharging.

4 Simulations and Discussion

In this paper, we performed comparative evaluation of appliances scheduling based on heuristic EMCU in order to achieve our desired objectives: Minimize the tradeoff between electricity cost and user comfort, the peak power consumption, and PAR. We tackled the challenge of RE integration and minimize the cost by trading/cooperating excess generation to neighboring consumers. The detailed description is as follows:

4.1 Electricity Cost Analysis for Different Scheduling Scheme

RTP profile of EP is midwest independent system (MISO) operator daily EP tariff taken from federal energy regulatory commission (FERC) where timeslots 30–45 (6–9 am) and timeslots 85–100 (5–8 pm) are on peak, timeslots 50–70 (10 am–2 pm) are shoulder peak, and the rest of the slots are off peak. The daily electricity cost of unscheduled and scheduled appliances for a single home is shown in the Fig. 1. In unscheduled the electricity cost is high during the timeslots 30–45 (6 am–9 am) and 85–100 (5–8 pm) because the consumer use more appliances in these timeslots which tends to lead high electricity cost of 1.25 and 1.15. We can conclude from these results even under RTP+IBR the consumers who do not adjust their power consumption appropriately may not actually benefit and in fact pay more. The maximum daily electricity cost per timeslot is reduced from 1.25 to 0.3 with the WDO which is 76% reduction in percent decrement. So, the WDO outperforms as compared to unscheduled and GA, BPSO scheduled as shown in Fig. 1 because as seen from the results it has most stable and regular pattern. Note 1 timeslot is equal to 12 min.

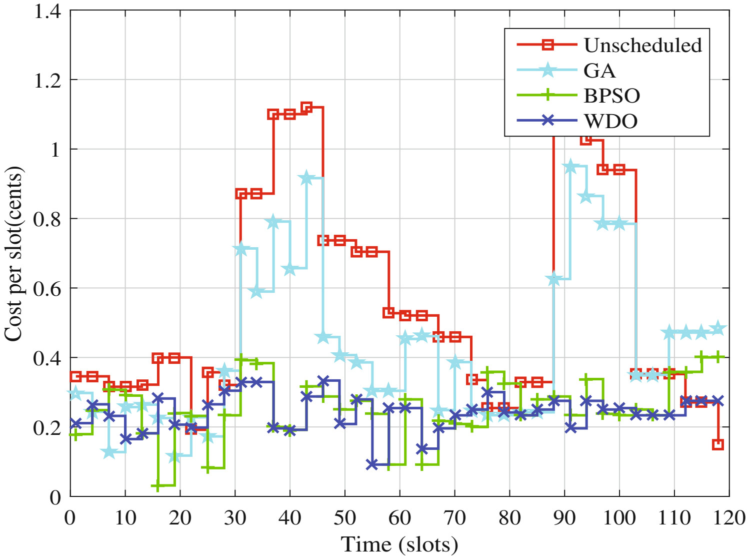


Fig. 1. Electricity cost per timeslot

4.2 Trade-Off Analysis of Electricity Cost and User Comfort

User comfort is related to both the electricity cost and appliances waiting time. The EMCU based scheduling of appliances with GA, BPSO, and WDO lead towards low electricity cost as compared to unscheduled because heuristic based EMCU is designed keeping in view the objective function, constraints, control parameters, operation time interval, scheduling time horizon, and appliances waiting time. Generally, the electricity cost and appliances waiting time are inversely related. So, the heuristic based EMCU tries to minimize the tradeoff of electricity cost and user comfort. In addition, by applying the user comfort constraint on the objective function the performance of scheduling techniques (GA, BPSO, WDO) enhanced in terms of user comfort and electricity cost. The electricity cost is high if appliances waiting time is zero and low when appliances waiting time is other than zero for all optimization techniques (GA, BPSO, WDO) as shown in Fig. 2. Performance of WDO is much better than other due to minimum effect of tradeoff.

4.3 PAR Performance Analysis

The relationship between unscheduled and scheduled load with respect to PAR are shown in the Fig. 3. The PAR minimization, emphasis on allocating appliances a time horizon to level peak loads of on peak hours to off peak hours. The PAR is reduced by EMCU based on GA, BPSO, and WDO scheduling the appliances by using DR incentive, RTP+IBR, and shifting the loads from high peak price to low peak price. The PAR of unscheduled, GA, BPSO, and, WDO

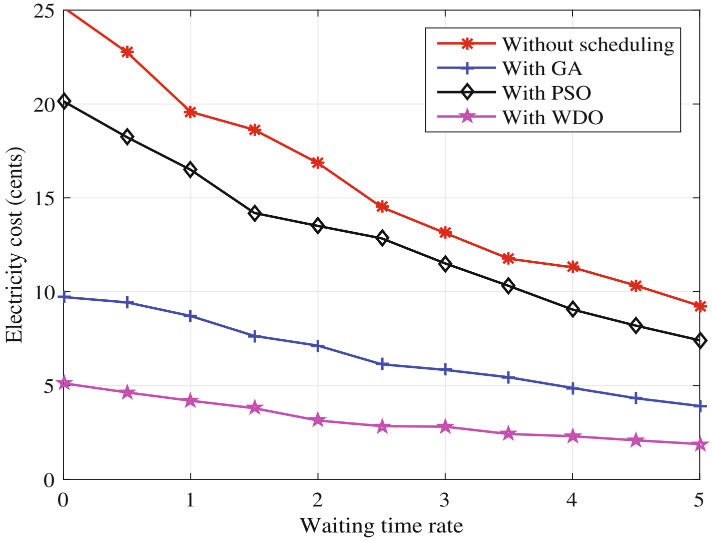


Fig. 2. Electricity cost and user comfort tradeoff

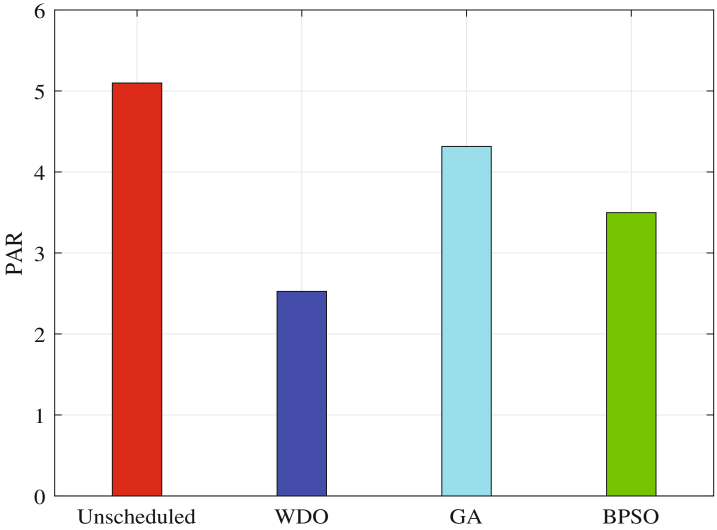


Fig. 3. PAR

scheduled load are 5.2, 4.7, 3.4, and 2.6 respectively. Their percent decrement is 15.6, 27.45, and 49.01 respectively.

5 Conclusion

In this paper, we proposed techniques for RE integration and DSM. A heuristic based EMCU is used to schedule appliances in order to achieve desired objective, and lyapnov optimization technique to optimally handle time varying and intermittent nature of RE by ESS and trading/cooperating energy exchange among neighboring consumers in order to increase the revenue and decrease the carbon emission. Simulation results show that our scheme is useful in terms of: electricity cost, PAR, and reducing the tradeoff between electricity cost and user comfort. Moreover, our solution can be useful for power grid designer in choosing the optimal combination of ESS and cooperation/trading in order to meet cost criterion.

References

1. Saad, W., et al.: Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications. IEEE Sig. Process. Mag. **29**(5), 86–105 (2012)
2. Logenthiran, T., Srinivasan, D., Shun, T.Z.: Demand side management in smart grid using heuristic optimization. IEEE Trans. Smart Grid **3**(3), 1244–1252 (2012)
3. Roselund, C., Bernhardt, J.: Lessons learned along Europes road to renewables. IEEE Spect., 4 May 2015

4. Liang, X., Bagen, B.: Probabilistic planning and risk analysis for renewable power generation system. In: Proceedings of CIGRE Canada Conference, Winnipeg, Manitoba, 31 August–2 September 2015
5. Hart, E.K., Jacobson, M.Z.: A Monte Carlo approach to generator portfolio planning and carbon emissions assessments of systems with large penetrations of variable renewables. Renew. Energy **36**(8), 22782286 (2011)
6. Khalid, A., et al.: Optimized home load management with reduced cost and peak to average ratio in smart grid with demand side management. Energies **10**(1), 1–28 (2016)
7. Lakshminarayana, S., Quek, T.Q.S., VincentPoor, H.: Cooperation and storage tradeoffs in power grids with renewable energy resources. IEEE J. Sel. Areas Commun. **32**(7), 1386–1397 (2014)
8. Member Sr, X.L.: Emerging power quality challenges due to integration of renewable energy sources. IEEE Trans. Ind. Appl. **9994**(c) (2016)
9. Li, T., Member, S., Dong, M., Member, S.: Real-time residential-side joint energy storage management and load scheduling with renewable integration. IEEE Trans. Smart Grid **3053**(c), 115 (2016)
10. Zhao, Z., Lee, W.C., Shin, Y., Member, S., Song, K.: An optimal power scheduling method for demand response in home energy management system. IEEE Trans. Smart Grid **4**(3), 13911400 (2013)
11. Rahim, S., et al.: Exploiting heuristic algorithms to efficiently utilize energy management controllers with renewable energy sources. Energy Build. **129**, 452–470 (2016)
12. Ullah, I., et al.: An incentive-based optimal energy consumption scheduling algorithm for residential users. Procedia Comput. Sci. **52**, 851–857 (2015)
13. Arun, S.L., Selvan, M.P.: Intelligent residential energy management system for dynamic demand response in smart buildings. IEEE Syst. J. **19379234**, 112 (2017)
14. Wang, L., Wang, Z., Yang, R.: Intelligent multiagent control system for energy and comfort management in smart and sustainable buildings. IEEE Trans. Smart Grid **3**(2), 605–617 (2012)
15. Ramchurn, S.D., Vytelingum, P., Rogers, A., Jennings, N.: Agent-based control for decentralised demand side management in the smart grid, p. 512 (2011)
16. Awais, M., et al.: An efficient genetic algorithm based demand side management scheme for smart grid. In: 2015 18th International Conference on Network-Based Information Systems (NBiS). IEEE (2015)

Cost and Comfort Based Optimization of Residential Load in Smart Grid

Fahim Ahmed[1], Nadeem Javaid[1](✉), Awais Manzoor[1], Malik Ali Judge[1], Fozia Feroze[1], and Zahoor Ali Khan[2]

[1] COMSATS Institute of Information Technology, Islamabad 44000, Pakistan
nadeemjavaidqau@gmail.com
[2] Higher Colleges of Technology, Fujairah Campus, Fujairah 4114, UAE
http://www.njavaid.com

Abstract. In smart grid, several optimization techniques are developed for residential load scheduling purpose. Preliminary all the conventional techniques aimed at minimizing the electricity consumption cost. This paper mainly focuses on minimization of electricity cost and maximization of user comfort along with the reduction of peak power consumption. We develop a multi-residential load scheduling algorithm based on two heuristic optimization techniques: genetic algorithm and binary particle swarm optimization. The day-ahead pricing mechanism is used for this scheduling problem. The simulation results validate that the proposed model has achieved substantial savings in electricity bills with maximum user comfort. Moreover, results also show the reduction in peak power consumption. We analyzed that user comfort has significant effect on electricity consumption cost.

1 Introduction

The electrical power grid is the gigantic man-made unit. It has often been considered as the complex engine ever developed. It comprises of synchronous and asynchronous generators, transformers, transmission lines, relays and switches, compensators, and controllers [1]. The vertical hierarchy of the electrical power system includes power generation, transmission, distribution, and consumption. The conventional grid is responsible for providing the electrical power from generation system to end user via transmission and distribution network. Conventional grid only focuses on one-way communication, vulnerable to malicious activities, lacks in rapid faults restoration, dominated by central generation, and distorted power quality. In a conventional grid, there exist a strong need to optimize several factors: unit commitment and generation planning, economic dispatch and state estimation, maintenance scheduling and dynamic security of the entire electrical network.

In this regard smart grid (SG) is considered as best possible solution to the aforementioned problems. The incorporation of information and communication technology (ICT) along with the integration of renewable energy sources (RESs) in conventional grid converts it into SG. The SG can be referred as a modern

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_56

two-way information and power flow system having the properties of self-healing, resilience and adaptability, and can predict the uncertainties. The interoperability for current and forth coming standards of devices are ensured in SG that are cyber-secured against threats [2].

Although all the components of SG are equally important, but the consumer's end that is commonly known as demand side captured special attention from many researchers. Demand side management (DSM) is one of the key components of the SG that aims at utilizing the available energy effectively and optimally. In DSM, load is categorized into three classes: residential, commercial and industrial sector. The residential sector is considered in this work because a major part of total energy is being consumed by this sector.

In [3], demand response (DR) is discussed to modify the power consumption pattern of consumers. Depending upon the pricing mechanism DSM uses different techniques to distribute the residential load over the time horizon. The techniques normally used: peak clipping, load shifting, valley filling, strategic conservation, strategic load growth and flexible load shape. The peak clipping and valley filling work on the principle of DLC, the utility is capable to turn on or off the consumer's devices remotely whenever needed [4]. In strategic conservation the load curve is reduced by using efficient devices at consumer's end, consequently the overall generation and demand is reduced. The strategic load growth is the phenomenon of introducing the load transcend the valley filling. In flexible load shape the consumers have elastic load that are available to control whenever the utility required. The load shifting technique most widely used in DSM is responsible to shift the consumer's load to off-peak hours [18].

In this paper, load shifting technique with day ahead bidding mechanism is used. Heuristic optimization technique is proposed for wide number of households appliances. Extensive simulations are conducted to validate the results. More detailed discussion is carried out in next section. The hierarchical order of the remaining paper is as follows. Section 2 contains the related work. Section 3 discusses the problem statement. In Sect. 4 system model is described in detail. Simulations and discussions are performed in Sect. 5. Section 6 concludes the work.

2 Related Work

In SG, DSM is responsible for efficient utilization of available energy. There exist different optimization techniques that can efficiently handle the energy consumption at consumer's end. Many researchers focused at both mathematical and heuristic optimization techniques which are capable to optimally scheduled the consumers load. In [5], an optimization technique: genetic algorithm (GA) is proposed, in which cost is taken as an objective function to be minimized. Yi et al. in [6] has proposed an opportunistic based optimal stopping rule (OSR) for scheduling of home appliances. Three appliances with real time pricing (RTP) are considered in this scheme. The proposed technique finds the optimal interval where prices are less than a predefined threshold while waiting time of appliances is taken under consideration. The results of the proposed technique show that consumption cost is reduced significantly with minimum user inconvenience.

Rasheed et al. in [7] has proposed an OSR based scheduling scheme for residential appliances with real time pricing environment. The proposed technique has successfully managed to reduce the electricity cost with reasonable waiting time. The appliances having short length of operation and high priorities have been scheduled earlier at the cost of less savings in electricity bills. Zhao et al. in [8] has developed an optimization technique called GA and used RTP combine with inclined block rate (IBR) pricing tariff. The main objective of the proposed technique is to reduce the electricity cost and PAR. In this way the proposed scheme has strengthened the stability of the entire grid. Rhaim et al. in [9] has analyzed the performance of three heuristic optimization techniques: GA, ant colony optimization (ACO) and binary particle swarm optimization (BPSO). The main objective of the proposed work is to reduce the electricity cost and peak to average ratio (PAR) while considering the RESs and storage system.

Althaher et al. in [10] has developed a scheme in which residential load is classified into four categories: deferrable, curtailable, thermal and critical loads. The main goal of the proposed scheme is to minimize the cost while taking care of comfort zone of the consumers in term of indoor temperature. The MINLP along with dynamic pricing scheme is used. In [11], Derakhshan et al. has proposed optimization techniques: teacher and learning based optimization (TLBO) and shuffled frog leap (SFL). In this model load is categorized into three classes: shiftable, sheddable and non sheddable loads. The proposed scheme aimed at minimizing the electricity cost. In this work three different pricing schemes are used: ToU, RTP and critical peak pricing (CPP). The results demonstrated that the proposed technique has successfully managed to reduce the consumption cost. In [12], Muralitharan et al. has proposed a multi objective evolutionary optimization technique that aimed to minimize the consumption cost while considering the waiting time of consumers. ToU pricing mechanism is used in the proposed scheme. The results of the proposed scheme have demonstrated the trade-off between cost and waiting time of consumers.

In [13], a scenario has been developed that aimed to maximize the user comfort and minimize the consumption cost. Ogunjuyigbe et al. has proposed an optimization technique that is capable to generate an optimal power consumption pattern, which offered maximum user comfort while keeping the predefined budget under consideration. In order to implement this scheme GA is used due to its flexibility and capability to handle non linearities. In [15], author has developed a novel concept of cost efficiency (CE): the ratio of total energy consumption benefits to the total electricity payments. CE is considered as an indicator for consumers to adjust their energy consumption pattern. The effects of DERs and service fee on CE have also been analyzed. The fractional programming technique along with day ahead pricing (DAP) and RTP is used in this scheme. The performance results showed that CE has increased with increasing DERs and decreased with increasing service fee.

In [14], a multi objective MINLP has been developed to optimize the residential energy consumption pattern. The main contribution of the Anvari-Moghaddam et al. has demonstrated through simulations that the proposed technique has successfully managed to reduce the electricity cost and energy consumption. Thermal and electrical comfort of the consumers has also been taken into consideration. All the aforementioned techniques performed well as per their objective functions. The problems are identified from recent research and an optimal solution is proposed which is discussed in coming section.

3 Problem Statement

In [15], Ma et al. developed a novel concept of cost efficiency-based scheduling mechanism for residential appliances. In this work cost is efficiently minimized by using fractional programming technique for day ahead bidding and real time pricing. For practical implementation, the author also considered and analyzed the effects of service fee and DERs on cost efficiency. In [16], author proposed a model based on large number of residences and appliances. The proposed model is then formulated to optimize the sum of overall satisfaction level of consumers in term of cost. PL-Generalized Benders technique is used for scheduling the residential load and protecting the private information of the consumers. The interval number optimization technique is proposed in [17] to handle residential load scheduling problem, thermostatically controlled and interruptible loads are considered in this scheme. Moreover, BPSO combined with integer linear programming is used for load scheduling. In this paper author aimed to minimize the cost while keeping the comfort (i.e.,in term of temperature) under consideration. Tolerance in degrees is taken as comfort of the consumers when thermostatically controlled load is addressed.

In [18], Logenthiran et al. proposed an evolutionary algorithm (EA) for scheduling of large number of appliances of different types. The research aimed

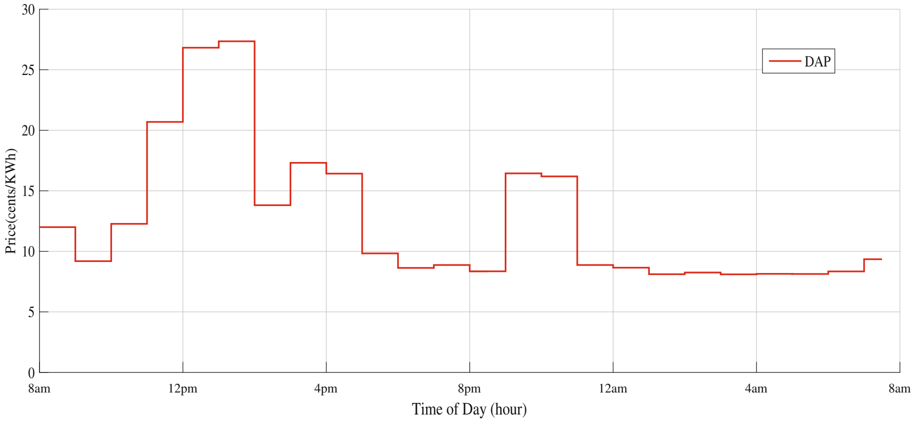


Fig. 1. Price signal

to minimize the overall cost of the user while reducing the peak power consumption. An objective curve is given by the utility, so that scheduler has to bring the final consumption curve as closer to the objective curve as possible. The day ahead load shifting pricing scheme is used for scheduling of large number of appliances. In aforementioned work, all the optimization techniques performed well while tackling the cost as an objective function, however, lack to consider the user comfort in their models. So, there is a strong need to incorporate this core problem in the optimization design. This paper develops a novel concept to maximize the user comfort and minimize the energy consumption cost along with the reduction of peak power. The combination of two heuristic optimization techniques: GA and BPSO, are used to overcome the existing problem in residential sector (Fig. 1).

4 System Model

The system model comprises of energy management controller, smart homes and communication networks.

4.1 Energy Management Controller

In this model, DSM focuses on efficient utilization of energy in residential sector. The power utility is directly connected to EMC and exchanges bidirectional information and unidirectional power flow in real time. The central EMC receives the price information from the power utility and performs the appropriate action. At the same time it contains the information from the consumer's end. It acts as a gate way between power utility and several homes.

4.2 Residential Consumers

In residential sector, the appliances subjected to control have low energy consumption ratings and short length of operation. There are 2604 controllable appliances available in this sector from 14 different types of appliances. All types of appliances have different energy consumption pattern and operation time. As in this area consumers have low priorities regarding the time when the energy has to be utilized, so more savings can be achieved in residential sector. The amount of incentives given to consumers depend on how much discomfort the consumer is willing to undergo. The power ratings of appliances and their length of operation are given in Table 1. So, the proposed technique is considered to manage electricity cost and user comfort along with peak consumption simultaneously. The parameters used in the proposed technique are given in Table 2.

4.3 Communication Network

The communication network includes wide area network (WAN), neighbourhood area network (NAN) and home area network (HAN). The residential appliances

Table 1. Appliances parameters

Appliance type	Power rating (kW)	Length of operation time (hour)	Number of devices
Dryer	1.2	4.0	189
Dish washer	0.7	3.0	288
Washing machine	2.0	2.5	268
Oven	1.3	3.0	279
Iron	1.0	2.0	340
Vacuum cleaner	2.0	1.7	158
Fan	0.2	24	288
Kettle	2.0	4.0	406
Toaster	0.9	3.0	048
Rice cooker	0.85	4.0	059
Hair dryer	1.5	2.0	058
Blender	0.3	1.5	066
Frying pan	1.1	1.5	101
Coffee maker	0.8	2.5	056
Total	-	-	2604

Table 2. Algorithm parameters

Variables	Values
Pc	0.9
Pm	0.1
Insite	1
Vmax	4
Vmin	-4
Population size	200
Maximum iterations	600

are connected to smart meter via HAN. The residential appliances share their information to the smart meter and then this information is forwarded to the central EMC. The smart meters of different homes are connected to the central EMC via NAN. Through NAN the collective and collaborative information is reached to the main EMC. The EMC exchanges the received information to the power utility via WAN. Through WAN the demand response and the price information is exchanged between power utility and the main EMC (Fig. 2).

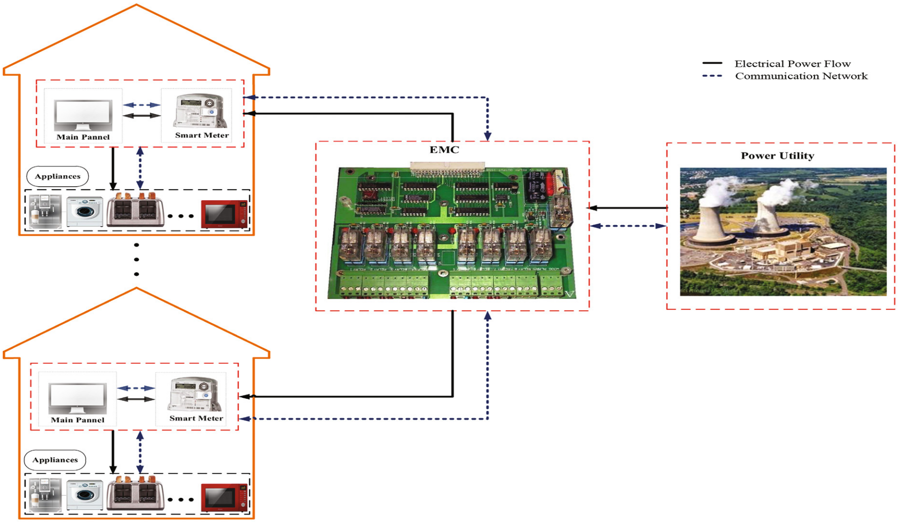


Fig. 2. System model

5 Simulations and Discussions

In this section the performance of GA alone and along with BPSO is discussed in detail. The former implementation of EA for residential load can be referred as; it has focused on electricity cost minimization along with peak power reduction. The results of this formerly implemented technique are shown in Table 3. While implementing the EA for scheduling of residential appliances, various factors are observed regarding cost minimization, efficient power consumption and peak reduction. The main objective of the algorithm is to minimize the users electricity bill while reducing the peak power. In this way user and utility acquire reasonable benefits: user in the form of reduction in electricity bill while utility in the form of cost which is incurred by using additional power plants at peak hours.

Electricity cost and peak power consumption is reduced to a considerable amount, EA has successfully managed to reduce the electricity cost by an amount of 4.97% while peak consumption is reduced by an amount of 18.275%. In this case load is shifted at off-peak hours where algorithm meets less electricity cost and consequently reduced the electricity bill and peak power consumption.

Table 3. Consumption cost and Peak load

Parameters	Unscheduled	Scheduled	Reduction (%)
Cost ($)	1581.9	1090.1	31.0955
Peak load (KW)	1706.3	1572.3	7.8532

The former implementation of EA reveals that, it has considered 1 h time slot and 12 h as maximum allowable delay for each appliance. Since, by increasing the delay of an appliance, more monetary benefits are achieved at end users. In order to transform this implementation, one hour time slot is converted into half an hour time slot. The modified scenario is implemented by using heuristic techniques: GA and BPSO, it is observed from Fig. 3 that with the incorporation of user comfort in term of waiting time, the performance parameters are also affected. It is shown (scheduled cost via GA) that the user achieved maximum monetary benefits as compared to that of unscheduled cost but compromised on convenience. Similarly, from scheduled cost with GAPSO, it is observed that user achieved comparatively less savings in electricity bills but with maximum comfort level. The maximum allowable delay of 5 h is considered. So, in this way electricity cost and user comfort both are efficiently addressed. The savings in electricity bills are decreased by 5%, this decrement in savings is due to the fact that electricity cost and user comfort are inversely proportional to each other. By increasing the user comfort, savings in electricity bills are decreased and vice versa. The tradeoff between user comfort and cost is obvious since without sacrificing the convenience user is incapable to achieve the reduction in cost.

Depending upon the consumption pattern, particularly in this case, Fig. 4 shows that PAR is reduced upto 29.367% by GAPSO, whereas GA reduced PAR by 7.8532. This is due to the fact that the proposed technique has successfully managed to optimally schedule the load over time horizon and shift the load from on-peak hours to off-peak hours while taking well care of the waiting time.

The results show that the proposed technique has tangibly outperformed the existing scheme. Throughout the simulations it has been observed that the electricity cost of the user and waiting time along with peak consumption are

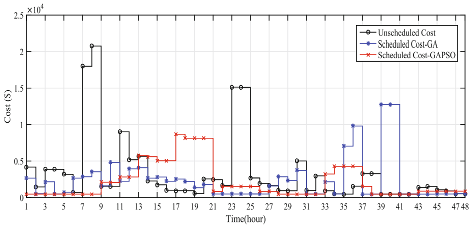


Fig. 3. Electricity cost

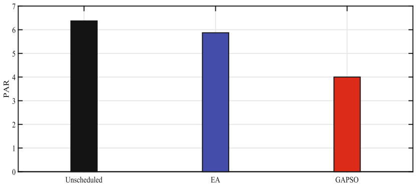


Fig. 4. PAR ratio

Table 4. Consumption cost and Peak load

Technique	Parameters	Unscheduled	Scheduled	Reduction (%)
GA	Cost ($)	1581.9	1090.1	31.0955
	Peak load (KW)	1706.3	1572.3	7.8532
GAPSO	Cost ($)	1581.9	1162.3	26.4997
	Peak load (KW)	1706	1205.3	29.367

optimally addressed. Consequently, the entire grid is protected from vulnerabilities, instabilities and outages, whereas the consumers utilized the energy efficiently with minimum electricity bills. The results are summarized in Table 4.

6 Conclusion

In this paper, a multi-residential load of different types is considered. The proposed technique is the combination of two heuristic algorithms. The main objective of the proposed technique is to reduce the consumption cost and maximize the user comfort while considering peak consumption into account. The residential power consumption behavior in time significantly affect the electricity cost. The proposed algorithm optimally distribute the residential load over the time horizon. Simulations results show that there exist a trade-off between electricity cost and user comfort in term of waiting time. The saving in electricity cost depends on how long consumer is willing to face the inconvenience. The overall performance of the proposed technique shows that consumer has achieved a significant reduction in electricity cost with a reasonable waiting time for residential devices.

References

1. Del Valle, Y., et al.: Particle swarm optimization: basic concepts, variants and applications in power systems. IEEE Trans. Evol. Comput. **12**(2), 171–195 (2008)
2. Momoh, J.: Smart Grid: Fundamentals of Design and Analysis, vol. 63. Wiley, Orlando (2012)
3. Gelazanskas, L., Gamage, K.A.A.: Demand side management in smart grid: a review and proposals for future direction. Sustain. Cities Soc. **11**, 22–30 (2014)
4. Maharjan, I.K.: Demand Side Management: Load Management, Load Profiling, Load Shifting, Residential and Industrial Consumer, Energy Audit, Reliability, Urban, Semi-urban And Rural Setting. LAP Lambert Academic Publ. (2010)
5. Miao, H., Huang, X., Chen, G.: A genetic evolutionary task scheduling method for energy efficiency in smart homes. Int. Rev. Electric. Eng. (IREE) **7**(5), 5897–5904 (2012)
6. Yi, P., et al.: Real-time opportunistic scheduling for residential demand response. IEEE Trans. Smart Grid **4**(1), 227–234 (2013)

7. Rasheed, M.B., et al.: Priority and delay constrained demand side management in real-time price environment with renewable energy source. Int. J. Energy Res. **40**(14), 2002–2021 (2016)
8. Zhao, Z., et al.: An optimal power scheduling method for demand response in home energy management system. IEEE Trans. Smart Grid **4**(3), 1391–1400 (2013)
9. Rahim, S., et al.: Exploiting heuristic algorithms to efficiently utilize energy management controllers with renewable energy sources. Energy Buildings **129**, 452–470 (2016)
10. Althaher, S., Mancarella, P., Mutale, J.: Automated demand response from home energy management system under dynamic pricing and power and comfort constraints. IEEE Trans. Smart Grid **6**(4), 1874–1883 (2015)
11. Derakhshan, G., Shayanfar, H.A., Kazemi, A.: The optimization of demand response programs in smart grids. Energy Policy **94**, 295–306 (2016)
12. Muralitharan, K., Sakthivel, R., Shi, Y.: Multiobjective optimization technique for demand side management with load balancing approach in smart grid. Neurocomputing **177**, 110–119 (2016)
13. Ogunjuyigbe, A.S.O., Ayodele, T.R., Akinola, O.A.: User satisfaction-induced demand side load management in residential buildings with user budget constraint. Appl. Energy **187**, 352–366 (2017)
14. Anvari-Moghaddam, A., Monsef, H., Rahimi-Kian, A.: Optimal smart home energy management considering energy saving and a comfortable lifestyle. IEEE Trans. Smart Grid **6**(1), 324–332 (2015)
15. Ma, J., et al.: Residential load scheduling in smart grid: a cost efficiency perspective. IEEE Trans. Smart Grid **7**(2), 771–784 (2016)
16. Moon, S., Lee, J.-W.: Multi-residential demand response scheduling with multiclass appliances in smart grid. IEEE Trans. Smart Grid (2016)
17. Wang, J., Li, Y., Zhou, Y.: Interval number optimization for household load scheduling with uncertainty. Energy Buildings **130**, 613–624 (2016)
18. Logenthiran, T., Srinivasan, D., Shun, T.Z.: Demand side management in smart grid using heuristic optimization. IEEE Trans. Smart Grid **3**(3), 1244–1252 (2012)

Efficient Utilization of HEM Controller Using Heuristic Optimization Techniques

Asif Khan[1], Nadeem Javaid[1](✉), Adnan Ahmed[1], Saqib Kazmi[1],
Hafiz Majid Hussain[2], and Zahoor Ali Khan[3]

[1] COMSATS Institute of Information Technology, Islamabad 44000, Pakistan
nadeemjavaidqau@gmail.com
[2] Center for Advanced Studies in Engineering (CASE), Islamabad 44000, Pakistan
[3] Higher Colleges of Technology, Fujairah Campus, Fujairah 4114, UAE
http://www.njavaid.com

Abstract. The performance and comparative analysis of home energy management controller using three optimization techniques; genetic algorithm (GA), enhanced differential evolution (EDE) and optimal stopping rule (OSR) has been evaluated in this paper. In this regard, a generic system model consisting of home area network, advanced metering infrastructure, home energy management controller, and smart appliances has been proposed. Price threshold policy and priority of appliance have also been considered to depict monthly and yearly average electricity bill savings and appliance delay using day-ahead real-time pricing (DA-RTP). Simulation results validate that all our proposed schemes successfully shifts the appliance operations to off-peak times and results in reduced electricity bill with reasonable waiting time.

1 Introduction

Smart grid (SG) uses new and advanced technologies including intelligent hardware, autonomous controllers, and robust software for the management of data along with two-way communication between consumers and power utilities to deliver energy in a reliable and efficient way. The main objectives of the SG are to improve the reliability, efficiency, and safety of the entire system [1]. The interactive feature of SG allows interaction among prosumers and participation of consumers in demand side management (DSM) programs. Demand response (DR) programs are in the form of financial incentives or other time-based rates which provide an ample opportunity for consumers of electricity in SG to play a vital role in shifting or reducing their usage of electricity during peak periods. DR has been considered as the most reliable and cost-effective solution to reduce peak demand and smooth the demand curve, under system stress [2].

DR offers motivation in the form of price signals to the consumers to shift or reduce their power demands. These price signals may be in the form of pricing like time-of-use (ToU), critical peak pricing (CPP), peak load pricing (PLP), real-time pricing (RTP), day-ahead pricing (DAP) or day-ahead real-time pricing

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_57

(DA-RTP). Integration of renewable energy sources in SG causes intermittent generation of electricity [3]. To deal with real-time scenarios and intermittent nature, RTP signal has been proposed. The drawback associated with RTP is that it requires continuous real-time communication between the utility and consumers which may cause network congestion and data loss problems. DA-RTP is an alternative to RTP based pricing scheme where predicted real-time prices are announced to customers beforehand and consumers are billed on this day-ahead price [2].

Appliance scheduling or load shifting has been considered as an optimization problem in many studies and solved through various techniques. Consumers load shifting in peak hours is done through appliances task or energy scheduling. In task scheduling appliances are switched on/off, while in energy scheduling power consumption of appliances are reduced and their length of operational time (LoT) extended when the system is under stress period [4]. Appliances load scheduling problem has been solved by mixed integer nonlinear programming (MINLP) [5]. However, many others have used heuristic algorithms like genetic algorithm (GA), ant colony optimization (ACO), binary particle swarm optimization (BPSO) [6]. Various studies have considered the scheduling problem as optimal stopping problem and proposed optimal stopping rule (OSR) for its solution [7,8].

In this paper, we have proposed DSM strategy which is based on load shifting technique during system stress period. Two heuristic population-based algorithms GA, enhanced differential evolution (EDE) have been evaluated along with OSR using DA-RTP. Extensive simulations were carried out to show the average monthly electricity bill savings and waiting time of appliances. The remaining paper is organized as follows. Section 2 describes state of the art work in SG and its subarea DSM along with motivation. Next, Section 3 describes system model, Section 4 discusses simulation results and finally, paper is concluded in Section 5.

2 Related Work and Motivation

Real time scenarios use RTP signals which require continues data transmission between utility and consumers. Ye *et al.* [7] proposed distributed and centralized scheduling algorithms to shift appliances from on-peak to off-peak hours in order to achieve a multiobjective function of cost minimization and user comfort through minimizing waiting time/delay in real time. Power constraints and OSR technique were used. The simulation results show a comparable performance of OSR when compared with linear optimization technique in terms of minimizing cost and waiting time. Rasheed *et al.* [8] also used OSR technique to reduce cost and maximize user comfort in real-time price environment. The users were categorized and autonomous home architecture was proposed using three algorithms to reduce cost and maximize user comfort.

The home energy management (HEM) architecture proposed in [9] has modeled the energy management problem as MINLP. The technique gives an accurate and effective solution, but, also requires a large amount of computational time. As the size of the problem increases it becomes difficult to handle the constraints and parameters in linear programming. To overcome the limitation of linear programming some studies have proposed population-based optimization algorithms which are widely applied in HEM controllers to reduce peak load along with electricity bill. Differential evolution (DE) algorithm was proposed by Storn and Price in 1995 [10] which was enhanced by Arafa *et al.* [11]. EDE was proposed to enhance the accuracy and convergence speed of DE. In literature, various HEM controllers have been proposed to reduce the electricity bill using various heuristic and optimization techniques [6,8]. Rahim *et al.* [6] designed HEM controller based on three heuristic algorithms ACO, BPSO and GA to reduce consumer bill, PAR and user discomfort. However, all these HEM controllers focused on appliances scheduling where power consumption period for appliances are shifted while dominant energy consumption cycles of an individual appliance are ignored.

In this paper, we have used optimization techniques which are suitable for complex problems and also requires less computational time and complexity. In nutshell, we have compared the simulation results of GA, EDE, OSR with the unscheduled scenario using DA-RTP pricing scheme. The major contribution of our work include cost and delay minimization using appliances priority.

3 Proposed System Model

The proposed system model consists of home area network, advanced metering infrastructure (AMI), home energy management controller and appliances with energy consumption pattern.

Our proposed system model as shown in Fig. 1 consists of the smart home which has been equipped with smart appliances and energy management controller (EMC). Communication facilities enable smart appliances to convey its status (on/off), usage reports to the EMC. Home area network provides a platform that enables communication among appliances, smart meter, and consumers.

AMI supports smart meters which receive RTP price signals in real time from the utility. The RTP signal is sent by the utility to EMC via neighbourhood area network. Based on RTP prices signal EMC makes intelligent decision to shift load from on-peak to off-peak hours. DR is the changes in electric usage patterns by consumers from their normal consumption in response to price changes from the utility. In order to save electricity bill, residential customers schedule appliances in time slots where prices are low. To achieve DR requires the availability of AMI and smart meters which enable bidirectional flow between utility and consumers.

The HEM controller (HEMC) installed in residence connects AMI and home area network (HAN) to enable bidirectional flow of communication between the two subdomains. The HEMC has been connected to all smart appliances through HAN. The HEMC has also information of power ratings of appliances, operational time, status, consumer defined comfort level, threshold and the priority of appliances. Based on the information HEMC makes an intelligent decision

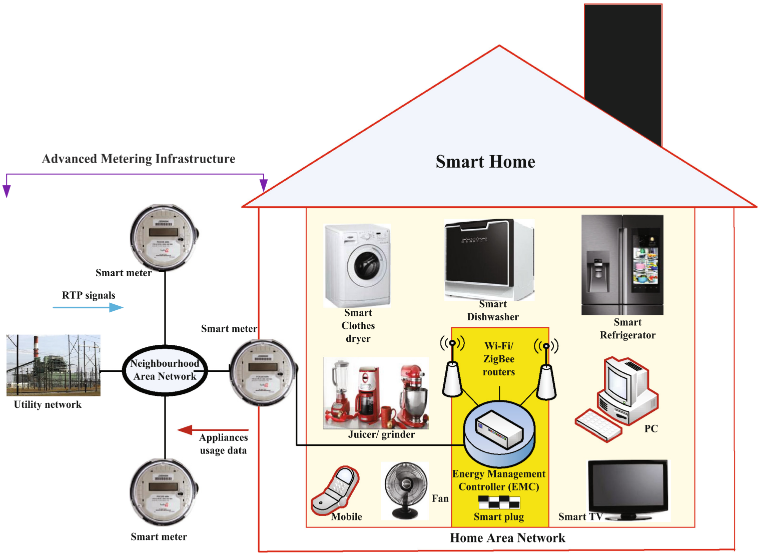


Fig. 1. Proposed system model

based on our proposed optimization techniques GA, EDE, and OSR, to schedule load not only considering cost minimization but also user comfort. Other legacy devices like lightings, fans etc. which do not have intelligence and communication abilities could be controlled and managed via smart plug installed in home area network. Smart plugs are intelligent devices enabled with communication capabilities to remotely on/off the appliances.

Appliances energy consumption patterns are essential to be provided to HEMC for load scheduling and energy management. Each appliance installed in the home has unique energy usage profile. We consider three appliances, clothes

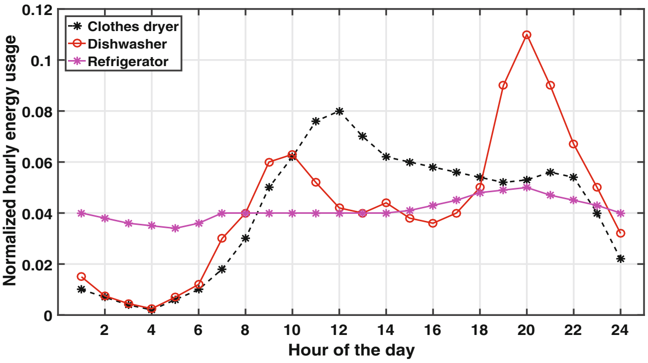


Fig. 2. Appliance energy profile data [12]

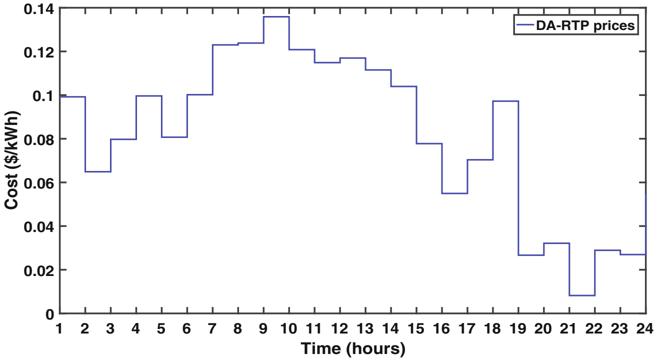


Fig. 3. DA-RTP signal

dryer, dishwasher and refrigerator for scheduling. Appliances usage profile in 24-hour time slots as been depicted in Fig. 2. The DA-RTP signal has been plotted in Fig. 3.

4 Results

Three different smart appliances clothes dryer, dishwasher, and refrigerator have been considered for scheduling. The appliance energy consumption profile [7] is given in the Table 1 and usage profile in Fig. 2. Due to randomness in population generation, we have plotted mean values of GA and EDE after ten iterations.

Table 1. Energy profile of appliances

Appliance	Total energy (kWh)	Cycle during (hours)	Peak power (kW)	Average power (kW)	Time factor (Mu)
Clothes dryer	3.0	0.75	6.0	3.0	0.001, 0.23
Dishwasher	1.4	1.75	1.18	0.8	0.001, 0.045
Refrigerator	2.1	24	0.574	0.089	0.0001, 0.0029

4.1 Clothes Dryer

We observe that clothes dryer is a suitable appliance for DR due to the facts that its duty cycles are short and requires high energy consumption and high peak demand.

The Table 2 shows simulation results summary of the appliance clothes dryer. Three different schemes EDE, GA and OSR have been applied to reduce the electricity cost with respect to time factor (priority). A large value time factor means high priority which will lead to shorter appliance delay. Figures 4 and 5

Table 2. Clothes dryer

Scheduling technique	Time factor	Average cost ($)	Difference	Percentage decrement in cost	Average delay hours
Unscheduled	-	235.25	-	-	-
EDE	0.001	87.46	147.79	62.82 %	7.80
	0.23	188.13	47.12	20.00 %	0.94
GA	0.001	100.33	134.92	57.35 %	7.24
	0.23	189.87	45.38	19.29 %	0.94
OSR	0.001	77.06	158.19	67.24 %	6.78
	0.23	146.97	88.28	37.52 %	2.69

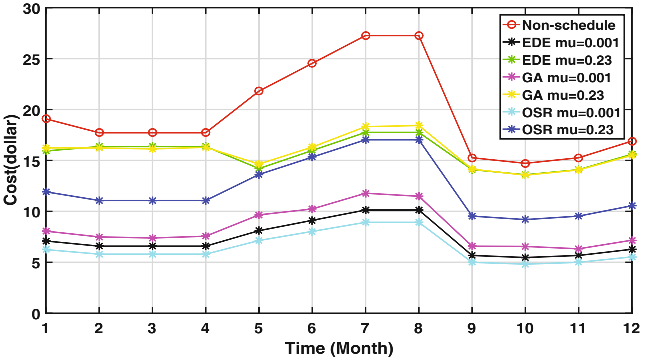


Fig. 4. Average monthly cost of clothes dryer

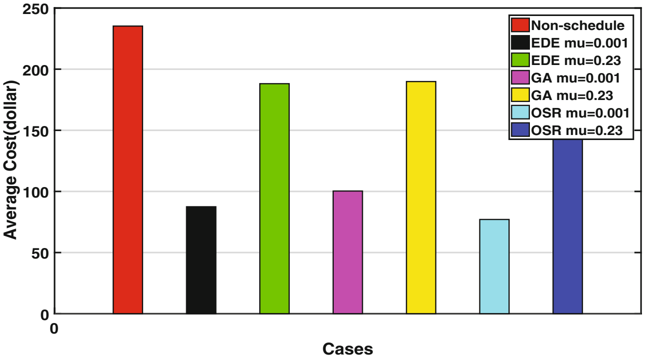


Fig. 5. Average cost of clothes dryer

shows average monthly and yearly cost of the clothes dryer. The average waiting delay of the clothes dryer is shown in Fig. 6. The unscheduled yearly average cost incurred by clothes dryer is $235.25. When the priority of the clothes dryer

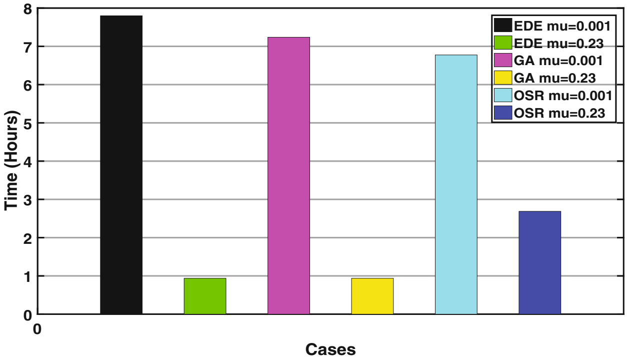


Fig. 6. Average delay of clothes dryer

is low i.e. 0.001 EDE, GA, and OSR have reduced the cost by 62.82%, 57.35% and 67.24% with an average delay of 7.80, 7.24 and 6.78 h respectively. When the priority of the clothes dryer is high i.e. 0.23, EDE and GA have reduced the cost by almost 20.00%, while OSR has reduced the cost by 37.52%. The average delay incurred by EDE, GA is same 0.94, while OSR has incurred a delay of 2.69 h respectively which has been shown in Table 2.

We also notice that there is a tradeoff between cost and delay with respect to time factor (priority). When clothes dryer priority is high the consumer has to bear maximum costs with least delays. GA with priority 0.023 has reduced minimum cost by $189.87. The EDE performed little well as compared to GA, since both beard same average delay of less than an hour. OSR has achieved better cost savings but at the expense of maximum delay of 2.69 h. When the priority of clothes dryer is low, EDE consumer has to bear a cost of $87.46 with an average delay of 7.8 h daily. At an average delay of 7.24 h, GA consumer has to bear an average cost of $100.33. However, OSR has performed best among all other schemes and the consumer has to bear an average cost of $77.06 with the least average delay of 6.78 h when priority is low.

4.2 Dishwasher

The simulation results of dishwasher have been summarized in the Table 3. The yearly unscheduled average cost of the dishwasher is $199.93. When priority is low (0.001) EDE, GA, and OSR have reduced the bill to $89.24, $108.67 and $77.78 bearing an average daily delay of 8.50, 7.52 and 8.33 h respectively. We noticed that consumer cost is reduced to the maximum with this least priority. However, when priority is increased to 0.045 the cost calculated through EDE, GA and OSR has been given as $147.53, $148.01 and $138.55. We observe that the consumer has to wait for less than 2 h in case of EDE and GA and 2.68 h for OSR.

The average monthly and yearly cost of the dishwasher are shown in 7 and 8. The performance of dishwasher shows that it can save extremely 30.70% cost

Table 3. Dishwasher

Scheduling technique	Time factor	Average cost ($)	Difference	Percentage decrement in cost	Average delay hours
Unscheduled	-	199.93	-	-	-
EDE	0.001	89.24	110.69	55.36 %	8.50
	0.045	147.53	52.40	26.21 %	1.61
GA	0.001	108.67	91.26	45.64 %	7.52
	0.045	148.01	51.92	25.96 %	1.50
OSR	0.001	77.78	122.15	61.09 %	8.33
	0.045	138.55	61.38	30.70 %	2.68

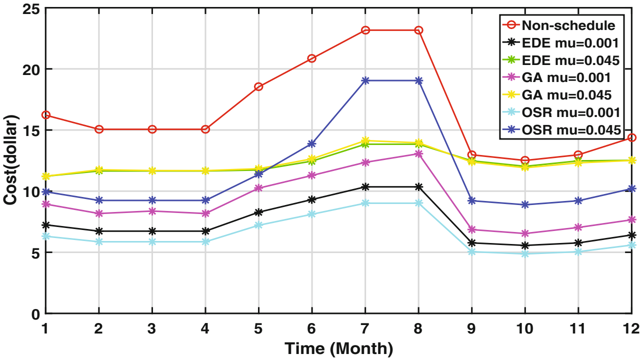


Fig. 7. Average monthly cost of dishwasher

by waiting less than three hours in case of OSR when priority is set to highest. When priority is set to lowest value the consumer can save the maximum cost of 61.09% while delaying appliance for 8.33 h daily as shown in the Table 3. We can see the similar effects of OSR in Fig. 7 where the cost has been reduced significantly by OSR with minimum priority. The simulation results in Fig. 8 show that when priority is set to minimum value i.e. 0.001, EDE, GA and OSR has decremented the average cost by 55.36%, 45.64%, and 61.09% respectively. In this case, we notice that EDE has the highest delay of 8.50 h, followed by OSR with a delay of 8.33 h and GA has a minimum delay of only 7.52 h has been shown in Fig. 9. When priority is set to maximum value i.e. 0.045 EDE and GA schemes have less than 2-hour delay and OSR has a delay of 2.68 h respectively.

4.3 Refrigerator

The simulation results of appliance refrigerator have been summarized in the Table 4. The unscheduled yearly average cost incurred by the refrigerator is $360.43. When the priority of the refrigerator is low i.e. 0.0001, EDE, GA,

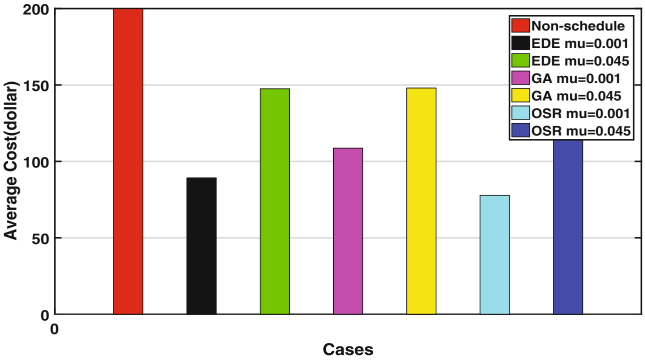


Fig. 8. Average cost of dishwasher

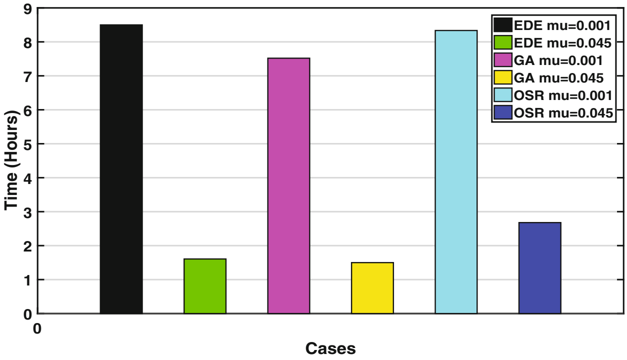


Fig. 9. Average delay of dishwasher

Table 4. Refrigerator

Scheduling technique	Time factor	Average cost ($)	Difference	Percentage decrement in cost	Average delay hours
Unscheduled	-	360.43	-	-	-
EDE	0.0001	340.54	19.89	5.5 %	10.00
	0.0029	357.23	3.2	0.89 %	3.00
GA	0.0001	344.39	16.04	4.45 %	5.75
	0.0029	356.75	3.68	1.02 %	3.00
OSR	0.0001	336.06	24.37	12.19 %	6.76
	0.0029	345.40	15.03	4.17 %	6.00

and OSR consumer has to pay a bill of $340.54, $344.39 and $336.06. When the priority of the refrigerator is high i.e. 0.0029, EDE, GA and OSR consumers have to pay electricity bill of $357.23, $356.75 and $345.40 respectively.

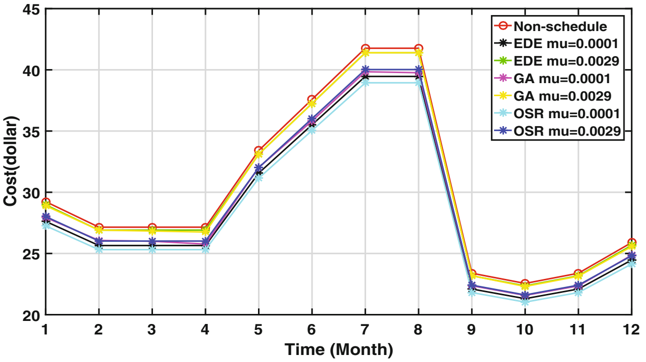


Fig. 10. Average monthly cost of refrigerator

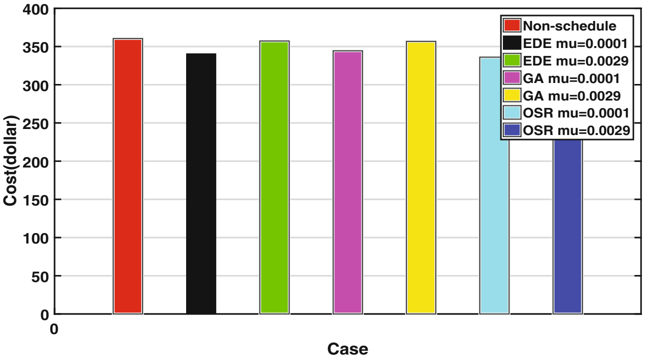


Fig. 11. Average cost of refrigerator

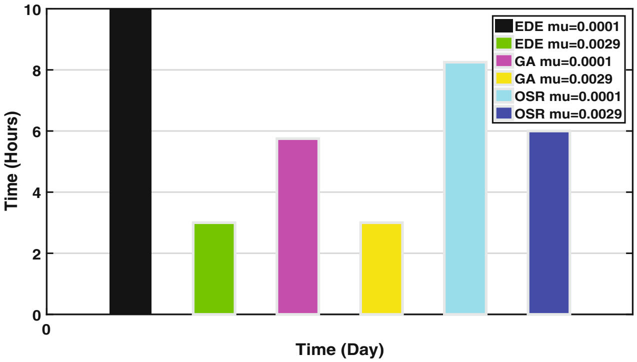


Fig. 12. Average delay of refrigerator

Figures 10 and 11 show average monthly and yearly cost savings of the refrigerator. The average waiting delay of the clothes dryer is shown in Fig. 12. When the priority of the refrigerator is low EDE, GA and OSR have reduced the cost by 5.5%, 4.45% and 12.19% with an average delay of 10, 5.75 and 6.76 h respectively. As shown in the Table 4 the maximum cost has been decreased by OSR. We also notice that the cost of the refrigerator has not been significantly decreased as compared to clothes dryer and dishwasher because only the ice-making and defrost phases of the refrigerator have been considered in scheduling. Both ice-making and defrost phases of the refrigerator account only small portion of its operation time.

When the priority of the refrigerator is set high, EDE and GA have reduced the cost 0.89% and 1.02%. OSR has reduced the cost by 4.17% but the delay has raised to 6 h which is doubled as compared with other schemes EDE and GA delay 3 h as been shown in the Table 4 and Fig. 11. Thus, it has been noticed that the priority has a minor effect on the refrigerator because increase or decrease in delay may not contribute in major cost savings as compared with other appliances.

5 Conclusion

In this paper, we proposed HEMC using GA, EDE, and OSR in order to minimize consumers electricity bill according to appliances priority. Three home appliances clothes dryer, dishwasher, and refrigerator were considered. The simulation results show the tradeoff between consumers electricity bill and user comfort. The increase in appliance priority causes a decrease in waiting time and increase in electricity cost. Among the three schemes, OSR has performed well and reduced the electricity cost more as compared to EDE and GA.

References

1. Tuballa, M.L., Abundo, M.L.: A review of the development of smart grid technologies. Renew. Sustain. Energy Rev. **59**, 710–725 (2016)
2. Vardakas, J.S., Zorba, N., Verikoukis, C.V.: A survey on demand response programs in smart grids: pricing methods and optimization algorithms. IEEE Commun. Surv. Tutorials **17**(1), 152–178 (2015)
3. Hossain, M.S., Madlool, N.A., Rahim, N.A., Selvaraj, J., Pandey, A.K., Khan, A.F.: Role of smart grid in renewable energy: an overview. Renew. Sustain. Energy Rev. **60**, 1168–1184 (2016)
4. Vardakas, J.S., Zorba, N., Verikoukis, C.V.: Performance evaluation of power demand scheduling scenarios in a smart grid environment. Appl. Energy **142**, 164–178 (2015)
5. Moon, S., Lee, J.-W.: Multi-Residential Demand Response Scheduling with Multi-Class Appliances in Smart Grid. IEEE Trans. Smart Grid (2016)
6. Rahim, S., Javaid, N., Ahmad, A., Khan, S.A., Khan, Z.A., Alrajeh, N., Qasim, U.: Exploiting heuristic algorithms to efficiently utilize energy management controllers with renewable energy sources. Energy Build. **129**, 452–470 (2016)

7. Yi, P., Dong, X., Iwayemi, A., Zhou, C., Li, S.: Real-time opportunistic scheduling for residential demand response. IEEE Trans. Smart Grid **4**(1), 227–234 (2013)
8. Rasheed, M.B., Javaid, N., Ahmad, A., Awais, M., Khan, Z.A., Qasim, U., Alrajeh, N.: Priority and delay constrained demand side management in real time price environment with renewable energy source. Int. J. Energy Res. **40**(14), 2002–2021 (2016)
9. Shirazi, E., Jadid, S.: Optimal residential appliance scheduling under dynamic pricing scheme via HEMDAS. Energy Build. **93**, 40–49 (2015)
10. Storn, R., Price, K.: Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997)
11. Arafa, M., Sallam, E.A., Fahmy, M.M.: An enhanced differential evolution optimization algorithm. In: 2014 Fourth International Conference on Digital Information and Communication Technology and Its Applications (DICTAP), Bangkok, pp. 216–225 (2014). doi:10.1109/DICTAP.2014.6821685
12. Iwayemi, A., Yi, P., Dong, X., Zhou, C.: Knowing when to act: an optimal stopping method for smart grid demand response. IEEE Netw. **25**(5), 44–49 (2011)

A Shadow Elimination Algorithm Based on HSV Spatial Feature and Texture Feature

Ranran Song[1,2], Min Liu[1,2](✉), Minghu Wu[1,2], Juan Wang[1,2], and Cong Liu[1,2]

[1] Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar Energy, Hubei University of Technology, Wuhan, People's Republic of China
Liu_Min@mail.hbut.edu.cn

[2] Hubei Power Grid Intelligent Control and Equipment Engineering Technology Research Center, Wuhan 430068, People's Republic of China

Abstract. In order to improve the accuracy of the detection and tracking task in the intelligent surveillance system, we propose a shadow elimination algorithm based on HSV spatial feature and texture feature. In this paper, firstly the background subtraction is used to obtain the motion area of the sequence image, where HSV feature is used to determine the threshold value of the shadow elimination which can be completely removed. Then the complete moving target is obtained by OR operator of combining the foreground which is extracted by OTSU and the result which is extracted by HSV. The algorithm is applied to several realistic scenario where exists various shadow. We compare our method with other traditional algorithm and report experimental results, both in terms of noise suppression and detection accuracy. The experimental results show that the proposed method has the better noise suppression and detection accuracy.

1 Introduction

In recent years, with the rapid development of image technology, there are many algorithms for shadow detection, which are divided into two categories: statistic shadow detection method and shadow detection method based on determination threshold [1–3]. Among them, the shadow-based detection method can be divided into parameter-based shadow detection method [4] and non-parametric shadow detection method [5]. The shadow detection method based on the threshold can be divided into model-based shadow detection method [6] and non-model-based shadow Detection method [7]. Stauder et al. uses the edge width of the image to detect penumbra, and points out that the edge width in the distinction between the target contour and shadow contours plays an important role of removing shadow [8]. Cucchiara et al. points out that the chromaticity information of the shadows which cast sonto the ground change within a certain range. They first proposes to detect motion shadows in the HSV color space [9].

HSV color space contains hue, saturation and value. And visual perception system of human is closely linked with it, so the moving targets and shade of gray are expressed more accurately. Traditional algorithm which uses HSV eliminates shadow by the luminance ratio between video frame and background frame. In this way,

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_58

however, moving targets are easily mistaken for shadow. This paper proposes to a shadow elimination algorithm based on HSV spatial feature and texture feature. The algorithm determines the threshold of the ratio which can basically eliminate shadow by HSV. OTSU [10] extracts the foreground which is obtained by background subtraction method, and we can get the part of moving objects. The complete moving target is obtained by OR operator of combining the foreground which is extracted by OTSU and the result which is extracted by HSV. The method obtains better result in the different standard shadow detection video test.

2 Shadow Detection of Traditional HSV Color Space

2.1 Background Subtraction Method

Lots of methods have been proposed for moving object detection, such as background subtraction method, optical flow method and frame difference method [11]. In the background subtraction method, it detects the moving object by subtracting the background from the current frame in the image sequence [12]. In this method, the image is compared with the background image in the pixels. If the difference value between the two pixels in the background and current frame is larger than the threshold, the pixels are regarded as the foreground, otherwise the pixels are regarded as the background. If the $k-th$ image frame is set to $f_k(x,y)$, and the background model is $B_k(x,y)$. The background subtraction method is represented as formula (1) and (2) in the follows,

$$D_k(x,y) = |f_k(x,y) - B_k(x,y)| \tag{1}$$

$$M_k(x,y) = \begin{cases} 0, D_k(x,y) < T \\ 1, D_k(x,y) \geq T \end{cases} \tag{2}$$

2.2 Model of the HSV Color Space to Removing Shadow

The shadow characteristics can be understood by the human visual perception system can be easily understand. On the one hand, the brightness values of the shadow is always lower than the brightness values background region which is projected by shadow; On the other hand, it connects to the object of projection which keeps motion state with shadow. As a result, the classic HSV color space which removes shadow uses HSV color information which contains the hue, saturation and value to eliminate the shadow. The discriminant formula for shadow elimination is shown in formula (3):

$$SP_k(x,y)\begin{cases} 1 & \alpha \leq \frac{I_K^V(x,y)}{B_K^V(x,y)} \leq \beta \wedge (I_K^S(x,y) - B_K^S(x,y)) \leq \tau_S \\ & \wedge \left| I_K^H(x,y) - B_K^H(x,y) \right| \leq \tau_H \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In the formula, $I_K^V(x,y)$ is brightness values of the current frame, $B_K^I(x,y)$ is brightness values of background frame, $I_K^H(x,y)$ is hue value of the current frame, $B_K^H(x,y)$ is hue value of the background frame, $I_K^S(x,y)$ is saturation value of the current frame and $B_K^H(x,y)$ is saturation value of the background frame. τ_S, τ_H is respectively saturation and chroma threshold, but they are not effective to the result of detecting shadow. α is related to the strength values of shadow and β is related to the strength of the light.

3 The Proposed Shadow Elimination Method

In the traditional algorithm, method of background subtraction extracts the foregrounds. Then in the foregrounds, the shadow is removed by HSV The detection method which based on color detects almost all shadow pixels. And the moving targets in some dark areas or some areas are similar with the background area in hue information, so they detect moving target as shadow. The methods used the color information can not achieve satisfactory results in the process of shadow detection.

In order to overcome the shortcomings of HSV color space model, the algorithm proposes OTSU method to extract a more complete foregrounds. Although it just obtains part of the moving target, shadow can be basically eliminated. Therefore, complete moving targets can be composed of two phase. The method proposed flow chart is shown in Fig. 1.

3.1 OTSU

According to the gray feature of image, the image is divided into the shadow and the target. If the shadow and the target have large variance, the two parts which make up the image have much discrepant. When the partial target is regarded as shadows or part of the shadow is regarded as the target, the discrepancy of two parts become smaller. Thus, the largest variance segmentation between classes can make the probability of misclassification which get minimum value.

The number of pixels with a gray scale value of i is n_i, and the total number of pixel is N. $p_i = n_i/N$ represents the pixel probability with a gray value of i. So there are $\sum_{i=0}^{L-1} p_i = 1$, and $\mu_T = \sum_{i=0}^{L-1} ip_i$ is the image total average. C_0 and C_1 respectively show two kinds of pixel group of $C_0 = [0,\cdots,k]$, $C_1 = [k+1,\cdots,L-1]$. The average of C_0 and C_1 respectively represent $\mu_0(k)$ and $\mu_1(k)$. When $\omega_0(k) = \sum_{i=0}^{k} p_i$, $\mu(k) = \sum_{i=0}^{k} ip_i$, the variance between class σ_B^2 can be obtained by formula (4):

$$\sigma_B^2(k) = \omega_0(k)\left[\frac{\mu(k)}{\omega_0(k)} - \mu_T\right]^2 + \omega_1(k)\left[\frac{\mu_T - \mu(k)}{1-\omega_0(k)} - \mu_T\right]^2 \quad (4)$$

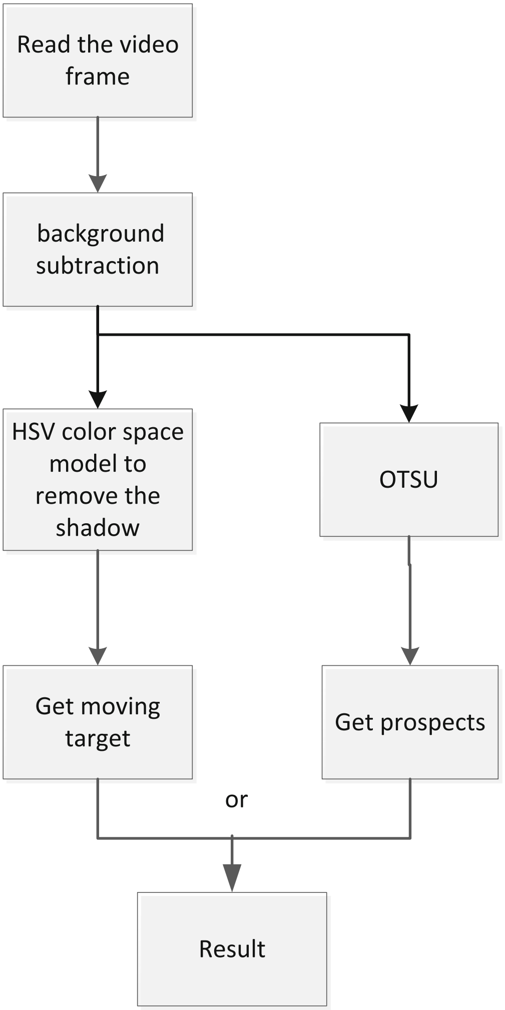


Fig. 1. The method flow chart in this paper

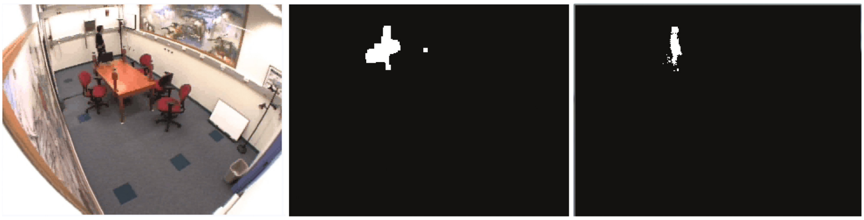
The optimal threshold k^* selection principle is formula (5):

$$\sigma_B^2(k^*) = \max_{0 \le k < L-1} \sigma_B^2(k) \tag{5}$$

4 Experimental Results

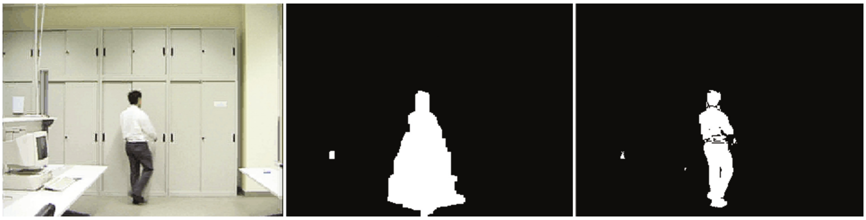
The proposed method are carried out in handling the video frame of size 320 × 240 on the computer with 2.70 GHz dual core CPU, 4 GB RAM. When the shadow is eliminated by HSV, the value of hue and saturation threshold selection is enough large and the brightness threshold which is set to 0.7 to 1 is the most balanced. The original image frames, extracted foreground and the results of the proposed method are represented respectively in the figure (a), (b), (c).

We can see from Figs. 2 and 3. The moving target can be detected accurately. Noise can be suppressed better.



(a) original image frames (b) extracted foreground (c) shadow elimination

Fig. 2. The 160th frame shadow elimination results in intelligentroom_raw.AVI



(a) original image frames (b) extracted foreground (c) shadow elimination

Fig. 3. The 155th frame shadow elimination results in laboratory_raw.AVI

In order to guarantee the reliability of the experimental results, we propose the shadow detection rate η and discriminant rate ξ as evaluation index of algorithm performance [13], and the sum of the two is averaged to further analyze its performance [14]. The concrete are defined as follows:

$$\eta = \frac{TP_s}{TP_s + FN_s} \times 100\% \quad \xi = \frac{TP_F}{FN_F + TP_F} \times 100\% \quad Avg = \frac{\eta + \xi}{2}$$

Among them, TP_S shows the number of detecting correctly shadow pixels, and FN_S shows the shadow pixel is mistaken for the number of foreground pixels. TP_F shows the number of detecting correctly foreground pixel, FN_F shows the foreground pixel is mistaken for the number of shadow pixels. We can find out TP_S, FN_S, TP_F and FN_F through ground truth in video frame.

As shown in Table 1, our method is compared with various shadow elimination algorithm, which include DNM1 algorithm and DNM2 algorithm in the paper [3]. The algorithm improves the shadow detection rate and average, and remains the shadow discriminant rate.

Table 1. Shadow elimination algorithm (%)

Test sequence	Test standard	DNM1	DNM2	This paper
Intelligent room	η	78.6	62.0	85.0
	ξ	90.3	93.9	92.3
	Avg	84.5	78.0	88.7
Laboratory	η	76.2	60.3	82.0
	ξ	89.8	81.5	92.4
	Avg	83	70.9	87.2

5 Conclusions

In this paper, the foreground in the video frame is extracted by the background subtraction method, and the shadows in the foreground are completely eliminated by the HSV color space model. The complete moving target is obtained by OR operator of combining the foreground which is extracted by OTSU and the result which is extracted by HSV. Our method is compared with other traditional algorithm and the experimental results is reported, both in terms of noise suppression and detection accuracy. The experimental results show that the proposed method has the better noise suppression and detection accuracy. The future work will study how to improve the real-time performance and stability, which is applied in intelligent video surveillance.

Acknowledgments. This research was supported by National Natural Science Foundation of China (61471162); Program of International science and technology cooperation (2015DFA10940); Science and technology support program (R & D) project of Hubei Province (2015BAA115); PhD Research Startup Foundation of Hubei University of Technology (BSQD13032); Open Foundation of Hubei Collaborative Innovation Center for High-Efficiency Utilization of Solar Energy (HBSKFZD2015005, HBSKFTD2016002).

References

1. Kang, J., Cohen, I., Medioni, G.: Tracking Objects from Multiple Stationary and Moving Cameras, pp. 31–35. The Institution of Electrical Engineers, England (2004)
2. Lee, B.J., Park, J.B., Jin, S.H., et al.: Intelligent Kalman filter for tracking a manoeuvring target. IEEE Proc. Radar Sonar Navig. **151**(6), 344–350 (2004)
3. Prati, A., Mikic, I., Cucchiara, R., et al.: Detecting moving shadows: algorithms and evaluation. IEEE Trans. Pattern Anal. Mach. Intell. **25**(7), 918–923 (2003)
4. Liu, Z., Huang, K., Tan, T.: Cast shadow removal in a hierarchical manner using MRF. IEEE Trans. Circuits Syst. Video Technol. **22**(1), 56–66 (2012)
5. Choi, J.M., Yoo, Y.J., Choi, J.Y.: Adaptive shadow estimator for removing shadow of moving object. Comput. Vis. Image Underst. **114**(9), 1017–1029 (2010)
6. Mc Feely, R., Glavin, M., Jones, E.: Shadow identification for digital imagery using colour and texture cues. IET Image Process. **6**(2), 148–159 (2012)

7. Fang, L.Z., Qiong, W.Y., Sheng, Y.Z.: A method to segment moving vehicle cast shadow based on wavelet transform. Pattern Recogn. Lett. **29**(16), 2182–2188 (2008)
8. Cucchiara, R., Grana, C., Piccardi, M., et al.: Improving shadow suppression in moving object detection with HSV color information. In: Proceedings of 2001 IEEE Intelligent Transportation Systems, pp. 334–339. IEEE (2001)
9. Cucchiara, R., Grana, C., Piccardi, M., et al.: Detecting moving objects, ghosts, and shadows in video streams. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1337–1342 (2003)
10. Wang, H.-Y., Pan, D.-L., Xia, D.-S.: A fast algorithm for two-dimensional Otsu adaptive threshold algorithm. Acta Automatica Sin. **33**(9), 968–971 (2007)
11. Li, Y., Sun, Z.-X., Yuan, B.: An improved method for motion detection by frame difference and background subtraction. J. Image Grap. **14**(6), 1162–1168 (2009)
12. Elgammal, A., Harwood, D., David, L.S.: Nonparametric background model for background subtraction. In: Proceedings of the Sixth European Conference on Computer Vision (2000)
13. Soh, Y.S., Lee, H., Wang, Y.: Invariant color model-based shadow removal in traffic image and a new metric for evaluating the performance of shadow removal methods. In: PRICAI 2006: Trends in Artificial Intelligence, pp. 544–552. Springer, Heidelberg (2006)
14. Joshi, A.J., Papanikolopoulos, N.P.: Learning to detect moving shadows in dynamic environments. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 2055–2063 (2008)

A Provably Secure Certificateless User Authentication Protocol for Mobile Client-Server Environment

Alzubair Hassan[1], Nabeil Eltayieb[1], Rashad Elhabob[2], and Fagen Li[1(✉)]

[1] School of Computer Science and Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
alzubairuofk@gmail.com, nabeil9@yahoo.com, fagenli@uestc.edu.cn
[2] School of Information and Software Engineering,
University of Electronic Science and Technology of China, Chengdu 611731, China
rashaduestc@gmail.com

Abstract. Based on mobile devices limitations, several user authentications and key exchange schemes have been proposed for mobile devices using identity-based public key cryptography (ID-PKC). However, these schemes suffer from key escrow problem. Moreover, they are not secure against impersonation attacks, and they can't achieve perfect forward secrecy. In this paper, a new user authentication and key exchange protocol for the mobile client-server environment is proposed. Certificateless public key cryptography (CL-PKC) and bilinear pairing are adopted in the proposed scheme. Our protocol solves the key escrow problem of identity-based public key cryptography. Also, it is secure against both adversaries type I and type II. Furthermore, the proposed protocol achieves perfect forward secrecy. We prove the security of our protocol in the random oracle model under the Computational Diffie-Hellman (CDH) problem. Hence, the proposed scheme is more suitable for the mobile devices environments.

1 Introduction

In recent years, handheld devices (i.e., cellular phones, smart phones, and PDAs) are widely used in the client-server applications. The handheld devices have storage and battery limitations. In addition, it's raised many security issues due to use it in the bank payment, online voting and online shopping. To solve these security issues, some schemes are designed for the handheld devices' applications based on traditional public key cryptography [8,17]. Unfortunately, these schemes had computational costs on the user side in practical applications [16,21,22,24].

In 1984, Shamir [19] proposed an identity-based public key cryptosystem (ID-PKC) which simplified the certificate management, compared with the traditional certificate-based public key systems. Again, the system in [19] has a major drawback as all the users' private keys are generated by the key generator

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_59

center (KGC), which in turn leads to the key escrow problem. Since the security of Shamir's system depends on the integer factorization problem, it is not easily realized in practical applications. Fortunately, Boneh and Franklin [4] proposed an efficient identity-based encryption scheme from the Weil pairing defined on elliptic curves. The security of their proposed scheme is based on the computational Diffie-Hellman problem. Since then user authentication protocol for the mobile client-server environment has been studied extensively.

2 Related Work

Recently, many identity-based cryptography schemes based on bilinear pairings have been proposed such as in [7,9–11,23]. These schemes haven't provided mutual authentication and key agreement. To overcome the weaknesses mentioned above, many schemes are proposed, which are based on bilinear pairings in the random oracle model. In 2010, Wu and Tseng [26] proposed a new user authentication and key exchange protocol. Their scheme is secure against impersonation attack, known session key, ID attack, and partial forward secrecy. In same year, Yoon and Yoo [27] proposed another user authentication and key agreement scheme for the mobile client-server environment to improve the performance. In 2012, He [12] proposed a new user authentication and key exchange protocol. He claimed that his scheme is secure against various known attacks. Therefore, his scheme is better than the schemes that mentioned above.

In 2015, Tsai and Lo [20] proposed a new ID-based authentication protocol. They claimed that their protocol is provably secured and less communication overhead at the mobile user side. In 2016, Wu et al. [25] found that Tsai and Lo's protocol [20] has several drawbacks. In order to solve Tsai and Lo' drawbacks, they proposed a new efficient and secure user authentication key agreement protocol. Their proposed protocol keeps user anonymity. Regarding solving the key escrow problem, which is the inherent issue of identity-based cryptography, Al-Riyami and Paterson [1] proposed a new paradigm called certificateless public-key cryptography (CL-PKC). Based on [1], some user authentication certificateless cryptography schemes, without providing mutual authentication, have been proposed in [12,14]. Therefore, it is authoritative for us to propose a provable secure user authentication and key agreement protocol based on CL-PKC.

In this paper, a certificateless user authentication and key agreement protocol using bilinear pairing is presented. We demonstrate that the proposed protocol does not require the use of certificates and does not have the built-in key escrow feature of identity-based public key cryptography (ID-PKC). We show that our protocol is secure against impersonation attack and a chosen identity attack. Also, it offers key agreement, mutual authentication, and perfect forward secrecy. We show that our protocol is more suitable for mobile client-server environment.

The rest of this paper is organized as follows. In Sect. 3, the preliminaries of bilinear pairings and security model are presented. In Sect. 4, our protocol is proposed. Security analysis of our protocol is presented in Sect. 5. In Sect. 6, performance analysis is demonstrated. Conclusions are given in Sect. 7.

3 Preliminaries

In this section, we present a bilinear pairing properties and the security model of our protocol.

3.1 Bilinear Pairings

Let $\mathbb{G}_1$ and $\mathbb{G}_2$ be the additive and multiplicative groups of the same prime order q, respectively. Let $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ be a bilinear pairing function, and P is the generator of $\mathbb{G}_1$. A bilinear pairing has the following properties [3,5]:

1. **Bilinearity:** $\forall a, b \in \mathbb{Z}_q^*$ and $\forall Q, P \in \mathbb{Z}_q^*$, $e(aQ, bP) = e(Q, P)^{ab}$
2. **Nondegeneracy:** There exists $Q, P \in \mathbb{G}_1$ such that $e(Q, P) \neq 1_{\mathbb{G}_2}$ where $1_{\mathbb{G}_2}$ the identity element of $\mathbb{G}_2$.
3. **Computability:** There exists an efficient algorithm to compute $e(Q, P)$ for $\forall Q, P \in \mathbb{G}_1$.

Computational Diffie-Hellman Problem (CDHP)
Given $(P, aP, bP) \in \mathbb{G}_1$ where $\forall a, b \in \mathbb{Z}_q^*$, it is hard to compute abP.

3.2 Security Model

In this section, we describe the capabilities of an adversary $\mathscr{A}_i\{i = 1, 2\}$ and list of security requirements for mutual authentication and key exchange. The strong schemes use CL-PKC must withstand the two types adversary, called Type I and Type II defined in [1]. The Type I adversary $\mathscr{A}_1$ can't access to the master key of the KGC, but he/she can replace the public key of the users. The type II adversary $\mathscr{A}_2$ has the master key of the KGC, but he/she cannot replace the public key of the users. Note that we define an instance k of participant U as Π_U^k. Here, game I and game II are defined as follows involve the following queries:

Game-I:
Setup: Inputting a security parameter k, the challenger C runs Setup algorithm. The system parameter *params* and the master key s can be obtained. Then *params* is sent to the adversary $\mathscr{A}_1$ while s is kept secret.
Probing: The adversary $\mathscr{A}_1$ can perform a polynomial bounded number of following queries for any identity ID_c except the challenged identity ID_j:
Extract partial private key: $\mathscr{A}_1$ is able to request the partial private key for any identity ID_c except the challenged identity ID_j. C computes the partial private key D_{ID_c} corresponding to the identity ID_c and returns to $\mathscr{A}_1$.
Extract private key: For any ID_c except the ID_j. C computes the private key corresponding to the ID_c and returns the private key to $\mathscr{A}_1$.
Request public key: Upon receiving a public key query for any ID_c. C computes the corresponding public key PK_{ID_c} and sends it to $\mathscr{A}_1$.
Replace public key: $\mathscr{A}_1$ can pick a new secret value x'_{ID_c} for any ID_c and compute the new public key corresponding to the value x'_{ID_c}. Then replace PK_{ID_c} with PK'_{ID_c}.

Send (Π_U^k, M) **query:** an adversary $\mathscr{A}_1$ can send a message M to C. if C receives M according to the proposed protocol, then makes the computation and response to adversary $\mathscr{A}_1$.
Reveal (Π_U^k, M) **query:** an adversary $\mathscr{A}_1$ can get a session key sk from C if has accepted, else; it returns a null.
Corrupt (U) **query:** in this query, an adversary $\mathscr{A}_1$ can issue the participant U and gets back its a private key.
Test (Π_U^k) **query:** an adversary $\mathscr{A}_1$ can send a single test query to the C flips an unbiased coin b. If $b = 1$, then it returns the session key, else it returns a random string. This query measures the semantic security of the session key sk.

In the end, $\mathscr{A}_1$ outputs a bit b' as its guess for b. The advantage of $\mathscr{A}_1$ is defined as $\mathrm{Adv}(\mathscr{A}_1) = |\Pr[b' = b] - 1/2|$, where $\Pr[b' = b]$ denotes the probability that $b' = b$. The limitation of the $\mathscr{A}_1$ in this game, $\mathscr{A}_1$ can't extract the private key for ID_j at any point. In addition $\mathscr{A}_1$ can't both replace the public key for the ID_j before the challenge phase and extract the partial private key for ID_j in some phase.

Game-II:
Setup: Inputting a security parameter k the challenger C runs setup algorithm. The system parameter *params* and the master key s can be obtained. Then *params* and s are sent to the adversary $\mathscr{A}_2$.
Probing: The adversary $\mathscr{A}_2$ can perform a polynomial bounded number of queries for any identity ID_c except the challenged identity ID_j. The adversary $\mathscr{A}_2$ can make extract private key query, request public Key query, send (Π_U^k, M) query, reveal (Π_U^k, M) query, corrupt (U) query and test (Π_U^k) query as same in game-I. Here, there are no partial private key queries and replace public key queries in this game, since $\mathscr{A}_2$ owns the master key s and runs the partial private key extraction D_{ID_c} by itself.

In the end, $\mathscr{A}_2$ outputs a bit b' as its guess for b. The advantage of $\mathscr{A}_2$ is defined as $\mathrm{Adv}(\mathscr{A}_2) = |\Pr[b' = b] - 1/2|$, where $\Pr[b' = b]$ denotes the probability that $b' = b$. The limitation of the $\mathscr{A}_2$ in this game, $\mathscr{A}_2$ can't replace public keys at any point and can't extract the private key for ID_j at any point.

For the key agreement property without authentication, the adversary $\mathscr{A}_i\{i = 1, 2\}$ can create these queries send, reveal, corrupt, and test [15]. Note that the adversary $\mathscr{A}_i\{i = 1, 2\}$ has abilities to create finite queries under adaptive chosen message attacks [6,15]. The reader can find more details, for additional mutual authentication and key exchange protocol security requirements in [3].

4 Proposed Protocol

This section shows that our protocol design is based on He at el.'s short signature [13]. Our protocol consists of two phase an initialization phase and user authentication key agreement phase. We describe our protocol phase as follows:

4.1 Initialization Phase

In this phase, we show the details of setup, extract partial private key, set private key and set public key of the proposed protocol. This phase is depicted in Fig. 1. The details communication steps are described as follows:

- **Setup:** In the following, we present the setup of our protocol. Here, the powerful server S is a Key Generator Center (KGC) to generate all needed keys and parameters. Given a security parameter k, the server S chooses two groups $\mathbb{G}_1$ and $\mathbb{G}_2$ with same prime order q and bilinear pairing $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ where P is a generator of $\mathbb{G}_1$. Then, the server chooses random number $s \in \mathbb{Z}_q^*$ as its master private key and computes its corresponding master public key $P_{pub} = sP$. Then the server chooses $x_{ID_s} \in \mathbb{Z}_q^*$ and computes $P_{ID_s} = x_{ID_s}P$ as it's public key. Next, server chooses five cryptographic secure hash function $H_1 : \{0,1\}^* \times \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{Z}_q^*$, $H_2 : \mathbb{G}_1 \times \{0,1\}^* \times \mathbb{Z}_q^* \times \mathbb{G}_1 \times \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{Z}_q^*$, $H_3 : \mathbb{G}_1 \times \{0,1\}^* \times \mathbb{Z}_q^* \times \mathbb{G}_1 \times \mathbb{Z}_q^* \rightarrow \mathbb{Z}_q^*$, $H_4 : \{0,1\}^* \times \mathbb{G}_1 \times \mathbb{G}_1 \times \mathbb{G}_1 \times \mathbb{Z}_q^* \times \mathbb{G}_1 \times \mathbb{Z}_q^* \rightarrow \mathbb{Z}_q^*$ and $H_5 : \{0,1\}^* \times \mathbb{G}_1 \times \mathbb{G}_1 \times \mathbb{G}_1 \times \mathbb{Z}_q^* \times \mathbb{G}_1 \times \mathbb{Z}_q^* \rightarrow \mathbb{G}_1$. Finally, the server publishes $\{\mathbb{G}_1, \mathbb{G}_2, q, e, P, P_{pub}, P_{ID_s}, H_1, H_2, H_3, H_4, H_5\}$ as public parameters.
- **Extract partial private key:** Given a master private key, public parameters and client's ID_c, the server computes $R_{ID_c} = r_{ID_c}P$, where $r_{ID_c} \in \mathbb{Z}_q^*$, $h_{ID_c} = H_1(ID_c, R_{ID_c}, P_{pub})$, $s_{ID_c} = (r_{ID_c} + h_{ID_c}s) \bmod q$ and sends the partial private key $D_{ID_c} = (s_{ID_c}, R_{ID_c})$ to the client. When the client receives D_{ID_c} from the server, the client verifies the validity of the D_{ID_c} by

$$s_{ID_c}P \overset{?}{=} R_{ID_c} + h_{ID_c}P_{pub} \tag{1}$$

If its holds, the client believes that the partial private key is valid.

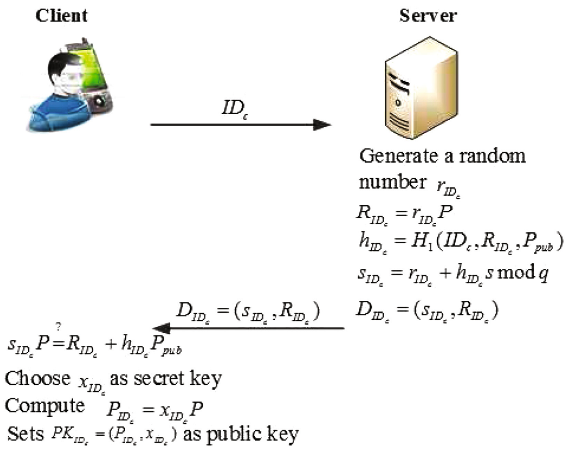

Client
Server
ID_c
Generate a random number r_{ID_c}
$R_{ID_c} = r_{ID_c}P$
$h_{ID_c} = H_1(ID_c, R_{ID_c}, P_{pub})$
$s_{ID_c} = r_{ID_c} + h_{ID_c}s \bmod q$
$D_{ID_c} = (s_{ID_c}, R_{ID_c})$
$D_{ID_c} = (s_{ID_c}, R_{ID_c})$
$s_{ID_c}P \overset{?}{=} R_{ID_c} + h_{ID_c}P_{pub}$
Choose x_{ID_c} as secret key
Compute $P_{ID_c} = x_{ID_c}P$
Sets $PK_{ID_c} = (P_{ID_c}, x_{ID_c})$ as public key

Fig. 1. Initialization phase

- **Set private key:** Given public parameters, the client with ID_c chooses random number $x_{ID_c} \in \mathbb{Z}_q^*$ as his/her secret value and computes $SK_{ID_c} = (x_{ID_c}, D_{ID_c})$ his/her private key.
- **Set public key:** Given public parameters, the client with ID_c computes $P_{ID_c} = x_{ID_c}P$ as his/her public key.

4.2 User Authentication and Key Agreement Phase

In this phase, the low-power client interacts with the powerful server as in Fig. 2. The details interactions steps are describes as follows:

1. The client randomly chooses an integer $r \in \mathbb{Z}_q^*$, computes $U = rP$, $k_1 = rP_{pub}$ and $k_2 = rP_{ID_s}$. Then the client sends (ID_c, U) to the server.
2. Upon receiving (ID_c, U), the server randomly chooses an integer $\alpha \in \mathbb{Z}_q^*$ and computes $k_3 = sU$, $k_4 = x_{ID_s}U$, $Auth = H_2(P_{pub}, ID_c, \alpha, U, k_3, k_4)$ and $h = H_3(P_{pub}, ID_c, \alpha, U, Auth)$. Finally, the server sends $(\alpha, Auth)$ to the client.
3. Upon receiving $(\alpha, Auth)$, the client first verifies if the $Auth = H_2(P_{pub}, ID_c, \alpha, U, k_1, k_2)$ holds. The client computes a common session key $sk = H_2(P_{pub}, ID_c, \alpha, U, k_1, k_2)$, sets k_{ID_c} and Q.

$$k_{ID_c} = H_4(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth) \qquad (2)$$

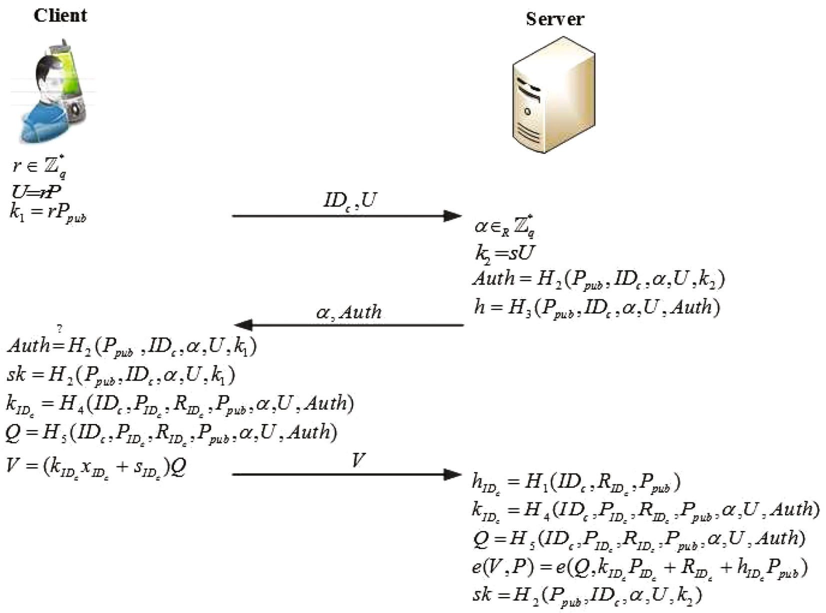

Client
Server
$r \in \mathbb{Z}_q^*$
$U=rP$
$k_1 = rP_{pub}$
ID_c, U
$\alpha \in_R \mathbb{Z}_q^*$
$k_2 = sU$
$Auth = H_2(P_{pub}, ID_c, \alpha, U, k_2)$
$h = H_3(P_{pub}, ID_c, \alpha, U, Auth)$
$\alpha, Auth$
$Auth \overset{?}{=} H_2(P_{pub}, ID_c, \alpha, U, k_1)$
$sk = H_2(P_{pub}, ID_c, \alpha, U, k_1)$
$k_{ID_c} = H_4(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth)$
$Q = H_5(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth)$
$V = (k_{ID_c}x_{ID_c} + s_{ID_c})Q$
V
$h_{ID_c} = H_1(ID_c, R_{ID_c}, P_{pub})$
$k_{ID_c} = H_4(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth)$
$Q = H_5(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth)$
$e(V, P) = e(Q, k_{ID_c}P_{ID_c} + R_{ID_c} + h_{ID_c}P_{pub})$
$sk = H_2(P_{pub}, ID_c, \alpha, U, k_2)$

Fig. 2. User authentication and key agreement phase

$$Q = H_5(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth) \tag{3}$$

Then, the client uses k_{ID_c} and Q to computes V. Finally, the client sends V to the server.

$$V = (k_{ID_c}x_{ID_c} + s_{ID_c})Q \tag{4}$$

4. Upon receiving V, the server computes $h_{ID_c} = H_1(ID_c, R_{ID_c}, P_{pub})$, $k_{ID_c} = H_4(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth)$ and $Q = H_5(ID_c, P_{ID_c}, R_{ID_c}, P_{pub}, \alpha, U, Auth)$ to verifies

$$e(V, P) = e(Q, k_{ID_c}P_{ID_c} + R_{ID_c} + h_{ID_c}P_{pub}) \tag{5}$$

If its holds, The server computes the common $sk = H_2(P_{pub}, ID_c, \alpha, U, k_3, k_4)$.

4.3 Correctness of Our Protocol

The correctness of our protocol is done as in [13] by verifying $e(V, P) = e(Q, k_{ID_c} P_{ID_c} + R_{ID_c} + h_{ID_c}P_{pub})$. Where $P_{ID_c} = x_{ID_c}P$, $s_{ID_c}P = R_{ID_c} + h_{ID_c}P_{pub}$ and $V = (k_{ID_c}x_{ID_c} + s_{ID_c})Q$. Then we have

$$e(V, P) = e((k_{ID_c}x_{ID_c} + s_{ID_c})Q, P) = e(Q, (k_{ID_c}x_{ID_c} + s_{ID_c})P)$$

$$= e(Q, (k_{ID_c}x_{ID_c}P + s_{ID_c}P)) = e(Q, k_{ID_c}P_{ID_c} + R_{ID_c} + h_{ID_c}P_{pub})$$

5 Security analysis

In this section, we show that the proposed protocol can perform the security requirement defined in Sect. 3.2 in random oracle model [2]. we use similar ways in [13,26], for the security proofs of Theorems 1, 2 and 3.

5.1 Client-to-server Authentication

In the following theorem, we demonstrate that an adversary $\mathscr{A}_i$ $\{i = 1, 2\}$ can't impersonate the client to communicate with the server under the computational Diffe-Hellman assumption.

In the random oracle model, let C_0 be an algorithm with an advantage ε_0 with in time t_0 to perform an adaptive chosen message and an identity attack to our proposed protocol. Using Lemma 1 in [6], it implies that there is an algorithm C for an adaptive chosen message attack and fixed identity attack which has the advantage $t \leq t_0$. Hence, our protocol is secure against a chosen identity attack.

Theorem 1. *we assume that an adversary $\mathscr{A}_i$ $\{i = 1, 2\}$ can break up the client-to-server authentication with a non-negligible ε advantage, and makes at most q_s and q_{H_i} $\{i = 1, 2, .., 5\}$ queries to the client/server respectively. Then there is challenger to solve the computational Diffie-Hellman problem.*

Proof. Our proof follows description in Lemmas 1 and 2.

Lemma 1. The proposed protocol is secure against the type I adversary $\mathscr{A}_1$ in the random oracle model if the CDHP is hard.

Lemma 2. The proposed protocol is secure against the type II adversary $\mathscr{A}_2$ in the random oracle model if the CDHP is hard.

This proof is omitted because of page limitation. Please contact the authors for the full version.

5.2 Key Agreement

In the following theorem, we demonstrate that our protocol provides key agreement under the computational Diffe-Hellman assumption.

Theorem 2. *Adopt that an adversary* $\mathscr{A}_i$ $\{i = 1, 2\}$ *can guess the value* b *in the Test-query with a non-negligible advantage* ε *and makes at most* q_s *and* q_{H_i} $\{i = 1, 2, .., 5\}$ *queries to the client/server respectively. Then there is challenger to solve the computational Diffie-Hellman problem.*

Proof. Our proof follows description in Lemmas 1 and 2.

Lemma 1. The proposed protocol is secure against the type I adversary $\mathscr{A}_1$ in the random oracle model if the CDHP is hard.

Lemma 2. The proposed protocol is secure against the type II adversary $\mathscr{A}_2$ in random oracle if the CDHP is hard.

This proof is omitted because of page limitation. Please contact the authors for the full version.

5.3 Sever-to-client Authentication

In the following theorem, we demonstrate that an adversary $\mathscr{A}_i$ $\{i = 1, 2\}$ cannot impersonate the server to interact with the client under the CDHP.

Theorem 3. *we assume that adversary* $\mathscr{A}_i$ $\{i = 1, 2\}$ *can break up the server-to-client authentication with a non-negligible advantage* ε*. Then there exists algorithm* C *to solve the CDHP with advantage* $\varepsilon^{'} \geq \varepsilon - q_c/2^k - q_c^2/q$*, where* q_c *is the maximum number of queries to the client.*

This proof is omitted because of page limitation. Please contact the authors for the full version.

6 Performance Analysis

In this section, we assess the performance of our scheme from computation cost, and security properties including mutual authentication, key agreement, resistance to forgery attack, perfect-forward-secrecy, key escrow problem and provable security. We compare our protocol with Wu and Tseng [26], He [12] and Tasi and Lo [20] schemes as shown in Tables 3 and 4. For convenience to evaluate the computational cost, we define some notations as follows:

T_e: The time of a bilinear map operation $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$.
T_M: The time of a scalar multiplication operation of $\mathbb{G}_1$.
T_i: The time for performing a modular inversion operation.
T_A: The time of an addition operation of $\mathbb{G}_1$.
T_H: The performing time of a one-way hash function.

We adopt qualitative analysis in Wu and Tseng [26], He [12] and Tasi and Lo [20] schemes for the execution time of each operation based on experimental results from [18]. In [18], the processors on the Philips HiPersmart card and Pentium IV offer maximum clock speeds of 36 MHz and 3 GHz, separately. Considering the security level of the Ate pairing system in [18], we use an elliptic curve $y^2 = x^3 + Ax + B$ over a finite field F_p, with $p = 512$ bits and a large prime

Table 1. Performance evaluation of our protocol

	Client	Server
Authentication phase	$5T_M + T_A + 4T_H$	$2T_e + 4T_M + 2T_A + 6T_H$

Table 2. Computation cost at client side and server side

	T_e	T_M	T_A	T_i	T_H
Server	3.16 ms	1.17 ms	<0.1 ms	<1 ms	0.01 ms
Client	0.38 s	0.13 s	<0.1 s	<0.01 s	<0.001 s

Table 3. Comparisons based computation costs

	[26]	[12]	[20]	Our protocol
Computation cost (Client)	$4T_M + T_A + 3T_H$	$3T_M + 3T_H + T_i$	$2T_M + 3T_H + T_i$	$5T_M + T_A + 4T_H$
Execution time (client)	0.533 s	0.472 s	0.266 s	0.754 s
Computation cost (Server)	$2T_e + 2T_M + T_A + 3T_H$	$T_e + 2T_M + 2T_A + 3T_H$	$T_e + 5T_M + 2T_A + 5T_H$	$2T_e + 4T_M + 2T_A + 6T_H$
Execution time (server)	8.77 ms	5.72 ms	9.26 ms	11.26 ms

Table 4. Comparisons based security Properties

	[26]	[12]	[20]	Our protocol
Mutual authentication	Yes	Yes	Yes	Yes
Key agreement	Yes	Yes	Yes	Yes
Resistance to forgery attack	Yes	Yes	Yes	Yes
Perfect forward-secrecy	No	No	Yes	Yes
No key escrow problem	No	No	No	Yes
provable secrecy	Yes	Yes	Yes	Yes

^{Yes} denotes that the scheme satisfies this security property
^{No} denotes that the scheme does not satisfy this security property

order $q = 160$ bits. The results of our analysis show that the execution time of the proposed protocol is still well suited for the mobile client-server environment. The point is that our protocol has neither key escrow problem since it is based on the CL-PKC (Tables 1 and 2).

7 Conclusion

We propose a new user authentication and key agreement protocol based on certificateless cryptosystem to solve the problem of certificate management of traditional public key cryptosystem and key escrow in identity public key cryptosystem. Based on the assumption of CDH, we prove the security of our protocol to provide secure user authentication, key agreement and mutual authentication in random oracle model. Furthermore, we show that our protocol can be applied to mobile client-server environment.

References

1. Al-Riyami, S.S., Paterson, K.G.: Certificateless public key cryptography. In: International Conference on the Theory and Application of Cryptology and Information Security, pp. 452–473. Springer (2003)
2. Bellare, M., Rogaway, P.: Random oracles are practical: a paradigm for designing efficient protocols. In: Proceedings of the 1st ACM conference on Computer and communications security, pp. 62–73. ACM (1993)
3. Boneh, D., Franklin, M.: Identity-based encryption from the weil pairing. In: Annual International Cryptology Conference, pp. 213–229. Springer (2001)
4. Boneh, D., Franklin, M.: Identity-based encryption from the weil pairing. SIAM J. Comput. **32**(3), 586–615 (2003)
5. Boneh, D., Lynn, B., Shacham, H.: Short signatures from the weil pairing. J. Cryptol. **17**(4), 297–319 (2004)
6. Choon, J.C., Cheon, J.H.: An identity-based signature from gap Diffie-Hellman groups. In: International Workshop on Public Key Cryptography, pp. 18–30. Springer (2003)
7. Das, M.L., Saxena, A., Gulati, V.P., Phatak, D.B.: A novel remote user authentication scheme using bilinear pairings. Comput. Secur. **25**(3), 184–189 (2006)

8. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. Inf. Theory **22**(6), 644–654 (1976)
9. Fang, G., Huang, G.: Improvement of recently proposed remote client authentication protocols (2006)
10. Giri, D., Srivastava, P.: An improved remote user authentication scheme with smart cards using bilinear pairings. IACR Cryptol. ePrint Arch. **2006**, 274 (2006)
11. Goriparthi, T., Das, M.L., Saxena, A.: An improved bilinear pairing based remote user authentication scheme. Comput. Stan. Interfaces **31**(1), 181–185 (2009)
12. He, D.: An efficient remote user authentication and key agreement protocol for mobile client-server environment from pairings. Ad Hoc Netw. **10**(6), 1009–1016 (2012)
13. He, D., Huang, B., Chen, J.: New certificateless short signature scheme. IET Inf. Secur. **7**(2), 113–117 (2013)
14. Hou, M.b., Xu, Q.l.: Secure certificateless-based authenticated key agreement protocol in the client-server setting. In: IEEE International Symposium on IT in Medicine & Education 2009, ITIME 2009, vol. 1, pp. 960–965. IEEE (2009)
15. Jakobsson, M., Pointcheval, D.: Mutual authentication for low-power mobile devices. In: International Conference on Financial Cryptography, pp. 178–195. Springer (2001)
16. Nam, J., Lee, J., Kim, S., Won, D.: DDH-based group key agreement in a mobile environment. J. Syst. Softw. **78**(1), 73–83 (2005)
17. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
18. Scott, M., Costigan, N., Abdulwahab, W.: Implementing cryptographic pairings on smartcards. In: International Workshop on Cryptographic Hardware and Embedded Systems, pp. 134–147. Springer (2006)
19. Shamir, A.: Identity-based cryptosystems and signature schemes. In: Workshop on the Theory and Application of Cryptographic Techniques, pp. 47–53. Springer (1984)
20. Tsai, J.L., Lo, N.W.: Provably secure and efficient anonymous id-based authentication protocol for mobile devices using bilinear pairings. Wirel. Pers. Commun. **83**(2), 1273–1286 (2015)
21. Tseng, Y.M.: Gprs/umts-aided authentication protocol for wireless lans. IEE Proc. Commun. **153**(6), 810–817 (2006)
22. Tseng, Y.M.: A secure authenticated group key agreement protocol for resource-limited mobile devices. Comput. J. **50**(1), 41–52 (2007)
23. Tseng, Y.M., Wu, T.Y., Wu, J.D.: A pairing-based user authentication scheme for wireless clients with smart cards. Informatica **19**(2), 285–302 (2008)
24. Wong, D.S., Chan, A.H.: Efficient and mutually authenticated key exchange for low power computing devices. In: International Conference on the Theory and Application of Cryptology and Information Security, pp. 272–289. Springer (2001)
25. Wu, L., Zhang, Y., Xie, Y., Alelaiw, A., Shen, J.: An efficient and secure identity-based authentication and key agreement protocol with user anonymity for mobile devices. Wirel. Pers. Commun. 1–17 (2016)
26. Wu, T.Y., Tseng, Y.M.: An efficient user authentication and key exchange protocol for mobile client-server environment. Comput. Netw. **54**(9), 1520–1530 (2010)
27. Yoon, E., Yoo, K.: A new efficient id-based user authentication and key exchange protocol for mobile client-server environment. In: 2010 IEEE International Conference on Wireless Information Technology and Systems (ICWITS), pp. 1–4. IEEE (2010)

Improved Online/Offline Attribute Based Encryption and More

Jindan Zhang[1,2,3](✉), Baocang Wang[4], and Xu An Wang[2,5]

[1] Xianyang Vocational Technical College, Xianyang, China
69957106@qq.com

[2] Guangxi Key Laboratory of Cryptography and Information Security,
Guilin University of Electronic Technology, Guilin, China
wangxazjd@163.com

[3] State Key Laboratory of Integrated Service Networks,
Xidian University, Xi'an, China

[4] School of Telecommunications Engineering, Xidian University, Xi'an, China
bcwang79@aliyun.com

[5] Key Laboratory of Cryptology and Information Security,
Engineering University of CAPF, Xi'an, China

Abstract. Attribute based encryption is a very useful primitive for scalable access control on the ciphertexts and has found broad applications, such as secure cloud storage etc. When this primitive is used by mobile phones, the computation cost is too heavy. So Hohenberger and Waters introduced the concept of Online/offline attribute based encryption and give a such concrete construction. In this paper, we give an improved construction based on their proposal. Compared with their proposal, our proposal needs 5 pairings instead of $2|I| + 1$ pairings, which is much more efficient than the original scheme. Furthermore, we generalize this technique to speed up the computation of multi-modular exponentiation, and thus also get an interesting result.

1 Introduction

Attribute based encryption is a cryptographic primitive for flexible controlling on decryption ability for the ciphertexts, such as secure cloud storage [5–8]. However, most of the attribute based encryption use the bilinear pairing, which is a heavy computation task. Thus if we use mobile phone to implement this primitive, the mobile phone will run out the energy in a very short time. Thus we need some advanced techniques to help the mobile phones to implement the task of attribute based encryption and decryption. Usually there are two ways for doing this: the first one is outsourcing the decryption of attribute based encryption to the cloud, which has received great attention these years, and many wonderful results have been achieved, such as [1–4,9]; the second one is to implement the encryption in the online/offline way, that is, when the mobile phone is in charging, it can implement the offline part of the encryption, which is a heavy

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_60

task, when the mobile phone is working without charging, it can implement the online part of encryption, which is a more easy part for encryption. Which can be implement in several seconds. These two ideas are very useful for widely application of attribute based encryption for our life.

In this paper, we concentrate on the second technique, online/offline attribute based encryption. We find that the HK proposal can even be improved again, we can reduce the number of bilinear pairings from linear with the attributes to constant ones, thus can further enlarge the time the mobile phone can live when doing such encryption.

We first review of HW's online/offline scheme and then we propose an improved one based on it. Then we generalize our technique to speed up the computation of multi-modular exponentiation, taking the vector commitments as an example, which is also an interesting work. Finally we give the conclusion.

2 Review of HW's Online/Offline ABE Scheme

Here we first review the concept and scheme of online/offline ABE. In PKC'14 [10], Hohenberger and Waters proposed an online/offline ABE scheme based on the unbounded KP-ABE scheme of Rouselakis and Waters [11].

1. Setup(λ, U). The setup algorithm takes as input a security parameter and a universe U of attributes. To cover the most general case, we let $U = \{0,1\}^*$. It then chooses a bilinear group $\mathbb{G}$ of prime order p, generators $g, h, u, w \in \mathbb{G}$. In addition, it chooses random exponents $\alpha \in Z_p$. The authority sets $MSK = (\alpha, PK)$ as the master secret key. It publishes the public parameters as:

$$PK = (\mathbb{G}, p, g, h, u, w, e(g,g)^\alpha)$$

   We assume that the universe of attributes can be encoded as elements in Z_p.
2. Extract$(MSK, (M, \rho))$. The extract algorithm takes as input the master secret key MSK and an LSSS access structure (M, ρ). Let M be an $l \times n$ matrix. The function ρ associates rows of M to attributes. The algorithm initially chooses random values $y_2, \cdots, y_n \in Z_p$. It then computes l shares of the master secret key as $(\lambda_1, \lambda_2, \cdots, \lambda_l) := M \cdot (\alpha, y_2, \cdots, y_n)^T$ (where T denotes the transpose). It then picks l random exponents $t_1, t_2, \cdots, t_l \in Z_p$. For $i = 1$ to l, it computes

$$K_{i,0} := g^{\lambda_i} w^{t_i}, K_{i,1} := (u^{\rho(i)} h)^{-t_i}, K_{i,2} = g^{t_i}$$

   and the private key is $SK = ((M, \rho), \{K_{i,0}, K_{i,1}, K_{i,2}\}_{i \in [1,l]})$.
3. Offline.Encrypt(PK)). The offline encryption algorithm takes in the public parameters only. Here we describe the basic system which assumes a maximum bound of P attributes will be associated with any ciphertext. The algorithm first picks a random $s \in Z_p$ and computes

$$key = e(g,g)^{\alpha s}, C_0 = g^s$$

Next for $j = 1$ to P, it chooses random $r_j, x_j \in Z_p$ and computes

$$C_{j,1} = g^{r_j}, C_{j,2} = (u^{x_j}h)^{r_j}w^{-s}$$

One can view this as encrypting for a random attribute x_j, where this will be corrected in the online phase. We remark that the work done in the offline phase is roughly equivalent to the work of the regular encryption algorithm in [11].

The intermediate ciphertext is $IT = (Key, C_0, \{r_j, x_j, C_{j,1}, C_{j,2}\}_{j\in[1,P]})$.

4. $\mathsf{Online.Encrypt}(PK)$). The online encryption KEM algorithm takes as input the public parameters, an intermediate ciphertext IT, and a set of attributes $S = (A_1, A_2, \cdots, A_{k\le P})$. For $j = 1$ to k, it computes $C_{j,3} := (r_j(A_j - x_j)) \bmod p$. Intuitively, this will correct to the proper attributes. It sets the ciphertext as:
$$CT = (S, C_0, \{C_{j,1}, C_{j,2}, C_{j,3}\}_{j\in[1,k]})$$
The encapsulated key is key. The dominant cost is one multiplication in Z_p per attribute in S.
5. $\mathsf{Decrypt}(SK, CT)$. The decryption algorithm in the KEM setting recovers the encapsulated key. It takes as input a ciphertext $CT = (S, C_0, \{C_{j,1}, C_{j,2}, C_{j,3}\}_{j\in[1,k]})$ for attribute set S and a private key $SK = ((M, \rho), \{K_{i,0}, K_{i,1}, K_{i,2}\}_{i\in[1,l]}$ for access structure (M, ρ). If S does not satisfy this access structure, then the algorithm issues an error message. Otherwise, it sets $I := \{i : \rho(i) \in S\}$ and computes constants $\omega_i \in Z_p$ such that $\Sigma_{i\in I}\omega_i \cdot M_i = (1, 0, \cdots, 0)$, where M_i is the i-th row of the matrix M. Then it then recovers the encapsulated key by calculating $key :=$
$$\Pi_{i\in I}(e(C_0, K_{i,0})e(C_{j,1}, K_{i,1})e(C_{j,2} \cdot u^{C_{j,3}}, K_{i,2}))^{\omega_i} = e(g, g)^{\alpha s}$$
where j is the index of the attribute $\rho(i)$ in S(it depends on i). This does not increase the number of pairing operations over [[11], Appendix C], although it adds $|I|$ exponentiations.

$\mathsf{Correctness}$. If the attribute set S of the ciphertext is authorized, we have that $\Sigma_{i\in I}\omega_i\lambda_i = \alpha$. Therefore, $Key =$:

$$\begin{aligned}
&\Pi_{i\in I}(e(C_0, K_{i,0})e(C_{j,1}, K_{i,1})e(C_{j,2} \cdot u^{C_{j,3}}, K_{i,2}))^{\omega_i}\\
&= \Pi_{i\in I}(e(g^s, g^{\lambda_i}\omega^{t_i})e(g^{r_j}, (u^{\rho(i)}h)^{-t_i})e((u^{x_j}h)^{r_j}\omega^{-s} \cdot u^{r_j(\rho(i)-x_j)}, g^{t_i}))^{\omega_i}\\
&= \Pi_{i\in I}(e(g^s, g^{\lambda_i}\omega^{t_i})e(g^{r_j}, (u^{\rho(i)}h)^{-t_i})e((u^{x_j}h)^{r_j}\omega^{-s} \cdot u^{r_j(\rho(i)-x_j)}, g^{t_i}))^{\omega_i}\\
&= \Pi_{i\in I}(e(g, g)e(g, \omega)^{st_i}e(g, u)^{-r_jt_i\rho(i)}e(g, h)^{-r_jt_i}e(g, u)^{\rho(i)r_jt_i}e(g, h)^{r_jt_i}e(g, \omega)^{-st_i})^{\omega_i}\\
&= \Pi_{i\in I}e(g, g)^{s\omega_i\lambda_i} = e(g, g)^{s\alpha}
\end{aligned}$$

2.1 Our Improved Online/Offline ABE Scheme

1. $\mathsf{Setup}(\lambda, U)$. The setup algorithm takes as input a security parameter and a universe U of attributes. To cover the most general case, we let $U = \{0, 1\}^*$.

It then chooses a bilinear group $\mathbb{G}$ of prime order p, generators $g, h, u, w \in \mathbb{G}$. In addition, it chooses random exponents $\alpha \in Z_p$. The authority sets $MSK = (\alpha, PK)$ as the master secret key. It publishes the public parameters as:

$$PK = (\mathbb{G}, p, g, h, u, w, e(g, g)^{\alpha})$$

We assume that the universe of attributes can be encoded as elements in Z_p.

2. Extract$(MSK, (M, \rho))$. The extract algorithm takes as input the master secret key MSK and an LSSS access structure (M, ρ). Let M be an $l \times n$ matrix. The function ρ associates rows of M to attributes. The algorithm initially chooses random values $y_2, \cdots, y_n \in Z_p$. It then computes l shares of the master secret key as $(\lambda_1, \lambda_2, \cdots, \lambda_l) := M \cdot (\alpha, y_2, \cdots, y_n)^T$ (where T denotes the transpose). It then picks l random exponents $t_1, t_2, \cdots, t_l \in Z_p$ and also a random exponent $T_1 \in Z_p$. It first computes $K_0 = g^{T_1}$, and for $i = 1$ to l, it computes

$$K_{i,0} := g^{\lambda_i} w^{t_i}, K_{i,1} := (u^{\rho(i)} h)^{-t_i}, K_{i,2} = (t_i - T_1) \bmod p$$

and the private key is $SK = ((M, \rho), K_0, \{K_{i,0}, K_{i,1}, K_{i,2}\}_{i \in [1,l]})$.

3. Offline.Encrypt$(PK))$. The offline encryption algorithm takes in the public parameters only. Here we describe the basic system which assumes a maximum bound of P attributes will be associated with any ciphertext. The algorithm first picks a random $s \in Z_p$ and computes

$$key = e(g, g)^{\alpha s}, C_0 = g^s$$

Next it first selects a random T_2 and computes $C_1 = g^{T_2}$, and for $j = 1$ to P, it chooses random $r_j, x_j \in Z_p$ and computes

$$C_{j,1} = (r_j - T_2) \bmod p, C_{j,2} = (u^{x_j} h)^{r_j} w^{-s}$$

One can view this as encrypting for a random attribute x_j, where this will be corrected in the online phase. We remark that the work done in the offline phase is roughly equivalent to the work of the regular encryption algorithm in [11].

The intermediate ciphertext is $IT = (Key, C_0, C_1, \{r_j, x_j, C_{j,1}, C_{j,2}\}_{j \in [1,P]})$.

4. Online.Encrypt$(PK))$. The online encryption KEM algorithm takes as input the public parameters, an intermediate ciphertext IT, and a set of attributes $S = (A_1, A_2, \cdots, A_{k \leq P})$. For $j = 1$ to k, it computes $C_{j,3} := (r_j(A_j - x_j)) \bmod p$. Intuitively, this will correct to the proper attributes. It sets the ciphertext as:

$$CT = (S, C_0, C_1, \{C_{j,1}, C_{j,2}, C_{j,3}\}_{j \in [1,k]})$$

The encapsulated key is key. The dominant cost is one multiplication in Z_p per attribute in S.

5. Decrypt(SK, CT). The decryption algorithm in the KEM setting recovers the encapsulated key. It takes as input a ciphertext $CT = (S, C_0, C_1, \{C_{j,1}, C_{j,2}, C_{j,3}\}_{j \in [1,k]})$ for attribute set S and a private key

$SK = ((M,\rho), K_0, \{K_{i,0}, K_{i,1}, K_{i,2}\}_{i\in[1,l]}$ for access structure (M,ρ). If S does not satisfy this access structure, then the algorithm issues an error message. Otherwise, it sets $I := \{i : \rho(i) \in S\}$ and computes constants $\omega_i \in Z_p$ such that $\Sigma_{i\in I}\omega_i \cdot M_i = (1, 0, \cdots, 0)$, where M_i is the i-th row of the matrix M. Then it then recovers the encapsulated key by calculating $key :=$

$$\Pi_{i\in I}(e(C_0, K_{i,0})e(C_1 \cdot g^{C_{j,1}}, K_{i,1})e(C_{j,2} \cdot u^{C_{j,3}}, K_0 \cdot g^{K_{i,2}}))^{\omega_i} = e(g,g)^{\alpha s}$$

where j is the index of the attribute $\rho(i)$ in S(it depends on i).

Correctness. If the attribute set S of the ciphertext is authorized, we have that $\Sigma_{i\in I}\omega_i\lambda_i = \alpha$. Therefore, $Key =:$

$$\begin{aligned}
&\Pi_{i\in I}(e(C_0, K_{i,0})e(C_1 \cdot g^{C_{j,1}}, K_{i,1})e(C_{j,2} \cdot u^{C_{j,3}}, K_0 \cdot g^{K_{i,2}}))^{\omega_i}\\
&= \Pi_{i\in I}(e(C_0, K_{i,0})e(g^{r_j}, K_{i,1})e(C_{j,2} \cdot u^{C_{j,3}}, g^{t_i}))^{\omega_i}\\
&= \Pi_{i\in I}(e(g^s, g^{\lambda_i}\omega^{t_i})e(g^{r_j}, (u^{\rho(i)}h)^{-t_i})e((u^{x_j}h)^{r_j}\omega^{-s} \cdot u^{r_j(\rho(i)-x_j)}, g^{t_i}))^{\omega_i}\\
&= \Pi_{i\in I}(e(g^s, g^{\lambda_i}\omega^{t_i})e(g^{r_j}, (u^{\rho(i)}h)^{-t_i})e((u^{x_j}h)^{r_j}\omega^{-s} \cdot u^{r_j(\rho(i)-x_j)}, g^{t_i}))^{\omega_i}\\
&= \Pi_{i\in I}(e(g,g)e(g,\omega)^{st_i}e(g,u)^{-r_jt_i\rho(i)}e(g,h)^{-r_jt_i}e(g,u)^{\rho(i)r_jt_i}e(g,h)^{r_jt_i}e(g,\omega)^{-st_i})^{\omega_i}\\
&= \Pi_{i\in I}e(g,g)^{s\omega_i\lambda_i} = e(g,g)^{s\alpha}
\end{aligned}$$

But note here

$$\begin{aligned}
&\Pi_{i\in I}(e(C_0, K_{i,0})e(C_1 \cdot g^{C_{j,1}}, K_{i,1})e(C_{j,2} \cdot u^{C_{j,3}}, K_0 \cdot g^{K_{i,2}}))^{\omega_i}\\
&= e(C_0, \Pi_{i\in I}K_{i,0}^{\omega_i})e(C_1, \Pi_{i\in I}K_{i,1}^{\omega_i})e(g, \Pi_{i\in I}K_{i,1}^{\omega_i C_{j,1}})e(\Pi_{i\in I}(C_{j,2} \cdot u^{C_{j,3}})^{\omega_i}, K_0)\\
&e(\Pi_{i\in I}(C_{j,2} \cdot u^{C_{j,3}})^{\omega_i K_{i,2}}, g)
\end{aligned}$$

which needs 5 pairings instead of $2|I| + 1$ pairings for the original scheme, and the original scheme needs $2|I|$ modular exponentiation while this scheme needs $5|I| + 2$ modular exponentiation, which is still more efficient than the original scheme.

3 Generalization

Here we generalize the above technique to the setting for modular exponentiation, which is a very usual operation in cryptographic primitives. We illustrate the new technique for speeding up multi-modular exponentiation via an improvement to [12]. First we review the CF vector commitment scheme.

3.1 CF's Vector Commitments

1. VC.KeyGen$(1^k, q)$. Let $\mathbb{G}, \mathbb{G}_T$ be two bilinear groups of prime order p equipped with a bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$. Let $g \in \mathbb{G}$ be a random generator. Randomly choose $z_1, \cdots, z_q \leftarrow Z_p$. For all $i = 1, \cdots, q$ set $h_i = g^{z_i}$. For all $i, j = 1, \cdots, q$, $i \neq j$ set $h_{i,j} = g^{z_iz_j}$. Set $pp = (g, \{h_i\}_{i\in[q]}, \{h_{i,j}\}_{i,j\in[q], i\neq j})$. The message space is $\mathcal{M} = Z_p$.

2. $\mathsf{VC.Com}_{pp}(m_1, \cdots, m_q)$. Compute $C = h_1^{m_1} h_2^{m_2} \cdots h_q^{m_q}$ and output C and the auxiliary information $aux = (m_1, \cdots, m_q)$
3. $\mathsf{VC.Open}_{pp}(m_i, i, aux)$. Compute

$$\Lambda_i = \prod_{j=1, j\neq i}^{q} h_{i,j}^{m_j} = (\prod_{j=1, j\neq i}^{q} h_j^{m_j})^{z_i}$$

4. $\mathsf{VC.Ver}_{pp}(C, m_i, i, \Lambda_i)$. If the following equations hold,

$$e(C/h_i^{m_i}, h_i) = e(\Lambda_i, g)$$

then outputs 1, otherwise output 0.
5. $\mathsf{VC.Update}_{pp}(C, m, m', i)$. Compute the updated commitment $C' = C \cdot h_i^{m_i - m}$. Finally output C' and $U = (m, m', i)$.
6. $\mathsf{VC.ProofUpdate}_{pp}(C, \Lambda_j, m', U)$. A client who owns a proof Λ_j, that is valid w.r.t. to C for some message at position j, can use the update information $U = (m, m', i)$ to compute the updated commitment C' and produce a new proof Λ'_j which will be valid w.r.t C'. We distinguish two cases:
   a. $i \neq j$. Compute the updated commitment $C' = C \cdot h_i^{m' - m}$ while the updated proof is $\Lambda'_j = \Lambda_j (h_i^{m'-m})^{z_j} = \Lambda_j h_{j,i}^{m'-m}$
   b. $i = j$. Compute the updated commitment as $C' = C \cdot h_i^{m'-m}$ while the updated proof remains the same as Λ_i.

3.2 Our Improved Algorithm

1. $\mathsf{VC.KeyGen}(1^k, q)$. Let $\mathbb{G}, \mathbb{G}_T$ be two bilinear groups of prime order p equipped with a bilinear map $e : \mathbb{G} \times \mathbb{G} \to \mathbb{G}_T$. Let $g \in \mathbb{G}$ be a random generator. Randomly choose $z_1, \cdots, z_q \leftarrow Z_p$. For all $i = 1, \cdots, q$ set $h_i = g^{z_i}$. Furthermore, choose a random $t \in Z_p$, computes $T = g^t, r_1 = z_1 - t \bmod p, r_2 = z_2 - t \bmod p, \cdots, r_q = z_q - t \bmod p$. Note here $h_i = g^t \cdot g^{r_i} = g^{t + z_i - t} = g^{z_i}$. For all $i, j = 1, \cdots, q$, $i \neq j$ set $h_{i,j} = g^{z_i z_j}$. Set $pp = (g, T, r_1, r_2, \cdots, r_q, \{h_{i,j}\}_{i,j \in [q], i \neq j})$. The message space is $\mathcal{M} = Z_p$.
2. $\mathsf{VC.Com}_{pp}(m_1, \cdots, m_q)$. Compute

$$C = h_1^{m_1} h_2^{m_2} \cdots h_q^{m_q} = (Tg^{r_1})^{m_1} \cdots Tg^{r_q})^{m_q})$$
$$= T^{(m_1 + m_2 + \cdots + m_q)} g^{r_1 m_1 + r_2 m_2 + \cdots + r_q m_q}$$

and output C and the auxiliary information $aux = (m_1, \cdots, m_q)$. Note here the committer needs only compute two modular exponentiations instead of q modular exponentiations.
3. $\mathsf{VC.Open}_{pp}(m_i, i, aux)$. Compute

$$\Lambda_i = \prod_{j=1, j\neq i}^{q} h_{i,j}^{m_j} = (\prod_{j=1, j\neq i}^{q} h_j^{m_j})^{z_i}$$

4. $\mathsf{VC.Ver}_{pp}(C, m_i, i, \Lambda_i)$. If the following equations hold,

$$e(C/h_i^{m_i}, h_i) = e(\Lambda_i, g)$$

   then outputs 1, otherwise output 0.
5. $\mathsf{VC.Update}_{pp}(C, m, m', i)$. Compute the updated commitment $C' = C \cdot h_i^{m_i - m}$. Finally output C' and $U = (m, m', i)$.
6. $\mathsf{VC.ProofUpdate}_{pp}(C, \Lambda_j, m', U)$. A client who owns a proof Λ_j, that is valid w.r.t. to C for some message at position j, can use the update information $U = (m, m', i)$ to compute the updated commitment C' and produce a new proof Λ'_j which will be valid w.r.t C'. We distinguish two cases:
   a. $i \neq j$. Compute the updated commitment $C' = C \cdot h_i^{m'-m}$ while the updated proof is $\Lambda'_j = \Lambda_j (h_i^{m'-m})^{z_j} = \Lambda_j h_{j,i}^{m'-m}$
   b. $i = j$. Compute the updated commitment as $C' = C \cdot h_i^{m'-m}$ while the updated proof remains the same as Λ_i.

4 Conclusion

In this paper, we consider the issue of implementing of online/offline of ABE for mobile devices with energy efficiency. We give an improvement to the HW's proposal. And we also generalize our technique to the setting of multi modular-exponentiation. However, we also note our results are very basic, there are many work need to do in the future, such as proving the security of the proposals in the formal model, and extending this technique to other settings.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 61572390), the 111 Project (No. B08038), and Guangxi Key Laboratory of Cryptography and Information Security (Grant No. GCIS201610).

References

1. Green, M., Hohenberger, S., Waters, B.: Outsourcing the decryption of ABE ciphertexts. In: Proceedings of the USENIX Security Symposim, San Francisco, CA, USA (2013)
2. Lai, J., Deng, R., Guan, C., Weng, J.: Attribute-based encryption with verifiable outsourced decryption. IEEE Trans. Inf. Forensics Secur. **8**(8), 1343–1354 (2013)
3. Li, J., Huang, X., Li, J., Chen, X., Xiang, Y.: Securely outsourcing attribute-based encryption with checkability. IEEE Trans. Parallel Distrib. Syst. (2013, in Press). doi:10.1109/TPDS.2013.27
4. Qin, B., Deng, R.H., Liu, S., Ma, S.: Attribute-based encryption with efficient verifiable outsourced decryption. IEEE Trans. Inf. Forensics Secur. **10**(7), 1384–1393 (2015)
5. Puzar, M., Plagemann, T.: Data sharing in mobile ad-hoc networks-a study of replication and performance in the MIDAS data space. Int. J. Space-Based Situated Comput. **1**(2/3), 137–150 (2015)

6. Petrlic, R., Sekula, S., Sorge, C.: A privacy-friendly architecture for future cloud computing. Int. J. Grid Util. Comput. **4**(4), 265–277 (2013)
7. Wang, Y., Du, J., Cheng, X., Liu, Z., Lin, K.: Degradation and encryption for outsourced PNG images in cloud storage. In. Int. J. Grid Util. Comput. **7**(1), 22–28 (2016)
8. Ye, X., Khoussainov, B.: Fine-grained access control for cloud computing. Int. J. Grid Util. Comput. **4**(2/3), 160–168 (2013)
9. Wang, X.A., Ma, J., Xhafa, F.: Outsourcing decryption of attribute based encryption with energy efficiency. In: Proceeding of the International Conference on P2P, Parallel, Grid, Cloud and Internet Computing 3PGCIC 2015, pp. 444-448. IEEE (2015)
10. Hohenberger, S., Waters, B.: Online/offline attribute-based encryption. In: Krawczyk, H. (ed.) PKC 2014. LNCS, vol. 8383, pp. 293–310. Springer, March 2014
11. Rouselakis, Y., Waters, B.: Practical constructions and new proof methods for large universe attribute-based encryption. In: Sadeghi, A.-R., Gligor, V.D., Yung, M. (eds.) ACM CCS 13, pp. 463–474. ACM Press, November 2013
12. Catalano, D., Fiore, D.: Vector commitments and their applications. In: Kurosawa, K., Hanaoka, G. (eds.) PKC 2013. LNCS, vol. 7778, pp. 55–72. Springer, February/March 2013

On the Security of a Cloud Data Storage Auditing Protocol IPAD

Xu An Wang[1,2(✉)], Xiaoshuang Luo[1], Jindan Zhang[3], and Xiaoyuan Yang[1]

[1] Key Laboratory of Cryptology and Information Security,
Engineering University of CAPF, Xi'an 710086, China
wangxazjd@163.com

[2] Guangxi Key Laboratory of Cryptography and Information Security, Guilin
University of Electronic Technology, Guilin, People's Republic of China

[3] State Key Laboratory of Integrated Service Networks, Xidian University,
Xi'an 710071, China

Abstract. Nowadays cloud data storage is a very important storage service for us, but to ensure the datum stored in the remote cloud server remains unmodified, we need a mechanism to check the datum's integrity, cloud data storage auditing protocol is such a mechanism, which has received great attention from researchers. Recently Zhang et al. proposed an efficient ID-based public auditing protocol called IPAD for the outsourced data by combing Waters signature and public auditing for the outsourced data. They claimed IPAD is the first ID-based auditing protocol for data integrity in the standard security model. But in this paper we show their proposal is not secure. Especially, the adversaries can easily generate tags for any file, which obviously break the unforgeability property of the cloud storage auditing protocol.

1 Introduction

In these days, cloud computation is a very hot research topic for its promising properties of cheap management cost for users, any where/any time access, and very scalable software and hard ware investigation [21]. However, before adapting cloud computation, data owners should ensure their data shall be secure and well protected [22–24]. Integrity is one of the most important security property for cloud storage. However when the data owners outsource their datum to the cloud, the datum is not controlled by the owners any more, how to ensure the datum has not been modified and changed? Cloud storage auditing protocol is such a mechanism. In 2007, the first provable data possession (PDP) scheme was proposed by Atenesis et al. [1,2]. Also the proof of retrievability protocol for cloud storage was proposed by Jules et al. [3] and Shacham and Waters [4]. Later many cloud auditing protocols with different properties have been proposed, such as cloud auditing protocols with dynamic updates [5–9], cloud auditing protocols with publicly verifiability [4,11], cloud auditing protocols with privacy-preserving [11–13], cloud auditing protocols with other interesting properties [14–18].

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_61

Recently, Zhang et al. [25] proposed an efficient ID-based public auditing protocol called IPAD for the outsourced data by combing Waters signature and public auditing for the outsourced data. They claimed IPAD is the first ID-based auditing protocol for data integrity in the standard security model. But in this paper we show their proposal is not secure. Especially, the adversaries can easily generate tags for any file, which obviously break the unforgeability property of the cloud storage auditing protocol.

2 Review of Zhang et al.'s IPAD Scheme

Here we first review Zhang et al.'s ID-based public auditing protocol in the standard model [25]. It consists the following six algorithms: Setup, Key-extract, TagGen, Challenge, Proof, Verifying, the details are given as follows:

1. Setup. For public key generator (PKG), it sets up the following system parameters. Given a security parameter k, it selects two multiplicative cyclic groups $(\mathbb{G}_1, \mathbb{G}_2)$ of prime order $q \geq 2^k$, let $e : \mathbb{G}_1 \times \mathbb{G}_1 \rightarrow \mathbb{G}_2$ be a bilinear map and $g \in_R \mathbb{G}_1$ be a generator of group $\mathbb{G}_1$. h, g_2 are two random generators of group $\mathbb{G}_1$. Let $H_1 : \{0,1\}^* \rightarrow Z_q$ and $H_v : \{0,1\}^* \rightarrow \{0,1\}^{n_v}$ be two collision-resistant hash functions, where $n_v \in Z$. And randomly choose $\alpha \in Z_q$ as master key of PKG and compute the corresponding public key $P_{pub} = g^{\alpha}$. Also randomly choose the following elements:
   a. $v' \in_R \mathbb{G}_1$
   b. $v_i \in_R \mathbb{G}_1$ for $i = 1, \cdots, n_v$. Let $V = \{v_i\}_{i \ in 1, \cdots, n_v}$
   Finally publish the following system parameters:
   $$Param = (\mathbb{G}_1, \mathbb{G}_2, q, e, v', V, g, h, H_1, H_v, g_2, P_{pub})$$
   At the same time, PKG secretly keeps his master secret key s.
2. Key Extraction. For a data user with identity ID_j, if it wants to register his identity ID_j to PKG, the following steps are executed:
   a. First it submits his identity ID to the PKG.
   b. PKG computes $\mathfrak{B}_j = H_v(ID_j)$. Let $\mathfrak{B}_j[i]$ be the i-th bit of $\mathfrak{B}_j$. Then define $V_j \subset \{1, \cdots, n_v\}$ be the set of indicies such that $\mathfrak{B}_j = 1$
   c. To produce the private key d_j of the data user with identity ID_j, PKG randomly choose $a_{u_j} \in_R Z_q$ to compute
   $$d_j = (d_{j1}, d_{j2}) = (g_2^{\alpha}(v' \prod_{i \in V_j} v_i)^{a_{u_j}}, g^{a_{u_j}})$$
3. TagGen. Given a data file M, the data user with identity ID_j splits M into n blocks such that each block has s sectors. Namely, $M = m_1 || \cdots || m_n$ and $m_i = m_{i1} || \cdots || m_{is}$. Then it chooses a random file name $Name$ from a sufficiently large domain Z_q^* and $s+1$ random values $r_0, r_1, \cdots, r_s \in_R Z_q$ to compute $u_i = g_2^{r}$ for each $0 \leq i \leq s$.

To produce file tag, it also does the following steps:

a. Compute $(pk_s, sk_s) \leftarrow \Sigma.KeyGen(1^k)$ to obtain a pair of public/private keys, where Σ is a secure signature algorithm.
b. Compute $\Phi = \Sigma.sign(sk_s, \tau_0)$ to obtain a signature on string τ_0, where $\tau_0 =$ "$Name||n||u_0||u_1||\cdots||u_s$"
c. For each data block m_i, $1 \leq i \leq n$, it computes

$$\omega_i = r_0 H_1(Name||i) + \sum_{j=1}^{s} r_j m_{ij}$$

d. The authentication tag on data block m is computed as

$$t_i = \ (t_{i,1} = (d_{j1})^{\omega_i} = g_2^{\varepsilon_1}(v' \prod_{i \in V_j} v_i)^{\varepsilon_2},$$
$$t_{i,2} = d_{j2}^{\omega_i} = g^{\varepsilon_2})$$

where $\varepsilon = \alpha \cdot \omega_i$, $\varepsilon = a_{u_j} \cdot \omega_i$ (Note that to produce a probabilistic signature, the data user with identity ID_j also can select $\hat{r} \in Z_q$ to compute

$$t_i = \ (t_{i1} = (d_{j1})^{\delta_i}(v' \prod_{i \in V_j} v_i)^{\hat{r}},$$
$$t_{i2} = d_{j2}^{\delta_i} \cdot g^{\hat{r}})$$

Finally, the data user sends the data file M together with all the authentication tag t_i, $1 \leq i \leq n$ to the cloud storage server, And delete the above random values $r_0, r_1, \cdots, r_s$, private key sk_s of signature algorithm $\sum$ and the local file M.

4. Challenge Phase. To check data integrity of the outsourced data, the auditor first verifies whether the signature ϕ is valid by invoking $\sigma.Verify(\phi, pk_s)$. If it is not, outputs 0 and terminates it. otherwise, the auditor parses τ to recover the file, name $Name$ and n as well as $u_0, u_1, \cdots, u_s$. Then it randomly chooses a l-element subset I of the set $[1, n]$ and a number $\rho \in Z_q$ to produce the following challenging message

$$Chall = \{\rho, I\}$$

and sends them the cloud storage server.

5. Prove. Upon receiving the challenging message $Chall = (I, \rho)$, the cloud storage server first produces a l-element set $Q = (i, \beta_i)$ where $i \in I, \beta_i = \rho^i \bmod q$, Then based on the outsourced data file $M = \{m_1, \cdots, m_n\}$ and authentication tags t_i, $1 \leq i \leq n$, it computes

$$\delta_1 = \prod_{i \in I} t_{i1}^{\beta_i}$$

$$\delta_2 = \prod_{i \in I} t_{i2}^{\beta_i}$$

and for $j = 1$ to s, it computes

$$\mu_j = \sum_{i \in I} \beta_i m_{ij}$$

Finally, the cloud storage server responds the auditor with the corresponding proof information $Prf = (\delta_1, \delta_2, \{\mu_j\}_{j=1,\cdots,s})$

6. Verifying. According to the responded proof information $Prf = (\delta_1, \delta_2, \{\mu_j\}_{j=1,\cdots,s})$, the auditor first computes

$$\hat{h} = \sum_{i \in I} \beta_i H_1(Name||i)$$

Then it verifies the integrity of data file by the following equation

$$e(u_0^{\hat{h}} \cdot \prod_{i=1}^{s} u_i^{\mu_j}, P_{pub}) e(v' \cdot \prod_{i \in V_j} v_i, \delta_2) = e(\delta_1, g)$$

If the above Equation holds, the auditor outputs $VerifyRst$ as accept; otherwise, output $VerifyRSt$ as reject.

3 Our Attack

Our attack shows that the adversary can forge tags for any new files.

- For the deterministic tag generation algorithm, the attack runs as the following:
  1. First the adversary can query on the data owner ID_t for the tag generation on different data blocks $(m_1, m_2, \cdots, m_n)$, he can get the following tags for $(m_1, m_2, \cdots, m_n)$:

$$\begin{aligned} t_{11} &= (d_{t1})^{r_0 H_1(Name||1) + \sum_{j=1}^{s} r_j m_{1j}}, \\ t_{12} &= (d_{t2})^{r_0 H_1(Name||1) + \sum_{j=1}^{s} r_j m_{1j}}, \\ &\cdots\cdots\cdots \\ t_{n1} &= (d_{t1})^{r_0 H_1(Name||n) + \sum_{j=1}^{s} r_j m_{nj}}, \\ t_{n2} &= (d_{t2})^{r_0 H_1(Name||n) + \sum_{j=1}^{s} r_j m_{nj}}, \end{aligned}$$

  2. Let $A_j = (d_{t1})^{r_j} (0 \leq j \leq n)$, the adversary can get the following equations:

$$t_{11} = (d_{t1})^{r_0 H_1(Name||1) + \sum_{j=1}^{s} r_j m_{1j}} = (A_0)^{H_1(Name||1)} (A_1)^{m_{11}} \cdots (A_s)^{m_{1s}} \quad (1)$$

$$\cdots\cdots\cdots$$

$$t_{n1} = (d_{t1})^{r_0 H_1(Name||n) + \sum_{j=1}^{s} r_j m_{nj}} = (A_0)^{H_1(Name||n)} (A_1)^{m_{n1}} \cdots (A_s)^{m_{ns}} \quad \text{(n)}$$

3. Note in the above equations, $H_1(Name||1), \cdots, H_1(Name||n), m_{11}, \cdots, m_{1s}, m_{n1}, \cdots, m_{ns}$ are all known to anyone including the adversary, thus he can compute
$$A_0, A_1, \cdots, A_s$$
by implementing linear transformation on the exponentials with high probability (in case the computation fails, the adversary can query on new files with different data blocks and compute again until succeeding).
4. Once the adversary get $A_0, A_1, \cdots, A_s$, he can compute tags for any file with name $Name*$ and $m^*_{11}, \cdots, m^*_{1s}, m^*_{n1}, \cdots, m^*_{ns}$ as following:
$$\begin{aligned} t_{11} &= (d_{t1})^{r_0 H_1(Name^*||1)+\sum_{j=1}^{s} r_j m^*_{1j}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{11}} \cdots (A_s)^{m^*_{1s}} \\ t_{12} &= (d_{t2})^{r_0 H_1(Name^*||1)+\sum_{j=1}^{s} r_j m^*_{1j}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{11}} \cdots (A_s)^{m^*_{1s}} \\ & \qquad \cdots\cdots\cdots \\ t_{n1} &= (d_{t1})^{r_0 H_1(Name^*||n)+\sum_{j=1}^{s} r_j m^*_{nj}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{n1}} \cdots (A_s)^{m^*_{ns}} \\ t_{n2} &= (d_{t2})^{r_0 H_1(Name^*||n)+\sum_{j=1}^{s} r_j m^*_{nj}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{n1}} \cdots (A_s)^{m^*_{ns}} \end{aligned}$$
5. Thus the adversary could generate tags for any file, which obviously break the un-forgeability property of the cloud storage auditing protocol.

- For the randomized tag generation algorithm, the attack runs as the following:
  1. First the adversary can query on the data owner ID_t for the tag generation on different data blocks $(m_1, m_2, \cdots, m_n)$, he can get the following tags for $(m_1, m_2, \cdots, m_n)$:
  $$\begin{aligned} t_{11} &= (d_{t1})^{r_0 H_1(Name||1)+\sum_{j=1}^{s} r_j m_{1j}}, \\ t_{12} &= (d_{t2})^{r_0 H_1(Name||1)+\sum_{j=1}^{s} r_j m_{1j}}, \\ & \qquad \cdots\cdots\cdots \\ t_{n1} &= (d_{t1})^{r_0 H_1(Name||n)+\sum_{j=1}^{s} r_j m_{nj}}, \\ t_{n2} &= (d_{t2})^{r_0 H_1(Name||n)+\sum_{j=1}^{s} r_j m_{nj}}, \end{aligned}$$
  2. Let $A_j = (d_{t1})^{r_j} (0 \le j \le n)$, the adversary can get the following equations:
  $$t_{11} = (d_{t1})^{r_0 H_1(Name||1)+\sum_{j=1}^{s} r_j m_{1j}} = (A_0)^{H_1(Name||1)}(A_1)^{m_{11}} \cdots (A_s)^{m_{1s}} \tag{1}$$
  $$\cdots\cdots\cdots$$
  $$t_{n1} = (d_{t1})^{r_0 H_1(Name||n)+\sum_{j=1}^{s} r_j m_{nj}} = (A_0)^{H_1(Name||n)}(A_1)^{m_{n1}} \cdots (A_s)^{m_{ns}} \tag{n}$$
  3. Note in the above equations, $H_1(Name||1), \cdots, H_1(Name||n), m_{11}, \cdots, m_{1s}, m_{n1}, \cdots, m_{ns}$ are all known to anyone including the adversary, thus he can compute
  $$A_0, A_1, \cdots, A_s$$
  by implementing linear transformation on the exponentials with high probability (in case the computation fails, the adversary can query on new files with different data blocks and compute again until succeeding).

4. Once the adversary get $A_0, A_1, \cdots, A_s$, he can compute tags for any file with name $Name*$ and $m^*_{11}, \cdots, m^*_{1s}, m^*_{n1}, \cdots, m^*_{ns}$ as following:

$$\begin{aligned} t_{11} &= (d_{t1})^{r_0 H_1(Name^*||1)+\sum_{j=1}^{s} r_j m^*_{1j}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{11}} \cdots (A_s)^{m^*_{1s}} \\ t_{12} &= (d_{t2})^{r_0 H_1(Name^*||1)+\sum_{j=1}^{s} r_j m^*_{1j}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{11}} \cdots (A_s)^{m^*_{1s}} \\ &\qquad \cdots\cdots\cdots \\ t_{n1} &= (d_{t1})^{r_0 H_1(Name^*||n)+\sum_{j=1}^{s} r_j m^*_{nj}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{n1}} \cdots (A_s)^{m^*_{ns}} \\ t_{n2} &= (d_{t2})^{r_0 H_1(Name^*||n)+\sum_{j=1}^{s} r_j m^*_{nj}} = (A_0)^{H_1(Name^*||1)}(A_1)^{m^*_{n1}} \cdots (A_s)^{m^*_{ns}} \end{aligned}$$

5. Thus the adversary could generate tags for any file, which obviously break the un-forgeability property of the cloud storage auditing protocol.

4 Conclusion

In this paper, we show a recent proposed cloud auditing protocol is not secure, the reason why their scheme is not secure is that, the tag generation algorithm is not secure, by querying many times of tag generation oracle, the adversary can easily forge new tags for any block. We point out this attack is a very basic result, we leave how to strengthen their scheme to be secure as our future work.

Acknowledgements. This work is supported by Nature Science Foundation of China (Grant NO. U1636114, 61572521) and Natural Science Foundation of Shaanxi Province (Grant No. 2016JQ6037) and Guangxi Key Laboratory of Cryptography and Information Security (Grant No. GCIS201610).

References

1. Ateniese, G., Burns, R.C., Curtmola, R., Herring, J., Kissner, L., Peterson, Z.N.J., Song, D.: Provable data possession at untrusted stores. In: Ning, P., di Vimercati, S.D.C., Syverson, P.F. (eds.) ACM CCS 2007, pp. 598–609. ACM Press, Alexandria (2007)
2. Ateniese, G., Burns, R.C., Curtmola, R., Herring, J., Kissner, L., Peterson, Z.N.J., Song, D.: Remote data checking using provable data possession. ACM Trans. Inf. Syst. Secur. **14**(1), 12 (2011)
3. Juels, A., Kaliski Jr, B.S.: PORS: proofs of retrievability for large files. In: Ning, P., di Vimercati S.D.C., Syverson P.F. (eds.) ACM CCS 2007, pp. 584–597. ACM Press, Alexandria (2007)
4. Shacham, H., Waters, B.: Compact proofs of retrievability. In: Pieprzyk, J. (ed.) ASIACRYPT 2008. LNCS, vol. 5350, pp. 90–107. Springer, Heidelberg (2008)
5. Shi, E., Stefanov, E., Papamanthou, C.: Practical dynamic proofs of retrievability. In: Sadeghi, A.R., Gligor, V.D., Yung, M. (eds.) ACM CCS 2013, pp. 325–336. ACM Press, Berlin (2013)
6. Cash, D., Küpçü, A., Wichs, D.: Dynamic proofs of retrievability via oblivious RAM. In: Johansson, T., Nguyen, P.Q. (eds.) EUROCRYPT 2013. LNCS, vol. 7881, pp. 279–295. Springer, Berlin (2013)

7. Wang, Q., Wang, C., Ren, K., Lou, W., Li, J.: Enabling public auditability and data dynamics for storage security in cloud computing. IEEE Trans. Parallel Distrib. Syst. **22**(5), 847–859 (2012)
8. Yang, K., Jia, X.: An efficient and secure dynamic auditing protocol for data storage in cloud computing. IEEE Trans. Parallel Distrib. Syst. **24**(9), 1717–1726 (2013)
9. Wang, B., Baochun, L., Hui, L.: Public auditing for shared data with efficient user revocation in the cloud. In: Proceedings of the 33th Conference on Information Communications (INFOCOM 2013), pp. 2750–2758. IEEE Press (2013)
10. Yuan, J., Yu, S.: Proofs of retrievability with public verifiability and constant communication cost in cloud. In: Proceedings of the 2013 International Workshop on Security in Cloud Computing, Cloud Computing, pp. 19–26 (2013)
11. Wang, C., Chow, S., Wang, Q., Ren, K., Lou, W.: Privacy-preserving public auditing for secure cloud storage. IEEE Trans. Comput. **62**(2), 362–375 (2013)
12. Yu, Y., Zhang, Y., Ni, J., Au, M., Chen, L., Liu, H.: Remote data possession checking with enhanced security for cloud storage. Future Gener. Comput. Syst. **52**, 77–85 (2014). doi:10.1016/j.future.2014.10.006
13. Yu, Y., Au, M.H., Ateniese, G., Huang, X., Susilo, W., Dai, Y., Min, G.: Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. IEEE Trans. Inf. Forensics Secur. **12**(4), 767–778 (2016). doi:10.1109/TIFS.2016.2615853
14. Zhu, Y., Hu, H., Ahn, G., Yu, M.: Cooperative provable data possession for integrity verification in multi cloud storage. IEEE Trans. Parallel Distrib. Syst. **23**(12), 2231–2244 (2012)
15. Halevi, S., Harnik, D., Pinkas, B., Shulman-Peleg, A.: Proofs of ownership in remote storage systems. In: Chen, Y., Danezis, G., Shmatikov, V. (eds.) ACM CCS 2011, pp. 491–500. ACM Press, Chicago (2011)
16. Zheng, Q., Xu, S.: Secure and efficient proof of storage with deduplication. Cryptology ePrint Archive, Report 2011/529 (2011). http://eprint.iacr.org/2011/529
17. Yuan, J., Yu, S.: Public integrity auditing for dynamic data sharing with multi-user modification. IEEE Trans. Inf. Forensics Secur. **10**(8), 1717–1726 (2015)
18. Yu, Y., Li, Y., Ni, J., Yang, G., Mu, Y., Susilo, W.: Comments on "public integrity auditing for dynamic data sharing with multi-user modification". IEEE Trans. Inf. Forensics Secur. **11**(3), 658–659 (2016)
19. Yuan, J., Yu, S.: PCPOR: public and constant-cost proofs of retrievability in cloud. J. Comput. Secur. **23**, 403–425 (2015)
20. Yuan, J., Yu, S.: Efficient public integrity checking for cloud data sharing with multi-user modification. In: Proceedings of the 33rd Conference on Information Communications (INFOCOM 2014), pp. 2121–2129. IEEE Press (2014)
21. Puzar, M., Plagemann, T.: Data sharing in mobile ad-hoc networks-a study of replication and performance in the MIDAS data space. Int. J. Space-Based Situated Comput. **1**(2/3), 137–150 (2015)
22. Petrlic, R., Sekula, S., Sorge, C.: A privacy-friendly architecture for future cloud computing. Int. J. Grid Util. Comput. **4**(4), 265–277 (2013)
23. Wang, Y., Du, J., Cheng, X., Liu, Z., Lin, K.: Degradation and encryption for outsourced PNG images in cloud storage. Int. J. Grid Util. Comput. **7**(1), 22–28 (2016)
24. Ye, X., Khoussainov, B.: Fine-grained access control for cloud computing. Int. J. Grid Util. Comput. **4**(2/3), 160–168 (2013)
25. Zhang, J., Li, P., Mao, J.: IPad: ID-based public auditing for the outsourced data in the standard model. Cluster Comput. **19**(1), 127–138 (2016). doi:10.1007/s10586-015-0511-3

LF-LDA: A Topic Model for Multi-label Classification

Yongjun Zhang[1,2(✉)], Jialin Ma[1,2], Zijian Wang[2], and Bolun Chen[1,2]

[1] Huaiyin Institute of Technology, Huaian, China
13511543380@139.com

[2] College of Computer and Information, Hohai University, Nanjing, China
zhjwang@hhu.edu.cn

Abstract. The textual data grows explosively with the advent of the era of big data, a significant portion of textual data is text documents labeled with multi-label such as the papers with keywords. Multi-label classification is a power technology to handle the multi-labeled textual data, but a huge room stays for improving the effect of multi-label classifying for textual data. This paper introduces *labeled LDA with function terms* (*LF-LDA*), a topic model that extracts noisy function terms from textual data to improve the performance of multi-label classification. The experimental result on *RCV1-v2* textual dataset shows that *LF-LDA* can outperform the other two state-of-art multi-label classifiers: *Tuned SVM* and *L-LDA* on both *Macro-F1* and *Micro-F1* metrics. The low variance also indicates *LF-LDA* is a robust classifier.

1 Introduction

With the development of Web technology, the data embedded in Web becomes massive and grows explosively. Among the massive data, text document is a major data organization structure, and it's important and valuable to analyze text documents to exploit the knowledges behind them. Taking academic papers as an example, to make papers easy to retrieve, the papers are expected to be organized properly, which means they should be categorized clearly and a paper may be associated with multiple classification labels.

Text classification can help to quickly search for the desired text data. However, the method of manual classification has been unable to adapt to the explosive growth of the amount of data, then automatic text categorization comes into play to address this problem. Traditionally, most of the textual data is associated with a single label. But nowadays, the multi-label textual data is more and more popular. Compared with the traditional single label text classification problem, the multilabel classification is more complicated and challenging, and is more sensitive to the label balance and the number of labels. At present, multi-label text classification is far from reaching the excellent classification performance, and a huge room stays for improvement.

Most of the multilabel classification algorithms make some extensions based on the single label classification algorithm, there are two extension ways [1]: *algorithm adaption* (*AA*) and *problem transformation* (*PT*). The extended algorithms based on *AA* improve the single label classification algorithm to make them can deal with multi-label

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_62

data, the representative algorithms include: *Multi-label decision tree* (*MLDT*) [2], *Multi-label k nearest neighbor* (*MLk NN*) [2, 3], improved *Adaboost algorithm* [4], improved algorithm based on *support vector machine* (*SVM*) and *neural network* [5–7]. On the other hand, the extended algorithms based on *PT* transform the multi-label classification task into several single label classification tasks. For example, *binary relevance* (*BR*) [8] trains a binary classifier for each label, and then classifies the test data with each binary classifier to determine final labels. *BR* algorithm is intuitive and easy to implement, but does not take the correlation between labels into account. Boutell et al. [9] then proposed *label powerset* (*LP*) algorithm to incorporate the correlation between labels. In *LP* algorithm, all the labels in the training text set are took as new labels to represent the correlations of labels. Tsoumakas et al. [10] goes further to propose the *random k-labelset* (*RAkLE*) algorithm to address the problem of excessive number of label sets in *LP* algorithm. In *RAkLE* algorithm, labels are randomly divided into groups, which are limited to the set of statistical labels, thus greatly reduced the number of labels. Some other famous extension algorithms of *PT* include: *classifier chains* (*CC*) [11], multi-label learning by exploiting label dependency (*LEAD*) [12] algorithm and conditional dependency network (*CDN*) [13].

Recently, supervised topic model has been developed as a new multi-label classification algorithm. Supervised topic model is a generation algorithm, it fits a joint probability distribution model for both text documents and labels from train documents, then the classifier is built in terms of estimated model parameters. The supervised topic models are based on the *Latent Dirichlet Allocation* (*LDA*) [14] model, they embed the label information into the standard LDA generation process to make they can identify the labels associated with text document. The representative supervised topic model includes: *supervised LDA* (*sLDA*) [15] model, *discriminative LDA* (*Disc LDA*) [16] model and *maximum entropy discrimination LDA* (*Med LDA*) model [17].

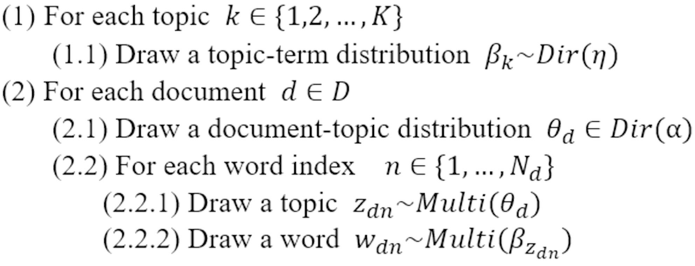
Labeled LDA (*L-LDA*) model [18] is an earlier supervised topic model for multi-label classification. Based on *LDA* model, *L-LDA* model introduces the label information by corresponding each label to a topic, then views each document as a probabilistic mixture of labels to generate its tokens. *L-LDA* does not change the structure of LDA model, it is easy to implement and can achieve good classification performance. However, in each text document there are some terms, such as ‘get’, ‘why’, ‘properly’, which are not clearly relevant to any label and make no contribution to classify, we call these terms *function terms*. In *L-LDA* model, function terms are considered to generate by labels, which runs count to the purpose of function terms and would decrease the classification effect.

In this paper, we develop a *labeled LDA with function terms* (*LF-LDA*) to separate function terms from label-associated terms. Compared to *L-LDA*, *LF-LDA* adds a function term topic component to model function terms, then documents are generated by either label topics or the shared function term topic, and each token in documents is associated with a hidden binary variable to identify it’s a topic term or a function term. Since the noise function term is filter out, the remaining label terms can help to identify the labels of document clearly. The multi-label classification experiment on the *RCV1-v2* dataset shows that our *LF-LDA* can improve the multi-label classification effect significantly.

2 LF-LDA

2.1 Review *LDA* and *L-LDA*

LDA is a powerful generative probabilistic model to model text corpora, it can be regarded as a mixture component model, in which each component is represented as a multinomial distribution over terms. The components, also called topics, are shared by all text documents. Each text document is associated with a multinomial distribution over topics, in this way *LDA* maps text documents to lower dimensional topic space. The probability generate process of *LDA* is shown in Fig. 1:



```
(1) For each topic $k \in \{1,2,\dots,K\}$
    (1.1) Draw a topic-term distribution $\beta_k \sim Dir(\eta)$
(2) For each document $d \in D$
    (2.1) Draw a document-topic distribution $\theta_d \in Dir(\alpha)$
    (2.2) For each word index $n \in \{1,\dots,N_d\}$
        (2.2.1) Draw a topic $z_{dn} \sim Multi(\theta_d)$
        (2.2.2) Draw a word $w_{dn} \sim Multi(\beta_{z_{dn}})$
```

Fig. 1. The probability generate process of LDA

Figure 2 illustrates the graph model representation of *LDA*:

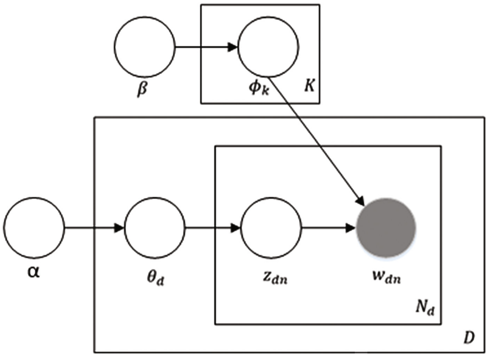


Fig. 2. The graph model representation of *LDA*

The notations listed in Figs. 1 and 2 are illustrated in Table 1.

Thomas L. Griffiths and Mark Steyvers [19] develops a *collapse Gibbs Sample* algorithm to inference the hidden parameters θ_d, z_{dn} and ϕ_k:

$$p(z_{dn} = k | \mathrm{z}_{-dn}, \boldsymbol{w}) \propto \frac{n_{-dn,k}^{w_{dn}} + \beta}{n_{-dn,k}^{(\cdot)} + v\beta} \frac{n_{-dn,k}^{(d)} + a}{n_{-dn,\cdot}^{(d)} + Ka} \tag{1}$$

Table 1. Notations of *LDA*

Notation	Meaning
K	The topic number
β	The parameter of *Dirichlet* prior distribution of topic-term distribution
ϕ_k	The topic-term distribution of topic k, where ϕ_{kv} is the probability of term v in topic k
D	The text document corpus
α	The parameter of *Dirichlet* prior distribution of topic-term distribution
θ_d	The document-topic distribution of document d, where θ_{dk} is the proportion of topic k in document d
N_d	The term count of document d
z_{dn}	The topic of the nth position in document d, it is drawn from the multinomial distribution $Multi(\theta_d)$ with a value range of from 1 to K
$\boldsymbol{w}_{dn}$	The nth term in document d

$$\hat{\phi}_{kv} = \frac{n_k^{(v)} + \beta}{n_k^{(\cdot)} + v\beta} \tag{2}$$

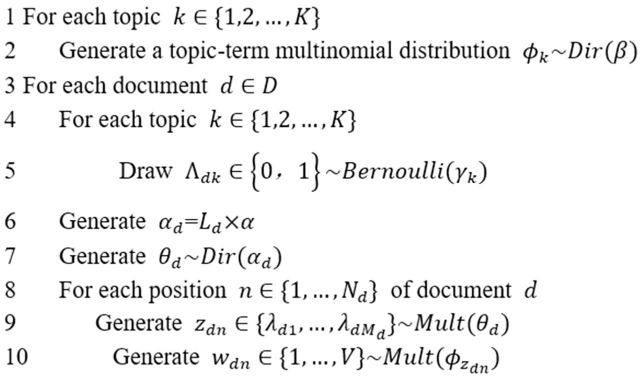
$$\hat{\theta}_{dk} = \frac{n_k^{(d)} + \alpha}{n_{\cdot}^{(d)} + k\alpha} \tag{3}$$

Table 2 lists the notations and their illustrations:

Table 2. Notations in formula (1), (2) and (3)

Notation	Illustration
$\mathbf{z}_{-dn}$	All the topic assignments for each term in the corpus, but not including the nth topic assignment in document d
$\boldsymbol{w}$	All the terms occurred in the corpus
$n_{-dn,k}^{w_{dn}}$	The count of term w_{dn} which has a topic assignment k in the corpus excluding the nth term of document d
$n_{-dn,k}^{(\cdot)}$	The count of all the terms having a topic assignment k in the corpus excluding the nth term of document d
$n_{-dn,k}^{(d)}$	The count of topic k in the document d excluding its nth position
$n_{-dn,\cdot}^{(d)}$	The term count of the document d excluding the nth position
$\hat{\phi}_k^{(v)}$	The estimated value of ϕ_{kv}, which is the probability of term v for given topic k
$n_k^{(v)}$	The count of term v which has a topic assignment k
$n_k^{(\cdot)}$	The count of topic k in the corpus
V	The size of term table
$\hat{\theta}_{dk}$	The estimated value of θ_{dk}, which is the proportion of topic k in document d
$n_k^{(d)}$	The count of topic k in document d
$n_{,}^{(d)}$	The term count of document d

LDA is an unsupervised learning model, thus it is unsuitable for multi-labeled corpora. Based on *LDA*, *L-LDA* defines a one-to-one correspondence between topics and labels to make it can apply to multi-label classification. Compared to *LDA*, *L-LDA* restricts θ_d to be defined only over the topics that correspond to the labels Λ_d of the document d to ensures that all the topic assignments are restrained to the document's labels. The generate process of L-LDA is illustrated in Fig. 3:



```
For each topic $k \in \{1,2,\ldots,K\}$
    Generate a topic-term multinomial distribution $\phi_k \sim Dir(\beta)$
For each document $d \in D$
    For each topic $k \in \{1,2,\ldots,K\}$
        Draw $\Lambda_{dk} \in \{0,\ 1\} \sim Bernoulli(\gamma_k)$
    Generate $\alpha_d = L_d \times \alpha$
    Generate $\theta_d \sim Dir(\alpha_d)$
    For each position $n \in \{1,\ldots,N_d\}$ of document $d$
        Generate $z_{dn} \in \{\lambda_{d1},\ldots,\lambda_{dM_d}\} \sim Mult(\theta_d)$
        Generate $w_{dn} \in \{1,\ldots,V\} \sim Mult(\phi_{z_{dn}})$
```

Fig. 3. The probability generate process of *L-LDA*

To explain the process more clearly, we assume a document *d* has labels {2,4}, then the Λ_d will be a binary vector $\Lambda_d = (0,1,0,1,0,0,0,0)^T$ (assuming the number of labels is 8, i.e. $K = 8$), in which only the *2*th component and the *4*th component is 1 to indicate the document *d* has labels {2, 4}. The projection matrix L_d extracts the components of the prior parameter vector α that corresponds to the labels of document *d* to α_d, it is formulated as:

$$L_{d,ij} = \begin{cases} 1 & \text{if } \Lambda_{dj} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

For the case of document *d* having labels {2,4}, L_d is a 2 × 8 matrix, in which only the elements $L_{d,12}$ and $L_{d,24}$ equal 1 and other elements equal 0. The projection matrix L_d makes the prior parameter α_d of θ_d only retain the components of α corresponding to the labels of *d*, i.e. $\alpha_d = (\alpha_2, \alpha_4)$. In this way, *L-LDA* restrains the topics of each document corresponding to its labels. The graph model representation of *L-LDA* is illustrated in Fig. 4:

Like *LDA*, *L-LDA* also uses the formula (1), (2) and (3) to infer parameters, but the topic z_{dn} restrains to the labels of the document *d*. To apply *L-LDA* to the multi-label document classification problem, a training set consisting of documents with multiple labels is used to fit a *L-LDA*, then a standard *LDA* with topic number *k* is applied to the test set, in which topic-term distributions $\phi_1, \ldots, \phi_k$ are fixed with the results inferred by *L-LDA*. The document's most likely labels can then be inferred by suitable thresholding of its posterior probability over topics θ_{dk}.

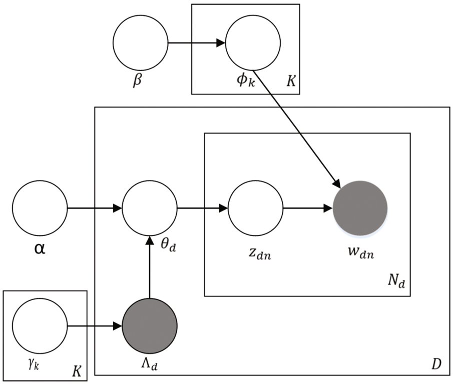
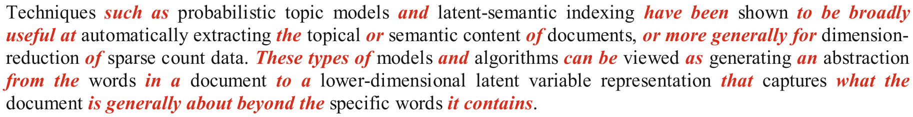


Fig. 4. The graph model representation of *L-LDA*

2.2 Our Proposed *LF-LDA*

The *LF-LDA* we proposed uses a specific component, which is a multinomial distribution over terms and shared by all the documents in the corpus, to handle the common function terms, Fig. 5 gives a live example of a document with function terms:



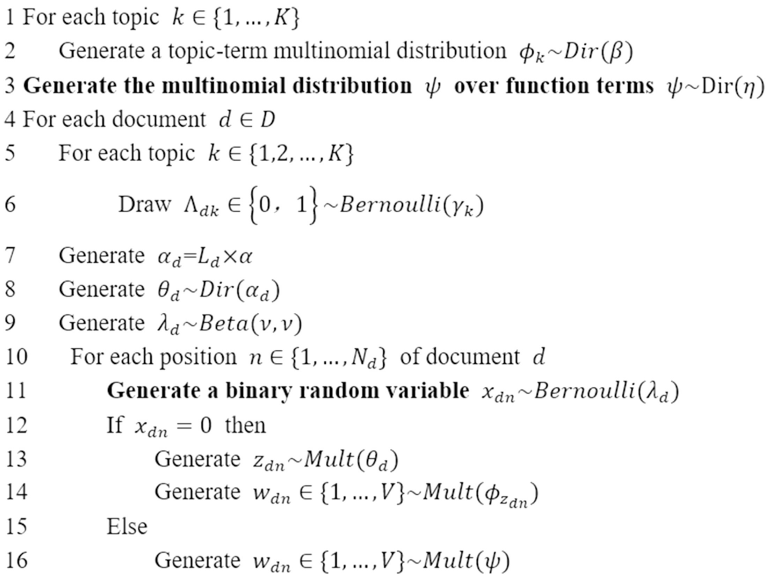
Techniques ***such as*** probabilistic topic models ***and*** latent-semantic indexing ***have been*** shown ***to be broadly useful at*** automatically extracting ***the*** topical ***or*** semantic content ***of*** documents, ***or more generally for*** dimension-reduction ***of*** sparse count data. ***These types of*** models ***and*** algorithms ***can be*** viewed ***as*** generating ***an*** abstraction ***from the*** words ***in a*** document ***to a*** lower-dimensional latent variable representation ***that*** captures ***what the*** document ***is generally about beyond the*** specific words ***it contains***.

Fig. 5. An example of document with function terms, the function terms are marked with bold italic font and red color

It can be seen clearly that the function terms are weakly relevant to the labels of documents, therefore they don't contribute to but disturb the recognition of labels. *LF-LDA* aims to separate the noisy function terms from label-terms, which strongly indicate the label of the document containing it.

The probability generate process of *LF-LDA* is illustrated in Fig. 6:

In step 3, the multinomial distribution ψ corresponding to the shared component is generated to model the function terms. A binary random variable x_{dn} is generated in step 11 to identify whether the current token is a function term $(x_{dn} = 1)$ or a topic term $(x_{dn} = 0)$. The graph model representation of *LF-LDA* is illustrated in Fig. 7:



```
For each topic $k \in \{1, \ldots, K\}$
    Generate a topic-term multinomial distribution $\phi_k \sim Dir(\beta)$
Generate the multinomial distribution $\psi$ over function terms $\psi \sim \mathrm{Dir}(\eta)$
For each document $d \in D$
    For each topic $k \in \{1,2, \ldots, K\}$
        Draw $\Lambda_{dk} \in \{0, 1\} \sim Bernoulli(\gamma_k)$
    Generate $\alpha_d = L_d \times \alpha$
    Generate $\theta_d \sim Dir(\alpha_d)$
    Generate $\lambda_d \sim Beta(\nu, \nu)$
    For each position $n \in \{1, \ldots, N_d\}$ of document $d$
        Generate a binary random variable $x_{dn} \sim Bernoulli(\lambda_d)$
        If $x_{dn} = 0$ then
            Generate $z_{dn} \sim Mult(\theta_d)$
            Generate $w_{dn} \in \{1, \ldots, V\} \sim Mult(\phi_{z_{dn}})$
        Else
            Generate $w_{dn} \in \{1, \ldots, V\} \sim Mult(\psi)$
```

Fig. 6. The probability generate process of *LF-LDA*

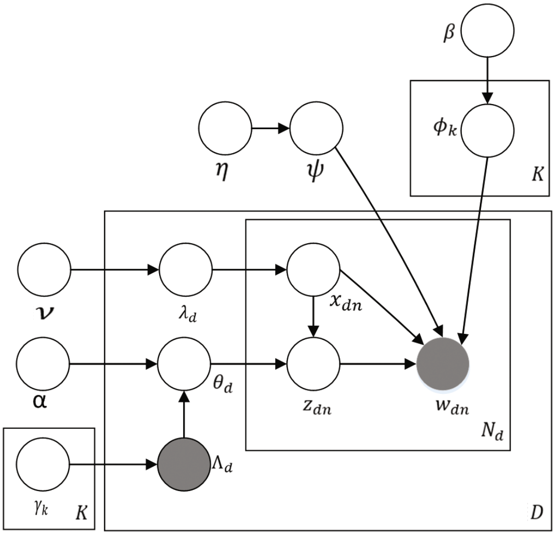


Fig. 7. The graph model representation of *LF-LDA*

3 Inference and Parameter Estimation

We exploit the *collapse Gibbs Sample* algorithm to infer the hidden variables x_{dn} and z_{dn}, the other hidden variables, λ_d, θ_d and ϕ_k, are integrated out, the sample formulas are:

$$\begin{aligned} p(\mathrm{x}_{dn} = 1 | \boldsymbol{x}_{-dn}, \boldsymbol{w}, \mathbf{z}_{-dn}) &\propto p(\mathrm{x}_{dn} = 1, w_{dn} = t | \boldsymbol{x}_{-dn}, \boldsymbol{w}_{-\boldsymbol{dn}}, \mathbf{z}_{-\mathrm{dn}}) \\ &= \frac{n^{(d)}_{-dn,x=1} + \upsilon}{n^{(d)}_{-dn,x=\cdot} + 2\upsilon} \frac{n^{(x=1)}_{-dn,t} + \eta}{n^{(x=1)}_{-dn,\cdot} + V\eta} \end{aligned} \tag{5}$$

$$\begin{aligned} p(\mathrm{x}_{dn} = 0, \mathrm{z}_{dn} = k | \boldsymbol{x}_{-dn}, \boldsymbol{w}, \mathbf{z}_{-dn}) &\propto p(\mathrm{x}_{dn} = 0, \mathrm{z}_{dn} = k, w_{dn} = t | \boldsymbol{x}_{-dn}, \boldsymbol{w}_{-\boldsymbol{dn}}, \mathbf{z}_{-\mathrm{dn}}) \\ &= \frac{n^{(d)}_{-dn,x=1} + v}{n^{(d)}_{-dn,x=\cdot} + 2v} \frac{n^{(k)}_{-dn,t} + \beta}{n^{(k)}_{-dn,\cdot} + v\beta} \frac{n^{(k)}_{-dn,k} + \alpha}{n^{(d)}_{-dn,\cdot} + M_d\alpha} \end{aligned} \tag{6}$$

Where $n^{(d)}_{-dn,x=1}$ is the number of $x = 1$ in the document d excluding the current position, $n^{(x=1)}_{-dn,t}$ is the number of the term t with $x = 1$ assignment to indicate that it's a function term in the corpus excluding the position n of document d, M_d is the number of labels of document d. The hyper parameters, i.e. $\beta, v\,\text{and}\,\eta$, are set a symmetric value.

The key prerequisite knowledge to understand the formulas (5) and (6) is the following properties of the conjugate distribution *Beta- Bernoulli* and *Dirichlet-Multinomial*:

Property 1. For a *Bernoulli* distribution $p(x|\mu)$ with a conjugate prior *Beta* distribution $Beta(\mu|\alpha)$, where $\alpha = [\alpha_1, \alpha_2]^T$, the posteriori probability of a new observation is:

$$\begin{aligned} p(x = 1|D, \alpha_1, \alpha_2) &= \int p(x = 1, \mu|D, \alpha_1, \alpha_2)d\mu \\ &= \int p(x = 1|\mu)\, p\,(\mu|D, \alpha_1, \alpha_2)\mathrm{d}\mu \\ &= \frac{n_1^{(D)} + \alpha_1}{N + \alpha_1 + \alpha_2} \end{aligned} \tag{7}$$

Property 2. For a *Multinomial* distribution $p(x|\theta)$ with a conjugate prior *Dirichlet* distribution $Dir(\theta|\alpha)$, where $\alpha = [\alpha_1, \ldots, \alpha_k]^T$ and $x \in \{1, .., K\}$, the posteriori probability of a new observation is:

$$\begin{aligned} p(x = k|D, a) &= \int p(x = k, \theta|D, \alpha_1, \alpha_2)d\theta \\ &= \int p(x = k|\theta)p(\theta|D, \alpha)\mathrm{d}\theta \\ &= \frac{n_k^{(D)} + \alpha_k}{N + \sum_{i=1}^{K} \alpha_i} \end{aligned} \tag{8}$$

where $D = \{x_1, \ldots, x_N\}$ is the observed data with $x_n \in \{1, \ldots, K\}$, $n_k^{(D)}$ is the number of $x_n = k$.

4 Experiments

We performed the multilabel classification experiment on the *RCV1-v2* corpus, which contains 800,000 text documents and 103 labels, i.e. $k = 103$ We used a subset of *RCV1-v2* named *LYRL2004 training set* as training set. To evaluate the classification effect, 5 data set were chosen randomly from another subset of *RCV1-v2* named *LYRL2004 test set*, each of which contains 10,000 text documents, the test results are averaged on all the 5 test data sets. We remove the stop words and the terms occurred less than 8 times, the hyper parameters are set as following: $\alpha = 50/K$, $\beta = v = \eta = 0.1$ A 500 iteration times is used to burn in the *Gibbs Sample* algorithm, then another 1000 iteration times are used to sample x_{dn} and z_{dn} to make the posteriori distribution $p(\boldsymbol{x}, \mathbf{z} \geq |\boldsymbol{w})$ converging to its actual distribution.

We compared our *LF-LDA* with another two state-of-art multi-label classifiers: *L-LDA* and *Tuned-SVM* (*T-SVM*) on both *Macro-F1* and *Micro-F1* metrics, the experiment result is shown in Table 3:

Table 3. The classification effect of *LF-LDA*, *L-LDA* and *T-SVM*

Classifier	*Macro-F1*	*Micro-F1*
LF-LDA	0.827 ± 0.006	0.803 ± 0.007
L-LDA	0.542 ± 0.017	0.538 ± 0.023
T-SVM	0.794 ± 0.013	0.789 ± 0.012

For the multi-label classification, the *F1* metrics are defined as:

$$Recall(d) = \frac{|L_d \cap L'_d|}{|L_d|} \tag{9}$$

$$Precision(d) = \frac{|L_d \cap L'_d|}{|L'_d|} \tag{10}$$

$$F_1(d) = \frac{2 \times Recall(d) \times Precision(d)}{Recall(d) \times Precision(d)} \tag{11}$$

where L_d is the output labels of document d, L'_d is the true labels of document d. *Macro-F1* is the average of the *F1* of all test documents, while *Micro-F1* is the *F1* of the whole test set.

According to Table 3, *LF-LDA* outperforms the other two classifiers on both *Macro-F1* and *Micro-F1*. *LF-LDA* and *T-SVM* have close classification performance, while the result of *L-LDA* is far from them. It might be due to the noise affect raised by function terms which are unsuitably associated with the labels of documents by *L-LDA*.

LF-LDA, on the contrary, uses a binary random variable x_{dn} to indicate whether the current token is a noisy function term or not, which makes it can recognize the labels of documents more precisely. It's noteworthy that the lowest variance of *LF-LDA* among the three classifiers on both *Macro-F1* and *Micro-F1* suggests it's a robust classifier.

5 Conclusion

Labeled-LDA associated each label with a topic to make it can apply to multi-label classification. The labels of each document are identified by the posterior distribution over topics, and each topic corresponding to a label is a component represented by a multinomial distribution over terms. However, there are many terms called function terms which are weakly or not relevant to any label, it's valuable to separate these terms from the other label-related terms. In essence, function terms are used for syntactic representations or facilitating the writing, so they are filled in each document and shared by all documents. We proposed the *LF-LDA*, which uses a shared component by all document to pick out these function terms from each document, to improve effect of the multi-label classification. The experimental result shows it is an efficient and robust multi-label classification model. Some further improvement to LF-LDA include: represent each topic as a multinomial distribution over both words and phrases rather than only words, take the correlation between topics into account.

Acknowledgments. This work was supported by the Chinese National Natural Science Foundation (Grant No.: 61602202), the Natural Science Foundation of Jiangsu Province, China (Grant No.: BK20160428), the Social Key Research and Development Project of Huaian, Jiangsu, China (Grant No.: HAS2015020) and the Graduate Student Scientific Research and Innovation Project of Jiangsu Province, China (Grant No.: 2015B38314).

References

1. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook, pp. 667–685 (2009)
2. Zhang, M., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn. **40**(7), 2038–2048 (2007)
3. Brinker, K., Hüllermeier, E.: Case-based multilabel ranking. In: IJCAI, pp. 702–707 (2007)
4. Schapire, R.E., Singer, Y.: BoosTexter: a boosting-based system for text categorization. Mach. Learn. **39**(2–3), 135–168 (2000)
5. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: NIPS, vol. 14 (2001)
6. Zhang, M.-L., Zhou, Z.-H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans. Knowl. Data Eng. **18**(10), 1338–1351 (2006)
7. Zhang, M.-L.: ML-RBF: RBF neural networks for multi-label learning. Neural Process. Lett. **29**(2), 61–74 (2009)
8. Tsoumakas, G., Katakis, I.: Multi-label classification: an overview. Int. J. Data Warehouse. Min. **3**(3), 1–13 (2006)

9. Boutell, M.R., et al.: Learning multi-label scene classification. Pattern Recogn. **37**(9), 1757–1771 (2004)
10. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. IEEE Trans. Knowl. Data Eng. **23**(7), 1079–1089 (2011)
11. Read, J., et al.: Classifier chains for multi-label classification. Mach. Learn. **85**(3), 333 (2011)
12. Zhang, M.-L., Zhang, K.: Multi-label learning by exploiting label dependency. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2010)
13. Guo, Y., Gu, S.: Multi-label classification using conditional dependency networks. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, no. 1 (2011)
14. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
15. Mcauliffe, J.D., Blei, D.M.: Supervised topic models. In: Advances in Neural Information Processing Systems (2008)
16. Lacoste-Julien, S., Sha, F., Jordan, M.I.: DiscLDA: discriminative learning for dimensionality reduction and classification. In: Advances in Neural Information Processing Systems (2009)
17. Zhu, J., Ahmed, A., Xing, E.P.: MedLDA: maximum margin supervised topic models for regression and classification. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM (2009)
18. Ramage, D., et al.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1. Association for Computational Linguistics (2009)
19. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl 1), 5228–5235 (2004)

Data Analysis for Infant Formula Nutrients

Qian Huang[1](✉), Chao Zhang[1], Feng Ye[1], Qi Wang[1], and Sisi Chen[2]

[1] College of Computer and Information,
Hohai University, Nanjing 211100, China
huangqian@hhu.edu.cn

[2] Nanjing Huiying Electronics Technology Corporation, Nanjing 211103, China

Abstract. With the development of the social economy and the improvement of the people's living standard, more and more categories of infant formulas are presented according to nutritional requirements and regional differences. For a specific family, nowadays it is usually quite difficult to make a quick decision. This manuscript firstly analyzes some infant formulas made in Canada, The Netherlands, Denmark, Ireland and Germany, and then outlines the special nutrients of each given kind of infant formula. Based on these observations, dataset construction and classification are discussed so that relational decisions can be made according to specific needs.

1 Introduction

In contemporary society, young parents in Asian countries are apt to feed their infants with products from Europe or America due to some nutritional, environmental and economic reasons. Accordingly, European and American companies propose many infant formulas that are suitable for Asian babies. And Asian companies also import some of the popular products to make it easier for purchasing by domestic families. However, since there are so many brands and versions, it is usually difficult for new parents to make a quick and rational decision on which one to buy.

Taking infant milk powder as an example, this manuscript firstly analyzes several famous brands from Europe and America, i.e. MeadJohnson [1], Friso [2], Abbott [3], Wyeth [4] and Aptawelt [5]; and then constructs a new dataset. After that, classification [6–9] and recommendation [10–12] can be further performed to help young Asian parents. In the rest of this paper, Sect. 2 analyzes nutrients data, Sect. 3 discusses experiments with classification and Sect. 4 draws the conclusion.

2 Nutrients Data Analysis

A preliminary survey on the infant milk powder market indicates that infant milk powder consists of *Energy*, *Proteins*, *Fats*, *Carbohydrates*, *Vitamins*, *Minerals* and so on. For every 100 g of some infant milk powder products, the approximate amounts of nutrients are listed in Table 1, where the first six columns are measured in kilojoules, grams, grams, grams, milligrams and grams, respectively. Note that *Stage* indicates the infant age measured in months. In general, stage 1 corresponds to infants from 0 to 12

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_63

Table 1. Approximate amounts of nutrients in every 100 g of some infant milk powder products

Energy	Proteins	Fats	Carbohydrates	Vitamins	Minerals	Stage	Product
2130	10.1	27	57	231.72	4.17	1	MeadJohnson A+ Canada
2090	12.9	26	54	228.40	7.90	2	
2100	10.5	26	58	276.65	1.65	1	MeadJohnson A+ the Netherlands
2000	15.5	21	57	258.81	2.13	2	
2122	10.6	27	57	269.76	1.72	1	MeadJohnson Enfinitas the Netherlands
1897	14.9	18.5	58	243.32	2.15	2	
2130	11	27	54.6	321.99	1.84	1	Frisolac Prestige the Netherlands
2000	15.5	21.5	54	328.28	2.41	2	
2119	11.29	27.4	52.2	244.54	1.70	1	Abbott Eleva Denmark
1983	16.1	20.9	53.6	90.34	2.38	2	
2249	11	29	57	174.30	1.60	1	Wyeth Illuma Ireland
2005	15.25	21.28	55.32	261.34	2.39	2	
2064	10.4	26.5	50.4	227.16	1.84	1	Aptamil Pronutra+ Germany
1950	15.3	20.8	51.4	214.00	2.53	2	

months, sometimes from 0 to 6 months; and stage 2 corresponds to infants from 6 to 12 months, sometimes from 6 to 18 months.

The first generation infant milk powder focused on energy by adding grains, soybean milk and sucrose. The second generation infant milk powder added proteins, plant oils, maltose, vitamins and minerals to simulate breast milk. In the third generation, Taurine, Docosahexaenoic Acid (DHA) and Arachidonic Acid (ARA) are used and the proportions of ingredients are optimized to make it closer to breast milk. Currently we are in the fourth generation, which specifies the proportions of amino acids and proposes the usage of structured fat and restrictions on flavors and sucrose. In Table 2, the “Fats” column in Table 1 is expanded to explicitly illustrate DHA (in milligrams), ARA (in milligrams), Structured Fats (in grams) and Other Fats (in grams).

2.1 Observations

It can be observed from Tables 1 and 2 that:

(1) Proteins, Fats and Carbohydrates constitute the major part of infant milk powder, from 83.7% to 97%.

(2) For the same brand, the stage 1 product always has more *Energy* and *Fats*, whereas the stage 2 product always has more *Proteins* and *Minerals*.

(3) ARA is generally larger than or equal to DHA, except for the stage 2 product of Frisolac Prestige the Netherlands.

Table 2. Expansion of Fats for infant milk powder

DHA	ARA	Structured Fats	Other Fats	Stage	Product
86	172	0	26.74	1	MeadJohnson A+
84	169	0	25.75	2	Canada
85	170	0	25.75	1	MeadJohnson A+ the Netherlands
80	160	0	20.76	2	
90	180	0	26.73	1	MeadJohnson Enfinitas the Netherlands
78	156	0	18.27	2	
100	120	0	26.78	1	Frisolac Prestige the Netherlands
100	21	0	21.38	2	
79	105	0	27.22	1	Abbott Eleva Denmark
60	79	0	20.76	2	
57	98	3.2	25.65	1	Wyeth Illuma Ireland
80.9	80.9	4	17.12	2	
83	83	0	26.33	1	Aptamil Pronutra+ Germany
65	65	0	20.67	2	

(4) Let $ADRatio$ = ARA/DHA. Then for MeadJohnson, $ADRatio \geq 2$; for Friso (Frisolac Prestige), $0.21 \leq ADRatio \leq 1.2$; for Abbott, $1.31 < ADRatio \leq 1.33$; for Wyeth, $1 \leq ADRatio < 1.72$; for Aptawelt (Aptamil Pronutra+), $ADRatio = 1$.

(5) Only Wyeth has structured fats.

Table 3. Dataset for infant milk powder nutrients

Energy	Proteins	ADRatio	OPO	Vitamins	Minerals	Stage	Brand
2130	10.1	2.00	0	231.72	4.17	1	MeadJohnson
2090	12.9	2.01	0	228.40	7.90	2	MeadJohnson
2100	10.5	2.00	0	276.65	1.65	1	MeadJohnson
2000	15.5	2.00	0	258.81	2.13	2	MeadJohnson
2122	10.6	2.00	0	269.76	1.72	1	MeadJohnson
1897	14.9	2.00	0	243.32	2.15	2	MeadJohnson
2130	11	1.20	0	321.99	1.84	1	Friso
2000	15.5	0.21	0	328.28	2.41	2	Friso
2119	11.29	1.33	0	244.54	1.70	1	Abbott
1983	16.1	1.32	0	90.34	2.38	2	Abbott
2249	11	1.72	1	174.30	1.60	1	Wyeth
2005	15.25	1.00	1	261.34	2.39	2	Wyeth
2064	10.4	1.00	0	227.16	1.84	1	Aptawelt
1950	15.3	1.00	0	214.00	2.53	2	Aptawelt

It is easy to distinguish Wyeth by (5). After that, MeadJohnson and Abbott can be correctly classified according to (4). As for Friso and Aptawelt, the amount of Vitamins can be analyzed to make a quick decision.

2.2 Design of Dataset

Based on the above observations, a dataset can be designed for infant milk powder nutrients, as shown in Table 3, where OPO, which is a kind of structured fat, is used to represent whether *Structured Fats* exist in Table 2.

3 Experimental Results

In this section, the R language [13] widely used in machine learning [14–16] is employed for data analysis. The following code gives an example for classification using a decision tree, which is constructed based on the data in Table 3.

R code for infant milk powder nutrients data analysis

```
library('rpart')
partykit.installed <- 'partykit' %in% rownames(installed.packages())
if (partykit.installed) {
  print("the partykit package is already installed, let's load it...")
}else {
  print("let's install the partykit package first...")
  install.packages('partykit', dependencies=T)
}
library('partykit')

f_in <- 'Data/infant.csv'  #infant.csv is the same as Table 3
D <- read.csv(f_in)
M_rpart <- rpart(Brand~., data = D, method = 'class',control
           = rpart.control(minsplit = 1, minbucket =  1, maxdepth = 4))
plot(as.party(M_rpart), main='decision tree using advanced
           parameter setting')
D_test <- D[1:10,]
y_test_prob <- predict(M_rpart, D_test)
print(y_test_prob)
y_test_label <- predict(M_rpart, D_test, type='class')
print(y_test_label)
accuracy_test <- sum(D_test$Brand==y_test_label) / length(y_test_label)
print(paste0('accuracy on the test data set is ', accuracy_test))
```

The three parameters in function rpart.control are explained below:

(1) minsplit: a node will be considered to be split unless its related training samples are more than minsplit.
(2) minbucket: the minimum number of training samples corresponding to leaf nodes.
(3) maxdepth: the maximum depth of the decision tree.

In our example, the dataset is quite small. Therefore, both minsplit and minbuckt are set to 1. The resulting decision tree using advanced parameter setting is depicted in

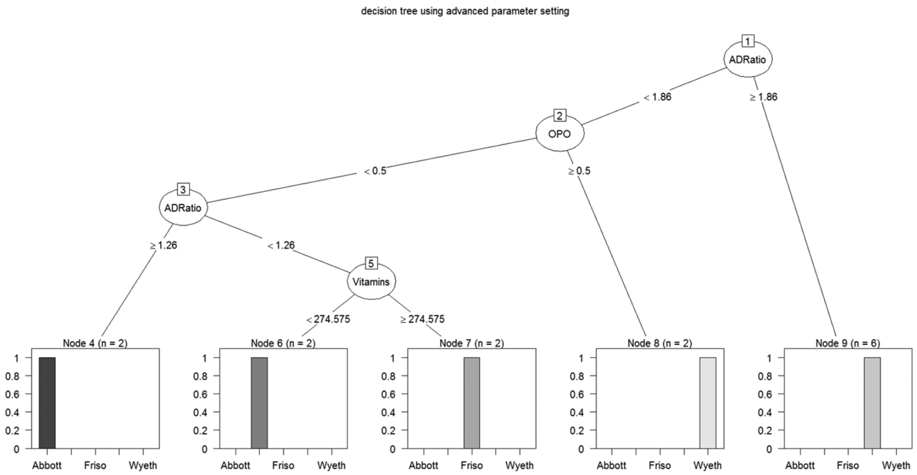


Fig. 1. Decision tree using advanced parameter setting (minsplit = minbucket = 1).

Fig. 1, where the *Brand* names are not fully displayed due to the limit of canvas. From left to right within a node, the five brands are Abbott, Aptawelt, Friso, MeadJohnson and Wyeth, respectively.

It can be seen that ADRatio is firstly used to identify MeadJohnson. For the other four brands, OPO is utilized to identify Wyeth. To further distinguish Friso, Abbott and Aptawelt, a combination of ADRatio and Vitamins is employed. Compared with the observations in Sect. 2.1, we can see that the construction of a decision tree is helpful to identify important features of a given dataset and recommend adequate items for us based on specific requirements.

Here, the result of print(y_test_label) is:

```
   Abbott  Aptawelt  Friso  MeadJohnson  Wyeth
1       0         0      0            1      0
2       0         0      0            1      0
3       0         0      0            1      0
4       0         0      0            1      0
5       0         0      0            1      0
6       0         0      0            1      0
7       0         0      1            0      0
8       0         0      1            0      0
9       1         0      0            0      0
10      1         0      0            0      0
```

The first ten rows are used as a test set, and the accuracy is:

```
[1] "accuracy on the test data set is 1"
```

This means that the above classification distinguishes all brands correctly. The summary of the trained model is as follows.

```
> print(M_rpart)
n= 14

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 14 8 MeadJohnson (0.14 0.14 0.14 0.43 0.14)
   2) ADRatio< 1.86 8 6 Abbott (0.25 0.25 0.25 0 0.25)
     4) OPO< 0.5 6 4 Abbott (0.33 0.33 0.33 0 0)
       8) ADRatio>=1.26 2 0 Abbott (1 0 0 0 0) *
       9) ADRatio< 1.26 4 2 Aptawelt (0 0.5 0.5 0 0)
        18) Vitamins< 274.575 2 0 Aptawelt (0 1 0 0 0) *
        19) Vitamins>=274.575 2 0 Friso (0 0 1 0 0) *
     5) OPO>=0.5 2 0 Wyeth (0 0 0 0 1) *
   3) ADRatio>=1.86 6 0 MeadJohnson (0 0 0 1 0) *
```

Note that if the calling to rpart.control is changed as

```
M_rpart <-rpart(Brand~., data = D, method = 'class',control
    = rpart.control(minsplit = 10, minbucket = 5, maxdepth = 4))
```

Then the resulting decision tree using advanced parameter setting is depicted in Fig. 2, from which we can see that this classification distinguishes MeadJohnson only.

Accordingly, the result of print(y_test_label) is:

```
   Abbott  Aptawelt  Friso  MeadJohnson  Wyeth
1    0.00      0.00   0.00            1   0.00
2    0.00      0.00   0.00            1   0.00
3    0.00      0.00   0.00            1   0.00
4    0.00      0.00   0.00            1   0.00
5    0.00      0.00   0.00            1   0.00
6    0.00      0.00   0.00            1   0.00
7    0.25      0.25   0.25            0   0.25
8    0.25      0.25   0.25            0   0.25
9    0.25      0.25   0.25            0   0.25
10   0.25      0.25   0.25            0   0.25
```

If we still use the first ten rows as a test set, then the accuracy is:

```
[1] "accuracy on the test data set is 0.8"
```

Note that the accuracy will be 1 if we use:

D_test <- D[1 : 6,]

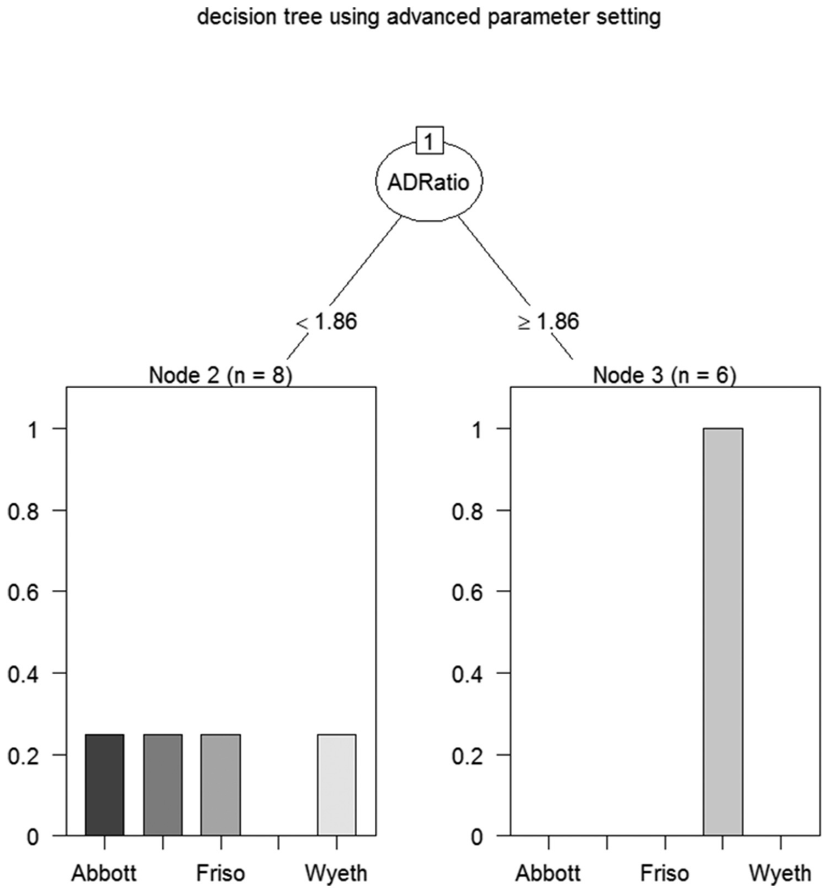


Fig. 2. Decision tree using advanced parameter setting (minsplit = 10, minbucket = 5).

4 Conclusion

This manuscript proposes a data analysis strategy for infant milk powder nutrients. Further classifications and recommendations can be performed as illustrated by the experiments.

It can be seen that if the dataset is large enough, outliers can also be detected by classification since we can specify the data for training and testing. Further recommendation can also be made based on the designed dataset and constructed decision tree.

Acknowledgments. This work is partly supported by the National Natural Science Foundation of China under Grant No. 61300122, 61502145 and 61602150, and the Fundamental Research Funds of China for the Central Universities under Grant No. 2013B01814.

References

1. MeadJohnson. www.meadjohnson.com
2. Friso. www.friso.com
3. Abbott. www.abbott.com
4. Wyeth. www.wyeth.com
5. Aptawelt. www.aptawelt.de
6. Mahesha, P., Vinod, D.: Support vector machine-based stuttering dysfluency classification using GMM supervectors. Int. J. Grid Util. Comput. **6**, 143–149 (2015)
7. Rodas, O., To, M.A.: A study on network security monitoring for the hybrid classification-based intrusion prevention systems. Int. J. Space-Based Situated Comput. **5**, 115–125 (2015)
8. Al-Kabi, M., Wahsheh, H., Alsmadi, I.M.: Polarity classification of Arabic sentiments. Int. J. Inf. Technol. Web. Eng. **11**, 32–49 (2016)
9. Wu, K., Kang, J., Chi, K.: Research on fault diagnosis method using improved multi-class classification algorithm and relevance vector machine. Int. J. Inf. Technol. Web. Eng. **10**, 1–16 (2015)
10. Zhao, Z., Lu, H., Cai, D., He, X., Zhuang, Y.: User preference learning for online social recommendation. IEEE Trans. Knowl. Data Eng. **28**, 2522–2534 (2016)
11. Shigeyasu, T., Nagamine, J.: DCR-MAC: a data channel recommending MAC for multi-channel WLANs toward the higher throughput performance. Int. J. Space-Based Situated Comput. **5**, 168–177 (2015)
12. Guo, L., Jin, B., Yu, R., Yao, C., Sun, C., Huang, D.: Multi-label classification methods for green computing and application for mobile medical recommendations. IEEE Access **4**, 3201–3209 (2016)
13. Ihaka, R., Gentleman, R.: R: a language for data analysis and graphics. J. Comput. Graphical Stat. **5**, 299–314 (1996)
14. Wu, Z., Lin, T., Tang, N.: Explore the use of handwriting information and machine learning techniques in evaluating mental workload. Int. J. Technol. Hum. Interact. **12**, 18–32 (2016)
15. Lin, H., Su, S., Wang, S., Tsai, S.: Influence of cognitive style and cooperative learning on application of augmented reality to natural science learning. Int. J. Technol. Hum. Interact. **11**, 41–66 (2015)
16. Gutierrez-Carreon, G., Daradoumis, T., Jorba, J.: Automatic composition of learning grid portlets: a comparison of syntactic and semantic approaches. Int. J. Grid Util. Comput. **1**, 308–315 (2009)

A Classification Method Based on Improved BIA Model for Operation and Maintenance of Information System in Large Electric Power Enterprise

Chong Wang[1], Qi Wang[2], Qian Huang[2](✉), and Feng Ye[2]

[1] Jiangsu Electric Power Information and Telecommunication Company,
Nanjing, China
[2] College of Computer and Information, Hohai University, Nanjing, China
huangqian@hhu.edu.cn

Abstract. As the integration of informatization and industrialization goes deeper in State Grid Jiangsu Electric Power Company, the lean management of O&M (operation and maintenance) of information system plays a more important role in the company's production and management. On the ground of a full investigation of the current information system of the company, this paper has improved the model of business impact analysis (BIA) and, based on which, proposed a new method to classify O&M of information system. As proved in our practices in the company, the proposed model and method are efficient in controlling the cost of optimizing the operation of the information system, raising the efficiency of resource utilization as well as in improving the O&M management.

1 Introduction

As the informatization strategy has been gradually implemented in State Grid Jiangsu Electric Power Company (hereafter referred to as "the company"), the information system has now covered all businesses of the company, where information becomes more integrated with businesses and the information is increasingly prominent as a supporter. A safe and stale operation of the information system is the foundation for the company to carry out its businesses. Now, the company has higher requirements for quick and steady information service, attributing to the application of new IT technologies such as cloud computing [1–3], internet of things [4–6] and big data [7], etc. In order to further improve the management of the information system, so as to guarantee a safe, stable and efficient operation, the company is now carrying out the research topic on classification service of the information system O&M.

2 Status Analysis

In China, classification [8] is a common method to manage information system. Under the precondition of guarantee both the reliability and the usability, this method can maximize the utilization of service resources. It is known that using classification to

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_64

optimize the allocation of the operational resources of information system is the most efficient management method. In the late 1990s, the classification management of information system O&M, as a part of ITIL, was introduced to Chinese firms, and, since then, it becomes more evolved. By far, similar schemes are used in famous Chines state-owned enterprises, including Bank of Communications and China Mobile, to manage information system O&M.

At present, all information systems of the company are operated and maintained in accordance with the highest standard, which, however, results in a low rate of resource utilization as more resources are invested for guarantee the information system O&M. This is because the only standard is used both for the core businesses, such as marketing a production that have higher requirements on O&M services, and for the non-control businesses and management information system that have lower requirements on O&M services. The O&M resources allocated to the core businesses are insufficient on the one hand, while some of the O&M resources allocated to the non-core businesses are wasted on the other hand. It is an urgent challenge to comprehensively mange all O&M resources and support multiple information systems, with giving priority to satisfy the O&M requirements during special time domains as well as during rush hours.

3 The Improved Information System Classification Model

3.1 Objective and Principle of the Model

Business impact analysis, abbreviated as BIA, is a method for evaluating the effect of service interruption on an organization's businesses. It has been widely used for O&M classification of information systems in many industries, including finance, communication and energy. According to the methodology of BIA and based on the actual conditions of the company's information system, this paper has improved the information system classification model and proposed a method for the classified O&M services, which realizes the management of the classified O&M services and can stabilize the operation of the information system, raising both the users' satisfaction and the O&M service value.

In the information system classification model proposed in this paper, the levels of the information systems are determined by evaluating the impacts of service interruption on the company's businesses, with consideration of the company's development strategy, informatization plan, regulatory compliance, user scale, internal impact, external impact, substitutability, correlation to other systems and supervision requirements. On this basis, measures for guarantee each level of the O&M services are worked out. The following principles are obeyed during our study.

Principle 1: business-oriented. In the information system classification model, the practical business demands shall be considered to evaluate the O&M service resources required for the operation of the system, and the classification shall be based on satisfying the business demands.

Principle 2: maximization of resource utilization. After satisfying the business demands of each information system, the model shall be reasonable to allocate and utilize the resources, so as to maximize the O&M service resources.

Principle 3: standardized classification. During the classification of the information systems, requirements as stated in the *Management Specifications of Informatization Standards for State Grid Corporations (Trail)* (No. 307, Information Technology (2010)) shall be followed to establish a set of normative and standard procedures, which is used to classify the O&M services of all the information system, and the classification shall also be reasonable and operable.

Principle 4: sustainability. The standard for classifying the O&M services shall be efficient, usable and improvable during a long term and shall be extended and modified according to objective requirements. In addition, it shall not be frequently changed for predictable problems or causes.

3.2 Content of the Model

Classification Settings

According the different impacts on business, the O&M services of the company's information system are classified into three levels: ordinary maintenance level (level 3), fundamental level (level 2) and guarantee level during rush hours (level 1). Considering that there are periods during which important events may be held or the company has special requirements on the information system, we set the fourth level: special guarantee level (special level), as shown in Table 1.

Table 1. Classification settings

Level	Name	Description
Level 1	Guarantee level during rush hours	Information system is very important. The impact of system interruption is serious
Level 2	Fundamental level	Information system is less important. The impact of system interruption is not very serious
Level 3	Ordinary maintenance level	Information system is not very important. The impact of system interruption is little
Special level	Special guarantee level	Important events may be held or there are special requirements on the information system

Procedure of the Model

The special procedure of the information system classification model is shown in Fig. 1. First of all, the model evaluates the business impacts of a business system during normal periods. Then, the model evaluates the business impacts of the business system during different business cycles. Finally, according to the standard of classifying business impacts, the model evaluates the service levels for the business system under all business cycles.

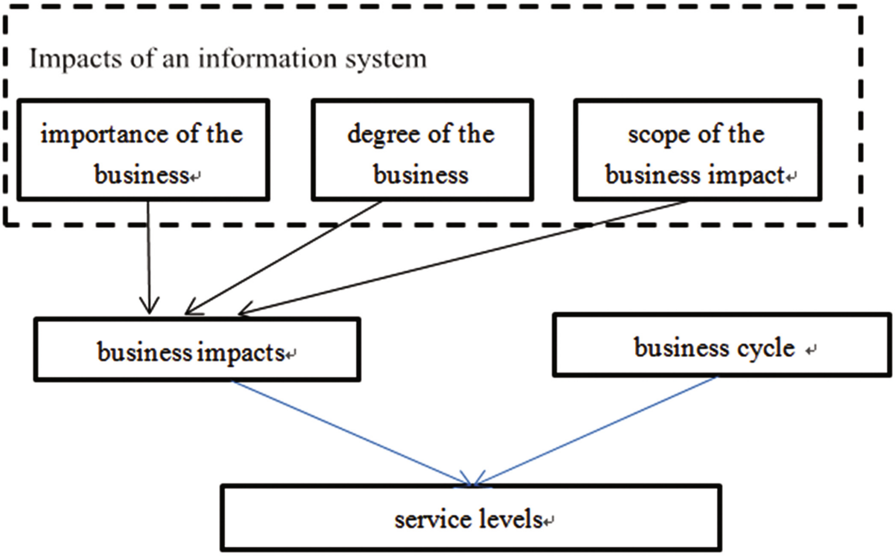

Impacts of an information system
importance of the business↵
degree of the business
scope of the business impact↵
business impacts↵
business cycle ↵
service levels↵

Fig. 1. Flow chart of the information system O&M classification model

First, the business impacts (Φ) of the business system during normal periods are evaluated. Next, the business impacts are evaluated during different business cycles (Ψ). Finally, according to the standard of classifying business impacts, the service levels (Λ) for the business system under all business cycles are evaluated.

Algorithm of the Model

As shown in Fig. 1, the information system for computing the business impact Φ includes three primary factors: importance of the business, degree of the business impact and scope of the business impact. Specifically, the importance of the business (weight 30%) consists of two secondary factors: the type of the information system and the impact of the system on other systems. The degree of the business impact (weight 35%), namely the degree of the impact of interruption on enterprises, society and nation, includes two secondary factors: external impact and the time for the system to tolerate the interruption [9]. The scope of the business impact (weight 30%), namely the range of the impact because of the interruption of the information system, also consists of two secondary factors: the service range and the service object of the information system. With considering practical conditions of the company, in addition, we set another secondary factor: the administrative requirements (weight 5%) from the competent authority, namely the administrative management requirements from a relevant department of our state or the company. Scores of above factors are listed in Table 2.

Table 2. Scores of the factors used for computing the business impact Φ

Number	Secondary factors	A	B	C	D
1	The type of the information system	100	70	40	0
2	The impact of the system on other systems	100	70	40	0
3	External impact and	100	70	40	0
4	The time for the system to tolerate the interruption	100	75	50	25
5	The service range	100	70	40	0
6	The service object	100	70	40	0
7	The administrative requirements	100	70	40	0

In Fig. 1, the business cycle Ψ is the product of the business cycle coefficient (u) and the business impact during a normal period. Table 3 shows the business cycle coefficients during different business cycles.

Table 3. Scores of the factors used for computing the business impact Φ

Number	Business cycles	Business cycle coefficient (u)
1	High	1.5
2	Normal	1
3	Low	0.5

The service level Λ is represented using a percentage score. Indexes of the business impact are weighted and a score is given to each index. Then, the business impact of the information system can be determined by analyzing and computing the indexes.

Step 1: computing the business impact Φ of the information system, as shown in Eq. (1).

$$\Phi = \sum_{k=1}^{n} \mathrm{K}^{*}\mathrm{KQ}^{*}\mathrm{FQ}, \tag{1}$$

where K is the score of a secondary factor; KQ is the weight of a secondary factor; and FQ is the weight of a primary factor.

Step 2: computing the impacts during different business cycles, as shown in Eq. (2).

$$\Lambda = \Phi * \mathrm{u}, \tag{2}$$

where u is the coefficient of the business cycle.

A higher score of the service level Λ means a higher level of the system. The scores and levels are defined as follows: the score greater than or equal to 70 means service level-1 (guarantee level); the score of 35～70 means service level-2 (fundamental level); and the score greater than or equal to 35 means service level-3 (ordinary maintenance level).

3.3 Application of the Model

In the case of the information system in the company, the scores of the seven secondary factors are B, C, B, D, B, C and D, respectively. The levels of the system can then be determined by using the improved information system classification model, as shown in Table 4.

Table 4. Computation of an actual system classification based on our model

Number	Secondary factors	Level	Scores	Impacts
1	The type of the information system	B	70	70*30%*50% = 10.5
2	The impact of the system on other systems	C	40	40*30%*50% = 6
3	External impact and	B	70	70*35% *50% = 12.25
4	The time for the system to tolerate the interruption	D	25	25*35%*50% = 4.38
5	The service range	B	70	70*35% *50% = 12.25
6	The service object	C	40	40*35%*50% = 7
7	The administrative requirements	D	0	
8	Impacts in normal period: 10.5 + 6 +12.25 + 4.38 + 12.25 + 7=52.38			35 < 52.38 < 70, Level 2
9	Impacts in high period: 52.38*1.5 = 78.57			78.57 > 70, Level 1
10	Impacts in low period: 52.38*0.5 = 26.19			26.19 < 35, Level 3

As indicated in the computation, the level-1 O&M service is required during the critical business cycles; the level-2 O&M service is required during shall be provided during normal business cycles, and; the level-3 O&M service is required during the periods when there are not much businesses.

4 Method for System Classification Service

The measures for O&M service are proposed according to the requirements of the company's information system O&M service and the requirements of the *Specifications for Operation and Maintenance of Information System in State Grid Corporations* (*Trail*) (No. 144, Security of Information Operation (2009)), with a full consideration of the information system itself and both hardware and software in the system.

The design of the O&M service at each level satisfies the highest requirements of all the information system under the current level. According to the volume of the businesses, the time of using the information system can be divided into: working period and non-working period. Therefore, more detailed measures can worked out to guarantee the service according to the different periods. Table 5 lists some level-1 (guarantee level) O&M service requirements.

Table 5. O&M service requirements

Number	Service periods	Measures for O&M service	Descriptions
1	Running	Monitor	7*24
2		Patrol	Twice a day
3		Backups [10]	Three times a day
4		Contingency	One hour
5	Overhaul	Scheduled maintenance	Standby unattended time
6		Failure response	5 min
7		Crash recovery	One hour
8		Fault feedback	3 days
9	Custom service	Response time	5 min
10		Process time	2 h

5 Conclusions

On the ground of a full investigation of the current conditions and requirements of the company's information system O&M, with consideration of the advanced ideas using for classifying information system O&M services in China, this paper has proposed the improved information system classification model and, based on which, put forward a new classification method. After been used in the company, this method has significantly raised the efficiency of using both human and material resources, greatly secured the operation and management of the information system and laid solid foundation for the company to promote the automation of the O&M.

Acknowledgments. This work is partly supported by the National Natural Science Foundation of China under Grant No. 61300122, 61502145 and 61602150, and the Fundamental Research Funds of China for the Central Universities under Grant No. 2013B01814.

References

1. Vaquero, L.M., Rodero-Merino, L., Caceres, J., Lindner, M.: A break in the clouds: towards a cloud definition. ACM SIGCOMM Comput. Commun. Rev. **39**, 50–55 (2008)
2. Yuriyama, M., Kushida, T.: Integrated cloud computing environment with it resources and sensor devices. Int. J. Space-Based Situated Comput. **1**, 163–173 (2011)
3. Mezghani, K., Ayadi, F.: Factors explaining is managers attitudes toward cloud computing adoption. Int. J. Technol. Human Interact. **12**, 1–20 (2016)
4. Chasaki, D., Mansour, C.: Security challenges in the internet of things. Int. J. Space-Based Situated Comput. **5**, 141–149 (2015)
5. Pencheva, E., Atanasov, I., Nikolov, A., Dimova, R., Ivanov, M.: An approach to data annotation for internet of things. Int. J. Inf. Technol. Web. Eng. **10**, 1–19 (2015)
6. Yang, K., Liu, S., Li, X., Wang, X.: D-S evidence theory based trust detection scheme in wireless sensor networks. Int. J. Technol. Human Interact. **12**, 48–59 (2016)

7. Honarvar, A., Sami, A.: Extracting usage patterns from power usage data of homes' appliances in smart home using big data platform. Int. J. Inf. Technol. Web. Eng. **11**, 39–50 (2016)
8. Mahesha, P., Vinod, D.: Support vector machine-based stuttering dysfluency classification using GMM supervectors. Int. J. Grid Util. Comput. **6**, 143–149 (2015)
9. Gotoh, Y., Yoshihisa, T., Taniguchi, H., Kanazawa, M.: A scheduling method for waiting time reduction in node relay-based webcast considering available bandwidth. Int. J. Grid Util. Comput. **2**, 295–302 (2011)
10. Stefanov, H., Jansen, S., Batenburg, R., Heusden, E.V., Khadka, R.: How to do successful chargeback for cloud services. In: Economics of Grids, Clouds, Systems, and Services, vol. 15, pp. 61–75. Springer, New York (2012)

A Model Profile for Pattern-Based Definition and Verification of Composite Cloud Services

Flora Amato[1], Nicola Mazzocca[1], Francesco Moscato[2](✉), and Fatos Xhafa[3]

[1] DIETI, University of Naples Federico II, Naples, Italy
{flora.amato,nicola.mazzocca}@unina.it
[2] DiSciPol, University della Campania Luigi Vanvitelli, Caserta, Italy
francesco.moscato@unicampania.it
[3] Department of Computer Science,
Technical University of Catalonia, Barcelona, Spain
fatos@cs.upc.edu

Abstract. Scientific community is now spending more and more efforts in defining and developing effective methodologies and technologies in order to easy design and development of Cloud solutions. In order to exploit the features of existing Cloud services and Resources Orchestration becomes a hot research topic. In this scenario, Cloud Designers promote reuse but a clear and simple design and verification methodology still misses in literature. In this scenario, a simple (UML-based) modelling profile and a Model-Driven Engineering methodology for Cloud-based Value Added Services are very appealing. In this work we define a modelling profile able to describe Orchestrated Cloud Services and Resources by means of Cloud Design Patterns and we show how Cloud Designer can use it both to ease composition and verification purposes.

1 Introduction

In the last years, Cloud Computing has emerged as the most widely used paradigm of parallel and distributed computing, as shown not only from the ever increasing efforts made by the scientific community in this field, but also for the interest shown by large companies and *Big Vendors* like Amazon, Microsoft or Google for Cloud-based solutions.

Besides classical Cloud architecture, services and resources, new paradigms are being proposed such as Cloud Federation [1] and Fog Computing [2]. However, admittedly these new paradigms have introduced more complexity in the design, verification and development of Cloud-based systems.

Essentially, two problems arise: one the one hand, in federated and multi-tenant Clouds, designers usually promote reuse of resources and services in order to create added valued services by composition. On the other hand, design constraints not only comprise performances and soundness of composite service, but also availability, security and other Quality of Services (QoS) parameters, as well as, of course, cost-efficient solution.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_65

Cloud resources management is of course one of the main problems to face with while Resource Orchestration [3,4] is envisaged as a key issue to solve the problem [5,6]. It consists in smart selection, deployment, monitoring and control of resources in single, multi-cloud and federated environments.

It should be recalled here that composition not only includes Resources Orchestration but first of all it involves services in all layers of Cloud Architecture (i.e. at least SaaS, PaaS and IaaS); in addition, Cloud users and designers should exploit good software engineering practices in order to achieve *enhanced* services. Therefore, these are some of the reasons motivating the recent research activities on Orchestration languages and techniques for Services composition, and on defining proper software engineering methodologies for improving quality of composite services.

New trends on cloud services design and management grew up recently leading to the definition of extended Design Patterns [7] for Cloud computing. Several recent works focus on mixing composition by Orchestration and patterns since many design patterns with different purposes can be described as complex workflow processes (and, of course, Orchestration languages are all based on definitions of workflow processes [8,9]).

Considering that it is still difficult for Cloud Developers to master languages, frameworks and technologies proposed in literature, the question is how to bridge the gap among common software engineering languages and practices, and the plethora of algorithms, tools, languages and frameworks (as well as different Cloud APIs) used for Cloud Services Orchestration.

In this paper we address this question through the connection of UML-based languages and Model Driven Engineering (MDE) techniques (in particular Model Transformation). More precisely, we describe a UML-based Modelling Profile able to address different Cloud Patterns used in composition. We show how profile is the starting point to define proper Model Transformation techniques able to solve some verification problem, as well as to provide a means to implement automatic generation of configuration and services composite Cloud Services [10,11]. The glue in MDE steps is a workflow language that we use to describe composition at all Cloud levels.

2 MetaMORP(h)OSy Profile for Cloud Patterns

In this work we improve the modelling profile for Cloud Systems developed in the MetaMORP(h)OSy Framework. Previous works [12,13] described the framework and the related methodology. In brief, it is based on a modelling profile (i.e. a meta-model) we used to define models based on Multi-Agent Systems (MAS). The framework describes MAS by using an extension of Beliefs, Desires, Intentions (BDI) logics [14]. From a designer perspective, the framework interface allows for definition of diagrams that are similar to UML diagrams (in fact, the modelling profile we defined is a UML meta-model).

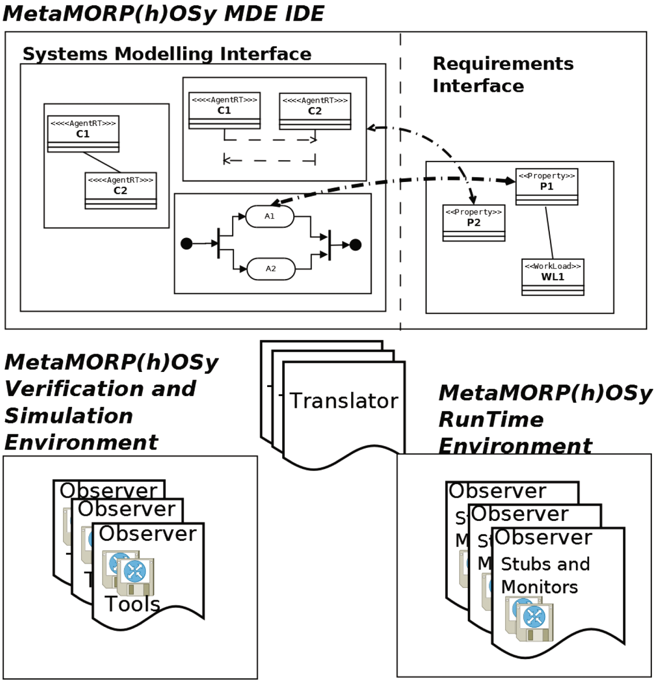


Fig. 1. MetaMORP(h)OSy framework

Figure 1 shows the main components of MetaMORP(h)OSY. The graphical interface of MetaMORP(h)OSy is based on Papyrus[1]. Hence, designers are able to produce UML-like diagrams with stereotypes defined by MetaMORP(h)OSY profile.

The modelling profile allows for both the description of systems and the definition of their requirements (in terms of QoS properties). A deeper description of Requirements profile can be found in [12,15,16]. In particular, the system enacts verification (of requirements on the model) activities when a requirement is linked to a component of the system model. Proper modules called Observers (whose structure is defined in the modeling profile too), select the best suitable *Translator* in the framework to execute proper Model Transformations [17] and to run proper tools to evaluate *QoS* properties on the model. Since MetaMORP(h)OSY covers all the life cycle of the whole system, it uses particular Observers at run-time for monitoring and testing purposes [18].

To facilitate reading, we describe here the part of the MetaMORP(h)OSy profile that enables Pattern-Based composition. Patterns are a widely used term in software engineering. They mainly have a *descriptive* nature in reporting good solutions to recurrent problems. For Cloud Patterns, several works exist in

[1] https://eclipse.org/papyrus/.

literature, both from commercial vendors [8,19,20], and from scientific communities[2] [21–23].

As outlined in [24,25], many of Cloud Patterns defined in the cited works share a common characteristics: they describe how to use and *compose* existing resources and services in order to achieve complex architectures or added value services, or to add particular features to existing services. We showed in [24,25] how the models on which composition relies is based on workflow graphs and languages. Here we want to show that it is in fact possible to build automatically the implementation skeleton of a composite Cloud Service by means of definition of Patterns on which it relies.

MetaMORP(h)OSY features this automatism by means of a modeling profile definition. Then, Model Transformation tools, provide generation of skeletons of composite services. MetaMORP(h)OSY modelling profile (Real Time Agent Modelling Language: RT-AML) already includes meta-models to describe structures of elements and their behaviours [26], by means of agent diagrams. They includes extended version of UML Class, Sequence and Activity diagrams (aforementioned works on MetaMORP(h)OSy contains more information on RT-AML profile).

In this work we present the RT-AML extensions to model composite Cloud Services by means of Patterns. Since the Patterns we are dealing with usually provide structural or architectural descriptions of composition, we extend RT-AML in order to build profiles to model each pattern we want to include in the framework. Patterns will be modelled by extension of UML Component and Deployment Diagrams. Since we are considering Cloud Systems, we have to manage resources at any layers of Cloud Architecture. Deployment Diagrams will help to define Cloud Federations, Zones and Resources. RT-AML requires the definition of Patterns Structures to extended Component Diagrams. Former RT-AML elements manage description of behaviours, structures and interactions of every cloud services.

To illustrate the idea, we describe here the structure of elements in Component and Deployment Diagrams used to model the *Multi-Datacenter* Pattern by Amazon AWS[3]. Figure 2 depicts the patterns: Users use it in order to increase data availability, and when they deal with distributed data (for such an example refer to [27]).

Amazon Web Services introduced the pattern, but its application is obviously more general. It describes the case when multiple data exist in different *Zones* (this is the term Amazon uses to identify autonomous Clouds in a federated or multi-cloud environment). A broker (Elastic Broker - ELB in the figure) accepts requests and manages data accesses and operations. Data can be distributed and/or replicated in different zones and eventually replication could exist in each zone. Users can define different policies for data allocation (i.e., warm replication, master-slave, etc.). Depending on replication and distribution policies, and on

[2] http://www.cloudpatterns.org/.

[3] http://en.clouddesignpattern.org/index.php/CDP:Multi-Datacenter_Pattern.

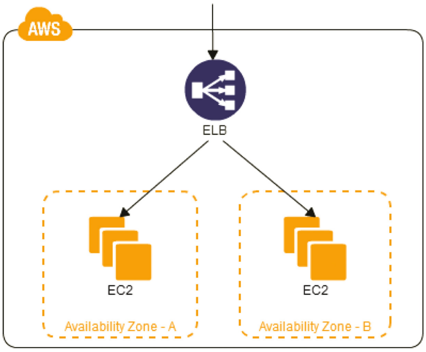


Fig. 2. Multi-datacenter pattern

zones QoS, the composite system improves performances [28–30] and availability. Obviously, proper services have to be developed in order to implement policies.

In order to describe this pattern, we need to define the distribution of data and a mechanism to describe zones. In addition, some elements in the profile must identify policies and roles of components services and resources. Note that simple elements in the profile specifying resources roles (like storage resources, SaaS Services, brokers etc.) were addressed in previous works.

Here we show the main elements in MetaMORP(h)OSY profile able to describe the structure of this pattern. First of all, we define a proper stereotype to identify the pattern structures. We use this stereotype (*Multi-Data-center Pattern*) later in the case study in order to address composition purposes.

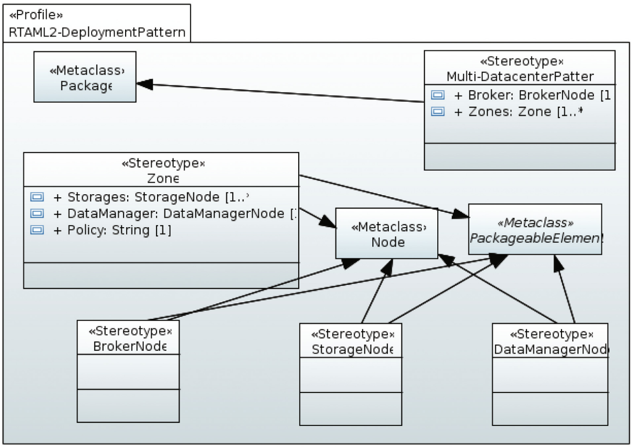


Fig. 3. Multi-datacenter deployment elements

Figure 3 shows the main elements in the modelling profile for definition of a set of services and resources organized as the aforementioned pattern. In order to apply stereotypes depicted in the figure, users must define a package in a Deployment Diagram where they apply the *Multi-Datacenter Pattern* stereotype. Then, they define Nodes delimiting *Zones*. Inside Zones users declare *StorageNodes* and *DataManagerNodes*. All Zones have a Node acting as *BrokerNode*. Properties in stereotypes (like Storages in Zone) allow for connection of sub-packaged elements.

For brevity's sake, we do not report properties linking Nodes with proper resources and services in other part of the profile.

3 Model Transformation

Model Transformation (MT) [17] is the key point on which relies MetaMORP(h)OSy. It allows for translation of a *model* expressed in a given language (called source language), into a different model expressed in the same or another language (the target language). In order to enable *automatic* transformation, both source and target language, as well as the transformation rules, have to be expressed in a formal way.

In order to verify designed composite Cloud Services [31], and to generate executable processes in real Cloud environments [27], MetaMORP(h)OSy needs to translate models at design stage into both analysable and executable models. Both targets need a formal description of the composite service in terms of a workflow.

MT hence works here in two stages:

- In the first stage, MetaMORP(h)OSy produces a workflow process from design models. This is a skeleton that will be eventually filled in with real resources and services;
- In the second stage, depending on the MT target, MetaMORP(h)OSy produces analysis models or run-time skeletons from the representation of the workflow process. Eventually it executes proper Observers to evaluate QoS properties on analysis models.

Hence, composition by patterns is based on the translation of design models into a workflow representation (a complex workflow graph) of composite service. We call Operational Flow Language (OFL) the language for workflow description of composition. Because of its formal definition, we can achieve MT algorithms to translate design models (that we have formally defined by a UML model profile) into a workflow graph (Operational Flow Graph: OFG) expressed in OFL.

In the second stage, we can use other MT algorithms to translate OFG both in analyzable models or run-time skeletons: target language of transformation depends on the goal we want to reach.

For example, if we are interested in the analysis of availability when we apply the pattern in Fig. 2 to a real distribution of data storages, we can translate a deployment diagram of the composite service first into the related OFG, and then we can further translate the OFG with information on real resources into a Fault Tree [32] in order to analyze availability.

At second stage MT depends on the patterns used for composition at first stage, namely, each pattern is related to a proper MT algorithms, in both stages MetaMORP(h)OSy works.

We report in the following the MT we use to translate the OFG of a Multi-Datacenter Pattern into a Fault Tree. The OFG alone does not contain enough information to discriminate between a distribution or replica semantics but the information that we generated the OFG from a Multi-Datacenter Pattern allows a proper understanding of semantics of OFL constructs in OFG. Figure 4 describes the overall procedure used for availability evaluation of pattern instances.

```
Input: OFL clustered Graph; Patterns Instances declarations
instances an empty Faultree FT;
expandCluster(cluster).
    create an event Ev in the FT for the cluster
    if cluster is a Sequence OR a Multiple Instance OR a Parallel with
    distributed semantics
        Create an OR port in the FT with Ev as output event
        foreach Sequence component comp:
            if comp is a simple Activity
                create a new Basic Event in the FT
                link the Basic Event to the OR port
            elseif comp is cluster nestCl
                expandCluster (nestCl)
                link expanded cluster top event to the OR port
            end elseif
        end foreach
    elseif cluster is a Parallel with replica semantics
        Create an AND port in the FT with Ev as output event
        foreach parallel branch branch:
            create a new Event
            link the Event to the AND port
            if branch is a simple Activity
                create a new Basic Event in the FT
                link the Basic Event to the AND port
            elseif branch is cluster nestCl
                expandCluster (nestCl)
                link expanded cluster top event to the AND port
            end elseif
        end foreach
    end elseif
    if cluster depends on Resource Cluster
        Create an OR port in the FT with Ev as output event
        expandCluster(Resource Cluster)
        link top event of expanded cluster to the OR port
    endif
end expandCluster
```

Fig. 4. Availability algorithm

For details of Workflow Patterns we refer to in the figure, like Sequences, Multiple Instances etc. the readers are referred to [33]. We show the results of the application to a real-case in Sect. 4.

4 A Case Study

Here we apply the profile methodology we sketched in Sect. 2 to define a simple application of multi-datacenter pattern, useful to replicate data in two different zones.

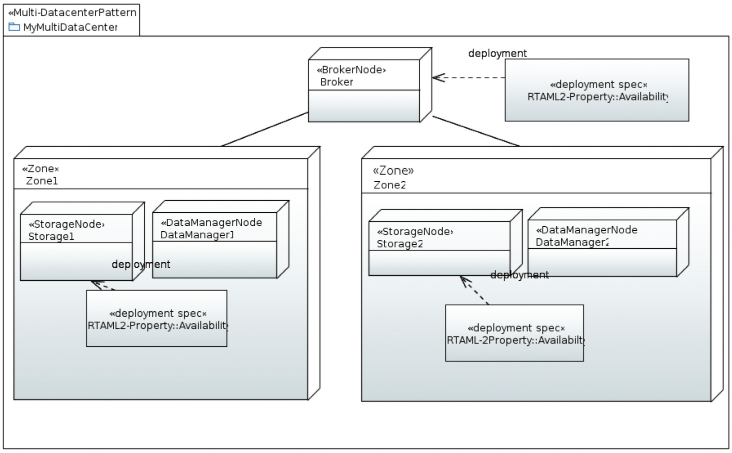


Fig. 5. Multi-datacenter deployment diagram

Figure 5 depicts the resulting model, where two zone nodes contain a data manager service and a storage resource. In each Zone these two elements are associated to proper agents implementing them (more details on modelling agents' structures in RTAML can be found in referred previous works). Properties on nodes, such as availability, can be defined too.

Now we can apply MT algorithms introduced in Sect. 3.

Figure 6 contains the OFG generated from the Deployment Diagram in Fig. 5 during a read operation. *BrokerNode* activity models actions of the ELB; *DM1* and *DM2* resumes the operations of Data Managers in different zones, while *St1* and *St2* are the storage nodes in the zones.

Notice that we define all information about resources and services in the deployment diagram.

The application of MT in Fig. 4 to the generated OFG, results in the Fault-Tree in Fig. 7.

Let us compare results from two solutions, when availability of components services and resources can vary.

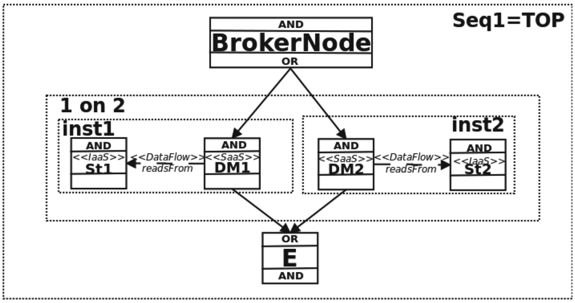


Fig. 6. Generated OFG

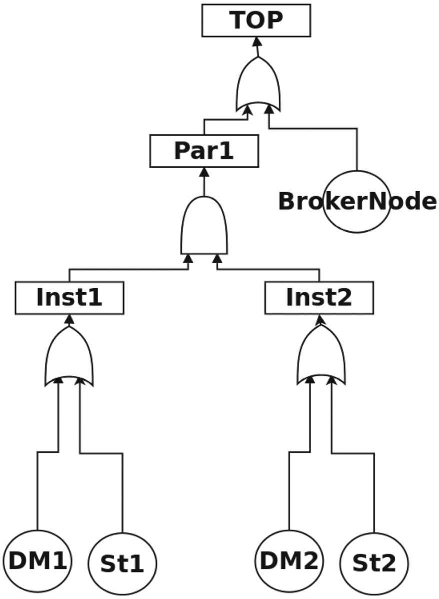


Fig. 7. Generated fault tree

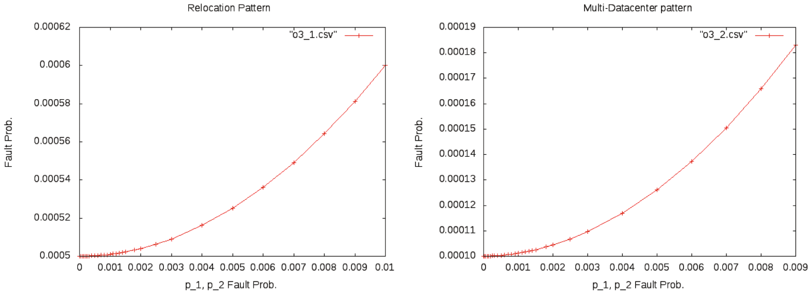


Fig. 8. Fault probability comparison, depending on virtual servers avail.

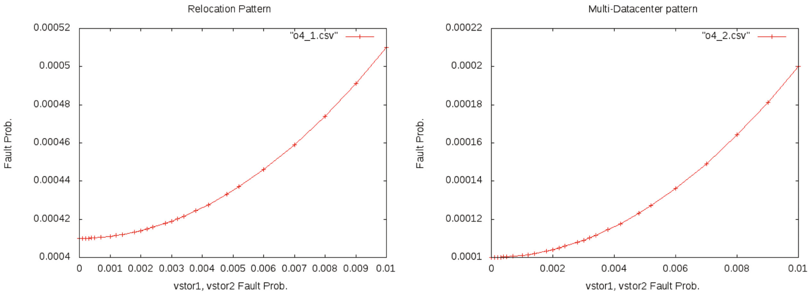


Fig. 9. Fault probability comparison, depending on virtual storage avail.

Comparisons of fault probability of composite service are reported in Figs. 8 and 9 when several fault probabilities of components services and resources vary. Except for parameters that vary, the fault probability of other elements are in Table 1.

Table 1. Default fault probability

p_1	p_2	prov1	prov2	vstor1	vstor2
0.000001	0.000001	0.00000001	0.00000001	0.00001	0.00001
Proxy	s	m	Copy	Deploy	
0.0001	0.00001	0.0001	0.0001	0.0001	

5 Conclusions and Future Works

This paper presents a model profile and a methodology to build and analyze composite Cloud Services and resources. It is based on MetaMORP(h)OSy framework, Cloud Pattern definition and model transformation techniques.

We enable Cloud Users to design their solutions by specifying the pattern they need. Pattern-based specification is then automatically translated into a workflow graph and then into analyzable models or in run-time skeletons.

We showed by a case study the application of the methodology to the Multi-Datacenter Patterns and we show how it allows for availability analysis in an fully automatic way.

Future works include the definition of model transformation algorithms for the study of other QoS properties.

References

1. Kurze, T., Klems, M., Bermbach, D., Lenk, A., Tai, S., Kunze, M.: Cloud federation. In: Proceedings of the 2nd International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING) (2011)
2. Bonomi, F., Milito, R., Zhu, J., Addepalli, S.: Fog computing and its role in the internet of things. In: Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, pp. 13–16. ACM (2012)
3. Wieder, A., Bhatotia, P., Post, A., Rodrigues, R.: Conductor: orchestrating the clouds. In: Proceedings of the 4th International Workshop on Large Scale Distributed Systems and Middleware, pp. 44–48. ACM (2010)
4. Liu, C., Mao, Y., Van der Merwe, J., Fernandez, M.: Cloud resource orchestration: a data-centric approach. In: Proceedings of the Biennial Conference on Innovative Data Systems Research (CIDR), pp. 1–8 (2011)
5. Ranjan, R., Benatallah, B., Dustdar, S., Papazoglou, M.P.: Cloud resource orchestration programming: overview, issues, and directions. Internet Comput. **19**(5), 46–56 (2015). IEEE
6. Feng, G., Buyya, R.: Maximum revenue-oriented resource allocation in cloud. Int. J. Grid Util. Comput. **7**(1), 12–21 (2016)
7. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-oriented Software. Pearson Education, Upper Saddle River (1994)
8. Wilder, B.: Cloud Architecture Patterns: Using Microsoft Azure. O'Reilly Media Inc., Sebastopol (2012)
9. Fehling, C., Leymann, F., Rütschlin, J., Schumm, D.: Pattern-based development and management of cloud applications. Future Internet **4**(1), 110–141 (2012)
10. Verma, A., Kaushal, S.: Deadline constraint heuristic-based genetic algorithm for workflow scheduling in cloud. Int. J. Grid Util. Comput. **5**(2), 96–106 (2014)
11. Zhu, X.D., Li, H., Li, F.H.: Privacy-preserving logistic regression outsourcing in cloud computing. Int. J. Grid Util. Comput. **4**(2–3), 144–150 (2013)
12. Moscato, F.: Model driven engineering and verification of composite cloud services in metamorp(h)osy. In: Proceedings of 6th, International Conference on Intelligent Networking and Collaborative Systems INCoS-2014. IEEE (2014)
13. Aversa, R., Martino, B., Moscato, F.: Critical systems verification in metamorp(h)osy. In: Bondavalli, A., Ceccarelli, A., Ortmeier, F. (eds.) SAFECOMP 2014. LNCS, vol. 8696, pp. 119–129. Springer, Cham (2014)
14. Wooldridge, M.: Agent-based software engineering. In: IEE Proceedings on Software Engineering, pp. 26–37 (1997)
15. Moscato, F., Amato, F., Amato, A., Aversa, R.: Model-driven engineering of cloud components in metamorp(h)osy. Int. J. Grid Util. Comput. **5**(2), 107–122 (2014)
16. Moscato, F., Amato, F.: Thermal-aware verification and monitoring of service providers in metamorp(h)osy. In: Proceedings of 6th International Conference on Intelligent Networking and Collaborative Systems INCoS-2014. IEEE (2014)
17. Mens, T., Van Gorp, P.: A taxonomy of model transformation. Electron. Notes Theor. Comput. Sci. **152**, 125–142 (2006). Proceedings of the International Workshop on Graph and Model Transformation (GraMoT 2005), Graph and Model Transformation (2005)
18. Di Domenico, D., Moscato, F.: Automatic monitor generation for cloud services, pp. 547–552 (2015)
19. Amazon Elastic Compute Cloud. Amazon web services. Accessed 9 Nov 2011

20. Microsoft Developer Network: Cloud Design Patterns: Prescriptive Architecture Guidance for Cloud Applications. Microsoft, New York (2014)
21. Fehling, C., Retter, R.: Cloud computing patterns (2011)
22. Fehling, C., Leymann, F., Retter, R., Schupeck, W., Arbitter, P.: Cloud Computing Patterns. Springer, Vienna (2014)
23. Di Martino, B., Cretella, G., Esposito, A.: Semantic and agnostic representation of cloud patterns for cloud interoperability and portability. In: Proceedings of the 5th IEEE International Conference on Cloud Computing Technology and Science (CloudCom), pp. 182–187 (2013)
24. Amato, F., Moscato, F.: Exploiting cloud and workflow patterns for the analysis of composite cloud services. Future Gener. Comput. Syst. **67**, 255–265 (2017)
25. Amato, F., Moscato, F.: Pattern-based orchestration and automatic verification of composite cloud services. Comput. Electr. Eng. **56**, 842–853 (2016)
26. Cicotti, G., Coppolino, L., D'Antonio, S., Romano, L.: Runtime model checking for SLA compliance monitoring and QOS prediction. J. Wirel. Mob. Netw. Ubiquit. Comput. Dependable Appl. **6**(2), 4–20 (2015)
27. Terzo, O., Ruiu, P., Bucci, E., Xhafa, F.: Data as a service (DaaS) for sharing and processing of large data collections in the cloud. In: Seventh International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), pp. 475–480. IEEE (2013)
28. Amato, F., Barbareschi, M., Casola, V., Mazzeo, A.: An FPGA-based smart classifier for decision support systems. Stud. Comput. Intell. **511**, 289–299 (2014)
29. Amato, F., Barbareschi, M., Casola, V., Mazzeo, A., Romano, S.: Towards automatic generation of hardware classifiers. In: Aversa, R., Kołodziej, J., Zhang, J., Amato, F., Fortino, G. (eds.) ICA3PP 2013. LNCS, vol. 8286, pp. 125–132. Springer, Cham (2013)
30. Spaho, E., Sakamoto, S., Barolli, L., Xhafa, F., Ikeda, M.: Trustworthiness in P2P: performance behaviour of two fuzzy-based systems for JXTA-overlay platform. Soft Comput. **18**(9), 1783–1793 (2014)
31. Bessis, N., Sotiriadis, S., Xhafa, F., Pop, F., Cristea, V.: Meta-scheduling issues in interoperable hpcs, grids and clouds. Int. J. Web Grid Serv. **8**(2), 153–172 (2012)
32. Hirel, C., Sahner, R., Zang, X., Trivedi, K.: Reliability and performability modeling using sharpe 2000. In: Haverkort, B.R., Bohnenkamp, H.C., Smith, C.U. (eds.) Computer Performance Evaluation. Modelling Techniques and Tools, vol. 1786. LNCS, pp. 345–349. Springer, Heidelberg (2000)
33. Moscato, F., Vittorini, V., Amato, F., Mazzeo, A., Mazzocca, N.: Solution workflows for model-based analysis of complex systems. IEEE Trans. Autom. Sci. Eng. **9**(1), 83–95 (2012)

A Routing Based on Geographical Location Information for Wireless Ad Hoc Networks

Yongqiang Li(✉), Zhong Wang, Qinggang Fan, Yanning Cai,
Yubin Wu, and Yunjie Zhu

Computer Staff Room, Xian Research Institute of High Technology,
Xian 710025, China
lyq200381@163.com

Abstract. This paper Propose the location-based routing algorithm forwarding with alternative copies. This routing algorithm considers both the location information and the velocity vector to establish the utility function, which uses a factor of speed to adapt to the change of the node density. And the forwarding strategy of alternative copies can guarantee the success of the transfer between adjacent nodes. The simulation results demonstrate preliminarily that the DTN routing algorithms designed in the paper performs well in terms to the delivery ratio, the delays and the network overhead.

1 Introduction

As an important part of the Intelligent Transportation System, the vehicle ad-hoc network is receiving more and more attention from all over the world. Vehicle ad-hoc network is a special category of the mobile Ad hoc networks, but its characteristics such as the high-speed nodes, the rapidly changing network topology and the unstable communication links, make it hard for the traditional routing protocols to meet the demand of the vehicle ad-hoc network. DTN [1], a new network with long delays and intermittently connected links, fits the vehicle ad-hoc network very well. Therefore, it's feasible to apply the DTN technology into the vehicle ad-hoc network.

2 Forwarding Algorithm of Backup Copy Based on Geographic Location

With the wide use of GPS, the forwarding algorithm based on geographic location information has greatly improved the communication effect of DTN network [2]. However, due to the factors such as the high speed of the node and the communication environment, the neighbor node at the edge of the communication radius of DTN is out of the communication range at the next moment. To avoid such a situation, it is necessary to use a backup copy [3, 4]. Therefore, this paper designs a Location Routing with Alternative Copies to improve the communication performance of DTN networks. This Location Aided Routing with Alternative Copies consists of two parts: forwarding strategy based on geographic location and forwarding strategy of backup copy.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_66

2.1 Forwarding Strategy Based on Geographic Location

1. **Distance**

When selecting the next hop, the nearest neighbor node is selected as the relay, which can shorten the forwarding path, reduce the number of hops, and save the end-to-end delay.

As shown in Fig. 1, the node A and the node D communication, the node B and the C are neighbor of the A node, in the choice of the next hop forwarding node, first calculate the node A, B, C and the distance between the destination node D $dist_{AD}$, $dist_{BD}$, $dist_{CD}$. the location information table Location Array, which is located in the node A maintenance, the position coordinates of the nodes are $A(x_A, y_A)$, B(x_B, y_B), C(x_C, y_C), D(x_D, y_D),Velocity vector $v_a(v_{xa}, v_{ya})$, $v_b(v_{xb}, v_{yb})$, $v_c(v_{xc}, v_{yc})$, $v_d(v_{xd}, v_{yd})$. Due to the timeliness of location information, meet location and Hello message neighbor node location node when compared with a certain deviation, therefore, in the calculation, according to the velocity vector and the time difference to predict the real-time location of nodes will be more accurate. Set (x'_A, y'_A) to the node A in the calculation of the real-time position, then, the x'_A, y'_A calculation as shown in formulas 1.1, 1.2:

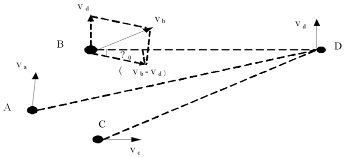


Fig. 1. Network node encounter

$$x'_A = x_A + (T_{now} - T_0) \times v_{xa} \tag{1.1}$$

$$y'_A = y_A + (T_{now} - T_0) \times v_{ya} \tag{1.2}$$

Among them, Tnow represents the current time, T0 represents the location information of the Hello message generation time. The real-time location of node B, C, D is calculated in turn (x'_B, y'_B), (x'_C, y'_C), (x'_D, y'_D), And according to formulas 1.3, 1.4, 1.5, we calculate the distance between each node and the destination node D.

$$dist_{AD} = \sqrt{(x'_A - x'_D)^2 + (y'_A - y'_D)^2} \tag{1.3}$$

$$dist_{BD} = \sqrt{(x'_B - x'_D)^2 + (y'_B - y'_D)^2} \tag{1.4}$$

$$dist_{CD} = \sqrt{(x'_C - x'_D)^2 + (y'_C - y'_D)^2} \tag{1.5}$$

Compare the distance between node A, B, C and destination node D. If $dist_{AD}$, $dist_{CD} > dist_{BD}$,indicating that the node B distance from the destination node is the Shortest, B will submit message to the destination node D, the spare time will be shorter, more probability, so the node A forwarding the message to a node B; if $dist_{BD}$, $dist_{CD} > dist_{AD}$ indicating that the node A than its neighbor node B, C distance closer to the destination node, A can not find suitable the next hop forwarding, so A will carry the message to wait until closer to the node.

2. **Velocity Vector**

In the routing algorithm based location, not only the location of the node can determine the forwarding path of the message, but also the speed vector [5] can have a great impact. Especially in the sparse node density environment, nodes often need to exercise after a period of time to meet a forwarding node, so we choose the node with the better the magnitude and direction of velocity as a forwarding node, message forwarding success rate will be higher.

In Fig. 1, the velocity vector difference of the B and the D on the BD line reflects a certain degree of the speed of the two node, the greater the speed, the higher the probability that the node will meet in the movement. The projection of the velocity vector difference on the line is represented by vs, and the formula is as follows:

$$\begin{aligned} vs(B) &= |v_b - v_d| \times \cos\theta_0 = |v_b - v_d| \times \frac{(v_b - v_d) \cdot \overrightarrow{BD}}{|v_b - v_d| \times \left|\overrightarrow{BD}\right|} \\ &= \frac{(v_{xb} - v_{xd}) \times (x_d - x_b) + (v_{yb} - v_{yd}) \times (y_d - y_b)}{dist_{BD}} \end{aligned} \tag{1.6}$$

Among them, θ_0 is the angle between the velocity difference of $(v_b - v_d)$ and the line connecting Band D; $dist_{BD}$ can be calculated by formula 1.4. calculating $vs(A)$, $vs(B)$, $vs(C)$ Respectively, and compare, the largest value of the node has the most forward advantage.

3. **Comprehensive Utility Function**

In order to comprehensively consider the effect of distance dist and velocity vector vs, this paper takes the utility function Util as the criterion, as shown in formulas 1.7, 1.8, and 1.9.

$$Util(B) = U\_location(B) + U\_speed(B) \times \alpha \tag{1.7}$$

$$U\_location(B) = \frac{dist_{AD} - dist_{BD}}{m} \tag{1.8}$$

$$U\_speed(B) = \frac{vs(B)}{240} \tag{1.9}$$

U_localtion is The utility of distance characterization of node, m as the communication radius of nodes due to $|dist_{AD} - dist_{BD}| < m$, the U_location range is (−1, 1), especially for the A, the node distance utility value U_location (A) = 0; U_speed denote the utility value of velocity vector of node, in vehicle network, the vehicle speed is generally 120 km/h so the range of U_speed is (−1, 1), keep the same order of magnitude and distance of utility value; α is the speed of impact factor, the speed had little effect in the dense nodes scene had a smaller value, in sparse environment, speed effect function change take a larger value.

In order to determine the specific value of the velocity influence factor α, we need to discuss the different density of the scene.

(1) node dense area
In the node intensive areas, in general, node based on distance has been able to meet to select the next hop node, unless its position after the encounter is not ideal, aided by the need for speed, but this probability is small.

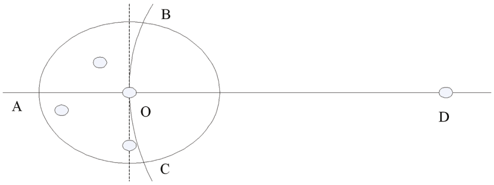


Fig. 2. Neighbor graph

In Fig. 2, assuming nodes were distributed randomly in the network, the node O to deliver the message to the destination node D, the communication radius node O is R, the distance between O and D is L, the number of neighbor nodes O is N, BOC is the circle that D is the center and OD is the radius of circle. If the speed reference is needed, the neighbor of the nodes O need to be distributed in the region BACO. The area of BACO is s, and the probability that the neighbor node falls in the BACO region is p:

$$s = \frac{\pi r^2}{2} - r^2 \arcsin\frac{r}{2l} + 2l^2 \arcsin\frac{r}{2l} - \frac{1}{4} r\sqrt{2l^2 - r^2} \tag{1.10}$$

$$\begin{aligned} p = \left(\frac{s}{\pi r^2}\right)^N = \left(\frac{\pi r^2}{2} - r^2 \arcsin\frac{r}{2l} + 2l^2 \arcsin\frac{r}{2l} - \frac{1}{4} r\sqrt{2l^2 - r^2}/\pi r^2\right)^N \\ \approx \left(\frac{1}{2} - \frac{1}{\pi}\arcsin\frac{r}{2l}\right)^N \end{aligned} \tag{1.11}$$

$$\alpha = p = \left(\frac{1}{2} - \frac{1}{\pi}\arcsin\frac{r}{2l}\right)^N \tag{1.12}$$

The velocity influence factor is similar to the probability p, so it can be represented by P. Because of r < L, so the value range of α is $(1/3)\mathrm{N} < \alpha < (1/2)\mathrm{N}$.

(2) sparse region
In the area with sparse nodes, the selection of the next hop node has a great dependence on the speed. According to the formula (1.12), when N is zero, the probability that the neighbor node falls in the BACO region is p = 1.

Therefore, when the node is often in the "isolated island" state without a neighbor node, the velocity factor can be used to characterize the.

$$\alpha = 1 \tag{1.13}$$

In summary, the nodes in the network according to the degree of density change perceived number of neighbor nodes around the node and the number of neighbor node is more, it may be considered that the area around the relatively dense nodes, the neighbor node number less, it may be considered that the area around the nodes are sparse. For this, each node can use NeighborNumber to record the number of neighbor nodes, and use LoneTime to record the time length of neighbor nodes. When the number of neighbor nodes is often in the state of zero, LoneTime is a buffer state greater than zero.

The process of forwarding algorithm based on geographical location is: when the node have forwarding task, first find the node of LocationArray maintenance, according to calculating and comparing itself with the neighbor node Util utility function, max (UtilA,UtilB,UtilC), if UtilA, indicates that other neighbor nodes are not suitable for forwarding. A storage node and carries the message, continue to wait for the next hop node, otherwise the message is forwarded to the neighbor node maximum value of Util.

2.2 Backup Copy Forwarding Strategy

The main idea of the backup copy forwarding strategy is that the node A keeps a spare copy in order to reduce the probability of transmission failure due to the rapid change of network topology. When the transmission is successful, the node A receives the confirmation information sent by the receiving node, and then deletes the backup copy. Otherwise, the node A will send the backup copy again.

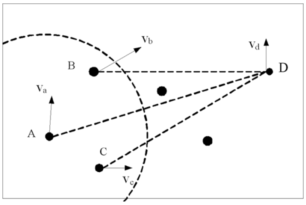


Fig. 3. Network node distribution

In order to realize the function of backup copy, it is necessary to introduce ACK response mechanism [6]. After the current one hop node receives the message successfully, it is necessary to send a ACK response message to the last hop node, so that the last hop has been successfully received. As shown in Fig. 3, the node B, the node C are A neighborhood, the node B utility value is better than the node A and C, the node A will be forwarded to the message B.

(1) if B successfully receives a forwarding from A, B immediately replies to a ACK response to A;
(2) A receives the ACK response, deletes the backup copy, completes the forwarding task;
(3) if A does not receive ACK responses to B, show that the forwarding failed after a t delay, A Util will recalculate the neighbor node, a copy forwarded to the optimal node C;
(4) repeat steps from 1 to 3, until the completion of the forwarding, node A will be removed from the backup copy.

2.3 Backup Copy Forwarding Algorithm Based on Geographic Location Information

Due to the geographical location of the clear direction of forwarding strategy and standby copy forwarding protection strategy in the submission process based on single packet copy theory has been able to ensure the transmission of network messages successfully, and multiple copies of the packets is likely to cause congestion, so DTN network, based on geographic location information backup copy forwarding algorithm LARAC the geographical position and the backup copy two forwarding strategy based on comprehensive [2]. In the process of LARAC algorithm, the location information of each node is shared by the geographic information sharing model (LSM) [7], and the location information is provided for the realization of the algorithm. When the node forwards the task, according to the position and velocity vector calculation and its neighbor nodes of the utility value of Until, when its Util value is maximum, no treatment, or choose the utility value of the largest neighbor node as the next hop forwarding node. After forwarding the message, if the node receives the ACK response message forwarding node to the next hop in a t delay in forwarding node will illustrate successful backup copy delete, complete the forwarding task, otherwise, the node recalculate the utility value and choose new next hop forwarding message, until success [8].

Compared to the advantages with the existing forwarding algorithm based on location information compared to the advantages of LARAC algorithm is: (1) make full use of the node position and velocity vector information forwarding node selection in the play in the next hop, utility value function; (2) to the density of the perception of the surrounding nodes, and the velocity factor the speed of adjustment, adaptive effects in different node density in the network, the utility function can adapt to the change of the environment; (3) using a single copy of the forwarding mechanism in general, ensure delivery success rate at the same time, save the cyber source, control network overhead, has a good control effect on network congestion.

3 Simulation Analysis

The simulation environment were extended with CMU wireless NS-2 (version 2.30), and were used to alternate copy location algorithm based on Forwarding (LARAC), alternate copy of two hop ACK confirm the forwarding algorithm based on Spray (2H) and Wait (SAW). We comparison these three algorithms. The simulation scene size is 2000 m * 1000 m, as shown in Fig. 4. Due to the different road sections, different time periods, the traffic density in the road traffic is also different, the experiments were carried out on the number of nodes for the 50 ~ 200 scene simulation, other specific parameters set as shown in Table 1.

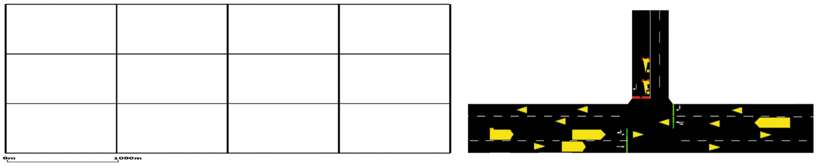


Fig. 4. Urban road SUMO scene settings

Table 1. Experimental parameters

The number of Node	50 ~ 200
packet size	1 KB
TTL	600 s
Link bandwidth	1 Mbps
buffer of node	1 M
Simulation time	3600 s
interval of message	(10,20) s
Node velocity	(0,20) m/s
Communication distance	150 m
Routing	LARAC, 2H, SAW

1. Submission rate

As can be seen from Fig. 5, in the urban road scene, the LARAC and 2H algorithm message delivery rate is significantly better than Spray and Wait. This is because the LARAC and 2H algorithm are forwarding algorithm based on location information, has a strong purpose, so the delivery efficiency is higher; and the Spray and Wait algorithm after spraying copy into the wait stage, the news spread has certain blindness, and when the destination node takes some time and probability, thus the delivery rate is low. Compared with the LARAC algorithm, 2H algorithm adds two hop ACK confirmation mechanism can effectively achieve the message in the back of the sparse nodes, increase the forwarding path of the message, so when the number of nodes is less than 100, 2H algorithm LARAC algorithm is slightly higher than the rate of delivery. When the number of nodes is greater than 100, the density of nodes becomes dense, and the

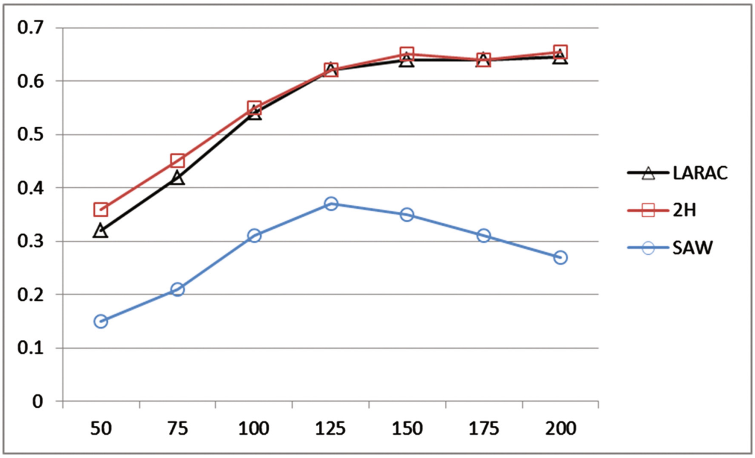


Fig. 5. Submission rate of urban roads

number of routing holes becomes less. The advantage of 2H algorithm is no longer obvious, so it has little difference with the LARAC algorithm.

From the overall trend, with the increase in the number of nodes, three algorithms of the delivery rate rising, when the number of nodes to 125 ~ 150, because the Spray and Wait algorithm for spraying congestion phenomenon began to copy more, network, delivery rate declined; while the LARAC and 2H algorithm by single copy and forwarding mode, in the network the copy number is less, the congestion phenomenon is not obvious, so the delivery rate is more stable.

2. Transmission delay

Transmission delay is the average value of the time delay for a message that has been successfully submitted. In Fig. 6, the transmission delay of LARAC and 2H algorithm is better than Spray and Wait algorithm. In the number of nodes is small, the LARAC algorithm on the mobile speed forwards the message to the next hop, and the 2H algorithm based on moving speed on the fallback mechanism of message, thus increasing the chance of forwarding message. Therefore, when the number of nodes is less than 100, the 2H algorithm has more advantages than LARAC, the transmission delay is the shortest, when the number of nodes is greater than 100, the network performance of the two is not the same, the transmission delay is similar.

With the increase of the number of nodes, the forwarding of messages in the network becomes smooth, and the transmission delay of the three algorithms decreases gradually. When the node number reaches 125, the Spray and Wait algorithm to produce congestion, the transmission delay increases gradually, while LARAC and 2H as the forwarding mode of single copy, will not produce congestion, so the transmission delay remains small value.

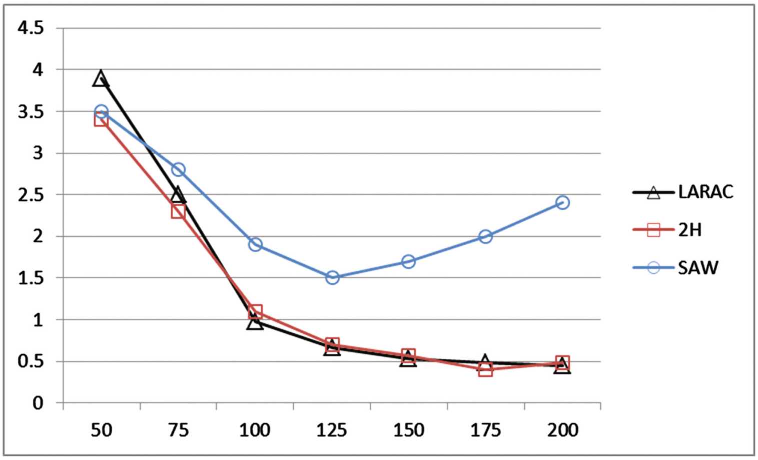


Fig. 6. Average delay of urban road

1. Network overhead

It can be concluded from the Fig. 7, three algorithms, Spray and algorithm of Wait network overhead, network overhead 2H and LARAC algorithm, which is composed of Spray and Wait is a multi copy forwarding algorithm, while LARAC and 2H is the essence of single copy forwarding algorithm decision. Compared with the LARAC algorithm, 2H algorithm can make the message back in the sparse area, increasing the number of message copies, so in the case of a small number of nodes, 2H algorithm network overhead than LARAC. With the increase of the number of nodes, the multi copy algorithm will cause congestion, so when the number of nodes is greater than 150,

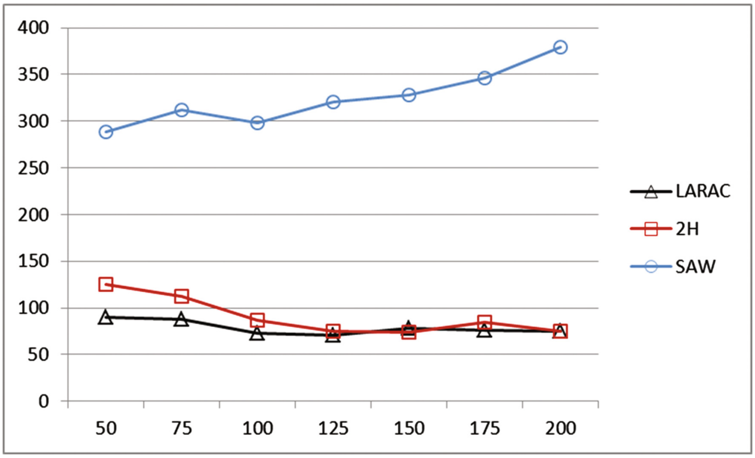


Fig. 7. Urban road network overhead

the network overhead of Spray and Wait algorithm is on the rise, but the 2H and LARAC algorithms are not affected.

It can be concluded from the above analysis results in complex scenes such as city road such a large number of communication nodes, heavier tasks in the Spray and Wait flooding algorithm has a lot of disadvantages, it is easy to cause network congestion. While LARAC and 2H single copy forwarding algorithm based on geographic location is the performance of good communication performance, not only the message delivery rate is higher, the average delay is short, but also save network overhead, can effectively solve the congestion problem in DTN network, strong applicability.

4 Conclusion

This paper proposes a backup copy forwarding algorithm based on location information, this Routing algorithm can perceive the changes of node density and adaptive adjustment of position and velocity vector effect on the next hop forwarding nodes in different environments, improve the message delivery rate, control the network overhead.

References

1. Pereira, P., Casaca, A., Rodrigues, J., et al.: From delay-tolerant networks to vehicular delay-tolerant networks. IEEE Commun. Surv. Tutorials **14**(4), 1166–1182 (2012)
2. Sok, P., Kim, K.: Distance-based PROPHET routing protocol in Disruption Tolerant Network. In: IEEE International Conference on ICT Convergence, pp. 159–164 (2013)
3. Park, H.S,, Jang, J.H,, Lee, S.H., et al.: Position-based DTN routing in metropolitan bus network. In: International Conference on Systems and Informatics, ICSAI, pp. 1449–1453. IEEE, Yantai (2012)
4. Cheng, P.C., Weng, J.T., Tung, L.C., et al.: GeoDTN + Nav: a hybrid geographic and DTN routing with navigation assistance in urban vehicular networks. Mob. Netw. Appl. **15**(1), 61–82 (2010)
5. Parvathi, P.: Comparative analysis of CBRP, AODV, DSDV routing protocols in mobile Ad-hoc networks. In: IEEE International Conference on Computing, Communication and Applications, pp. 1–4 (2012)
6. Jingfeng, X.: Advanced PROPHET routing in delay tolerant network. In: IEEE International Conference on Communication Software and Networks, pp. 411–413 (2009)
7. Miraldaa, C., Ilira, S., et al.: A simulation system based on ONE and SUMO simulators: performance evaluation of different vehicular DTN routing protocols. J. High Speed Netw. **23** (1), 59–66 (2017)
8. Prusty, A.R., Nayak, A.: A hybrid multi-hop mobility assisted heterogeneous energy efficient cluster routing protocol for Wireless Ad hoc Sensor Networks. J. High Speed Netw. **22**(4), 265–280 (2016)

Cyber-Attack Risks Analysis Based on Attack-Defense Trees

Wenjun Sun[1(✉)], Liqun Lv[1], Yang Su[1], and Xu An Wang[1,2]

[1] Department of Electronic Technology,
Engineering University of the People's Armed Police Force,
Xi'an, Shaanxi, China
sunwenjun94@163.com

[2] Xidian University, Xi'an, Shaanxi, China

Abstract. Considering the lack of theoretical analysis for systems under complicated attacks, a framework was proposed to analyze attack risks based on attack-defense trees. The attack period was divided into attack phase and defense phase and metrics was defined. First, action nodes were constructed by collecting system vulnerabilities and capturing invasive events, and defense strategies were mapped to defense nodes in the tree structure. Besides, formal definitions were given and attack-defense tree with metrics was constructed using ADTool and relevant algorithms. In addition, concepts of ROA (Return on attack) and ROI (Return on Investment) were introduced to analyze system risk as well as to evaluate countermeasures. Finally, a risk analysis framework based on attack-defense trees was established and numerical case was given to demonstrate the proposed approach. The result showed that the framework could clearly describe the practical scenario of the interaction between attacks and defenses. The objective of risk analysis and countermeasures evaluation could be achieved.

1 Introduction

Cyber-attacks are becoming one of the main threats of cyber security of critical infrastructures (CI) and information systems since the last decade [1]. Recent cybercrimes and cyber espionages have shown that stealthy and sophisticated attacks, such as advanced persistent threats (APT) will do great harm to information systems. For example, the famous security corporation RSA suffered from the compromise of private key server; Google e-mail servers were infiltrated and intercepted and the clients' information was leaked. Great economic and reputational damage came with such cyber-attacks [2].

Considerable countermeasures have been taken for the sake of information systems security. However, most current defending techniques based on border protection are of little effect faced with targeted and complicated attacks because they mainly focus on one-shot known types [3]. But to improve information protection, the interaction between attackers and defenders must be considered.

In this paper, a risk framework based on attack-defense trees to analyze the cyber-attack risks by calculating the benefits of both sides is proposed. Several metrics

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_67

were defined as quantitative analysis. ROA (Return on attack) and ROI (Return on Investment) were introduced to illustrate the impact of taking relative countermeasures towards attacks. Besides, algorithms of how to generate attack-defense trees were given and ADTool [4, 5] was made use of as well. At last, the approach was demonstrated through a numerical case.

The remainder of the paper is as follows. In Sect. 2 we summarize related work on modeling attack and defense with tree structure. Our own framework is declared in Sect. 3. Application and numerical illustrations are depicted in Sect. 4. Finally, we discuss our results and draw conclusions.

2 Related Work

Attack tree has been widely utilized to systemically analyze attacks risks, which can implicitly illustrate the attack path. The concept of attack tree model was first introduced by Schneier [6]. In [7], the attack tree model was extended by adding attack scenarios and profiles. However, attack tree only works from the perspective of attackers and is complicated in visualization. To show the effect of defense mechanism, Edge et al. proposed protection trees from the perspective of defenders [8]. In [9], Bistarelli et al. proposed the defense tree model. But neither the protection tree nor defense tree is able to be employed without attack tree. To solve this problem, Roy et al. introduced attack countermeasure tree to combine attack and defense yet it's too complicated to be realized [10].

In [11, 12], Kordy et al. proposed attack-defense tree (ADTree) which combines attack tree and defense tree to one structure. ADTree describes the interactions between attacker and defender and the iterative counteraction for after the actions of both. Therefore, it can clearly show the system risks before and after the implementation of countermeasures towards specific domains. For the convenience of application, Kordy et al. later proposed tree construction tool namely ADTool to generate ADTree. By numerating system risks due to vulnerabilities and attack success possibility, the ADTree can be well applied to practical cases such as vehicle network [13] and CPS network [14] hence we employ it as the foundation of our analysis.

3 Modeling with ADTree

3.1 Attack-Defense Tree Model

In an attack-defense tree model, the scenario is divided into the attack phase and defense phase and the properties are abstracted as nodes. The targeted node of the attacker is the root of the tree and to complete his compromise, the attacker has to start exploiting from the leaf node and move progressively layer by layer until managing the invasion of the root node. Meanwhile, the defenders have to take countermeasures relative to each node in the attack path to keep the attacker from continuing his move. Attention that during each move of both sides there is a cost of move. To better understand the model, the formal definition of ADTree is as the following.

Definition 1. The ADTree is a triad $ADT = (N, E, R)$. $N = (N_a, N_b)$ is the set of nodes the tree while N_a represents the set of attack nodes which is also the property node of the system compromised and N_d represents the set of defense nodes which, in other words, represents the defense countermeasures. We also define $Pa(N)$ as the parent node set of N. $E = \left(N_i, N_j\right)$ represents the edge between N_i and N_j. $R = (AND, OR)$ is defined as the relations of attacks. In this paper, the basic relation operators are "AND" and "OR", which mean that the attacker/defender has to complete all his attack/defense to move on to the higher layer and the attacker/defender just needs to complete at least one respectively.

An instance is given in Fig. 1 to illustrate the structure of ADTree. Notice that the circular nodes are the attack nodes and the rectangular nodes represent the defense nodes. Corresponding defense countermeasures are depicted as the dotted line. Child nodes with arc represent *AND* operation nodes while those without represent OR operation nodes.

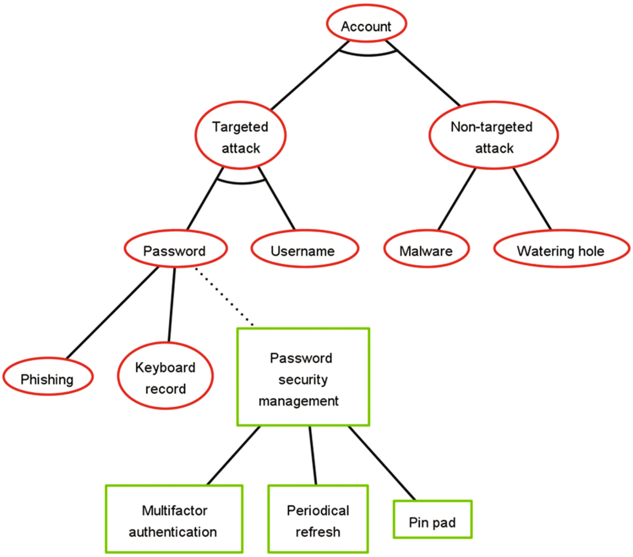


Fig. 1. An instance of attack-defense tree. This shows the possible attack path of illegally obtaining accounts and corresponding countermeasures towards password safety

3.2 Risk Analysis Framework with ADTree

Based on the ADTree theory and some concepts in [14], we establish the risk analysis framework by introducing several metrics. Our goal is to evaluate the risks resulting from potential attacks and the effects of countermeasures undertaken.

Step 1. Understanding system vulnerability
Attackers always take good advantage of vulnerabilities to exploit information system. It cannot be denied that some cyber-attacks, APT attacks for example, utilizes unidentified vulnerabilities such as 0 day vulnerabilities, but most do not. Common

system vulnerabilities could be found on the lists of CVE (Common Vulnerabilities and Exposures) and defenders can score them with CVSS (Common Vulnerabilities Scoring System)[1].

Step 2. Gathering attack information
After understanding system vulnerabilities, corresponding countermeasures should be made and attack path should be predicted according to the occurrence probability and extent of damage. Defense cost need to be taken into consideration as well. Attack information such as attack target, attack nodes, attack success probability, attack/defense cost and impact loss could be obtained from detection of attacks and vulnerability scanning. Besides, attack behavior database is also reference which needs regular updates. For the sake of convenience, definitions of such information are as follows.

Definition 2. Attack success probability p_i: the possibility of successfully committing an attack through risk $i(i = 1, 2, \ldots, m)$ which ranges from 0 to 1.

Definition 3. Attack cost $c_i \in (0, \infty)$: the resource required to commit an attacks, including human resource and physical resource needed.

Definition 4. Defense cost $d_i \in (0, \infty)$: the resource required to undertake countermeasures, capital of purchasing and employing security equipment and human resource included.

Definition 5. Potential loss l_i: the potential loss that may be resulted from attacking through risk i and can be divided into 1 to 10 levels according to the severity.

Step 3. Constructing ADTree
After completing step 1 and step 2, it is necessary to construct ADTree model for attack and defense

The risk i is composed of atom attacks numbering $1, 2, \ldots, n$ and can be expressed as the son nodes of one attack node. For the simplicity of calculation and comparison, monetary unit is introduced as a measure of attack costs and protection costs. Human resource consumed can be regards as monetary units, such as 100 *dollars* per hour. Assuming that the attacker employs his attack through risk i at time t, the defender shall undertake responding measures at time $t + 1$ after monitoring the attack. Therefore, the values of c_i, d_i and l_i shall change as a result of defense action. t is regarded as the attack time and $t + 1$ as the defense time. First, the expressions of metrics at attack time are as shown in Table 1. Notice that under the different relations of "AND" and "OR", the expressions differ.

Extents of system risk can be reflected in p_i and l_i. The greater p_i and l_i are, the more risks the system is facing. Besides, attack cost matters and rational attackers tend to choose the attack profile which costs less. As a consequence, system risk assessment metrics $r_i(t)$ can be expressed as

[1] Available at https://www.first.org/cvss.

Table 1. Metric equation during attack phase (t)

	AND	OR
Attack success probability	$p_i(t) = \prod_{k=1}^{n} p_k(t)$	$p_i(t) = 1 - \prod_{k=1}^{n} (1 - p_k(t))$
Attack cost	$c_i(t) = \sum_{k=1}^{n} c_k(t)$	$c_i(t) = \frac{\sum_{k=1}^{n} p_k(t) \times c_k(t)}{\sum_{k=1}^{n} p_k(t)}$
Potential loss	$l_i(t) = \frac{10^n - \prod_{k=1}^{n} (10 - l_k(t))}{10^{n-1}}$	$l_i(t) = Max_k^n(l_k(t))$

$$r_i(t) = \frac{p_i(t) \cdot l_i(t)}{c_i(t)} \tag{1}$$

Now that the basic metrics have been defined, it is important to construct ADTM (ADTree with metrics). The key algorithm pseudocode is as follows.

```
Algorithm 1 ADTM_generation( $ADT, metrics, r$ )
input: $ADT = (N_a, E, R)$ , $metrics = (p, c, l)$
output: $ADTM = (N_a, E, R, p, c, l, r)$
init $ADTM$ ;    /* Initialize each metric of $ADTM$ empty */
set $N_a, E, R$ ;    /* Copy each attack node, edge and relation of $ADT$ to
$N_a, E, R$ */
for (each leaf node $i$ in $ADT$ )
    set $p_i, c_i, l_i$ ;
end for (03)
for(each parent node $j$ )
    if ( $metrics(j) == null$ && $metrics(childnode(j))$ != $null$ )
        compute $p_j, c_j, l_j$ ;
        compute $r_i$ ;
        $j = parentnode(j)$ ;
    else if ( $k == childnode(j)$ && $metrics(k) == null$ )
        $j = k$ ;
    end if (07)
end for (06)
return $ADTM$
```

Step 4. Countermeasures implementation
The defender implements corresponding countermeasures to counter with attacks or to diminish the possibility of potential attacks hence the attack cost, defense cost and potential loss are not as the same as what they are at t. It is difficult to determine the value of attack success probability p_i as it changes as the attack-defense environment.

First, for the convenience of analysis, assuming that p_i keeps stable during the time interval $[t, t+1]$ namely $p_i(t) = p_i(t+1)$.

Define the increment of attack cost due to the implementation of defense actions as $\Delta c_k(t)$ at $t+1$. Theoretically, $\Delta c_k(t)$ is proportional to the value of defense cost. With the scale factor λ, the incremental equation is as follows:

$$\Delta c_k(t) = \lambda \times d_k(t) \tag{2}$$

λ is influenced by security strategy, security operation and personnel training. Meanwhile the potential loss can be updated at $t+1$

$$l_i(t+1) = \alpha \times l_i(t) \tag{3}$$

where $\alpha = 1 - \varphi$ represents surplus factor as a representative of the vulnerability rate that cannot be repaired due to the capability constraints of defenders. The formulation of φ is defined as

$$\varphi(t+1) = \frac{N_g(t+1)}{N_c(t+1)} \tag{4}$$

where N_g represents the number of vulnerability that can be repaired through undertaking countermeasures and N_c represents the number that cannot. From the equations above, metrics at $t+1$ can be derived as numerated in Table 2.

Table 2. Metric equation during defense phase $(t+1)$

	AND	OR
Defense cost	$d_i(t+1) = \sum_{k}^{n} d_k(t+1)$	$d_i(t+1) = \frac{\sum_{k=1}^{n} p_k(t) \times d_k(t+1)}{\sum_{k=1}^{n} p_k(t)}$
Attack cost	$c_i(t+1) = \sum_{k=1}^{n} c'_k(t)$	$c_i(t+1) = \frac{\sum_{k=1}^{n} p_k(t) \times c'_k(t)}{\sum_{k=1}^{n} p_k(t)}$
Potential loss	$l_i(t+1) = \sum_{k}^{n} \alpha \times l_k(t)$	$l_i(t+1) = Max_{k=1}^{n}(\alpha \times l_i(t))$

Step 5. Risk analysis
In order to evaluate system risk, the concepts of *ROA* (Return on Attack) and *ROI* (Return on Investment) are defined as follows.

Definition 6. *ROA*: the expected return rate of the attacker after his investment on the attacks. Its formulation is

$$ROA(t+1) = \frac{p_i(t) \times l_i(t+1)}{c_i(t+1)} \tag{5}$$

Definition 7. *ROI*: the expected return rate of the defender after his investment on the defense actions for the system security. Its formulation is

$$ROI = \frac{\Delta ALE}{CI} \tag{6}$$

In (6), ΔALE is the differential of loss resulting from the attacker after and before the implementation of countermeasures, expressed as $ROA(t+1) - ROA(t)$. While CI is the countermeasures cost of defenders which can also be represented as $d(t+1)$. The reason to define as this is to associate ROA and ROI to evaluate the effects of countermeasures. Consequently, (6) is turned to

$$ROI = \frac{ROA(t+1) - ROA(t)}{d(t+1)} \tag{7}$$

Now it's necessary to update Algorithm 1 to generate UADTM (updated ADT with metrics). The key algorithm pseudocode is as follows.

```
Algorithm 2. UADTM_generation( $ADTM$ , $N_d$ , $d$ )
input: $ADTM = (N_a, E, R, p, c, l, r)$ , $N_d$ , $d$
output: $UADTM = (N_a, N_d, E, R, p, c, l, ROA, ROI)$
init $UADTM$ ;
set $N_a, E, R, p, c, l, r$ ;
if( $defense\_state[i] == True$ )
    insert $N_{di}$ ;
    set $d_i$ ;
end if (03)
for(each defense node $i$ )
   compute $d_i$ ;
end for (07)
for(each parent node $j$ )
   if( $childnode(j) \in N_d$ )
        update $metrics(j)$ ;
    else if( $updates(metrics(childnode(j))) == True$ )
        update $metrics(j)$ ;
    end if (11)
end for (10)
compute $ROA$ , $ROI$ ;
return $UADTM$
```

Considering that system risk can be represented with attack utility, it's reasonable that ROA and $r(t)$ have the same expression to simplify the analysis. Therefore, the risk value is substituted by ROA in Algorithm 2.

From the perspective of the attacker, the goal is to maximizing ROA while minimizing the attack cost; while for the defender, the goal is to maximizing ROI while keeping the defense cost to the least level. Therefore, the defender shall consider how to minimize ROA and for the attacker, on the contrary, is to minimize ROI.

4 Risk Analysis Framework

In this section, a framework towards network attacks will first be established according to the metrics and definitions above, as shown in Fig. 2. Then numerical illustrations are given as a demonstration.

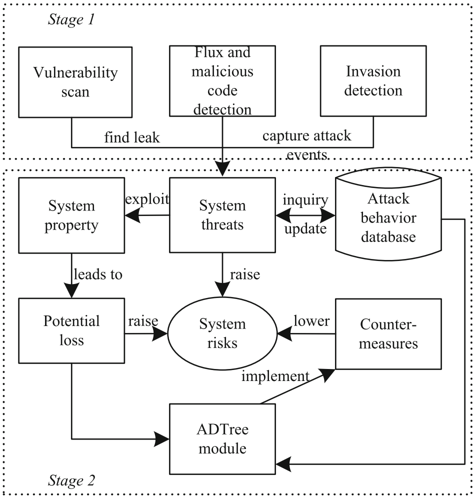


Fig. 2. Framework of system risk analysis based on ADTree including two stages

4.1 Framework Construction

Based on the approach given, the framework of network risk analysis could be established as follows. The process is composed of the system risk understanding and the construction of ADTree.

Stage 1. Understanding system risks

The main task in this phase is to collect system vulnerability and attack information detected.

First, network properties shall be modeled and assigned values. Then techniques such as vulnerability scanning, flux monitoring and malware detection are utilized to understand the risk information. Besides, potential attack path could be illustrated and the loss shall be estimated thru the inquiry of attack behavior database. Risks can also be scored referring to CVSS.

Stage 2. Establishing attack-defense tree analysis framework

After gathering the necessary information, relative metrics before and after the implementation of countermeasures need to be taken into consideration. Based on the five steps proposed, analysis framework could be established through the following three steps.

(1) ADTree construction. By employing ADTool, input the values of the parameters at t and construct the tree based on Algorithm 1 proposed to generate ADTM.
(2) Countermeasures undertaken. System metrics change at the defense time $t + 1$ and need to be updated and generate UADTM based on Algorithm 2.
(3) Risk evaluation. After generating ADTM and UADTM, values of *ROA* and *ROI* of each node need to be calculated as the reference.

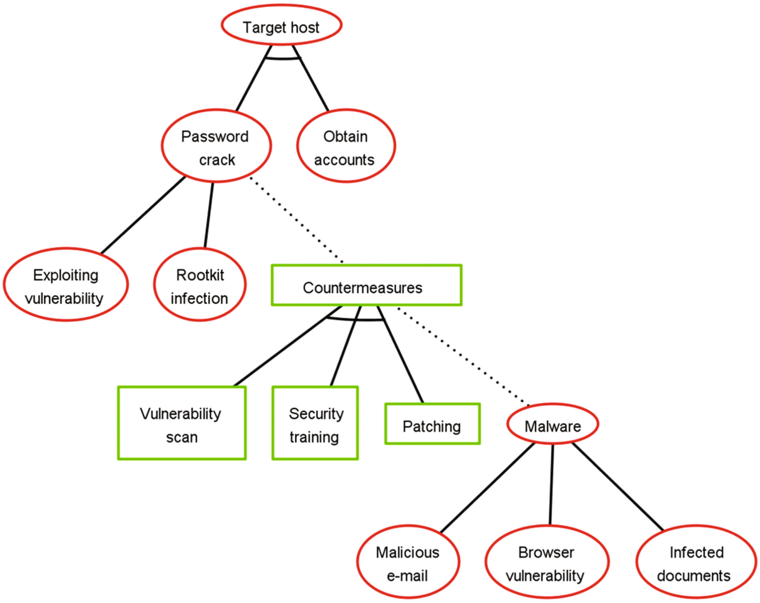


Fig. 3. ADTree structure for modeling attack path and corresponding countermeasures

4.2 Numerical Illustrations

In this section, numerical illustrations are given to demonstrate the framework proposed. As for the possible attack path, we consider Night Dragon attack [15], one example of APT attacks, whose goal is to infect target hosts, install remote control tools, establish stealthy transfer tunnel and steal confidential documents. The ADTree of the attack and some defense actions are shown in Fig. 3.

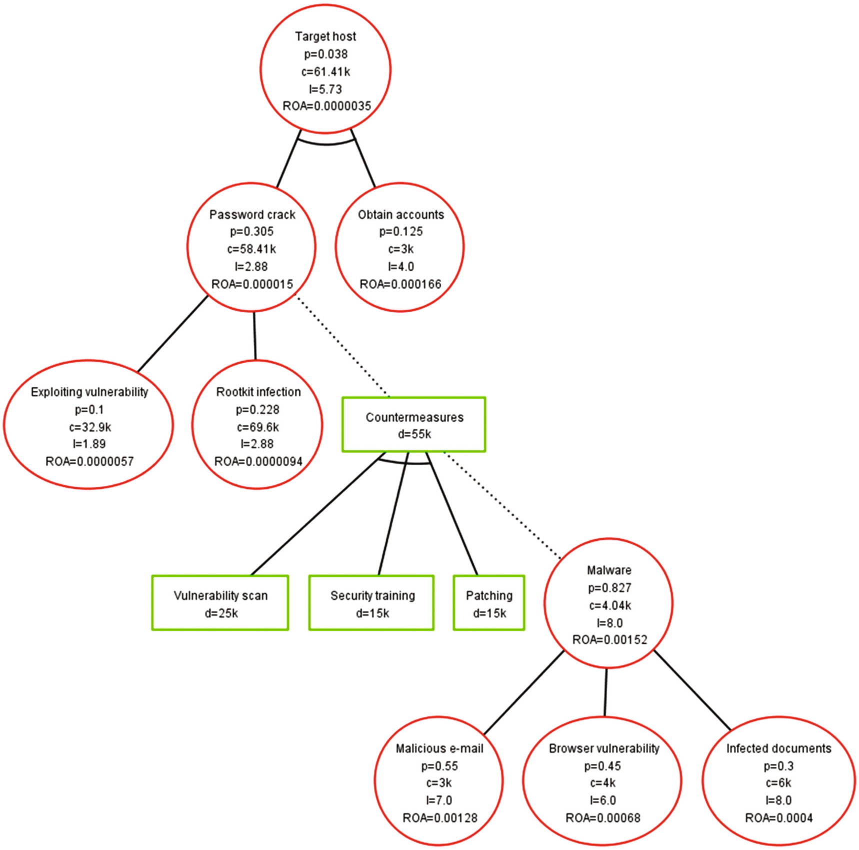


Fig. 4. ADTree structure for modeling attack path and corresponding countermeasures with metrics. This structure is in the case of $\lambda = 0.5, \alpha = 0.3$

In the case of $\lambda = 0.5, \alpha = 0.3$, the updated metrics are shown in Fig. 4. The calculation of each metric has been defined in the previous page. As is shown in Fig. 4 and Table 3, when the defender implements countermeasure worth 55 k dollars toward the node of *password crack*, the attack cost increment is 22.81 k dollars. Attention that for the convenience of analysis, attack success possibility p is assumed to remain unchanged. The loss impact drops from 7.78 to 5.53 and the values of *ROA* diminishes by 54.55%. It can be inferred that by taking specific defense actions, the risk of

Table 3. Metric values and variations of the password crack node

Metrics	Values before defense	Values after defense	Variations
ROA	0.0000077	0.0000035	−54.55%
p	0.038	0.038	–
c	38.6 k	61.41 k	+59.09%
l	7.78	5.73	−26.35%
d	0	55 k	–

password cracked reduces by 54.55%. From Fig. 4, it can also be inferred that both the attacker and defender can learn from the interaction of attack-defense. Considering the persistent characteristics of current cyber-attacks, the process can be derived iteratively between the attacker and defender. The closer is the attack node to the root node, the more the corresponding defense cost is while defense cost comes to the least on the leaf nodes. This illustrates that countermeasures should be implemented as soon as the attack has been detected. Besides, attacks might be deterred if the attack cost is too high while the return on attack is little as a consequence of defense actions.

5 Conclusion

Considering the interaction of the attacker and defender, a framework of tree structure to evaluate the system risks caused by network attacks was established based on the theory of attack-defense tree. By constructing ADTree for specific attack-defense scenario and calculating the values of return on attack, the risks of specific attack before and after the implementation of defense actions can be compared quantitatively. The paper also suggests that the defender should take defense measures as soon as possible once the detection of attacks. In addition, taking specific countermeasures may possibly deter attackers as a result of the increase of attack costs and decline in return. In the future work, optimal strategy will be studied instead of the just given statics. Besides, attackers are assumed to be rational to choose the least cost route in this paper. Behaviors of irrational attackers and specific scenarios will also be studied in the future work to extend the proposed framework.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (61572521).

References

1. Bencsáth, B., Pék, G., Buttyán, L., Felegyhazi, M.: The cousins of stuxnet: Duqu, flame, and gauss. Future Internet **4**(4), 971–1003 (2012)
2. Virvilis, N., Gritzalis, D.: The big four-what we did wrong in advanced persistent threat detection? In: 2013 Eighth International Conference on Availability, Reliability and Security (ARES), pp. 248–254. IEEE, September 2013
3. Laszka, A., Johnson, B., Grossklags, J.: Mitigating covert compromises. In: International Conference on Web and Internet Economics, pp. 319–332. Springer, Heidelberg, December 2013
4. Kordy, B., Kordy, P., Mauw, S., Schweitzer, P.: ADTool: security analysis with attack–defense trees. In: International Conference on Quantitative Evaluation of Systems, pp. 173–176. Springer, Heidelberg, August 2013
5. Gadyatskaya, O., Jhawar, R., Kordy, P., Lounis, K., Mauw, S., Trujillo-Rasua, R.: Attack trees for practical security assessment: ranking of attack scenarios with ADTool 2.0. In: International Conference on Quantitative Evaluation of Systems, pp. 159–162. Springer International Publishing, August 2016

6. Schneier, B.: Attack trees. Dobb's J. **24**(12), 21–29 (1999)
7. Moore, A.P., Ellison, R.J., Linger, R.C.: Attack modeling for information security and survivability. Carnegie-Mellon Univ Pittsburgh PA Software Engineering Inst (No. CMU-SEI-2001-TN-001) (2001)
8. Edge, K.S., Dalton, G.C., Raines, R.A., Mills, R.F.: Using attack and protection trees to analyze threats and defenses to homeland security. In: IEEE Military Communications Conference, MILCOM 2006, pp. 1–7. IEEE, October 2006
9. Bistarelli, S., Fioravanti, F., Peretti, P.: Defense trees for economic evaluation of security investments. In: The First International Conference on Availability, Reliability and Security, 2006, ARES 2006, pp. 8–pp. IEEE, April 2006
10. Roy, A., Kim, D.S., Trivedi, K.S.: Attack countermeasure trees (ACT): towards unifying the constructs of attack and defense trees. Secur. Commun. Netw. **5**(8), 929–943 (2012)
11. Kordy, B., Mauw, S., Radomirović, S., Schweitzer, P.: Foundations of attack–defense trees. In: International Workshop on Formal Aspects in Security and Trust, pp. 80–95. Springer, Heidelberg, September 2010
12. Kordy, B., Mauw, S., Radomirović, S., Schweitzer, P.: Attack–defense trees. J. Logic Comput. **24**, 55–87 (2012). exs029
13. Du, S., Li, X., Du, J., Zhu, H.: An attack-and-defence game for security assessment in vehicular ad hoc networks. Peer-to-peer Netw. Appl. **7**(3), 215–228 (2014)
14. Ji, X., Yu, H., Fan, G., Fu, W.: Attack-defense trees based cyber security analysis for CPSs. In: 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 693–698. IEEE, May 2016
15. Wueest, C.: Targeted Attacks Against the Energy Sector. Symantec Security Response, Mountain View (2014)

Multi-focus Image Fusion Method Based on NSST and IICM

Yang Lei(✉)

Department of Electronics Technology,
Engineering University of Armed Police Force, Xi'an 710086, China
kwwking321@163.com

Abstract. Multi-focus image fusion is a classic issue in the field of image processing. How to extract and fuse the in-focus information from the source images into the single one is the key to resolving the above problem. As a novel multi-resolution analysis tool, non-subsampled shearlet transform (NSST) not only has better information capturing ability, but also owns a comparatively lower computational complexity compared with non-subsampled contourlet transform (NSCT). Intersecting cortical model (ICM) is the third generation of artificial neural network, and it can be viewed as the improved version of pulse-coupled neural network. The superiority of ICM lies in that it has much fewer parameters and better function mechanism. In this paper, a novel method for multi-focus image fusion based on NSST and improved ICM is presented. On the one hand, NSST is responsible for decomposing source images and reconstructing sub-images. On the other hand, ICM is used to complete the coefficients selecting of sub-images. Experimental results demonstrate that the proposed method has better performance compared with the current typical ones.

1 Introduction

Image fusion has been a research hotspot in the field of computer vision because of its better visual performance and visual information. At present, image fusion technology has been paid more and more attention in the world, and has been widely used in many fields, such as medical imaging [1] and remote sensing [2].

In the field of image fusion, multi-focus image fusion is a classical problem. Typically, we get at least two different images from the imaging sensor scene, but the focus and the focal area may differ from each other. For example, in image A, the left region is in focus, and the corresponding region in the image B is the focus. Therefore, it is necessary to extract the focus regions from the source images and fuse them into a single image to guarantee the results with excellent visual performance. In the past few years, a variety of fusion methods [3–30] have been proposed for multi-focus images.

Generally, these methods can be classified into two types, including the spatial domain (SD) and the transform domain (TD). The pioneer in SD is the weighted technique (WT) [3], and the final fusion image is estimated by weighted compromise between the pixels of the same spatial location in the corresponding input. WT can resist the noise in the input to some extent, but it always leads to the decrease of the contrast of the fused image. In addition, principal component analysis (PCA) [4] and

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_68

independent component analysis (ICA) [5] theory have been used for image fusion as well. However, the methods based on PCA and ICA put forward high requirements for component selection. Recently, as the third generation of artificial neural network (ANN), pulse coupled neural network (PCNN) [6], as well as its wide version, for example, intersecting cortical model (ICM) [7] has been proposed and widely used in image fusion problem. Essentially, ICM is an improved version of the traditional PCNN model. The above two models are able to simulate the process of biological pulse excitation and capture the intrinsic information of the source image. Compared with PCNN, ICM is more efficient because of its simple mechanism and fewer parameters. Unfortunately, little is known about the application of ICM.

On the other hand, there are some methods based on TD. Discrete wavelet transform (DWT) is considered as an ideal fusion technique. However, further studies show that DWT still has its inherent limitations. First of all, it is only good at capturing the point wise singularity, but the poor performance of edge expression. Second, it captures only limited directional information only along vertical, horizontal and diagonal directions [8]. In order to overcome the shortcomings of wavelet transform, an extensive series of improved models have been proposed, such as quaternion wavelet transform (QWT) [9], ridgelet transform (RT) [10], curvelet directional transform (CDT) [11], quaternion curvelet transform (QCT) [12], contourlet transform (CT) [13] and shearlet transform (ST) [14]. However, the performance of the proposed model is severely limited because there is no shift invariance property introduced by the down sampling process. CT is an extension of the translation invariance, that is, the non-subsampled contourlet transform (NSCT) [1, 15] has been exploited, but its computational complexity is higher compared to the above MSA technique. Easley *et al.* presented an improved version of the saint called non-subsampled shearlet transform (NSST) [16] which has not only the ability to capture the feature information of the input image, but also much lower computational resources with NSCT.

In this paper, a new technique and the improved ICM NSST for multi focus image fusion. The core idea includes three stages. First of all, the smoking area is used to decompose the source image into a series of sub images. Secondly, an improved ICM is proposed for the fusion of sub images. Finally, the final fusion image is reconstructed by inverse NSST.

The rest of this article is organized as follows. The improved ICM is presented in the second section, followed by a fusion framework for multi focus images in section third. The experimental results of correlation analysis are reported in section fourth. The conclusion is summarized in the fifth section.

2 Improved Intersecting Cortical Model

2.1 Basic ICM

As the third generation of artificial neural network, ICM is a model of cat primary visual cortex, which is formed by a large number of neurons. An intersecting cortical neuron commonly denoted by N_{ij} consists of three parts: receptive field, modulation and pulse generator. The corresponding discrete mathematical expressions of the basic ICM can be described as follows:

$$F_{ij}[n] = fF_{ij}[n-1] + S_{ij} + W_{ij}\{Y[n-1]\} \tag{1}$$

$$Y_{ij}[n] = \begin{cases} 1 & if\ F_{ij}[n] > \theta_{ij}[n-1] \\ 0 & else \end{cases} \tag{2}$$

$$\theta_{ij}[n] = g\theta_{ij}[n-1] + hY_{ij}[n] \tag{3}$$

Different from PCNN, the basic intersecting cortical neuron N_{ij} receives input signals via external sources by only one channel in the receptive field namely the feeding input F_{ij} which corresponds to Eq. (1). If F_{ij} is larger than θ_{ij}, then the neuron N_{ij} will be activated and generate a pulse, which is characterized by $Y_{ij} = 1$, else $Y_{ij} = 0$. The above course corresponds to Eq. (2). According to Eq. (3), the dynamic threshold θ_{ij} will decline with the iterative number n increasing. However, if $Y_{ij} = 1$, θ_{ij} will immediately raise with the function of V_θ whose value is relatively large, so that the behavior of firing of N_{ij} will stop at once and $Y_{ij} = 0$. Later, if θ_{ij} reduces to be equal to or less than F_{ij}, N_{ij} will fire again and make an impulse sequence. On the other hand, the relations between N_{ij} and its surrounding neurons exist, therefore, if N_{ij} is activated, those neurons having similar gray values around it may also be activated at the next iteration. The result is an auto wave expanding from an active neuron to the whole region.

The ICM used for image fusion is a single layer two-dimensional array of laterally linked pulse coupled neurons. The number of neurons in ICM is equal to that of pixels in the input image, and all neurons in the network are considered to be identical. There exists a one-to-one correspondence between the image pixels and network neurons. Commonly, S_{ij}, the gray value of each pixel, is directly referred to as the external stimulus of N_{ij}.

Apart from the parameters mentioned above, there are still several ones required explaining. f and g are the magnitude scaling terms. W_{ijkl} is the linking matrix corresponding to F channel.

2.2 Improved ICM

Figure 1 shows the structure of an improved ICM neuron whose corresponding discrete mathematical expressions can be described as:

$$F_{ij}[n] = S_{ij} + W_{ij}\{Y[n-1]\} \tag{4}$$

$$Y_{ij}[n] = \begin{cases} 1 & if\ F_{ij}[n] \geq \theta_{ij}[n-1] \\ 0 & else \end{cases} \tag{5}$$

$$T_{ij}[n] = \begin{cases} n, \quad if\ Y_{ij} = 1 & for\ the\ first\ time \\ T_{ij}[n-1], & else \end{cases} \tag{6}$$

$$\theta_{ij}[n] = \theta_{ij}[n-1] - \Delta + hY_{ij}[n] \tag{7}$$

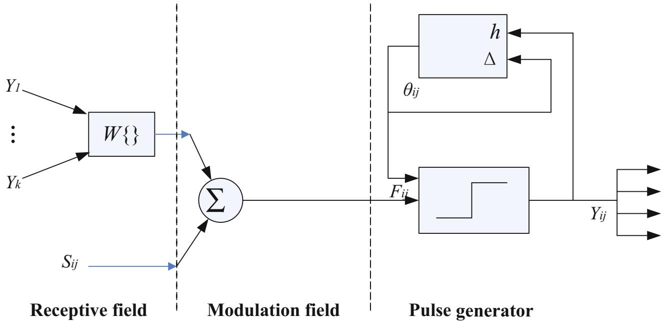


Fig. 1. The basic model of improved ICM neuron

An improved ICM neuron commonly denoted by N_{ij} is composed of three units: the receptive field, the modulation field, and the pulse generator. As shown in Fig. 1, N_{ij} receives input signals F_{ij} from other surrounding neurons via the matrix W_{ij} and external input S_{ij} in the receptive field, which corresponds to Eq. (4). Equation (5) indicates that F_{ij} is then compared with the dynamic threshold θ_{ij} to decide the value of the output Y_{ij}. If F_{ij} is larger than θ_{ij}, then the neuron N_{ij} will be activated and generate a pulse, which is characterized by $Y_{ij} = 1$, else $Y_{ij} = 0$. In Eq. (6), T is the time matrix, the iterative number n can be determined adaptively according to the intensity distribution of pixels in images. There are several aspects required to be noted: (1) T_{ij} will keep invariable if N_{ij} does not fire all the time; (5) if N_{ij} fires for the first time, T_{ij} will be set as the ordinal value of corresponding iteration; (6) once N_{ij} has already fired, T_{ij} will not alter again. Its value will be saved as the ordinal value of iteration, during which N_{ij} fired for the first time, and even N_{ij} may fire later. Once all pixels have fired, the whole iteration process is over, and the value of the largest element in T is the total iteration times. According to Eq. (7), Step Δ is a positive constant. θ_{ij} will decline with the iterative number n increasing. h is usually set as a relatively large value to ensure that the firing times of N_{ij} will not exceed one at most.

3 Fusion Framework of Multi-focus Images

We take the fusion process of two multi-focus images for instance. Suppose that two source images respectively denoted by A and B have been already accurately registered. F is the final fused image. The concrete steps of the fusion algorithm can be described as follows:

(a). Decompose the source images A and B into a pair of low-frequency sub-images $\left\{A_K^0, B_K^0\right\}$ and a series of high-frequency sub-images $\left\{A^{l_k}, B^{l_k}\right\}$ via NSST, where K denotes the level number of multi-scale decompositions and l^k is the stage number of multi-directional decompositions at the k^{th} level, $1 \leq k \leq K$.

(b). Select fused coefficients for each sub-image from A and B via improved ICM.

(b1). The coefficients of the sub-images from A and B are respectively converted into the external inputs to stimulate the corresponding improved ICM.

(b2). Initialize $T_{ij}^{0,l_k}[0] = Y_{ij}^{0,l_k}[0] = 0, \theta_{ij}^{0,l_k}[0] = \max\left\{\left|C_K^{0.l_k}\right|\right\}$ at the same time, let each pixel be inactivated. Where $\left|C_K^{0.l_k}\right|$ denotes the absolute value of each sub-image coefficient;

(b3). Compute $T_{ij}^{0,l_k}[n], Y_{ij}^{0,l_k}[n], \theta_{ij}^{0,l_k}[n]$ according to Eqs. (4)-(7).

(b4). Implement Step (b3) iteratively until all neurons have been activated, namely all of the elements in T amount to 1. Then the fused coefficients of low-frequency sub-images and high-frequency ones can be respectively chosen as follows:

$$f_0^K = \begin{cases} A_0^K, & if\ T_{ij,A}^0 \leq T_{ij,B}^0 \\ B_0^K, & if\ T_{ij,A}^0 > T_{ij,B}^0 \end{cases}, \quad f_{ij}^{l_k} = \begin{cases} A_{ij}^{l_k}, & if\ T_{ij,A}^{l_k} \leq T_{ij,B}^{l_k} \\ B_{ij}^{l_k}, & if\ T_{ij,A}^{l_k} > T_{ij,B}^{l_k} \end{cases} \tag{8}$$

(c). Reconstruct the fused image F by using an inverse NSST.

4 Experimental Results and Analysis

In this section, several illustrative experiments are done to demonstrate the effectiveness of our proposed method. The simulation is conducted in MATLAB 2013a on a PC with Intel Core i7/2.6 GHz/4G. Concretely speaking, this section consists of two parts: methods introduction and performance evaluation. To begin with, the proposed method and several current ones with their parameters settings are briefly introduced. Then, in order to test the effectiveness and efficiency of the proposed method, performance evaluation is implemented in succession from two aspects covering subjective visual effect and objective evaluation criteria.

4.1 Methods Introduction and Parameters Setting

The parameters setting of the proposed method is as follows: W = [0.707 1 0.707; 1 0 1; 0.707 1 0.707], Δ = 15, h = 500. The level of multi-scale decompositions is set as 4, and the number of direction from coarser to finer scale is set as 6, 10, 18 and 18, respectively. The size of the shearing window is set to be 3, 5, 9, and 9. The size of the neighborhood is 3×3. Note that it is not necessary for us to modify the parameters manually during the following simulation experiments. For simplicity, we term the proposed method M8.

In addition, seven current methods are utilized to compare with M8 in this section, which are Shearlet-based method (M1) [14], NSST-based method (M2) [31], LC-PCNN-based method (M3) [32], PCNN-based method (M4) [33], and SW-NSCT-PCNN-based method (M5) [34], Wavelets-based method (M6) [35], and Contourlet-Transform-based method (M7) [36]. In order to make the comparison reliable, the parameters settings in M1 $\sim$ M7 are still implemented according to the content in references [14, 31–36]. If you want to get more details, please refer to related references mentioned above.

In order to testify the superior performance of M8, extensive fusion experiments with a pair of images have been performed. The related source images with the size of 512 × 512 have been already accurately registered. The source images used in the following experiments cover 256 gray levels. Subjective evaluation system can be adopted to provide direct comparisons. However, it is easily prone to be affected by lots of personal factors, such as eyesight level, mental state, even the mood, and so on. As a result, it is very necessary for us to evaluate the fusion effects based on both subjective vision and objective quality assessment. In this paper, seven objective quality metrics are chosen as follows.

(a) Information entropy (IE)

IE directly reflects the amount of average information in fused image. The larger the IE is, the more abundant the information amount of the fused image is.

$$IE = -\sum_{i=0}^{L-1} P(i)\log_2 P(i) \tag{9}$$

Where $P(i)$ indicates the probability of pixels whose gray value amount to i over the total image pixels.

(b) Standard deviation (SD)

SD expresses the extent of deviation between the gray values of pixels and the average one of the fused image. In a sense, the quality of the fused image is in direct proportion to the value of SD.

$$SD = \sqrt{\frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}(f(i,j) - \frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}f(i,j))^2} \tag{10}$$

Where f denotes the final fused image whose size is $m \times n$.

(c) Average Grads (AG)

AG indicates the clarity extent and small details of the image. Similar to SD, better fusion effects can be commonly obtained with the AG value increasing.

$$AG = \frac{1}{m \times n}\sqrt{\left(\sum_{i=1}^{m}\sum_{j=2}^{n}|f(i,j) - f(i,j-1)|\right)^2 + \left(\sum_{i=2}^{m}\sum_{j=1}^{n}|f(i,j) - f(i-1,j)|\right)^2 /2} \tag{11}$$

(d) Root mean square error (RMSE)

RMSE is used to describe the difference extent between the fused image and the ideal standard fused image. The smaller RMSE is, the better the fused performance is.

$$RMSE = \sqrt{\frac{1}{m \times n}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(f(i,j) - S(i,j)\right)^2} \tag{12}$$

Where S denotes the ideal standard fused image.

(e) Peak signal-to-noise ratio (PSNR)

Here, we assume two points. On the one hand, the difference between the fused image and the ideal standard fused image is the noise. On the other hand, the information is the ideal standard fused image. PSNR reflects the level of noise restraints. Accordingly, the quality of the fused image is in direct proportion to the value of PSNR.

$$PSNR = 10\,\log\frac{255^2}{RMSE^2} \tag{13}$$

(f) Mutual information (MI)

MI can be used to reflect fused effects and measure the relativity between two or more images. The larger MI is, the more abundant information the fused image contains.

$$MI = \frac{\sum_{i=0}^{L-1}\sum_{k=0}^{L-1} P_{A,F}(i,k)\log\frac{P_{A,F}(i,k)}{P_A(i)P_F(k)} + \sum_{j=0}^{L-1}\sum_{k=0}^{L-1} P_{B,F}(j,k)\log\frac{P_{B,F}(j,k)}{P_B(j)P_F(k)}}{IE\_A + IE\_B} \tag{14}$$

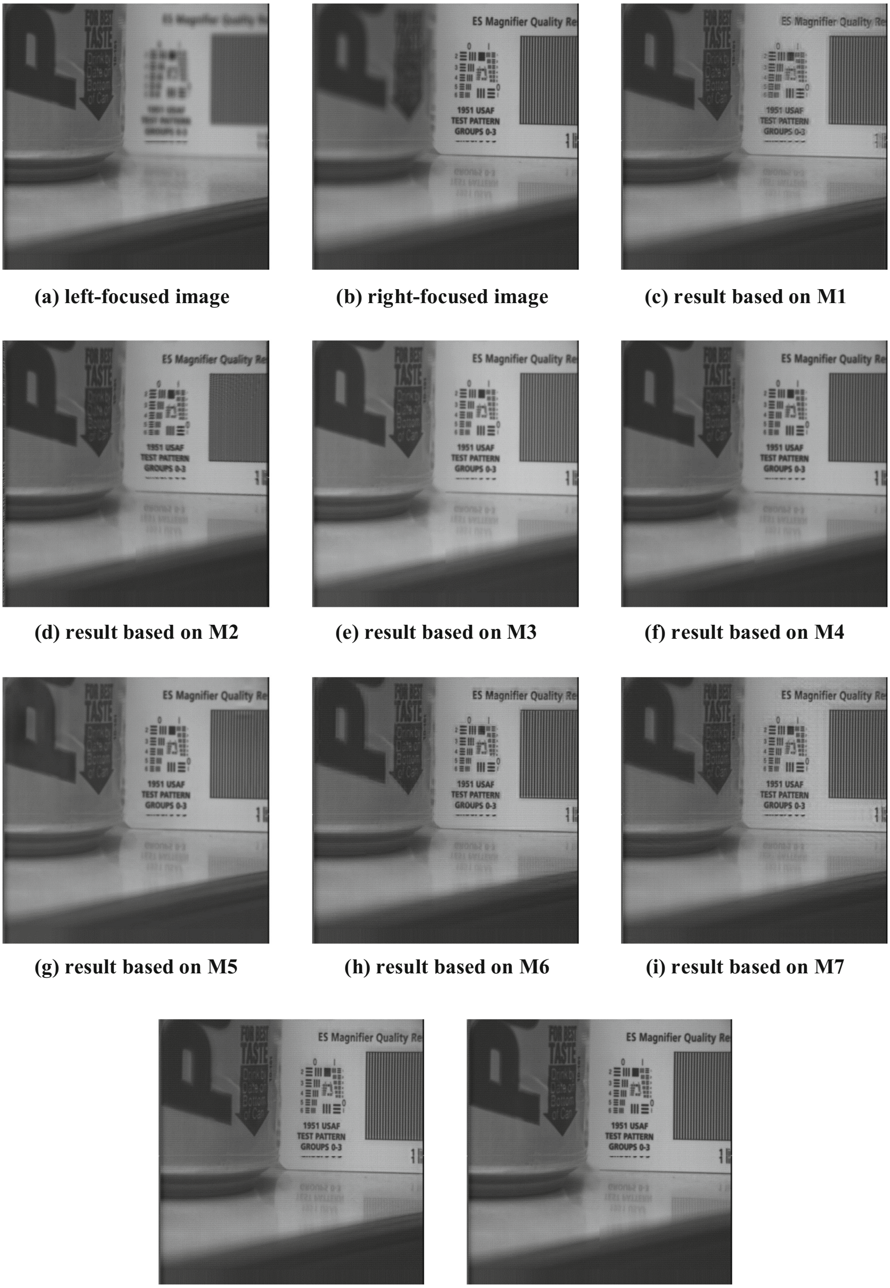
$P_{A,F}(i,\ k)$ and $P_{B,F}(j,\ k)$ are the normalized gray histogram between source image A and the fuse image F, the normalized gray histogram between source image B and F, respectively.

(g) Structural similarity index (SSIM index)

SSIM index describes the similarity of two inputting images. Large value shows both inputs are more similar.

4.2 Comparative Experiments of Multi-focus Image Fusion

Experimental results of multi-focus image fusion are shown in Fig. 2. Figure 2(a) and (b) are the corresponding source images. As obviously described in Fig. 2(a), the left area is in focus and holds a higher definition level than the right area, which is out of focus. On the contrary, the left area in Fig. 2(b) owns a lower definition level than that the right area owns. The fused images based on M1–M8 are shown in Fig. 2(c)-(j). Figure 2(k) shows the ideal standard fused image which is produced artificially. As seen in Fig. 2(c)-(j), the main information and characteristics of the two source images are fused by the methods M1–M8, and each of them has a good visual performance. However, by careful comparison, it is not difficult for us to find that the disparities among the eight methods still exist. The fused images based on M1 and M2 both have a



(a) left-focused image

(b) right-focused image

(c) result based on M1

(d) result based on M2

(e) result based on M3

(f) result based on M4

(g) result based on M5

(h) result based on M6

(i) result based on M7

(j) result based on M8

(k) ideal standard fused image

Fig. 2. Multi-focus source images and fused images based on M1-M8

comparatively low contrast compared with other methods; moreover, the Gibbs phenomena emerge in Fig. 2(c), (h) and (i). On the other hand, although the fused images based on M3–M5 have much better contrast effects than those based on M1, M2, M6 and M7, unfortunately, high clarity and proper contrast level are never combined simultaneously in any fused image of them. As revealed in Fig. 2(j), the fused image based on M8 is of higher clarity and reasonable contrast level, and has obvious details and edges. The results of objective evaluation based on the eight methods are listed in Table 1. The values of six metrics of M8 are the best. With regard to SD, M8 is only inferior to M3. It means that the fused effects based on the proposed method are much better than the other seven methods on the whole in terms of objective results, which is also nicely consistent with the visual effects.

Table 1. Comparison of the fusion methods for multi-focus images

Method	IE	SD	AG	RMSE	PSNR	MI	SSIM
M1	6.9649	43.3340	19.1551	2.3242	40.8052	0.7822	0.9041
M2	7.0172	42.8655	20.1918	2.2744	40.9933	0.8013	0.9196
M3	7.1544	**48.7633**	19.5401	2.1625	41.4317	0.8752	0.9709
M4	7.0057	45.5636	19.6717	2.1569	41.4542	0.8891	0.9763
M5	7.1309	47.2311	19.3970	2.2639	41.0337	0.8267	0.9455
M6	7.1278	32.2800	7.1388	4.4674	35.1297	0.6321	0.7596
M7	7.1177	32.1248	6.8711	4.9588	34.2233	0.6059	0.7257
M8	**7.1827**	47.7482	**24.3728**	**1.4528**	**43.0371**	**0.8973**	**0.9832**

5 Conclusions

In this paper, a new technique for image fusion based on NSST and improved ICM is introduced. Experimental results demonstrate that the proposed method has obvious superiorities over current typical ones. The optimization of the proposed method will be the focus in our future work.

Acknowledgements. The authors thank all the reviewers and editors for their valuable comments and works. The work was supported in part by the National Natural Science Foundations of China under Grant 61309008 and 61309022, in part by Natural Science Foundation of Shannxi Province of China under Grant 2014JQ8349, in part by Foundation of Science and Technology on Information Assurance Laboratory under Grant KJ-15-102, and the Natural Science Foundations of the Engineering University of the Armed Police Force of China under Grant WJY-201414.

References

1. Yang, Y., Que, Y., Huang, S., Lin, P.: Multimodal sensor medical image fusion based on type-2 fuzzy logic in NSCT domain. IEEE Sens. J. **16**, 3735–3745 (2016)
2. Ghahremani, M., Ghassemian, H.: Remote sensing image fusion using ripplet transform and compressed sensing. IEEE Geosci. Remote Sens. Lett. **12**, 502–506 (2015)

3. Burt, P.J., Kolcznski, R.J.: Enhanced image capture through fusion. Proc. Conf. Comput. Vis. **1**, 173–182 (1993)
4. Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., Benediktsson, J.A.: Model-based fusion of multi- and hyperspectral images using PCA and wavelets. IEEE Trans. Geosci Remot. Sen. **53**, 2652–2663 (2015)
5. Mitianoudis, N., Stathaki, T.: Optimal contrast correction for ICA-based fusion of multimodal images. IEEE Sens. J. **8**, 2016–2026 (2008)
6. Broussard, R.P., Rogers, S.K., Oxley, M.E., Tarr, G.L.: Physiologically motivated image fusion for object detection using a pulse coupled neural network. IEEE Trans. Neur. Net. **10**, 554–563 (1999)
7. Kinser, J.M.: Simplified pulse-coupled neural network. Proc. Conf. Appl. Arti. Neur. Net. **1**, 563–567 (1996)
8. Ali, F.E., El-Dokany, I.M., Saad, A.A., El-Samie, F.E.A.: Curvelet fusion of MR and CT images. Progr. Electromagn. Res. C **3**, 215–224 (2008)
9. Pertuz, S., Puig, D., Garcia, M.A., Fusiello, A.: Genaration of all-in-focus images by noise-robust selective fusion of limited depth-of-field images. IEEE Trans. Image Process. **22**, 1242–1251 (2013)
10. Do, M.N., Vetterli, M.: The finite ridgelet transform for image representation. IEEE Trans. Image Process. **12**, 16–28 (2003)
11. Candes, E.J., Donoho, D.L.: Curvelets: a surprisingly effective non-adaptive representation for objects with edges. Stanford University, CA (1999)
12. Cao, L., Jin, L., Tao, H., Li, G., Zhang, Z., Zhang, Y.: Multi-focus image fusion based on spatial frequency in discrete cosine transform domain. IEEE Signal Process. Lett. **22**, 220–224 (2015)
13. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multi-resolution image representation. IEEE Trans. Image Process. **14**, 2091–2106 (2005)
14. Miao, Q.G., Shi, C., Xu, P.F., Yang, M., Shi, Y.B.: A novel algorithm of image fusion using shearlets. Opt. Commun. **284**, 1540–1547 (2011)
15. Bhatnagar, G., Wu, Q.M.J., Liu, Z.: Directive contrast based multimodal medical image fusion in NSCT domain. IEEE Trans. Multimedia **15**, 1014–1024 (2013)
16. Easley, G., Labate, D., Lim, W.Q.: Sparse directional image representation using the discrete shearlet transforms. Appl. Comput. Harmon. Anal. **25**, 25–46 (2008)
17. Abdullah, A., Omar, A.J., Inad, A.A.: Image mosaicing using binary edge detection algorithm in a cloud-computing environment. Int. J. Inf. Technol. Web. Eng. **11**, 1–14 (2016)
18. Sathiyamoorthi, V.: A novel cache replacement policy for web proxy caching system using web usage mining. Int. J. Inf. Technol. Web. Eng. **11**, 1–13 (2016)
19. Sylvaine, C., Insaf, K.: Reputation, image, and social media as determinants of e-Reputation: the case of digital natives and luxury brands. Int. J. Technol. Human Interact. **12**, 48–64 (2016)
20. Wu, Z.M., Lin, T., Tang, N.J.: Explore the use of handwriting information and machine learning techniques in evaluating mental workload. Int. J. Technol. Human Interact. **12**, 18–32 (2016)
21. Kong, W.W., Lei, Y., Ren, M.M.: Fusion method for infrared and visible images based on improved quantum theory model. Neurocomputing **212**, 12–21 (2016)
22. Kong, W.W., Wang, B.H., Lei, Y.: Technique for infrared and visible image fusion based on non-subsampled shearlet transform and spiking cortical model. Infrared Phys. Technol. **71**, 87–98 (2015)

23. Kong, W.W., Lei, Y., Zhao, H.X.: Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. Infrared Phys. Technol. **67**, 161–172 (2014)
24. Kong, W.W., Liu, J.P.: Technique for image fusion based on NSST domain improved fast non-classical RF. Infrared Phys. Technol. **61**, 27–36 (2013)
25. Kong, W.W., Lei, Y.J., Lei, Y., Zhang, J.: Technique for image fusion based on non-subsampled contourlet transform domain improved NMF. Sci. China Ser. F-Inf. Sci. **53**, 2429–2440 (2010)
26. Kong, W.W., Lei, Y., Ma, J.: Virtual machine resource scheduling algorithm for cloud computing based on auction mechanism. Optik **127**, 5099–5104 (2016)
27. Kong, W.W., Lei, Y., Zhao, R.: Fusion technique for multi-focus images based on NSCT-ISCM. Optik **126**, 3185–3192 (2015)
28. Kong, W.W.: Technique for image fusion based on NSST domain INMF. Optik **125**, 2716–2722 (2014)
29. Kong, W.W., Lei, Y.: Technique for image fusion between gray-scale visual light and infrared images based on NSST and improved RF. Optik **124**, 6423–6431 (2013)
30. Kong, W.W., Lei, Y.: Multi-focus image fusion using biochemical ion exchange model. Appl. Soft Comput. **51**, 314–327 (2017)
31. Cao, Y., Li, S.T., Hu, J.W.: Multi-focus image fusion by nonsubsampled shearlet transform. In: Proceedings of IEEE 6th International Conference on Image and Graphics, vol. 1, pp. 17–21 (2011)
32. Miao, Q.G., Wang, B.S.: A novel image fusion algorithm based on local contrast and adaptive PCNN. Chin. J. Comput. **31**, 875–880 (2008)
33. Wang, Z.B., Ma, Y.D., Gu, J.S.: Multi-focus image fusion using PCNN. Pattern Recogn. **43**, 2003–2016 (2010)
34. Yang, S.Y., Wang, M., Lu, Y.X.: Fusion of multiparametric SAR images based on SW-nonsubsampled contourlet and PCNN. Sig. Process. **89**, 2596–2608 (2009)
35. Chiorean, L., Vaida, M.F.: Medical image fusion based on discrete wavelet transform using Java technology. In: Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, vol. 1, pp. 55–60 (2009)
36. Cai, W., Li, M., Li, X.Y.: Infrared and visible image fusion scheme based on contourlet transform. In: Proceedings of the ICIG 2009 5th International Conference on Image and Graphics, vol. 1, pp. 516–520 (2009)

Pilot Contamination Elimination in Massive MIMO Systems

Rui-Chao Hu(✉) and Bing-He Wang

Department of Information Engineering,
Engineering University of CAPF, Xi'an 710086, China
2388411722@qq.com

Abstract. The pilot contamination problem has been the primary limitation of massive multiple input multiple output (MIMO) systems. To improve it, in this paper, we propose a dynamical pilot assignment algorithm based on the priority of user location. First, we obtain the formulation of signal to interference plus noise power ratio (SINR) in uplink channel through minimum mean square error (MMSE) mechanism. Second, an objective function of SINR is defined together with constraint condition of real distance, based on which optimal value (OV) could be achieved. Third, we propose a novel cellular classification algorithm, that is, area with better channels adopts random pilot assignment scheme, and others use the novel algorithm. Last, the proposed algorithm is compared with the traditional algorithm. The results show that the proposed algorithm can effectively reduce the influence of pilot contamination on the communication performance and improve the system SINR and the system capacity.

1 Introduction

With the development of society and the exponential growth of data traffic, the existing communication technology can not meet the needs of people. In order to cope with the challenges of mass mobile terminals to the communications network, Each country competing to study the 5th generation of mobile communication technology (5G) [1, 2]. 5G plans to enter the commercial operation in 2020, compared with 3G and 4G technology, 5G technology is not only the speed of the upgrade, but also to improve the system capacity and improve the spectral efficiency of a qualitative leap, the typical characteristics of "high, speed, low latency" that can provide higher bandwidth and allow more terminal access to meet future communications requirements for high-speed data streams [3–5].

Therefore, in order to make full use of bandwidth resources, improve the spectrum utilization, to achieve low-power green communication, massive multiple input multiple output system (MIMO) technology came into being [6, 7]. As a key technology for the 5G physical layer, massive MIMO is equipped with large-scale antennas at the base station side, making it much larger than the number of single-antenna mobile terminals that can serve simultaneously [8]. Antenna information theory proves that the information transmission capacity of the channel capacity of the communication system with the end of the wireless communication link and the simultaneous use of multiple

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_69

antennas is significantly improved compared with the traditional single antenna system [9]. Therefore, massive MIMO technology can improve the peak rate and band efficiency of the system without increasing the bandwidth, so as to improve the transmission performance of wireless links and meet the needs of high-speed wireless data services and the rapid growth of users. According to the principle of probability statistics, when the number of base station side antennas is much larger than the number of single antenna users, the base station to each user's channel will tend to be orthogonal. As a result, inter-cell interference in adjacent cells will tend to disappear, and huge array gain will be able to effectively improve the SINR of each user, enabling the simultaneous scheduling of more users at the same time-frequency resources [10, 11]. However, massive MIMO technology to achieve the above goal is to accurately obtain the channel state information (CSI) as the premise, the base station side of the antenna only to obtain accurate CSI, in order to carry out effective data transmission, so as to achieve the purpose of improving system capacity. However, in the actual massive MIMO system, because the number of single-antenna mobile terminals (MT) in the cell is huge, in order to make the communication of each user not interfere with each other, it is necessary to ensure the orthogonality of each user's pilot sequence, That is, the number of orthogonal pilot sequences should be greater than or equal to the number of cells and the number of MT in the cell, the product is huge and the requirements for the communication equipment are extremely high. Therefore, it is inevitable to use the same pilot sequence to achieve normal communication, which leads to the pilot signal interference that pilot contamination has become the bottleneck of normal communication [12]. In the literature [12] demonstrates the effect of pilot contamination on system performance. In the non-cooperative multi-cell scenario, the interference between the noise and the cell is negligible with the increase of the number of base stations. However, the signal interference caused by the same pilot sequence between the cells still exists, which seriously affects the massive MIMO system communication performance; In the literature [13], it is assumed that the intra-cell pilot is orthogonal and the inter-cell pilot is fully multiplexed, and the closed expression of mean square error (MSE) is deduced. It is concluded that the length of the pilot sequence τ has little effect on the closed expression of MSE as the number of antennas M increases, and the uplink pilot power control method is proposed to improve the multiuser reachability and rate performance of each cell. But the article assumes that in the case of a one-dimensional plot, there is no discussion of the multidimensional cell scenario; In the literature [14] proposed inter-cell cooperative communication transmission scheme, the pilot transmission slot of the target cell and the data transmission time slot of the adjacent interference cell are shifted, so as to avoid the interference of the pilot information and the data information. The massive MIMO system, the pilot and data information is huge, how to accurately offset the time slot to send, especially in the fast moving scene, the implementation of more difficult; In the literature [15], according to the traditional method of randomly assigning the pilot sequence, the algorithm is proposed to continuously assign the pilot sequence according to the channel quality. Compared with the traditional random assignment pilot sequence, the performance is improved. But the article did not further classification of the district, so the system communication performance is limited.

Based on the literature [15] pilot scheduling scheme, this paper proposes an improved pilot sequence allocation method. Firstly, the channel estimation is carried out by using the least mean square error method. Then, the detection signal is received by the matched filter (MF) at the receiving end, and then the SINR expression. It is concluded that the SINR is mainly limited by the large-scale coefficients of the channel when the base station side antenna tends to infinity, and the deduced conclusion satisfies the probability statistics principle. In order to solve the problem of system pilot pollution, a mathematical model with distance as the constraint condition and SINR as the objective function is established. The convex optimization method is used to solve the optimal value of the objective function, and then the cell is classified according to the user area level to achieve the user within the district intelligent allocation of pilot sequences. The theoretical and simulation results show that this method has obvious effect in improving the SINR and system capacity, and can effectively reduce the influence of pilot contamination, and has good theoretical value and significance.

2 System Model

The massive MIMO multi-cell multi-user time-division duplex system, which is composed of L positive hexagonal cells, each cell number is $1, 2, \cdots, L$, each cell is composed of a base station with M antenna and K single antenna users, to meet the conditions of $K \leq M$. In order to facilitate the analysis of the problem, so that the middle of the cell as the target cell, the cell within the pilot sequence is completely orthogonal, the inter-cell pilot sequence is fully multiplexed. As shown in Fig. 1.

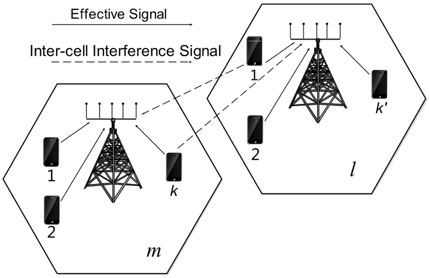


Fig. 1. Massive MIMO system model

The system channel vector is $\boldsymbol{H}_{lmp} \in \boldsymbol{C}^{M \times 1}$, which represents the channel vector of the p-th user in the m-th cell to the base cell of the l-th cell. The expression is:

$$\boldsymbol{H}_{lmp} = \eta_{lmp} \sqrt{\xi_{lmp}} \tag{1}$$

In the formula (1), η_{lmp} represents the large scale fading coefficient in the channel, which indicates the slow change of the mean value of the received signal over a certain period of time with the propagation distance and the environment [16]. ξ_{lmp} represents the small-scale fading factor in the channel, satisfying $\xi_{lmp} \sim CN(0, \boldsymbol{I}_M)$, and characterizes the rapid fluctuation of the received signal after a short or short distance [16]. Assuming the pilot sequence $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_k]^{\mathrm{T}} \in \boldsymbol{C}^{k\times\tau}$, the pilot sequence length is τ, and the pilot sequences of the users in the cell are orthogonal to each other, i.e., $\boldsymbol{\Psi}\boldsymbol{\Psi}^{\mathrm{H}} = \boldsymbol{I}_K$. In the uplink pilot transmission phase, the user in the cell sends the pilot signal to the respective service base station. If the base station l is the target base station, the received pilot signal is:

$$\boldsymbol{Y}_l = \sum_{l=1}^{L}\sum_{p=1}^{K} \boldsymbol{H}_{lmp}\boldsymbol{\psi}_{lp} + \boldsymbol{N}_l^{\mathrm{UL}} \tag{2}$$

In the formula (2), $\boldsymbol{N}_l^{\mathrm{UL}}$ is the additive white Gaussian white noise received by the base station l, which is a $M \times \tau$-dimensional matrix, satisfying the independent identically distributed and $\boldsymbol{N}_l^{\mathrm{UL}} \sim CN\left(0, \delta_N^2\right)$. $\boldsymbol{\Psi}_{lp} \in \boldsymbol{C}^{l\times\tau}$ is the pilot signal transmitted by the user p in the l-cell and satisfies $\boldsymbol{\Psi}_{lp}\boldsymbol{\Psi}_{lp}^{\mathrm{H}} = \tau$. After the base station side completes the estimation of the uplink channel, the downlink channel is the conjugate transpose of the uplink channel according to the reciprocity of the TDD system model channel. Assuming that $\boldsymbol{H}_{ml}$ is the channel vector of the m-cell user to the base station of l and satisfies $\boldsymbol{H}_{ml}^{\mathrm{DL}} = H_{ml}^{\mathrm{H}}$, The signal from the base station received by the user in the l-th cell is:

$$\boldsymbol{Y}_l^{\mathrm{DL}} = \sqrt{\rho_f}\sum_{l=1}^{L}\sum_{p=1}^{K} \boldsymbol{H}_{ml}^{\mathrm{H}} x_{mp} + \boldsymbol{N}_{mp}^{\mathrm{DL}} \tag{3}$$

In Eq. (3), ρ_f represents the average SINR of the downlink data, $\boldsymbol{N}_{mp}^{\mathrm{DL}}$ satisfies the independent identically distributed and $\boldsymbol{N}_{mp}^{\mathrm{DL}} \sim \left(0, \delta_{mp}^2\right), x_{mp}$ is the data vector sent by the base station of m to the user of p. For the massive MIMO systems, accurate access to channel state information determines the receiver to correctly detect and resume the transmit signal. In this section, the channel is estimated by the least mean square error method. According to the MMSE estimation criterion, the cost function formula is [16]

$$J_{\mathrm{MMSE}} = E\left(\boldsymbol{H}_l - \hat{\boldsymbol{H}}_l\right)\left(\boldsymbol{H}_l - \hat{\boldsymbol{H}}_l\right)^{\mathrm{H}} \tag{4}$$

In Eq. (4), the partial derivation of $\hat{\mathbf{H}}_l$, when the function formula is zero, the MMSE estimation result is:

$$\hat{\boldsymbol{H}}_l^{\mathrm{MMSE}} = Y_l\left(\boldsymbol{\Psi}^{\mathrm{H}} R_{\boldsymbol{H}_l\boldsymbol{H}_l}\boldsymbol{\Psi} + \sum_{l\neq m}^{L} \boldsymbol{\Psi}^{\mathrm{H}} R_{\boldsymbol{H}_l\boldsymbol{H}_m}\boldsymbol{\Psi} + \delta_N^2 \boldsymbol{I}_\tau\right)^{-1} \boldsymbol{\Psi}^{\mathrm{H}} R_{\boldsymbol{H}_l\boldsymbol{H}_l} \tag{5}$$

In Eq. (5), $R_{\boldsymbol{H}_l\boldsymbol{H}_l}$ represents the autocorrelation coefficient of the channel, and $R_{\boldsymbol{H}_l\boldsymbol{H}_m}(l \neq m)$ represents the cross-correlation coefficient of the channel. During the uplink transmission, the receiver uses the matched filter to obtain the detected signal as:

$$\hat{Y}_l = \hat{\boldsymbol{H}}_l Y_l = \left(\boldsymbol{H}_l + \sum_{m \neq l}^{L} \boldsymbol{H}_m + \frac{\boldsymbol{N}_l \boldsymbol{\psi}^{\mathrm{H}}}{\tau}\right)\left(\sum_{l=1}^{L}\sum_{p=1}^{K} \boldsymbol{H}_{lmp}\boldsymbol{\psi}_{lp} + \boldsymbol{N}_l\right) \tag{6}$$

The resulting signal to interference ratio of the uplink is:

$$\mathrm{SINR}_u = \frac{M^2 \eta_{ll}^2 \delta_x^2}{M^2 \delta_x^2 \sum\limits_{m \neq l}^{L} \eta_{im}^2 + \delta_n^2 \left(M_{\eta_{ll}} + M \sum\limits_{m \neq l}^{L} \eta_{lm} + \frac{\delta_n^2}{\delta_\psi}\right)} \tag{7}$$

Because in the massive MIMO system, the number of base station antennas is large, to meet $M^2 \gg M$, so when the number of base station antenna M tends to infinity, the uplink SINR limit:

$$\lim_{M \to \infty} \mathrm{SINR}_u = \frac{\eta_{ll}^2}{\sum\limits_{m \neq l}^{L} \eta_{lm}^2} \tag{8}$$

In Eq. (8), η_{ll} represents large-scale fading in the cell, and η_{lm} represents the large-scale fading coefficient between the cells. It can be seen that the SINR of the channel will be mainly limited by the large-scale fading coefficient as the number M of the base station increases gradually.

3 Intelligent Assignment Pilot Sequence Scheme Based on User Area Location Priority

In the massive MIMO systems, large-scale fading coefficients can be constructed for the model [8]:

$$\eta_{lmp} = \frac{\zeta_{lmp}}{\left(r_{lmp}/R\right)^{\alpha}} \tag{9}$$

In the formula (9), ζ_{lmp} represents the shadow fading and satisfies the lognormal distribution, that is, $10\lg\left(\zeta_{lmp}\right) \sim C(0, \delta_{shadow})$ and r_{lmp} represent the geometric distance between the p-th user of the m-th cell and the base station of the l-th cell, R is the cell radius, α is path loss factor. In order to study the convenience of the problem, it is assumed that the massive MIMO system has 7 cells, each cell has 8 users, the intermediate cell is the target cell, and the communication is carried out under the visual condition. Taking the base station distance of the user in the cell into the intermediate cell as the constraint condition, the system signal to interference ratio is the objective function, and the optimal value of the objective function is solved.

$$f\left(r_{lmp}\right) \lim_{M\to\infty} \mathrm{SINR}_u, \left(r_{lmp} > 0\right)\ s.t.\ f\left(r_{lmp}\right)_{\min} \tag{10}$$

In the Eq. (10), $\lim_{M\to\infty} \mathrm{SINR}_u$ indicates that the number of antennas configured in the target cell tends to be infinite when the number of antennas in the target cell approaches infinity, and $f\left(r_{lmp}\right)_{\min}$ represents the channel poor area in the target cell. The existence and uniqueness of $\left(r_{lmp}, f\left(r_{lmp}\right)_{\min}\right)$ are proved below.

Proof: By the function expression $f\left(r_{lmp}\right)$ know:

$$\begin{aligned} f\left(r_{lmp}\right) &= \lim_{M\to\infty} \mathrm{SINR}_u \\ &= \frac{\eta_{ll}^2}{\sum\limits_{m\neq l}^{L}\eta_{lm}^2} = \frac{\sum\limits_{m=1}^{L}\sum\limits_{p=1}^{K}\left[\frac{\zeta_{lmp}}{\left(r_{lmp}/R\right)^3}\right]^2}{\sum\limits_{m\neq l}^{L}\sum\limits_{p=1}^{K}\left[\frac{\zeta_{lmp}}{\left(r_{lmp}/R\right)^3}\right]^2} = \frac{\sum\limits_{p=1}^{7}\left[\frac{1}{\left(r_{llp}/R\right)^3}\right]^2}{\sum\limits_{m\neq l}^{7}\sum\limits_{p=1}^{8}\left[\frac{1}{\left(r_{llp}/R\right)^3}\right]^2} \\ &= \frac{\sum\limits_{p=1}^{7} r_{llp}^{-6}}{\sum\limits_{m\neq l}^{7}\sum\limits_{p=1}^{7} r_{lmp}^{-6}} \geq 6 \end{aligned}$$

(If and only $r_{lmp} = R$ is the equation holds, and that the minimum value is unique). It can be seen that the interference ratio of the target cell and the interference cell in the overlapping area of the hexagonal circumscribed circle is the smallest, and the ratio of the signal to noise ratio is the smallest. As shown in Fig. 2.

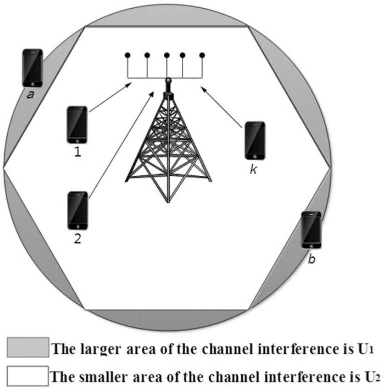

a
1
k
2
b
▨ The larger area of the channel interference is U1
□ The smaller area of the channel interference is U2

Fig. 2. Smart pilot assignment scheme in massive MIMO systems based on priority of user location

In the round, the user in the shaded area, such as user a, user b, the interference is relatively serious, because in this overlapping area, the user is far from the target cell base station, the user to the base station pilot signal fading more serious, but, The user of this area is relatively close to the neighboring cell base station, the pilot signal between users is not orthogonal, the base station can not obtain the channel state information accurately, resulting in more serious interference, seriously affecting the communication performance.

Assume that U_1 is the shaded area in the cell and U_2 is the hexagonal region in the cell. Based on the principle of user priority allocation, the U_1 precedence is higher than U_2, and the orthogonal pilot sequence is preferentially assigned to the U_1 area user.

In order to obtain the maximum signal-to-noise ratio in the shaded area where the interference is more severe, the pilot signals are arranged in descending order: $\boldsymbol{\Psi}_{\mathrm{D}}^{\mathrm{DOWN}} = [\psi_1, \psi_2, \cdots, \psi_k]$, satisfying $\psi_1 \geq \psi_2 \geq \cdots \geq \psi_k \geq 0$. Also in accordance with the quality of the channel will be the user in ascending order: $\mathrm{U}_1^{\mathrm{UP}} = [u_1, u_2, \cdots, u_k]$, to meet $0 \leq u_1 \leq \cdots \leq u_{k-1} \leq u_k$. The users $u_l(l = 1, 2, 3, \cdots, k)$ and $u_m(m = a, b, \cdots, k')$ (where $u_m \in \mathrm{U}_1, u_l \in \mathrm{U}_2$) in the target cell are defined by the distance between the user and the base station in the target cell, respectively,

$$u_l = r_{llp}, p = 1, 2, 3, \cdots, k \tag{11}$$

$$u_m = r_{llp'}, p' = a, b, \cdots, k' \tag{12}$$

Under the above conditions, Fig. 2 shows a scheme for intelligently assigning pilot sequences based on user region priorities. Assuming that the base station is aware of the large-scale fading coefficient of each user, the base station can locate the user's position through the large-scale fading coefficient, using the Eq. (9), where the relationship between the large-scale fading coefficient and the distance satisfies $\eta_{lmp} \rightarrow r_{lmp}^{-\alpha}$. By comparing the distance between the user and the base station, the cell users are divided into U_1 and U_2. After the classification of the community users, and then consider the community user pilot classification problem. The user allocates the pilot sequence intelligently within the $\sqrt{3}R/2 \leq r \leq R$ range, and the remaining users in the cell randomly allocate the remaining pilot sequences. Define the random variable $\mathrm{N} = \mathrm{U}_1$ the number of regional users, use the allocation algorithm proposed in this paper to calculate the value of N, the convergence condition is that each user assigned to the respective pilot sequence, if the number of users within U_1 is N_1, the target cell The first N_1 pilot signals arranged in descending order of the pilot signal strength are sequentially assigned to the first N_1 subscribers arranged in ascending order of the channel quality, and the area U_2 with better channel quality is randomly assigned to the remaining user. The specific allocation scheme is shown in the table.

```
Input: Pilot sequence: $\psi_1, \psi_2, \cdots, \psi_k$ , $k = 1, 2, \cdots, K$
(1) Pilot assignments to users of the target cell
for  k=1:K
    if  $|d| \in \left\{ \sqrt{3} / 2R, \mathrm{R} \right\}$
        $\mathrm{U} \in \mathrm{U}_1$
    else
        $\mathrm{U} \in \mathrm{U}_2$
    end
end
(2) $\mathrm{U}_1$ area users in ascending order of interference intensity, $\mathrm{U}_1^{\mathrm{up}} : [u_{11}, u_{12}, \cdots, u_{1K}]$
$\mathrm{U}_2$ area users are randomly arranged: $[u_{21}, u_{22}, \cdots, u_{2K}]$
(3) $\mathrm{U}_1$ area user pilot assignment
for  k=1:N
    $\psi_k$ : $\psi_i \rightarrow u_{1i} (i = 1, 2, \cdots, K)$
end
$\mathrm{U}_2$ area user pilot assignment
for  k=N+1:K
    $\psi_k$ : $\psi_i \rightarrow u_{2j} (i = j或i \neq j.i, j = 1, 2, \cdots, K)$
end
Output: Interference cell user pilot $\psi_k$ , $k = 1, 2, \cdots, K$
```

4 Experimental Simulation and Result Analysis

This section uses the Monte Carlo method to simulate the scheme of assigning the pilot sequence based on the user region priority. Simulation conditions: L is a hexagonal cell composed of cellular communication network. Each cell consists of a base station and K single antenna users that are located at the center of the M antenna. Surrounded by other communities in the middle of this area for the analysis of communication performance of the target cell [4, 16]. The number of cells is L, The cell radius is R, the number of antennas configured at the base station side of the cell is M, the number of users in each district is K, the path loss factor is α, the logarithmic shadow fades is δ_{shadow}, the critical distance value is $\sqrt{3}/2R \leq r \leq R$, the pilot sequence length is τ. The channel model is quasi-static Rayleigh fading channel, the system parameters in Table 1.

Table 1. System simulation parameters

L	R	M	K	α	δ_{shadow}	r	τ
7	500	$40 \leq M \leq 500$	8	3	8 dB	$\sqrt{3}/2R \leq r \leq R$	16

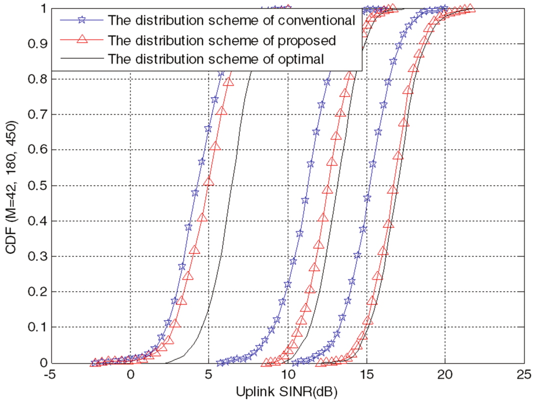


Fig. 3. CDF vs. SINR in uplink training

Figure 3 is the simulation of the probability of the uplink signal to noise ratio. It can be seen from Fig. 3 that the increase in the number of base station antennas can effectively improve the upstream SINR of the system. When the number of base station antennas is M = 42, the priority scheme is improved by about 0.5 dB and the optimal pilot scheme is about 2 dB higher than the traditional pilot scheme. This is because the number of antennas in the area is less and the number of dynamic users in the region is less. In the shadow area of Fig. 2, the interference of the channel area is less, and the antenna at the base station can better estimate the channel state information. However, as the number of base station end days increases, such as M = 180, and the exponential growth of dynamic users, the channel interference in the shaded area in Fig. 2 is large, and the pilot information received by the base station is not only from the target cell There are nearby users in the neighborhood.

In the traditional MIMO, because the number of antennas is low and the spatial reuse rate is low, the beamforming is not concentrated, so it can not accurately judge whether the pilot signal is sent by the local user or other cell users. However, the base station uses large-scale MIMO technology, Although it can reduce the interference of the pilot information within the cell, but the inter-cell pilot interference still exists [11], the traditional pilot allocation scheme, without taking into account the regional area of the user channel quality of the actual situation, the district users priority Level equal treatment, so the system letter to noise ratio to enhance the effect is not obvious. In this paper, we fully consider the above problems, so the proposed algorithm is verified by mathematical analysis and simulation: the system uplink signal to interference and noise ratio of traditional pilot distribution scheme has improved significantly, about the optimal pilot distribution scheme on the basis of 0.5 dB of the gain. When the number of antennas continues to increase, such as M = 450, the proposed frequency distribution algorithm and the optimal pilot distribution algorithm curve almost coincide, better reflects the superiority of the proposed algorithm.

Figure 4 is a simulation of the number of base station antennas and the user spectrum efficiency curve. As can be seen from the figure, whether the traditional pilot scheme or the user priority intelligent pilot distribution scheme, the increase in the number of antennas in the base station side of the cell will increase the efficiency of single-user spectrum, and it can be seen that when the number of base station antennas The rate of growth of single-user spectrum efficiency is less than the growth rate of base station-side antennas. This is because the number of antennas is small, the number of antennas is limited to the user spectrum efficiency growth bottleneck, but when the number of antennas is high, the pilot pollution is limited to the user spectrum efficiency growth bottleneck.

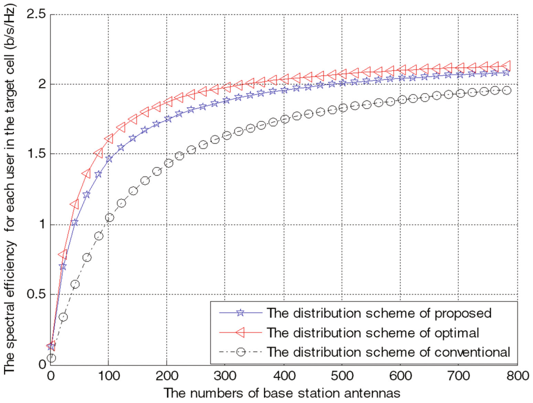


Fig. 4. Spectrum efficiency vs. the numbers of base station antennas

The traditional pilot allocation scheme has not limited the interference of the actual signal, the spectrum efficiency is limited and quickly reached the saturation, which seriously affected the system throughput. In this paper, the algorithm is proposed to improve the orthogonality of the user's pilot signal by taking into account the actual reality of the adjacent cell pilot pollution and the actual interference of the users in different areas of the cell, and then classifying the users according to the cell area priority. Single-user spectrum efficiency is significantly improved. It can be seen from the figure that when the number of base station antennas is M = 128, the proposed pilot frequency allocation algorithm is the most obvious, about 0.5bps/Hz, and the number of days continues to increase. The spectral efficiency variation curve of the algorithm is close to the optimal pilot distribution algorithm curve.

As can be seen from the above analysis, the class I cell user U_1 (shadow area) intelligently assigns the pilot sequence based on the user region priority, and the class II cell user U_2 (non-shaded area) randomly assigns the pilot sequence. The results of MATLAB software show that the proposed scheme can effectively improve the uplink

SINR of the system and the single-user spectrum efficiency in the case of $\sqrt{3}R/2 \leq r \leq R$ (R is the radius of the cell), also significantly improved, further confirming the accuracy of theoretical analysis.

5 Conclusion

In this paper, a pilot sequence scheduling method for mitigating pilot pollution of large-scale MIMO systems is proposed based on the scheme of intelligent allocation of pilot sequences based on user region priority. Under the premise of satisfying the normal communication of each user in the cell, the distance between the user and the base station is the constraint condition, the system signal to interference ratio is the objective function, and the corresponding mathematical model is established to deduce the adjacent distance of the channel quality. In this way, the cell is divided into two types according to the channel quality, the channel quality is better allocated to the pilot sequence, the channel quality is poorly distributed in the region, and the communication between the neighboring cells is reduced. The proposed frequency sequence scheduling scheme can effectively improve the system capacity and single-user spectrum efficiency under the condition of satisfying the quality of service of the user, and reduce the influence of pilot pollution on the communication performance of large-scale MIMO system.

References

1. Andrews, J.G., Buzzi, S., Wan, C., Hanly, S.V.: What will 5g be? IEEE J. Sel. Areas Commun. **32**(6), 1065–1082 (2014)
2. Liu, L., Al-Dubi, A., Ali, S., Zhu, D.: Special issue on ubiquitous computing and future communication systems. Future Gener. Comput. Syst. **39**(39), 1–2 (2014)
3. Shah, S.C., Park, M.S.: An energy-efficient resource allocation scheme for mobile ad hoc computational grids. J. Grid Comput. **9**(3), 303–323 (2011)
4. Boccardi, F., Heath, R.W., Lozano, A., et al.: Five disruptive technology directions for 5G. IEEE Commun. Mag. **52**(2), 74–80 (2014)
5. Chihlin, I., Rowell, C., Han, S., et al.: Toward green and soft: A 5G perspective. IEEE Commun. Mag. **52**(2), 66–73 (2014)
6. Björnson, E., Larsson, E.G., Marzetta, T.L.: Massive MIMO: ten myths and one critical question. IEEE Commun. Mag. **54**(2), 114–123 (2015)
7. Larsson, E.G., Edfors, O., Tufvesson, F., et al.: Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**(2), 186–195 (2014)
8. Lu, L., Li, G.Y., Swindlehurst, A.L., et al.: An overview of massive MIMO: benefits and challenges. IEEE J. Sel. Top. Sign. Process. **8**(5), 742–758 (2014)
9. Rusek, F., Persson, D., Lau, B.K., et al.: Scaling up MIMO: Opportunities and challenges with very large arrays. Mathematics **30**(1), 40–60 (2012)
10. Jose, J., Ashikhmin, A., Marzetta, T.L., et al.: Pilot contamination and precoding in multi-cell TDD systems. IEEE Trans. Wirel. Commun. **10**(8), 2640–2651 (2011)
11. Kitagami, S., Kaneko, Y., Kiyohara, R., Suganuma, T.: Autonomic load balancing for m2m communication with long-polling. Int. J. Space-Based Situated Comput. **3**(1), 45–54 (2013)

12. Marzetta, T.L.: Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans. Wirel. Commun. **9**(11), 3590–3600 (2010)
13. Wang, H., Wang, Y., Huang, Y., et al.: Pilot contamination reduction in very large MIMO multi-cell TDD systems. J. Sign. Process. **29**(2), 171–180 (2013)
14. Fernandes, F., Ashikhmin, A., Marzetta, T.L.: Inter-cell interference in noncooperative TDD large scale antenna systems. IEEE J. Sel. Areas Commun. **31**(2), 192–201 (2013)
15. Zhu, X., Wang, Z., Dai, L., et al.: Smart pilot assignment for massive MIMO. IEEE Commun. Lett. **19**(9), 1644–1647 (2015)
16. Li, J., Guo, T., Wu, G.: Mobile communication systems, pp. 99–109. Xi'an University Press, Xi'an (2012)

Improved Leader-Follower Method in Formation Transformation Based on Greedy Algorithm

Yan-Yu Duan(✉), Qing-Ge Gong, Zhen-Sheng Peng, Yun Wang, and Zhi-Qiang Gao

Department of Information Engineering,
Engineering University of PAP, Xi'an, China
1530189605@qq.com

Abstract. A method based on the leader-follower method is proposed for formation transformation in large-scale mass incidents. The greedy algorithm is introduced to realize regional division and leader matching problem in target formation by constructing a distance matrix, and to calculate the distribution of followers. In order to solve the problem of path conflict without error feedback, collision detection and collision avoidance are proposed, which effectively avoids motion failure. Experiment of transforming the line formation into wedge-shaped formation is simulated, and the result shows that the proposed formation transform method is feasible and can effectively improve the efficiency of formation transformation.

Keywords: Computer application · Formation transformation · Follower-leader · Greedy algorithm · Collision avoidance

1 Introduction

Formation transformation is a vital task for emergency-dealing. Timely and effective formation transformation can not only enhance the resistance to external forces attack and improve the ability to maintain the robustness of the formation [1, 2], but also strengthen the deterrent force. The group formation control technology is widely applied for the transformation design, exercise and simulation verification in large-scale mass incidents, which can quickly and effectively show a better visual effect and improve the quality and efficiency of the design and training.

This paper studies on the group formation control based on the complex unknown environment. Group formation control refers to the process of multiple moving objects to maintain a certain predefined formation or transform into a new formation under the constraints of the environment and their own rules. Group formation control methods can be divided into four parts: the formation of the constraint shape, the layout of each object in the formation, the pairing among the objects in the formation transformation, and the collision detection and collision avoidance in the motion [3].

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_70

2 Greedy Algorithm Based on Leader-Follower Method

Combined with the leader and greedy algorithm [4, 5], in the moving process of the formation towards the target point, obstacle avoidance is achieved and the geometric pattern remains unchanged; the obstacle avoidance and formation maintenance can be carried out simultaneously or successively. The merits of the formation can be judged by its performance indicators [4], as follows:

(1) The path length ratio, that is, the ratio of the average length of the whole team's moving path to the shortest straight-line distance between the starting point and the target point; the smaller the value, the better the effect.
(2) The formation retention rate, the proportion of the members located in the desired position.
(3) The running time, the time to reach the target position as a whole.

2.1 Formation Maintenance Based on Leader-Follower Algorithm

The basic idea of leader-follower algorithm [1] for formation maintenance is that in a group formation of multiple moving objects, one or more are designated as leaders, the remaining objects as the followers of the leader, which follow the leader's position and direction at a certain distance, thus maintaining a variety of formations [6]. In the leader-follower control structure, the group formation structure can be divided into parallel structure and series structure, as shown in Fig. 1.

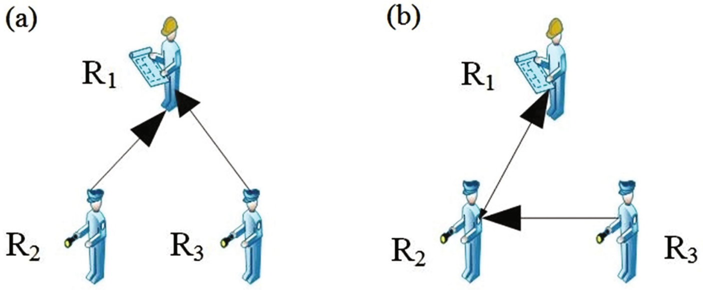


Fig. 1. Parallel structure (a) and series structure (b).

Leader is the core of the formation affecting the movement of the whole group, and the leader's state, speed, direction and other information are shared and visual for the followers. In series structure, each follower determines its position according to the $l-\varphi$ controller of the neighboring leader; in parallel structure, each follower's position is determined only by the $l-l$ controller of the only leader. Combined with LFS (Leader-to-Formation Stability) theory [3], the stability of two basic formation forms are analyzed and compared. It is concluded that the stability of parallel formation with single leader is higher than that of series formation. Since the formation transformation of troops involves in multiple leaders, this paper uses parallel structure to conduct the transformation.

Advantages of the leader-follower algorithm [7, 8] is that simply a given act or trajectory of the leader can control the subsequent behavior of the whole team, so the formation control problem can be simplified as an independent tracking problem, and each follower only needs to obtain the status information of the leader, which greatly simplifies the formation cooperation. The main drawback of this method is that there is no explicit formation feedback in the system. For example, if the leader moves too fast, the follower is likely to be lost. Another drawback is that if the leader were to fail, the whole formation would not be maintained. Therefore, the greedy algorithm is introduced to carry on the conditional feedback control method [7]. The positional relationship between the leaders and the followers is described as follows:

$$\begin{aligned} x_{Follower} &= x_{Leader} + x_{Team} \\ y_{Follower} &= y_{Leader} + y_{Team} \end{aligned}. \tag{1}$$

The desired position of each follower in the formation is derived according to the task, and combining with the environment information, the control variable is generated based on the control strategy, and the formation information is fed back to the leader. In the normal movement, the feedback information does not influence the leader's movement, and the follower adjusts its own speed to maintain the formation [9]. Only is a follower on the edge of the communication range, the leader would slow down to make the backward follower catch up.

The flow chart of the formation maintenance algorithm is shown in Fig. 2.

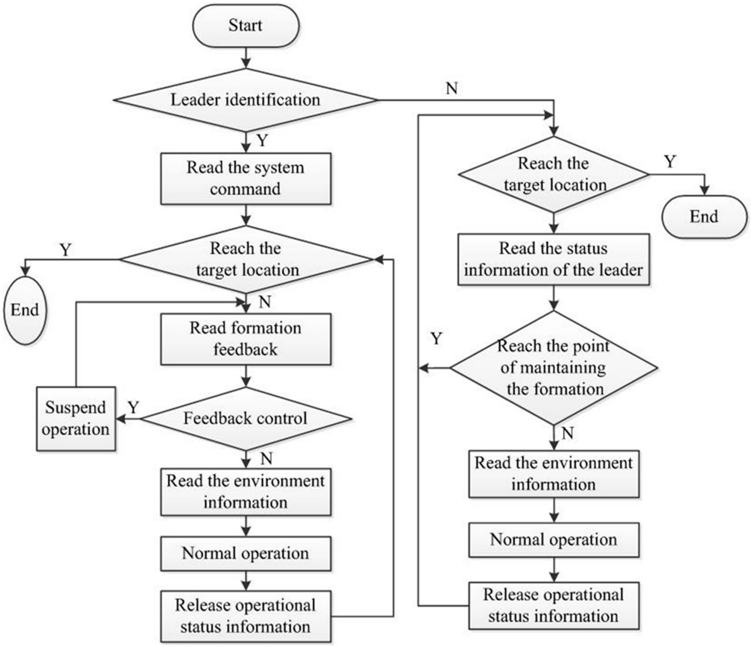

Start
Leader identification
N
Y
Read the system command
Reach the target location
Y
End
N
Read formation feedback
Suspend operation
Y
Feedback control
N
Read the environment information
Normal operation
Release operational status information
Reach the target location
Y
End
Read the status information of the leader
Y
Reach the point of maintaining the formation
N
Read the environment information
Normal operation
Release operational status information

Fig. 2. The flow chart of the formation maintenance algorithm.

2.2 Greedy Algorithm

The greedy algorithm [4, 10] starts from a certain initial solution of the problem, and obtains the optimal choice in the current state through a series of greedy choices, and gradually approaches the given target and then gets the optimal value as fast as possible. The algorithm stops when a certain step in the algorithm can no longer proceed. The greedy algorithm adopts the method of constructing the optimal solution step by step. At each stage, a seemingly optimal decision is made (under certain criteria).

The solving steps of greedy algorithm are as followed:

Start from a certain initial solution of the problem;

While (a step forward towards a given target based on greedy strategy) do

Find a solution element of the feasible solution;

A feasible solution is obtained by the combination of all elements.

Based on the leader-follower algorithm, the region of the target formation is divided. In order to get the optimal position distribution of the followers in the formation [11], the greedy algorithm is introduced, based on which a distance matrix is constructed and the matching problem of each leader in the target formation is realized, and then the distribution of each member in the target formation is obtained according to the location of the leader.

3 Modeling and Simulation

In computer simulation theory, the position of each virtual member is determined by coordinates. In this study, the emergency site is simulated as a two-dimensional plane. Each point on the plane has one unique coordinate (x, y), and the drop point of each virtual member corresponds to one coordinate point on the plane. The position change of the member can be regarded as the change of the coordinate value, and the interval between the coordinate points indicates the distance between the members. When the formation needs to be transformed, the forward direction of each member can be calculated based on the source coordinate and target coordinate.

In this paper, the field where the emergency-dealing formation locates is a flat site, which can be regarded as a two-dimensional plane. The moving increments Δx and Δy in x and y coordinate directions represent the movement direction of the virtual members. According to the value of Δx and Δy continuously recorded with some certain frequency, the route of the virtual members can be obtained. With the route transformation algorithm, the whole transformation process can be simulated to visually show the detailed trajectory of each member's route transformation during the formation transformation.

Because of the reality demands of emergency-dealing, each team member must remain in the group aggregation after the formation transformation. Therefore, this paper divides the transformed formation into polygon areas according to the number of the leaders. Firstly, the location nearest to the centroid of each region is taken as the target position of the leader; secondly, the leader is arranged to each region using the greedy algorithm; and then followers are matched to the leader by greedy algorithm so that location mapping is achieved; finally, the collision that may occur during the movement is detected and avoided.

3.1 Formation Control with Greedy Algorithm Based on Conditional Formation Feedback

In order to coordinate the obstacle avoidance and formation maintenance, when a member of the formation is in the state of obstacle avoidance, greedy algorithm is used so that other members can replace him to maintain normal operation. With a parallel structure, there is no relationship between followers while maintaining the formation. These features make the formation control framework greatly flexible. In the movement, if the obstacle avoidance path is relatively long, leading to the follower out of the leader's communication range, there will be a departure. Formation feedback is introduced to overcome this shortcoming, but causing obstruction in the whole movement. And when multiple followers mistakenly stuck in the obstacle avoidance state at the same time, the system will be locked as the leader formation feedback is stop. Therefore, conditional formation feedback [4] is introduced, as shown in Fig. 3.

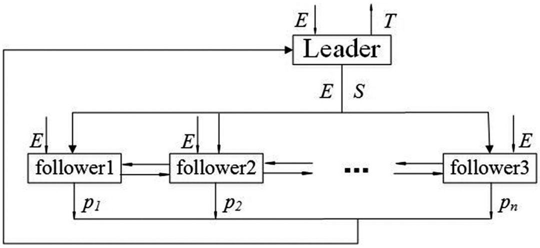


Fig. 3. The leader-follower control structure of the formation conditional feedback.

A two-dimensional plane is simulated, and n leaders are simulated as points on the plane. A matrix is introduced, representing the distance between the source position and the target position of each leader. So the problem of finding the shortest path will be transformed into a problem of finding the smallest number in a matrix of $n \times n$, which is noted as $L_{n \times n} = (a_{ij})$. Where, i indicates the nth leader in the source position; j indicates the position of the i-th leader in the target formation; a_{ij} represents the distance from the source position to the target position of the i-th leader. During the transformation, all the virtual members run reasonably so that the entire formation can be transformed into a certain target formation from an initial one. The positions of the virtual members are determined both in initial and target formation, and the movement routes of the members are uncertain. Assuming that the distance between the initial and target landing point of the virtual member is defined as a route, the shortest route priority algorithm is to find an ideal route that is the shortest one of all the routes without considering any conflict [12]. A virtual member's post move means running along the route from the initial landing point to the virtual landing point of the target formation. Based on the distance from the source position to the target position, an distance matrix of $n \times n$

$$\begin{pmatrix} 26 & 15 & 12 & 11 & \cdots \\ 37 & 20 & 14 & 21 & \cdots \\ 19 & 12 & 8 & 24 & \cdots \\ 9 & 13 & 11 & 23 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix} \rightarrow \begin{pmatrix} 26 & 15 & 11 & \cdots \\ 37 & 20 & 21 & \cdots \\ 9 & 13 & 23 & \cdots \\ \vdots & \vdots & \vdots & \end{pmatrix} \tag{2}$$

is obtained. The change in this distance matrix indicates that the shortest path is $a_{33} = 8$ according to greedy algorithm.

Combined with the conditional feedback mechanism, simulation is conducted using the greedy algorithm. To form a diamond-shaped formation as an example, the position code is as follows:

```
position_x = [leader_xleader_x-40leader_x-40leader_x-80];
position_y = [leader_yleader_y-10leader_y + 10leader_y].
```

First calculate the distance between the i-th follower and the leader, and then the distance between the i-th follower and each position of the target geometric formation, and finally find the nearest location from the i-th follower, which is the target point of the i-th follower in the diamond-shaped formation. Figure 4 shows the process of the formation movement simulation mentioned above.

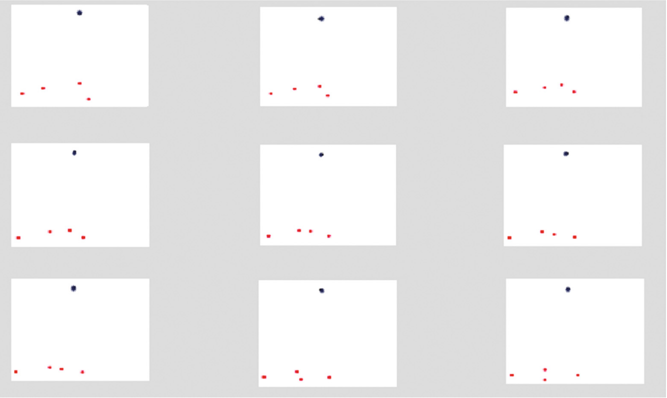


Fig. 4. Initial formation display process.

Figure 4 shows that the formation control with greedy algorithm based on conditional feedback can effectively maintain the formation in obstacle avoidance.

3.2 Collision Prediction

As mentioned above, the position of each member in the target formation is obtained with conditional feedback based on greedy algorithm. However, during the transformation of the entire formation, there will inevitably be a situation where two or more members being at one same position at a certain moment, which is called collision problem. If this problem is not dealt with, the entire formation will be stagnant [13]. Therefore, the collision prediction should be conducted throughout the transformation.

In this paper, x represents one emergency-dealing member, $L(x)$ represents the current position of x, $V(x)$ represents the speed of x, and $W(x)$ the width of the area (a fixed value in this paper). Assuming that two adjacent members are a and b, and $L_r = L(a) - L(b), V_r = V(a) - V(b)$, where L_r represents the relative position of a and b, V_r represents the relative velocity of a and b, a possible collision event between a and b satisfies:

$$L_r^2 + 2 \times L_r \times V_r \times t + V_r^2 \times t^2 = \left(W(a) + W(b) + \varepsilon^2\right) \tag{3}$$

where ε is the safe distance of a and b. If there is no solution or just one unique solution, then there is no predictive collision between a and b. If there are two solutions t_1 and $t_2 (t_1 < t_2)$, then there is a collision that will happen immediately (that is, need to avoid collision) if $t_2 < 0$; and if $t_1 \geq 0$, the collision will occur after t_1. The collision time can be denoted by t_p uniformly, then $L_a = L(a) + V(a) \times t_p$, $L_b = L(b) + V(a) \times t_p$, where L_a and L_b represents the position of a and b after the time of t_p respectively, and then:

(1) $(L_a - L_b) \times V(a) < 0$ represents rear collision,
(2) $(L_a - L_b) \times V(a) > 0$ and $V(a) \times V(b) < 0$ represent front collision,
(3) $(L_a - L_b) \times V(a) > 0$ and $V(a) \times V(b) \geq 0$ represent rear collision,
(4) $\|V(b)\| = 0$ represents static collision.

The above four types of collisions are proposed to design a local collision avoidance algorithm.

3.3 Collision Avoidance

The collision avoidance designed in this paper is mainly achieved by the speed change. Speed is mainly determined by magnitude and direction, so collision avoidance of two members can be realized by changing the direction, magnitude, or both at the same time. Direction change is divided into left and right; speed change is divided into acceleration and deceleration. Since the deceleration of a is equivalent to the acceleration of b, acceleration rules can be achieved through the deceleration rules. The geometrical explanations of the left, the right and the speed-shift avoidance are shown in Figs. 5, 6 and 7 respectively.

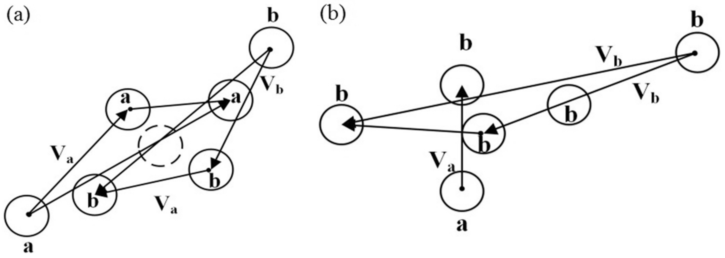


Fig. 5. Left avoidance: (a) front collision; (b) (static) collision.

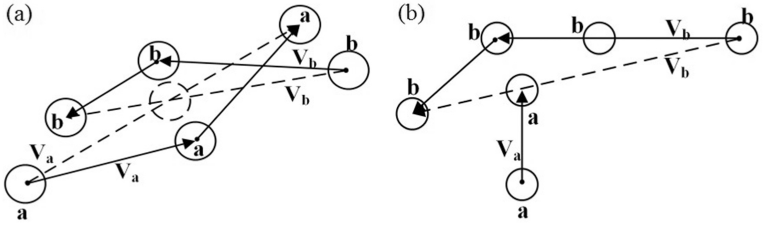


Fig. 6. Right avoidance: (a) front collision; (b) (static) collision.

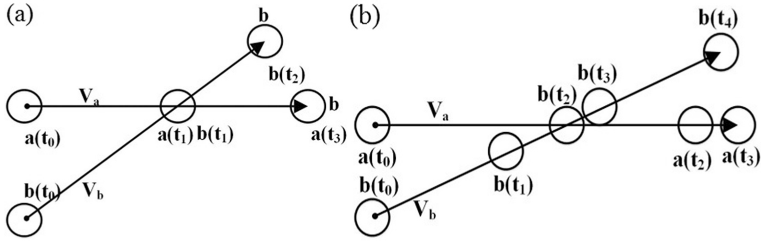


Fig. 7. Speed-shift avoidance: (a) collision at t_1 $(V_a < V_b)$; (b) a reaches the collision point far before b at the speed of $V_a(V_a > V_b)$.

4 Simulation Experiment and Result Analysis

In the process of dealing with emergencies, the formation is required to maintain a certain geometric shape according to the specific task requirements, and continue to transform in the development of the situation to better accomplish the task [9]. Typical basic simple formation includes line formation, diamond-shaped formation, wedge-shaped formation, the inner circle, the outer circle and so on [6]. In order to verify the feasibility of the proposed method, this paper simulates the transformation of the line formation into the wedge formation, and the transformation effect is shown in Fig. 8.

This experiment is implemented in the Matlab environment. Firstly, 100 source formation positions are given and divided into 10 groups, and each point located in the

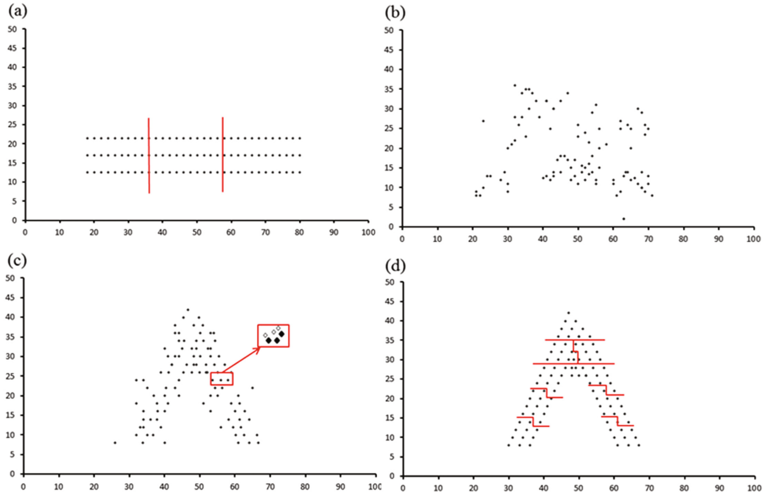


Fig. 8. The diagram of the transformation effect of the line formation to the wedge formation: (a) original formation (line); (b) transformation 1 (c) transformation 2 (d) target formation (wedge-shaped).

centroid position is the leader. And then, 100 target formation positions are given and divided into 10 regions by polygon, and the point closest to the center of each region is the target position of the leader. According to the greedy algorithm, the mapping coordinates of the leader of the initial formation to the target formation are obtained. Finally, the mapping coordinate of each team member is obtained by using the greedy algorithm. In the movement, collision avoidance is achieved according to the obstacle avoidance rule mentioned above (as shown in transformation 2).

Figure 8 shows that the formation begins to aggregate at transformation 1, that is to say, the leader groups emerge; and the target formation starts to emerge at formation 2. During the whole process of the transformation, there is always a gap between the points, and no collision occurs. From the perspective of the transformation, the path obtained by greedy algorithm is feasible, yet not optimal.

5 Conclusion

Based on the leader-follower algorithm, for formation transformation in emergencies,, this paper tries to get the position mapping of the team members from the original formation to the target formation by greedy algorithm. And in the process of path selection, collision detection and collision avoidance are realized by geometric constraint mechanism. The transformation of the line formation to the wedge-shaped formation is simulated by Matlab. The simulation results indicate that the decision-making process is distributed at each level, and the introduction of the greedy algorithm can reduce the information transmission, shorten the state probability of decision-making time, and greatly enhance decision-making adaptability under the dynamic uncertain condition. The method proposed in this paper is feasible for the formation transformation, and can improve the efficiency of formation design and exercise in emergencies. In the future, the formation transformation of small-scale units will be studied.

Acknowledgments. This work was supported by Shanxi Provincial Natural Science Foundation of China Youth (2015JQ6224) Colored Nodes and Network Complexity Measure Research.

References

1. Zhang, Z.Y., Zhang, R.B., Xin, L.: Hybrid formation control of multi-robot. J. Univ. Posts Telecommun. 38–41 (2008)
2. Wen, R., Li, D.W., Luan, X.F.: Path planning of robot based on ant colony algorithm. Comput. Digit. Eng. 20–22 (2012)
3. Consolini, L., Morbidi, F., Prattichizzo, D.: Leader–follower formation control of nonholonomic mobile robots with input constraints. Automatica **44**(5), 1343–1349 (2008)
4. Singh, J.A.: A greedy algorithm for task scheduling and resource allocation problems in cloud computing. Int. J. Res. Dev. Technol. Manag. Sci. Kailash **21**(1), 1–17 (2014)
5. Giryes, R.: A greedy algorithm for the analysis transform domain. Neurocomputing **173**, 278–289 (2016)
6. Lin, J.L., Hwang, K.S., Wang, Y.L.A.: Simple scheme for formation control based on weighted behavior learning. IEEE Trans. Neural Netw. Learn. Syst. **25**(6), 1033–1044 (2014)
7. Sharma, S., Jerripothula, S., Mackey, S.: Immersive virtual reality environment of a subway evacuation on a cloud for disaster preparedness and response training. In: 2014 IEEE Symposium on Computational Intelligence for Human-like Intelligence (CIHLI), pp. 1–6. IEEE (2014)
8. Wang, B., Wang, J., Zhang, B.: Leader-follower consensus for multi-agent systems with three-layer network framework and dynamic interaction jointly connected topology. Neurocomputing **207**, 231–239 (2016)
9. Peng, Z., Wen, G., Rahmani, A.: Leader-follower formation control of nonholonomic mobile robots based on a bioinspired neurodynamic based approach. Rob. Auton. Syst. **61**(9), 988–996 (2013)
10. Pan, Q.K., Ruiz, R.: An effective iterated greedy algorithm for the mixed no-idle permutation flowshop scheduling problem. Omega **44**, 41–50 (2014)
11. Ou, M., Du, H., Li, S.: Finite-time formation control of multiple nonholonomic mobile robots. Int. J. Robust Nonlinear Control **24**(1), 140–165 (2014)
12. Ma, T., Tang, B., Wang, Y.: The simulated greedy algorithm for several submodular matroid secretary problems. Theory Comput. Syst. **58**(4), 681–706 (2016)
13. Niewiadomska-Szynkiewicz, S.: Simulation-based design of self-organising and cooperative networks. Int. J. Space-Based Situated Comput. doi:10.1504/IJSSC.2011.039108

A Kind of Improved Hidden Native Bayesian Classifier

Yang Wenshuai(✉), Zhao Hongxu, and Gao Zhiqiang

Department of Information Engineering,
Engineering University of PAP, Xi'an, China
yangwswjgd@163.com

Abstract. In modern times, the number of sensing image is increasing on explosive speed. Human's ability on data analyzing and information accessing, however, has not caught up with this growth model, which requires our efforts to develop image mining technology, so that we can get what we need in short time. Classification and prediction method are two important contents on remote sensing image analyzing and information mining, as well as the focus of research. This paper has been written around the automation and intelligence of remote sensing image information obtaining and has done study on the remote sensing image data mining theory and technology. Besides, we have put forward an improved method on the Bayesian algorithm, having received a good effect.

Keywords: Remote sensing · Image · Data mining · Native Bayes

1 Introduction

Remote sensing is a kind of rising technology, which originates from 1960s. This technology can be used to detect the electromagnetic wave, visible light and infrared radiation of the target reflected from the distance. With the rapidly development and widely application of the remote sensing technology, the images are increasing greatly. How to get the effective information quickly and accurately has become a realistic problem. As far as I'm concerned, I think that taking measures to improve the accuracy and efficiency of data mining will undoubtedly become the best foothold.

Data mining (DM) is a process for people to extract the implicit information from the large, incomplete, noisy, fuzzy and random data. It is said that the information is unknown but potentially useful. However, the classification is the most important content of data mining. The main purpose of the classification is to build a model firstly and then to use this model for classification. Bayesian network, firstly proposed by R. Howard and J. Matheson in 1981, can integrate a priori knowledge and sample information. Besides, it has the ability to express uncertainty, which means that it is often used to solve the classification problem.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_71

2 The Naive Bayesian Classification

The Native Bayesian classification is a kind of statistical classification method, which is based on the theorem of Bayesian. Compared with the other classification algorithms, this kind of Bayesian classification has a higher accuracy rate. The model of it can be represented by this figure (Fig. 1).

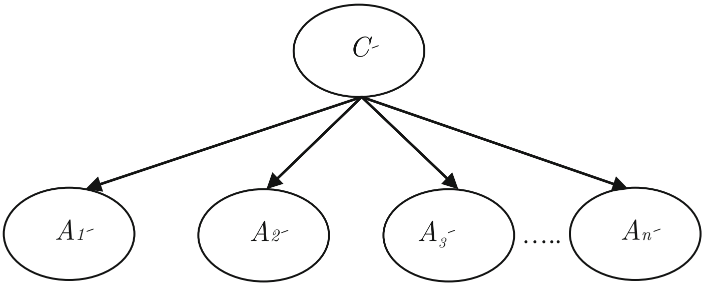


Fig. 1. The model of the Native Bayes classification

The naive Bayesian classification assumes that the value of each attribute is independent to each other, which is called the conditional independence. The specific algorithm:

(1) Each sample is represented as an n-dimensional feature vector: $X = (X_1, X_2, \ldots, X_n)$. And this kind of representing method can describe the n- measurement of different N attributes: $A_1, A_2, \ldots, A_n$.

(2) Suppose that there are m categories: C_1, C_2,……, C_m. And make x as the sample data, which is given unknown (no label for category). Classification will predict that X belongs to the category which has the highest posterior probability. In other words, the Naive Bayesian classification assigns the samples unknown to the C_i category, and it belongs to the largest category only if it is under the condition that $P(C_i|X) > P(C_j|X)$, $1 < j < m$, $j \neq i$. Besides, in this example, the $P(Ci|X)$, is called as the maximum a posteriori hypothesis. Maximize this according to the Bayes theorem:

$$P\left(C_i|X\right) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{1}$$

Since that the $P(x)$ is a constant for all categories, we only need maximize the product: $p(x|c_i)$ $P(c_i)$.

(3) On the other hand, if the prior probability of each category is unknown, it is generally assumed that these categories are equal probabilities, which means that $P(C_1) = P(C_2) = \ldots = P(C_n)$. And according to this, we can only maximize the $P(C_i|X)$. Otherwise, maximize this product: $P(X|C_i)$ $P(C_i)$.

(4) For the data set with many attributes that has been given, the calculation of the *P* $(X|C_i)$ is very expensive. In order to reduce the computation overhead of $P(X|C_i)$, we can make a conditional independence assumption. We suppose that it has been given sample category label, and every attribute value is independent of each other, which means that there are no dependence relationships between any two attributes. Then:

$$P(X|C_i) = \prod\nolimits_{k=1}^{n} P(X_k|C_i) \tag{2}$$

Probability $P(X_1|C_i)$, $P(X_2|C_i)$, $P(X_n|C_i)$ can be estimated by training samples.

(1) If $\mathrm{A_k}$ is a categorical attribute, the probability value $P(X_k|C_i) = Si_k/S_i$ is the number of the class *Ci*'s training samples with the value of number X_k on the property A_k. And besides, Si is the number of training samples of the class C_i.
(2) If A_k is a continuous value attribute, it is generally assumed that the property obeys the Gauss distribution. As a result, we can obtain this equation: $P(X_k|C_i) = g(X_k, U_C, C)$. And the value of number (A_k) of C_i's training sample should be given. The $g(X_k, C, C)$ is a kind of Gauss density function. The *Uc* is the average value and the *C* is the standard deviation.
(3) When classifying the certainly unknown sample *X*, $P(X|C_i)P(C_i)$ for each class of *Ci* should be computed. The sample X will be assigned to the class C_i only under the condition that $P(X|C_i)\ P(C_i) > P(X|C_j)\ P(C_j)$, $1 < j < m, j \neq i$. In other words, *X* will be assigned to the category named C_i, which has the largest $P(C_i|X)$. Besides, $i = max\{P(X|C_j)P(C_j),\ i = 1,2,\ldots\ N\}$.

3 The Hidden Naive Bayesian Classifier

In fact, the presenting of the Hidden native Bayesian classifier can be great progress. It looks like the TAN classifier in the structure. The accuracy of the classification can be greatly increasing by this means. But it doesn't mean that people can't find any way to make the result more accurate. Later, the description of this method will be given.

3.1 Brief Introduction of Hidden Naive Bayes Classifier

The so-called hidden naive Bayesian classifier, which is based on Naive Bayesian classifier, adds a hidden parent node for each attribute. This node can represent the dependency relationship among the attribute one with all the other attributes. Besides, it makes the calculation more convenient, including that it can lower the dependency relationship among the attributes. A hidden Naive Bayesian model has been put forward in Zhang's article. And by explaining this model and experiment, he has drawn the conclusion that this method can be much better than the Naive Bayesian model and the tree-extended Naive Bayesian model. Based on the article [3, 4], the schematic diagram of the structure has been shown in Fig. 2.

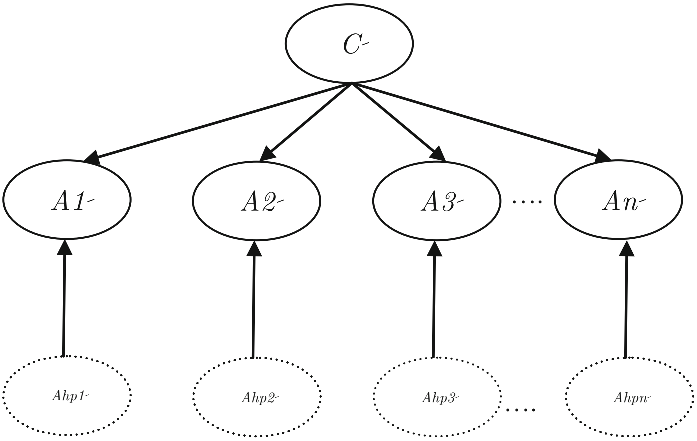


Fig. 2. Hidden Naive Bayes classification model

This figure has clearly expressed the relationships among C, A_i and A_{hpi}. In this figure, C represents the class node and A_i represent the attribute nodes, while A_{phi} represent the hidden parent nodes. Making a thorough analysis on this figure, we can easily obtain its joint distribution.

$$P(c)\prod\nolimits_{i=1}^{n}\left[p\left(A_i|A_{hpi},C\right)\right] \tag{3}$$

In this formula

$$p\left(A_i|A_{hpi},C\right)=\sum\nolimits_{j=1,j=i}^{n}w_{ij}*P(A_i|A_j,C) \tag{4}$$

$$\sum\nolimits_{j=1,j\neq i}w_{ij}=1 \tag{5}$$

The formulas (1), (2) and (3) are the key point of this method. Properly analyzing them, it is not difficult to find that the hidden parent nodes A_{hpi} are the weighted sum. This weighted sum represents the dependency relationship between the attribute A_i and all the other individual attributes.

The mutual information is used to define the weight here.

$$W_{ij}=\frac{Ip(A_i;A_j|C)}{\sum_{j=1,j=i}^{n}Ip(A_i;A_j|C)} \tag{6}$$

And in the formula, the significance of the element Ip(Ai;Aj|C) should be stated.

$$Ip(A_i; |C) = \sum\nolimits_{a_i,a_j,c} log \frac{p(a_i, a_j|c)}{p(a_i|c)p(a_j|c)} \tag{7}$$

What the weight meaning is the degree of dependence between every two attributes. The degree of dependence is directly proportional to the value. In other words, the greater the degree of dependence is, the greater the weight. On the contrary, the value of the weight will be smaller.

3.2 Evaluations of the Hidden Naive Bayesian Classifier

The hidden Naive Bayesian classifier tries to represent dependencies between one attribute node and the other attributes by introducing a hidden attribute for each parent node. This method can avoid the complexity of learning the Bayes optimal network structure, making full use of the dependencies between each attribute node. However, it is easy for us to find that the value of the weight (W_{ij}) can be very important from the two formulas. The significance of the value of the weight is that it is very important in the process of the construction of hidden Naive Bayesian model. Once the error of the weight is too big, the classification accuracy will be greatly affected.

In the hidden naive Bayes classifier, the authors have applied the mutual information method on the definition of the weights. Although many advantages this method has, there are still many problems waiting to be solved.

For example, analyzing from the point of view of computational methods, the feature selection algorithm of the mutual information tends to select rare words when the feature words are equal in a certain category. However, the characteristics with higher frequency can make for of the classification.

What should be stated is that the existence of this disadvantage will be conducive to the accuracy of classification. It is obvious that there will be a higher error if this Bayes classifier is used in the image data mining of remote sensing, especially in the military and other industries which require higher precision of data. As a result, the consequence will be difficult to be estimated. In order to improve the accuracy of the Hidden Bayes classifier, an improved model has been proposed in this paper. This model bases on the hidden naive Bayes classifier, which is called the implicit hidden naive Bayes classifier with improved mutual information.

4 The Implicit Naive Bayes Classifier with Improved Mutual Information

Basing on the information theory, the mutual information is a kind of variables, which can reflect the correlation between the two random variables. The two random variables refer to a_i and a_j here.

Something should be explained about the formula (7). The values of the *P(ai, aj*|*c)* represent the probabilities when a_i, a_j and c exist at the same time. The values of the

$p(a_i|c)$ are equal to the probabilities when a_i and c exist at the same time, and so as to the values of the $p(a_j|c)$.

Generally speaking, the bigger the value of $Ip(A_i;A_j|C)$ is, the stronger the connection will be among A_i, A_j and C.

4.1 The Basic Ideas

In the literature [1], Liu Song's group put forward a method to improve the mutual information, but this method only gives an idea. In other words, we ought to do something to improve it if we need solve a practical problem. This paper combines the disadvantage of the mutual information method to define the weight, the influence of the mutual information method to define the weight classification, with the classification accuracy of hidden Naive Bayesian classifier. And it gives a kind of improved hidden Naive Bayesian classifier whose weight is determined by the mutual information. Later, this method that improves the mutual information will be explained in detail.

To improve the mutual information, the weight difference factor X is introduced to reflect the importance of positive and negative correlation. For 'X', there is a formula, which can be its definition.

$$X = \frac{p(a_i, a_j|c) - p(a_i|c)p(a_j|c)}{p(a_i|c)p(a_j|c)} \tag{8}$$

And in consideration of X, the new improved mutual information formula can be calculated.

$$Ip(A_i;A_j|C) = \sum\nolimits_{a_i,a_j,c} X * log\frac{p(a_i, a_j|c)}{p(a_i|c)p(a_j|c)} \tag{9}$$

The analysis of the above formula is as follows.

On the one hand, if the limits that $p(a_i, a_j|c) > p(a_i|c)p(a_j|c), log\frac{p(a_i,a_j|c)}{p(a_i|c)p(a_j|c)} > 0$ are satisfied, every element such as X and $Ip(A_i;A_j|C)$ is positive value. That is to say that the formula (2) reflects the positive correlation degree between the attributes and the categories. On the other hand, if the limits aren't satisfied, it reflects the degree of negative correlation. All in all, it improves the mutual information weight attribute of A_i and A_j in the data.

4.2 The Algorithm Flow

The algorithm flow of the process is relatively simple. And later, the detail algorithm will be given.

```
    Algorithm: the pseudocode of the implicit hidden naive Bayes
classifier with improved mutual information
    Input: the training data set $D$
    For each $C'$s value $c$
    Computing the prior probability $P\ (c)$ for data set $D$
    For each attribute $A_i$, $A_j$
    For each assignment $a_i$, $a_j$ and $c$ for $A_i$, $A_j$ and $C$

    Calculated the $p\ (a_i, a_j|c)$ from$D$

    Calculated $X$
    For each attribute $A_i$, $A_j$
    Calculate the conditional mutual information

        $Ip\ (A_i; A_j|C) = \sum_{a_i, a_j, c} X * log \frac{p\ (a_i, a_j|c)}{p(a_i|c)p\ (a_j|c)}$

    For for each attribute $A_i$
    Calculate the attribute weighted sum which is associated
        $W_i = \prod_{nj=1,\ j \neq i} Ip\ (Ai; Aj|C)$
    For each attribute $A_j$, where $j \neq I$
    Calculate the weight between every two attribute

        $W_{ij} = Ip\ (A_i; A_j|C)\ W_i$
Procedure in testing
   Input: the sample $E=(a_1, a_2, \ldots a_n)$
   Step1  for $c=1$ to $C$
   Step2    for$i=1$ to $n$
Calculate the $P(A_i / A_{hpi}, C)$ according to the formula(4)
End for
End for
   Step3 Return $C(E) = argmaxP(c)\prod_{i=1}^{n} P(a_i|a_{p/i}, c)$
```

4.3 Test Procedures and Results

In order to prove that this Bayes classifier can be more effective, we select the UCI database as the sample of the experiment. What must be explained is that it is put forward for the machine learning by the University of Californialvine. With 335 datasets, it has become the most common one of standard testing.

The datasets we selected have been listed in the table. Besides, there are eight kinds of datasets in total, which can illustrate this problem. It is composed of four parts, such as the name, the number of samples, the number of attributes and the number of classes.

Each dataset in this table has its own characteristics. What make each of them unique are mainly the number of samples and the number of attributes. In other words,

Table 1. Datasets in UCI database

Name	Number of samples	Number of attributes	Number of classes
Abalone	4177	8	3
Automobile	205	26	3
Contraceptive	1473	9	2
Cylinder	512	39	3
Housing	506	14	3
Servo	167	4	2
University	285	17	2
Water	527	38	2

all of the datasets listed in the Table 1 are undoubtedly representative. By the implicit hidden naive Bayes classifier with improved mutual information, we classify the datasets. As a result, the degree of classification's accuracy has been listed in Table 2.

Table 2. Experimental results on accuracy

Name	HNB	NHNB
Abalone	0.5130	0.5181
Automobile	0.9207	0.9241
Contraceptive	0.9483	0.9491
Cylinder	0.7996	0.8002
Housing	0.9127	0.9140
Servo	0.7099	0.7067
University	0.9367	0.9372
Water	0.8965	0.8978

For ease of description and reading, we use the NHNB to replace the long-winded statement which is the implicit hidden naive Bayes classifier with improved mutual information. It is obvious that the NHNB has a stronger ability on classification. However, the promotion can't suit for each dataset. And it has been calculated that the average value of the NB is 0.8297, while another is 0.8309.

5 Summary

The author write this paper starting with the problem that it is very difficult for us to manage the remote sensing image data mining. And it has put forward an improved hidden Naive Bayesian classification with a higher classification accuracy, whose weight is determined by the mutual information. Although it is a little more complicated than the hidden Naive Bayesian classification, the results of the experiments show that the classification accuracy of the method can be higher. Besides, the accuracy of the classification is positively correlated with the accuracy of the data mining. However, condition limited, this paper only gives an improved method. So how to apply this method to the remote sensing image data mining will be the focus of future research.

References

1. Song, L., Dexian, Z., Research on the improvement of the mutual information based on the weight difference and the category correlation. Inf. Sci. **7**, 1998–2000 (2014)
2. Jinghui, L., Xiaogang, Z., et al.: An improved hidden naive Bayesian algorithm. Micro Comput. **7**, 1654–1658 (2013)
3. Liang-xiao, J., Zhang, H., Zhi-hua, C.: A novel Bayes model: hidden naive Bayes. IEEE Trans. Knowl. Data Eng. **21**, 1361–1371 (2009)
4. Zhang, H., Jiang, L., Su, J.: Hidden naive Bayes. In: Proceedings of the Twentieth National Conference on Artificial Intelligence, pp. 919–924 (2005)
5. Goyal, A., Bonchi, F., Lakshmanan, L.V.S.: Learning influence probabilities in social networks. In: International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, pp. 241–250. DBLP, February 2010
6. Cui, P., Jin, S., Yu, L., et al.: Cascading outbreak prediction in networks: a data-driven approach. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 901–909 (2013)
7. Chen, H., Hu, Y., Lian, Z., Jia, H., Wang, X.A.: INCIM: an additively homomorphic encryption over large message space. **10**, 82–102 (2015)
8. Wu, K., Kang, J., Chi, K.: Research on fault diagnosis method using improved multi-class classification algorithm and relevance vector machine, pp. 1–16
9. Honarvar, A.R., Sami, A.: Extracting usage patterns from power usage data of homes' appliances in smart home using big data platform, pp. 39–50
10. Bendekkoum, S., Boufaïda, M., Seinturier, L.: An approach based on service components for adapting web-oriented applications. Int. J. Inf. Technol. Web Eng. **11**, 1–21 (2016)

Study of a Disaster Relief Troop's Transportation Problem Based on Minimum Cost Maximum Flow

Zhen-Sheng Peng[1,2(✉)], Qing-Ge Gong[2], Yan-Yu Duan[2], Yun Wang[1,2], and Zhi-Qiang Gao[1,2]

[1] Key Laboratory of Cloud Computing and Big Data of PAP,
Xi'an 710086, China
pzsyjsgldd@163.com

[2] Department of Information Engineering, Engineering University of PAP,
Xi'an 710086, China

Abstract. In rescue missions, time is life, and only when the army arrives the first time around can risk to people's lives and property be minimized. Therefore, not only does a troop's transportation require reasonable dispatching but there is also a need to consider the time consumption problem of arriving at the location. Therefore, two factors – the number and time of transportation for disaster relief troops – are especially important. First, this study makes an in-depth analysis of the problem of troop dispatching in rescue and relief work, and proposes a network flow model of deploying troops thereof, thus making it a minimum cost maximum flow problem. Second, it defines the priority path to the existing minimum cost maximum flow algorithm; after joining the priority queue, the improved algorithm is more applicable to the study of a disaster relief troop's transportation problem. Finally, experiments are done concretely on examples based on real life troop data. The results show that the model can effectively support the disaster relief troop's transportation problem. The improved algorithm can effectively avoid small capacity and time-consuming deteriorated roads, and its time complexity is lowered from the original $O(n^2)$ to $O(n)$ on a path selection judgment. The algorithm results can provide scientific reference for a disposal of contingency plans when danger occurs.

Keywords: Rescue and relief work · Force transmission · Minimum cost maximum flow · Decision supporting

1 Introduction

A troop's transportation, as a military activity, is conducted in order to complete specific tasks. Therefore, whether we can meet the task demand is the basic standard to measure whether the troop's transportation is reasonable [1–4]. A troop's transportation and general goods transportation present obvious differences in the decision-making objectives. General goods' transportation decision goal is mainly economic efficiency, but a troop's transportation is influenced by many factors. In a rescue mission, time is life, and only when the army arrives the first time around can the risk to people's lives

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_72

and property be minimized. Thus, a study to determine how many troops can be sent to a disaster point in the shortest time has a very important strategic meaning and practical value, in order to work out the emergency plan the first time around, when danger occurs, and thus improve the efficiency of the emergency rescue.

2 Model of Rescue and Relief Troop Dispatching Network

In the task of troops deployed in rescue and relief work, the location and force's demand information of the disaster point will be provided first by intelligence, and the basic information of the road can be provided according to geographical command information systems [5–8]. The command post plans the number of troops dispatched to the save point and the force transmission path to reach the disaster areas on time.

The network flow model of troops deployed in rescue and relief work is composed of five parts: command post, save point, sending node, disaster points, and intelligence. Therefore, the disaster relief troop's transportation problem is described as the following: i is the number of save points to transport forces to k, which is the number of disaster points. That go through j, the number of sending points on the premise of making the selected route, and the flow of the route meeting the demand of delivery to make the troop's dispatching total time cost the shortest. The network flow model of troops deployed in rescue and relief work is shown in Fig. 1.

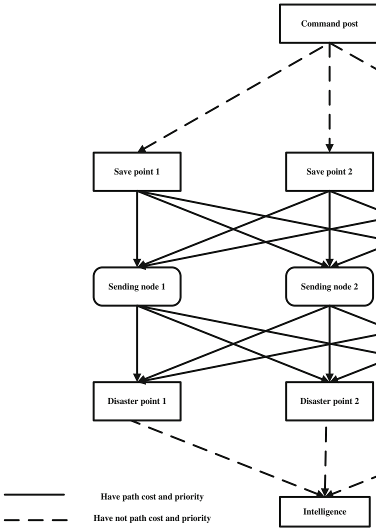


Fig. 1. The network flow model of troops deployed in rescue and relief work

In order to facilitate the problem and improve the generality of the model, this study assumed the following problems:

(1) The path distance from the save points to the sending nodes and from the sending nodes to disaster points are known;
(2) The force's demand of each disaster point is known or predictable;
(3) The information of the related roads are known;
(4) The sending time per force unit of different types of ways can be estimated;
(5) A sending node can get the troops from multiple save points and a disaster point's strength requirements can be provided by multiple sending nodes;
(6) Each road's force limit is known.

Based on the above assumptions, the model parameters and variables are defined as follows:

(1) i denotes a save point, I is the set of save points, such that $i \in I$;
(2) j denotes a sending node, J is the set of sending nodes, such that $j \in J$;
(3) k denotes a disaster point, K is the collection of disaster points, such that $k \in K$;
(4) Q_{ij} denotes the number of the troop that save points i sends to sending node j (unit: platoon);
(5) Q_{jk} denotes the number of the troop that sending nodes j sends to the disaster point k (unit: platoon);
(6) c_{ij} denotes the biggest force quantity of a save point i sent to a sending node j (unit: platoon);
(7) c_{jk} denotes the biggest force quantity of a sending node j sent to disaster points k (unit: platoon);
(8) T_{ij} denotes sending time per force unit from the save point i to sending node j (unit: min/platoon);
(9) T_{jk} denotes sending time per force unit from the sending node j to the disaster point k (unit: min/platoon);
(10) A_i denotes the supply capacity of the save point i (unit: platoon);
(11) M_j denotes the max force capacity of the sending node j (unit: platoon);
(12) D_k denotes the strength of aggregate demand of disaster points k (unit: platoon).

The purpose of the troop's deployed network flow problem in rescue and relief work is solved by the objective function T under the condition of meeting the force demand of all the disaster points to achieve the minimum total time $\sum_{\alpha \in \mathrm{E}} t_{\alpha} \cdot f_{\alpha}$ (defined below). The linear constraint model is as follows:

$$\boldsymbol{min}\ T = \sum_{\alpha \in \mathrm{E}} t_{\alpha} \cdot f_{\alpha}$$

$$\mathbf{s.t.}\left\{A_i = \sum\nolimits_{j=1}^{J} Q_{ij}\right. \tag{1}$$

$$\mathbf{s.t.}\left\{\sum\nolimits_{i=1}^{I} Q_{ij} = \sum\nolimits_{k=1}^{K} Q_{jk}\right. \tag{2}$$

$$\mathbf{s.t.}\left\{\sum\nolimits_{i=1}^{I} Q_{jk} \le M_j\right. \tag{3}$$

$$\mathbf{s.t.}\left\{\sum_{i=1}^{I} Q_{jk} = D_k\right. \tag{4}$$

$$\mathbf{s.t.}\left\{Q_{ij} \le c_{ij}\right. \tag{5}$$

$$\mathbf{s.t.}\left\{Q_{jk} \le c_{jk}\right. \tag{6}$$

(1) The constraint (1) is for the supply capacity.
(2) The constraint (2) is the balance constraint of forces in and out of the number of sending nodes. It is said that the amount of force sent to sending node j is equal to the force shipped from sending node j.
(3) The constraint (3) is the limited capacity of the sending nodes. The quantity transport from the save point to sending node j must be less than the maximum capacity of sending node j.
(4) The constraint (4) is the demanding constraint. The force transport quantity from all the sending nodes to the disaster point k must meet the force needs of the disaster point k.
(5) The constraints (5) and (6) are the limited capacities for each road.
(6) All of the parameters involved in this algorithm are non-negative parameters.

3 The Troops Dispatching the Algorithm in Rescue and Relief Work is Based on a Path Priority Limit

First, the basic concepts involved in this algorithm are defined as follows:

Definition 1 (Force − Time network diagram): $G = (V, E, c)$ is a directed graph with weightings, having the command post v_s and intelligence v_t. t is defined as a non-negative function base on E. $\forall \alpha = (v_i, v_j)\text{or}(v_j, v_k) \in E$, $t(\alpha)$ states the sent time per force unit on arc α, marked $t(\alpha) = t_{ij}\text{or } t_{jk}$, said network D is the forces-time network diagram, marked $D = (V, E, c, t)$.

Definition 2 (Force flow diagram): $W(c_\alpha, f_\alpha)$ is a directed graph with weightings, c_α, f_α are max force capacity, and the actual force flow on arc α is the $W(c_\alpha, f_\alpha)$ force flow diagram;

Definition 3 (Time network diagram): $W(t_\alpha)$ is a directed graph with weightings, t_α is the time of force transport spent on arc α, said $W(t_\alpha)$ is the time network diagram.

Definition 4 (Path set): Given a forces-time network diagram $D = (V, E, c, t)$ and a feasible flow f, L_k is defined as a path of feasible flow, $L_k = \{\alpha | \alpha \in E, k \in N\}$.

Definition 5 (Path priority): Each arc in the forces-time network diagram $D = (V, E, c, t)$ have priority $P_\alpha = (c * t)/\theta$, θ are the parameters of road types, dimensionless. Path priorities $P_{L_k} = \sum_{\alpha \in L_k} P_\alpha$, the priority parameter as shown in Table 1.

Table 1. Parameter of road type

Road type	θ
Highway	1
Forest	2
River-way	3
Turbulence	4
Swamp	5

Definition 6 (Path time): Set a path L_k, the path time is the sum of products of force and the force transport time, marked the total time as $T(k) = \sum_{\alpha \in L_k} t_\alpha \cdot f_\alpha$.

Definition 7 (Path minimum differential forces): Set a path L_k, for $\forall \alpha \in L_k$, marked $\mu = \min\{c_\alpha - f_\alpha\}$ as the path minimum of differential forces.

Definition 8 (Time remaining diagram): Set forces-time network diagram $D = (V, E, c, t)$ reached the maximum flow. Figure of the vertex is the same with the forces-time network diagram, having both directions of the arc, called the time remaining diagram, marked $W(r_\alpha^{\pm})$. r_α is the weight of the arc, its value is defined as:

$$r_\alpha^+ = \begin{cases} t_\alpha & f_\alpha < c_\alpha \\ +\infty & f_\alpha = c_\alpha \end{cases} \quad r_\alpha^- = \begin{cases} -t_\alpha & f_\alpha > 0 \\ +\infty & f_\alpha = 0 \end{cases}$$

When the arc power is infinite, the arc is not shown on the time remaining diagram.

Use the improved minimum cost maximum flow algorithm to deal with the model. In the solving process of the algorithm do the following. First, maximize the troops through the choice of path and calculation. Second, get the optimal time by adjusting and testing it step by step. Finally, achieve the shortest time, and the largest force. Steps of the algorithm are as follows:

(1) Obtain the force flow diagram by the forces-time network diagram.
(2) Using depth search to all paths in time network diagram, from the command post to the intelligence. Then enqueuing the path according to the path time from short to long. When the paths time are the same, enqueuing the path should be according to the path priority from high to low.
(3) Path dequeue and step to (4) if the path has a saturated arc. Otherwise, update the force and time according to the path minimum differential forces based on the force flow diagram. Subsequently, the path should be more than one saturated arc at least. Mark the new saturated arc.
(4) If the queue is not empty, repeat steps (3). If the queue is empty, the network gets the maximum forces flow.
(5) Drawing out the time remaining diagram according to the updated forces-time network diagram. Looking for the sum of the negative time loop, and adjust the forces-time network diagram according to the founded circuit.
(6) Repeat (5) until you could not find the adjust loop. Then time is the shortest, and the network flow achieves the shortest time and the biggest force.

Algorithm process is shown in Fig. 2.

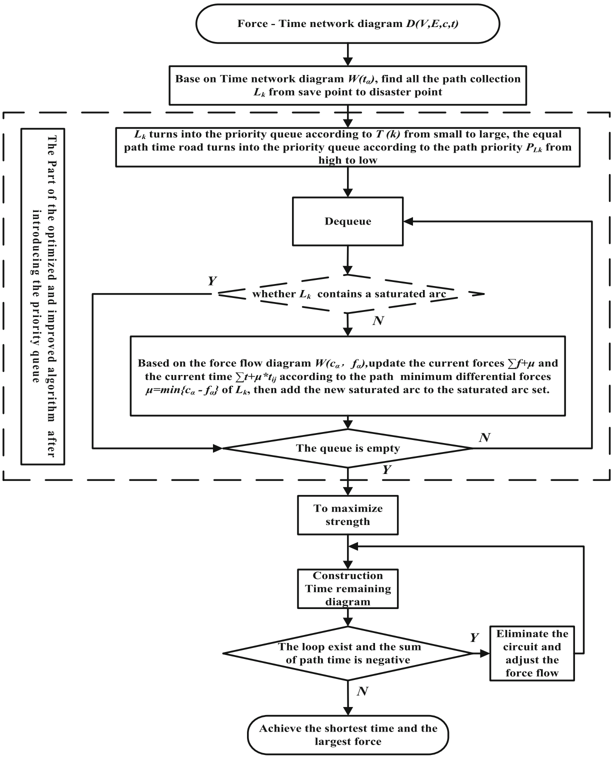

Force - Time network diagram $D(V,E,c,t)$
Base on Time network diagram $W(t_\alpha)$, find all the path collection L_k from save point to disaster point
The Part of the optimized and improved algorithm after introducing the priority queue
L_k turns into the priority queue according to T (k) from small to large, the equal path time road turns into the priority queue according to the path priority P_{Lk} from high to low
Dequeue
Y
whether L_k contains a saturated arc
N
Based on the force flow diagram $W(c_\alpha, f_\alpha)$,update the current forces $\sum f+\mu$ and the current time $\sum t+\mu*t_{ij}$ according to the path minimum differential forces $\mu=min\{c_\alpha - f_\alpha\}$ of L_k, then add the new saturated arc to the saturated arc set.
N
The queue is empty
Y
To maximize strength
Construction Time remaining diagram
The loop exist and the sum of path time is negative
Y
Eliminate the circuit and adjust the force flow
N
Achieve the shortest time and the largest force

Fig. 2. The flow chart of the troop's dispatching algorithm in rescue and relief work

Path priority does not exist in the original algorithm, which will result in the uncertainty sort of same cost (time) path. This leads to an unclosed minimum cost flow of the maximum flow network and also leads to number increase of the subsequent adjustable rings. Therefore, the algorithm efficiency is low. In addition, each time, the original algorithm marks the path that will traverse through all the paths to determine whether they contain a saturated arc and then determine whether all paths are marked. The time complexity of this process is $O(n^2)$. The algorithm is introduced into the priority queue only when it needs to determine whether the current path contains a saturated arc and to determine whether a queue is finally empty. The time complexity of this process is $O(n)$. The algorithm has a certain improvement in overall speed. The introduction of the priority queue can effectively circumvent the path problems like small capacity, time consumption, and bad road conditions through the following example.

4 Example Analysis of Troops Dispatching in Rescue and Relief Work

This example takes an actual disaster relief mission as an example: first, the intelligence is that there are two disaster points. This task has two suitable sending nodes for troops transit and scheduling. The command post decided that the rescue mission had two save points to finish together. Various parameters of the network flow model of troops deployed in rescue and relief work are given in Table 2.

Table 2. Parameters of the rescue and relief troop's example

Parameter types	Parameter	Value	Unit
c_{ij}	$c_{11}, c_{12}, c_{21}, c_{22}$	7, 3, 5, 6	Platoon
c_{jk}	$c_{11}, c_{12}, c_{21}, c_{22}$	6, 4, 6, 7	Platoon
T_{ij}	$T_{11}, T_{12}, T_{21}, T_{22}$	4, 2, 3, 3	min/platoon
T_{jk}	$T_{11}, T_{12}, T_{21}, T_{22}$	1, 2, 1, 1	min/platoon
A_i	A_1, A_2	7, 8	Platoon
M_j	M_1, M_2	12, 9	Platoon
D_k	D_1, D_2	5, 7	Platoon

According to the information provided by the Table 2 and the drawing of the forces-time network diagram. The numbers (t, f, c) besides the arc are time, flow, and capacity respectively. As shown in Fig. 3. The calculation steps are specific as follows:

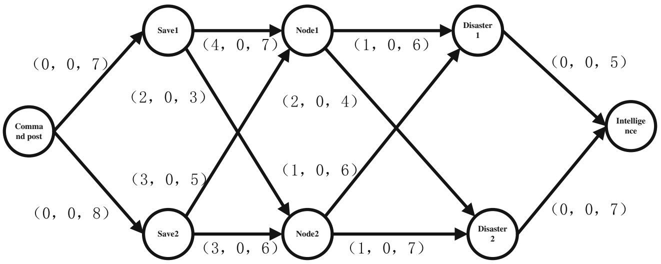


Fig. 3. Forces-time network diagram

(1) Structure the time network diagram based on the forces-time network diagram, as shown in Fig. 4.
(2) Using depth search to all the paths L_k in the time network diagram $W(t_\alpha)$ from the command post to the intelligence, and calculates the path priority based on Table 1. The result is shown in Table 3.

The path is in the sequence arrangement according to the path time $T(k)$ from small to large, and the path priority from high to low.

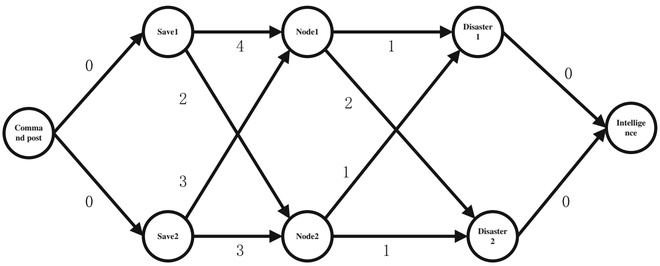


Fig. 4. Time network diagram

Table 3. Calculation process of path priority

Path	Road type	c	t	P_{α}
save1-node1	Highway	7	4	28
save1-node2	River-way	3	2	2
save2-node1	Highway	5	3	15
save2-node2	Forest	6	3	9
node1-disaster1	Swamp	6	1	1.2
node1-disaster2	Turbulence	4	2	2
node2-disaster1	Forest	6	1	3
node2-disaster2	River-way	7	1	2.3

L_0: Initial zero feasible flow, $T(0) = 0$
L_1: Command post-Save point1-Sending node2-Disaster point1-Intelligence, $T(1) = 3$, $P_{L_1} = 2 + 3 = 5$
L_2: Command post-Save point1-Sending node2-Disaster point2-Intelligence, $T(2) = 3$, $P_{L_2} = 2 + 2.3 = 4.3$
L_3: Command post-Save point2-Sending node1-Disaster point1-Intelligence, $T(3) = 4$, $P_{L_3} = 15 + 1.2 = 16.2$
L_4: Command post-Save point2-Sending node2-Disaster point1-Intelligence, $T(4) = 4$, $P_{L_4} = 9 + 3 = 12$
L_5: Command post-Save point2-Sending node2-Disaster point2-Intelligence, $T(5) = 4$, $P_{L_5} = 9 + 2.3 = 11.3$
L_6: Command post-Save point1-Sending node1-Disaster point1-Intelligence, $T(6) = 5$, $P_{L_6} = 28 + 1.2 = 29.2$
L_7: Command post-Save point2-Sending node1-Disaster point2-Intelligence, $T(7) = 5$, $P_{L_7} = 15 + 2 = 17$
L_8: Command post-Save point1-Sending node1-Disaster point2-Intelligence, $T(8) = 6$, $P_{L_8} = 28 + 2 = 30$

(3) According to the path queue order in (2). Dequeueing the path L_k and judging whether it contains a saturated arc will decide whether to adjust the force of this path. The result of the calculation includes the path, path time, path minimum

differential forces, current total force, current total time, new saturated arc, queue length, and the next saturated arc — a total of 8 indexes. The specific calculation process is shown in Table 4.

Table 4. Calculation process

L_k	$T(k)$	μ	Current total force	Current total time	New saturated arc	Queue length	Next saturated arc
L_1	3	3	3	9	(Save1,Node2)	7	L_2
L_3	4	2	5	17	(Disaster1,Intelligence)	5	L_4
L_5	4	6	11	41	(Command,Save2), (Save2,Node2)	3	L_6, L_7
L_8	6	1	12	47	(Disaster2,Intelligence)	0	Team is empty

(4) According to the calculation results of Table 4, we get the new forces-time network diagram. The maximum flow network is close to the shortest time network, as shown in Fig. 5.

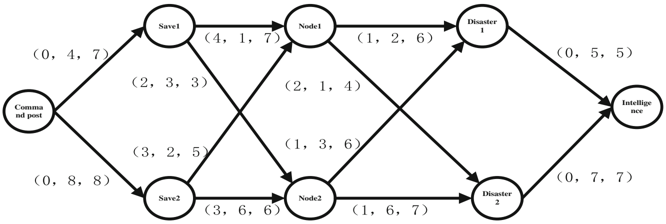


Fig. 5. Forces-time network diagram in largest forces

(5) According to Fig. 4, we draw the time remaining diagram shown in Fig. 6. Check the loop starting from the command post in the time remaining diagram. There is: 【command post-save point1-sending node1-disaster point1-sending node2-save point2-command post】, 【command post-save point1-sending node1-disaster point2-sending node2-save point2-command post】, 【command post-save point1-sending node1-save point2-command post】 three circuits of 1, 2, 1 total weights respectively. There is no circuit with negative values.

We delete the command post and its arc from the time remaining diagram and then check the loop from save point 1. It finds that there is no circuit with negative values. Next, we delete save point 1 and its arc from the time remaining diagram, and then check out save point 2 in the time remaining diagram, we can find that still there is no circuit with negative values. If we delete save point 2 and its arc from the time remaining diagram and then check the sending node 1, we find the existing loop sending node 1 − disaster point 1 − sending point 2 − disaster point 2 − sending node 1, where the total weight is – 1, as shown in Fig. 7.

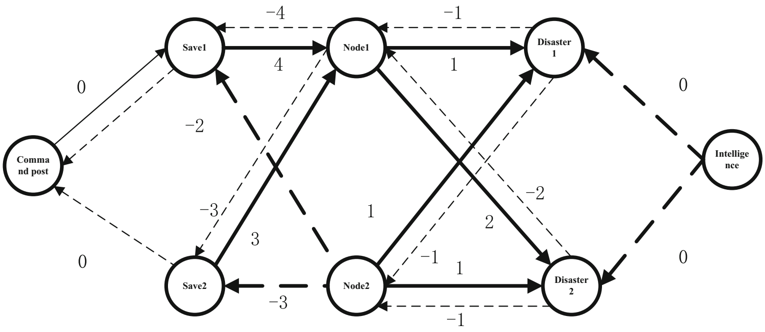


Fig. 6. Time remaining diagram

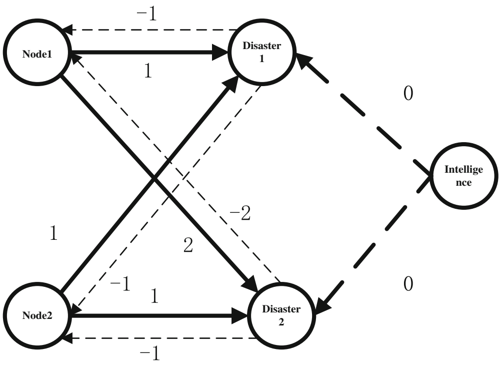


Fig. 7. Time remaining diagram with the sending node 1 as starting point

As we adjust the forces-time network diagram along with the loop in the time remaining diagram, we then get a new forces-time network diagram, as is shown in Fig. 8. After the adjustment, the total time is $T = 47 + (1 - 1 + 1 - 2) * 1 = 46(min)$

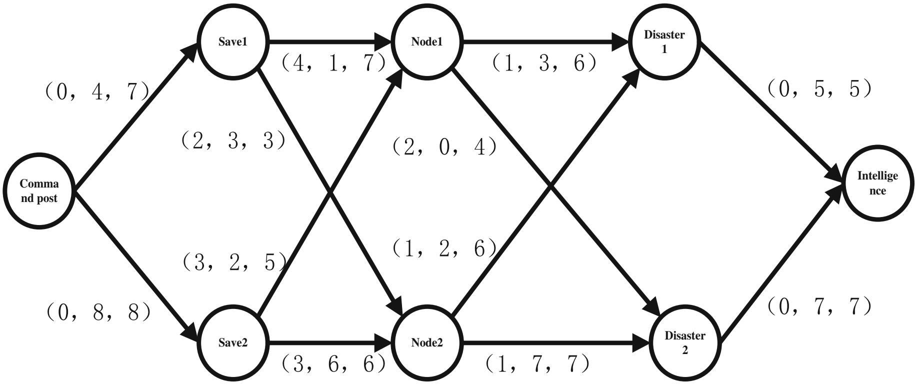


Fig. 8. Forces-time network diagram after inspection and adjustment

There is no circuit with negative values in the new time remaining diagram of the updated forces-time network diagram. The adjustment is down. The force flow, time, and road information of the obtained network are shown in Table 5.

Table 5. The information of the shortest time and largest force network

Parameter types	Parameter	Value (platoon)	Parameter	Value (min/platoon)	Passing time (min)	Road type
Q_{ij}, T_{ij}	Q_{11}	1	T_{11}	4	4	Highway
	Q_{12}	3	T_{12}	2	6	River-way
	Q_{21}	2	T_{21}	3	6	Highway
	Q_{22}	6	T_{22}	3	18	Forest
Q_{jk}, T_{jk}	Q_{11}	3	T_{11}	1	3	Swamp
	Q_{12}	0	T_{12}	2	0	Turbulence
	Q_{21}	2	T_{21}	1	2	Forest
	Q_{22}	7	T_{22}	1	7	River-way

In the empirical analysis of the results, you can find the choice of the road is good when the force flow is small and the passing time is short in a road of high-risk turbulence, low capacity, and a high time-consuming marsh area. It proves that the troop's dispatching algorithm in the rescue and relief work is based on the path priority limit, and can efficiently avoid small capacity roads that are time-consuming and represent high risk.

The above information provides the following reference for disposal plans in rescue and relief work, as shown in Table 6.

Table 6. Troop's dispatch scheme in rescue and relief

Save point	Force	Divide forces	Path	Road type	Task time	Send comments
save point1	4 platoon	1 platoon	sending node1-disaster point1	Highway-swamp	5 min	Walking turn vehicle
		3 platoon	sending node2-disaster point2	River-way-river-way	9 min	Speedboats
save point2	8 platoon	2 platoon	sending node1-disaster point1	Highway-swamp	8 min	Walking turn vehicle
		2 platoon	sending node2-disaster point1	Forest-forest	8 min	Motorized travel
		4 platoon	sending node2-disaster point2	Forest-river-way	16 min	Walking turn speedboats

The whole dispatching task uses 12 platoon forces. The total time consumption of the path is 46 min. The deadline of task arrival in the regional is 16 min after task begins.

5 Conclusion

The heart of the dispatch problem is to use the shortest time to send forces to the task region under the condition of meeting the requirements of strength; that is, in effect, finding the best dispatch scheme to make the task time the shortest. The troop's dispatching algorithm in rescue and relief work is based on the priority queue introduced into the concept of path priority on the basis of the original algorithm. In the process of path selection and the markers' judgment, the time complexity of this algorithm descended from the original time complexity of $O(n^2)$ to $O(n)$, which gets closer to the shortest time flow network compared with the original algorithm, shortening the time of finding an adjustable ring and thus improving the overall efficiency of the algorithm.

Through the analysis of the actual case, it can be found that the introduction of the priority queue can effectively circumvent the path, which is small capacity, time consuming and has a precarious-condition road. The network flow model of troops deployed in rescue and relief work and the optimized algorithm provided a strong basis of theoretical reference to formulate response plans. Thus, forcing a delivery plan and transmission mode helps it achieve the maximization of troops in the shortest delivery time, which has a very important strategic meaning and practical value for effective implementation in rescue and relief work.

Acknowledgments. This work was supported by Shaanxi Provincial Natural Science Foundation of China Youth (2015JQ6224) Colored Nodes and Network Complexity Measure Research.

References

1. Lu, B., Dai, X.: Quality evaluation on troops transport capacity based on interval grey numbers multi-attribute decision-making method. In: Software Engineering and Knowledge Engineering: Theory and Practice, pp. 901–908 (2012)
2. Kemball-Cook, D., Stephenson, R.: Lessons in logistics from Somalia. Disasters **8**(1), 57–66 (1984)
3. Rathi, A.K., Solanki, R.S., Church, R.L.: Allocating resources to support a multicommodity flow with time windows. Logistics Transp. Rev. **28**(2), 167–188 (1992)
4. Haghani, A., Oh, S.C.: Formulation and solution of a multi-commodity, multi-modal network flow model for disaster relief operations. Transp. Res. Part A Policy Pract. **30**(3), 231–250 (1996)
5. Vygen, J.: On dual minimum cost flow algorithms (extended abstract). Math. Meth. Oper. Res. **56**(1), 101–126 (2002)
6. Olabarriaga, S.D., Breeuwer, M., Niessen, W.J.: Minimum cost path algorithm for coronary artery central axis tracking in CT images. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003, pp. 687–694. Springer, Heidelberg (2003)
7. Ying, X., Huang, X.: Study on the critical side of transportation network for the emergency material base on the condition of minimum cost & maximum flow. In: Information Engineering and Applications, pp. 692–697 (2012)

8. Gao, J., Yang, J.F.: A new algorithm of mincost-maxflow. J. Yuncheng Univ. **3**, 23–26 (2016)
9. Weiwei, K., Wang, B., Lei, Y.: Technique for infrared and visible image fusion based on non-subsampled shearlet transform and spiking cortical model. Infrared Phys. Technol. **71**, 87–98 (2015)
10. Zabukovec, A., Jaklič, J.: The impact of information visualisation on the quality of information in business decision-making. Int. J. Technol. Hum. Interact. **11**(2), 61–79 (2015)
11. Lapo, M., et al.: Business intelligence system design and its consequences for knowledge sharing, collaboration, and decision-making: an exploratory study. Int. J. Technol. Hum. Interact. **11**(4), 1–25 (2015)
12. Yerra, R.V.P., Rajalakshmi, P.: Effect of relay nodes on end-to-end delay in multi-hop wireless ad-hoc networks. In: International Conference on Advanced Information Networking and Applications Workshops, pp. 343–348. IEEE (2013)
13. Baert, A., et al.: Data replication optimisation in grid delivery network. Int. J. Grid Util. Comput. **1**(4), 287–295 (2009)
14. Mondol, M.A.S., Akbar, M.M.: A new approach to schedule workflow applications for advance reservation of resources in grid. Int. J. Grid Util. Comput. **5**(3), 165–182 (2014)

A Cascading Diffusion Prediction Model in Micro-blog Based on Multi-dimensional Features

Yun Wang[1(✉)], Zhi-Ming Zhang[2], Zhen-Shen Peng[1], Yan-Yu Duan[1], and Zhi-Qiang Gao[1]

[1] Engineering University of CAPF, Xian, China
841880118@qq.com, 1090398464@qq.com,
pzsyjsgldd@163.com
[2] Department of Information Engineering,
Engineering University of CAPF, Xian, China
zhmzhang@sina.com

Abstract. Micro-blog, as a kind of weak relationship network, strengthens the communication among the bloggers, and propagates instant information in the social network. With the explosive growth of information flow in social network, researchers have a growing realization that it is essential to accurately predict the cascading diffusion of a message, which is of paramount importance to applications like public opinion monitoring, viral marketing and outbreaks detection. Although there have been extensive previous works on diffusion prediction, what kind of factors affects the information diffusion most and how to predict the propagation process are the focusing issues all the time. This paper analyzes the information dissemination and forwarding mechanism in the social network. In particular, we extract main features from multiple dimensions including node attributes, message content characteristics and the topology relation between nodes. Based on these features, this paper proposed a cascades diffusion model to predict the propagation process. Besides, we quantitatively evaluated the weights of the features in the proposed model by a stochastic gradient descent algorithm. We evaluate the proposed method on Sina micro-blog dataset. The experimental results show that the proposed method outperforms the other common models in precise prediction.

Keywords: Micro-blog · Cascading model · Diffusion prediction · Diffusion process · Multi-dimensional features

1 Introduction

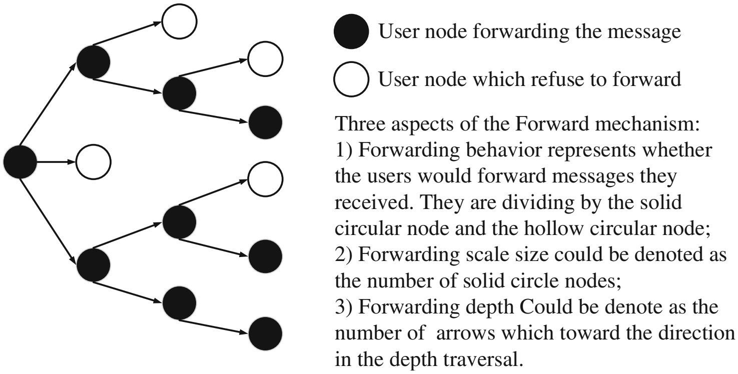
There are millions of people in process of passing information in social network every moment worldwide. The flow of information not only promotes the development of society, but also influences and promotes the cultural, economic, political and other fields [1]. Though there are many plenty of papers on information propagation models, most of them focus more on the importance of node inside the propagation patterns [2]. We not only focus on the features and behaviors of the various parts in the process of

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_73

model, but also make the optimization of the propagation model based on additional dimensional features. At the beginning, we mainly introduce the following two aspects as follows.

1.1 Forwarding Mechanism

According to the in-degree relation between the nodes, information forwarding path developed by layers of nodes' forwarding is shown in Fig. 1. Micro-blog network is formed based on the relation between bloggers. Research on the information forwarding mechanism is to obtain the result of three aspects: (1) forwarding behavior prediction; [3] (2) forwarding scale size; (3) forwarding depth. And the forwarding behavior prediction can be regarded as a binary classification problem. We build a model to calculate the probability of the bloggers' forwarding behavior by extracting the features from above three aspects.



● User node forwarding the message

○ User node which refuse to forward

Three aspects of the Forward mechanism:
1) Forwarding behavior represents whether the users would forward messages they received. They are dividing by the solid circular node and the hollow circular node;
2) Forwarding scale size could be denoted as the number of solid circle nodes;
3) Forwarding depth Could be denote as the number of arrows which toward the direction in the depth traversal.

Fig. 1. Forwarding mechanism.

1.2 Model of Information Dissemination

In the current social network research, there are two basic and classical model of information dissemination process: the independent cascade model (IC model) and the linear threshold model (LT model). As for the IC model, the edge between nodes represents the transmission probability $P\ (u_i,\ v_j)$, as shown in Fig. 2 of Sect. 3. At the beginning of the propagation, some nodes are in the initial activation state, which are called the seed node, and we define them as the set of S_0. In the nth step of the propagation process, the node u_i which was activated in last step will activate its adjacent node v_j with a probability of $P\ (u_i,\ v_j,)$. And each of the probabilities between different nodes is independent in any step of the propagation process.

As for the linear threshold model (LT model), it concentrates on the receiver nodes, and each receiver has a threshold defined as $\theta \in [0,\ 1]$. The edge between nodes represents the influence how much the node v is influenced by node u, which is defined

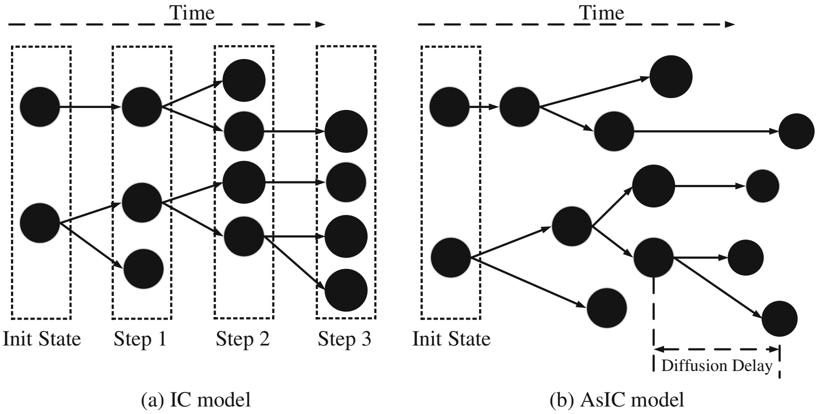


Fig. 2. IC model (a) and AsIC Model (b).

as weight W_{uv}. In the process of transmission, v would be activated when the total influence from the adjacent nodes of v exceed θ_v.

However, the information dissemination process is continuous and asynchronous with different transmission delay in reality. So the Saito K, Kimura M, Ohara K et al. [4]. improve the above model, and preliminarily propose an optimized asynchronous model—the asynchronous independent cascade model (AsIC model) and the asynchronous linear threshold model (AsLT model), as shown in Fig. 2 of Sect. 3. A representative work by A Goyal et al. [5] put forward two influence models under discrete time and continuous time. Although it proves that the evolution of the influence between nodes is attenuated with time, it does not involve the influence of node attributes on the formation and propagation. And the reference [6] proposes a data driven approach based on Orthogonal Sparse Logistic Regression (OSLOR) method, however, it doesn't figure out the common characteristics of these nodes discovered by data driven approach. [6] This paper is enlightened by the previous works, and extracts relative features to propose an optimized AsIC model to predict the diffusion process.

2 Notations and Problem Statement

We presume there is a directed social network $G\ (V,\ E)$ including a set of nodes V described as $V = \{v_1,\ v_2 \ldots\ v_n\ \}$, and a set of directed edge E described as $E = \{e_1,\ e_2 \ldots\ e_n\ \}$. Figure 2 shows the IC model and AsIC model in propagation progress of information. [7] As for the difference, the former assumes that the transmission delay defined as $\tau\ (u,\ v)$ is identical, and the propagation process is discrete and synchronous along the time axis. However, the latter is based on the continuous and asynchronous transmission delay. This paper based on the AsIC model proposes a modeling approach in regard to a fine-grained propagation probability $P\ (u,\ v,\ c)$ and transmission delay $\tau\ (u,\ v,\ c)$ [8].

3 Model Framework

In micro-blog network, the information propagation process can be decomposed into multiple information dissemination cases [9], defined as $m^{k}_{u,v,tu,tv}$, which includes several line segments. And each end of line segment connects one of the three types of entities involving sender node entities S, receiver node entities R and content entities C. The relationship between the entities is shown in Fig. 3.

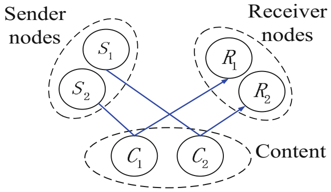


Fig. 3. Relationship of three entities in information diffusion process.

3.1 Features Extraction

As indicated in Table 1. This paper takes the micro-blog platform as an example to extract the main features as follows:

Table 1. The main features extracted from three aspects.

Three aspects	Main features	Abbreviations
Node attributes	The influence of the node	Influ(u)
	The authority of the node	Auth(u)
	The activity of the node	Act(u)
	The willing of the node	Will(u)
Content attributes	The sentiments of content	Senti(c)
	Contains URL link	URL(c)
	The number of Tags	Tags(c)
Edge attributes	Interest similarity	Sim-i(u,v)
	Structural similarity	Sim-s(u,v)
	Forwarding interest	Sim (u,v)

3.1.1 Node Attributes ϕ_U

(1) **Sender Node Attributes** ϕ_S. For the receiving node,the features are as follows:

The Influence of the Node. The influence of node could be calculates as the sum of the influence of the posted message within a period of time.

$$Influ(u) = \lg(\#RT + 1).$$

The Authority of the Node. We use the ratio of in-degree and out-degree to represent node authority, which is calculated as the rate of the fans number and the blogger's attention number.

$$Auth(u) = \lg\left(\frac{\#followers}{\#focus} + 1\right).$$

The Activity of the Node. The number of daily posted messages denote the activity of the node.

$$Act(u) = \lg\left(\frac{\#posts}{\#days} + 1\right).$$

(2) **Receiver Node Attributes ϕ_R**. The feature is referred to as the willing of receiver node, which reflects whether the blogger is willing to forward a message.

The Will of the Node. The ratio of the number of forwarding and the number of posting could indicate the will of the node.

$$Will(u) = \lg\left(\frac{\#forward}{\#post} + 1\right).$$

3.1.2 Content Attributes ϕC

In addition to the extracted features from the node, the features extracted from the message content also have a great impact on the information dissemination.

The Sentiments of Content ϕ_{SC}. This paper divides message content into the positive intention or negative intention based on binary classification methods, and extracts the emotional features of contents by sentiment analysis tool ROST, and standardize the value into the range of [−1, 1].

$$Senti(u) = \frac{\#positive - \#negative}{\#positive + \#negative}.$$

Contains URL Link(#URL#) ϕ_{URL}. The URL link typically represents a deep expression of the content, which is based on the micro-blog message content.

$$URL(u) = \begin{cases} 1, including\#URL\#; \\ 0, without\#URL\#. \end{cases}$$

The Number of Tags(#Tags#) ϕ_{Tag}. The contents with plenty of tags are more likely to attract the attention of other readers.

$$Tags(u) = \begin{cases} 1, including\#Tags\#; \\ 0, without\#Tags\#. \end{cases}$$

3.1.3 Edge Features ϕE

The Relation Between the Sender and Receiver Nodes $\phi_{S,R}$

Interest Similarity. We can record a blogger's messages in a document, and construct a vector based on the TF - IDF value of the words. The cosine of messages contents vector and blogger's documents vector could represent the blogger's interest in the forwarding of message content:

$$Sim - i(u, v) = \cos(\theta) = \frac{\mathbf{U} \cdot \mathbf{V}}{|\mathbf{U}| \times |\mathbf{V}|}.$$

Structural Similarity. The Jaccard distance between two neighbor-node sets indicates the similarity of the structure:

$$Sim - s(u, v) = \left| \frac{N(u) \cap N(v)}{N(u) \cup N(v)} \right|.$$

(1) **The Relationship Between the Receiver Node and Contents** $\phi_{C,R}$

Forwarding Interest. The cosine of messages contents vector and blogger's documents vector could represent the blogger's interest in the forwarding of message content:

$$Sim(u, c) = \cos(\theta) = \frac{\mathbf{C} \cdot \mathbf{V}}{|\mathbf{C}| \times |\mathbf{V}|}.$$

3.2 Model Construction

Based on the defined extracted features sets, we construct the relation among the features, the propagation probability function and the propagation delay function in this section. First of all, we combine the sender and receiver nodes attributes and the edge features into three groups of feature vectors.

$$\Phi_U = \begin{pmatrix} \mathbf{\Phi}_s \\ \mathbf{\Phi}_r \end{pmatrix}, \Phi_C = \begin{pmatrix} \mathbf{\Phi}_{sc} \\ \mathbf{\Phi}_{URL} \\ \mathbf{\Phi}_{Tag} \end{pmatrix}, \Phi_E = \begin{pmatrix} \mathbf{\Phi}_{s,r} \\ \mathbf{\Phi}_{c,r} \end{pmatrix}.$$

The basic function f_a *(u, v, c)* is indicated as a linear combination of eigenvectors as follows:

$$f_a(u,v,c) = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1^T \boldsymbol{\Phi}_U + \boldsymbol{\alpha}_2^T \boldsymbol{\Phi}_C + \boldsymbol{\alpha}_3^T \boldsymbol{\Phi}_E. \tag{1}$$

The transmission probability *P (u, v, c)* can be represented as a Bayesian logistic function:

$$p(u,v,c) = \frac{1}{1 + \exp\{-f_a(u,v,c)\}}. \tag{2}$$

As for the formula, $\boldsymbol{\alpha}_0$ is a constant, $\boldsymbol{\alpha}_1^{\mathrm{T}}$ is the weight of the sender node feature vectors, $\boldsymbol{\alpha}_2^{\mathrm{T}}$ is the weight of the receiver node feature vectors, and $\boldsymbol{\alpha}_3^{\mathrm{T}}$ is the weight of the edge feature vectors. The higher weight represents that the corresponding feature vector has more influence on the transmission probability *P (u, v, c).*

The transmission delay *τ (u, v, c)* is indicated as a linear combination of eigenvectors: $\boldsymbol{\phi}_U$, $\boldsymbol{\phi}_C$, $\boldsymbol{\phi}_E$, as follows:

$$\tau(u,v,c) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1^T \boldsymbol{\Phi}_U + \boldsymbol{\beta}_2^T \boldsymbol{\Phi}_C + \boldsymbol{\beta}_3^T \boldsymbol{\Phi}_E. \tag{3}$$

And we make a set of the above weights as follow:

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T), \boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T).$$

There are three kinds of representation methods refer to the time decay factor, which are exponential model, power law model and Rayleigh model. In the reference paper [10], the exponential model is used to study the propagation probability and ability of information dissemination between nodes. Literature [11] analyzed the function mechanism of microblog-spread with the help of social physics methods of convex mirror and resonance based on the Rayleigh model. And the paper [12] compares the information propagation rates between these three models. In this paper, the exponential model is used to describe the attenuation law of propagation probability with Δt.

Within the time period from t_u to t_v, the probability, that the node v is infection by node u, is denoted as $f((v,\ t_v)|(u,\ t_u);\ \boldsymbol{\alpha},\boldsymbol{\beta})$. Thus the spread probability density function as follows:

$$F((v,t_v)|(u,t_u);\boldsymbol{\alpha},\boldsymbol{\beta}) = \int_{t_u}^{t_v} f((v,t)|(u,t_u);\boldsymbol{\alpha},\boldsymbol{\beta})dt.$$

As information c^k spreads in the directed social network *G (V, E)*, Node u is infected to forward information c^k at time t_u. However, as for the node $v \in N(u)$, the probability that the node v is not infected by node u without forwarding the information c^k until the moment t_v is called the survival probability [13], as follows:

$$S((v,t_v)|(u,t_u);\boldsymbol{\alpha},\boldsymbol{\beta}) = 1 - F((v,t_v)|(u,t_u);\boldsymbol{\alpha},\boldsymbol{\beta}).$$

Within time window [0, *T*], the information c^k spreads in the directed social network *G (V, E).* And we create a set including all of the infected nodes (or called activated nodes) and infection processes $D^k = \{(v_1, t_1),(v_2, t_2)\ldots(v_{n(k)}, t_{n(k)})\}$. The sign *par(v)* represents the parent nodes of node *v*. All of the infected nodes forwarding the information c^k are denoted as $B^k(k) = \{v_i|(v_i, t_i) \in D^k, t_i < t\}$. Before the moment *t*, it is possible that the infected nodes forwarding information c^k affected by node *v* form a set as $Q^k(v) = N(v) \cap B^k(t)$.

At the moment *t*, the transmission probability, [14] that the node *v* is infected by the parent nodes, calculated as $f((v, t_v)|(u, t_u); \boldsymbol{\alpha}, \boldsymbol{\beta})$. And before the moment *t*, the transmission probability, that the node v is not infected by the parent nodes, can be calculated as follow:

$$\prod_{\omega \in Q^k(v)\backslash par(v)} S((v, t_v)|(u, t_u); \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{4}$$

As for the target information cascade M^k and the target node $(v, t_v) \in D^k$, $m^k_{u,v,tu,tv} \in M^k$, node *v* is only infected by node *u* at the moment t_v, meanwhile it is not infected by the other neighbors, of which the transmission probability as follow: [15]

$$f((v, t_v)|(u, t_u); \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot \prod_{\omega \in Q^k(v)\backslash par(v)} S((v, t_v)|(\omega, t_\omega); \boldsymbol{\alpha}, \boldsymbol{\beta})$$

Thus the probability of the occurrence of information cascade M^k as follow:

$$f(M^k|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{(v,t_v) \in D^k} f((v, t_v)|(u, t_u); \boldsymbol{\alpha}, \boldsymbol{\beta}) \cdot \prod_{\omega \in Q^k(v)\backslash par(v)} S((v, t_v)|(\omega, t_\omega); \boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{5}$$

The occurrence probability of a set of the information cascade $M = \{m^1, m^2 \ldots m^k\}$ as follow:

$$\begin{aligned} f(M|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{1 \le k \le K} f(M^k|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{1 \le k \le K} \prod_{(v,t_v) \in D^k} f((v, t_v)|(u, t_u); |\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \cdot S((v, t_v)|(\omega, t_\omega); \boldsymbol{\alpha}, \boldsymbol{\beta}). \end{aligned} \tag{6}$$

The maximum parameters $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, which can satisfy the global likelihood probability calculation is the ultimate solution to the prediction model, as follows:

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}} - \lg f(M|\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{7}$$

The formula includes the following two vectors:

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{0,} \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T)^T, \boldsymbol{\beta} = (\boldsymbol{\beta}_{0,} \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T)^T. \tag{8}$$

3.3 Solution Method

The parameter set is defined as $\boldsymbol{\omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, and the parameters equations are defined as:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\lg f(M|\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Theorem 1. Objective function $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has continuous partial derivatives of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in real field R.

Satisfying the theorem 1, the solution of formula (7) can be solved based on the stochastic gradient descent. [14, 15] The stochastic gradient descent is an iteration algorithm. The iterative method for parameter $\boldsymbol{\alpha}$ based on stochastic gradient algorithm as follows:

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} - \boldsymbol{\alpha} \cdot \frac{\partial L(\omega^{(t)})}{\partial \boldsymbol{\alpha}^{(t)}}. \tag{9}$$

As for the objective function $L(\boldsymbol{\alpha},\boldsymbol{\beta})$, the gradient calculation of $\boldsymbol{\alpha}$ is as follows:

$$\frac{\partial L(\omega^{(t)})}{\partial \boldsymbol{\alpha}^{(t)}} = -\sum_{1 \le k \le K} \sum_{(v,t_v) \in D^k} \left\{ \frac{\partial \lg p(\omega, v, c^k)}{\partial \boldsymbol{\alpha}^{(t)}} + \sum_{\omega \in Q^k(v) \backslash par(v)} \left(\frac{\partial p(\omega, v, c^k)}{\partial \boldsymbol{\alpha}} \cdot \frac{-(1 - \exp\{-\tau(\omega, v, c^k)(t_v - t_\omega)\})}{1 - p(\omega, v, c^k)(1 - \exp\{-\tau(\omega, v, c^k)(t_v - t_\omega)\})} \right) \right\}. \tag{10}$$

Calculated by the same process, the gradient descent calculation of $\boldsymbol{\beta}$ can be solved. By this way, we could effectively evaluate the optimal weights of the features in the proposed model.

4 Experiment and Evaluation

4.1 The Experimental Data and Method

Through a strong Sina micro-blog API base, we ultimately grab the bloggers' information as research dataset, [16] which is shown in Table 2.

Table 2. Statistics of the initial dataset C

Dataset property	Content value
Period	2016-03-01–2016-03-30
Bloggers	37,496
Messages	About 1,500,000
Links	912,775
Information cascades	2,190
Average length of the information cascade	11.3

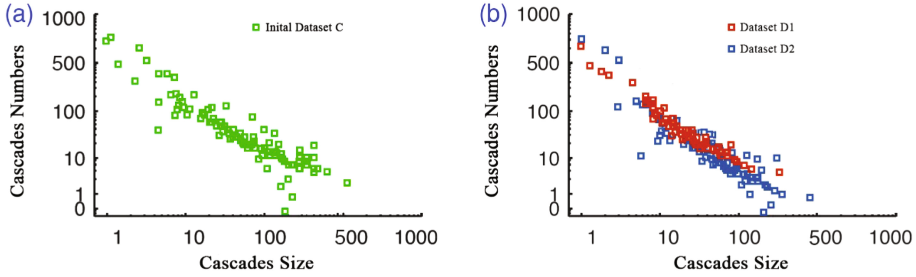


Fig. 4. Statistics of the initial dataset (a), and compare between sparse dataset and dense dataset (b)

The propagations of these information form a set of information cascade *C*, containing a total of 2,190 information cascade, and the average length of the information cascade is 11.3. Figure 4(a) descripts its scale and size. The first dataset D_1 (also known as the dense dataset) directly uses the preprocessed raw data. The second dataset D_2 (also known as the sparse dataset) is randomized on the basis of *D1*, as seen in Fig. 4(b). For each of those datasets, our experiment is performed by 4-fold cross-validation method [17].

This experiment is divided into two parts. The first part of the experiment is to solve the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, satisfying:

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \boldsymbol{\alpha}_3^T), \boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \boldsymbol{\beta}_3^T).$$

In order to evaluate the effect of the multi-dimensional features cascades diffusion model proposed in this paper, the following propagation model is chosen as the reference:

(1) Asynchronous independent cascade model (AsIC model)
(2) Linear threshold model (LT model)

This paper uses the precision at different recall point (called as *PRP*, *P*) as the evaluation criteria, such as *P@1*, *P@2*, *P@5*, *P@10* and the average precision (called as *MAP*). For message content c^k, *P@1* reflects the accuracy that the model predicts an infected node of message c^k. And P@2 reflects the accuracy that the model predicts two infected nodes of message c^k. Likewise, we conclude the signification of the *P@5* and *P@10*.

4.2 Experimentation Results

4.2.1 Analysis of Transmission Probability and Transmission Delay

Figure 5 shows the weights of features extracted in the propagation model, of which the values are standardized into the interval of [0, 1]. Among all features, the weight of structural similarity feature *Sim-s(u,v)* between nodes ranks first, followed by the forwarding Interest feature *Sim(u,v)* between the receiver node and contents. The influence is also an important feature influencing the forwarding behavior.

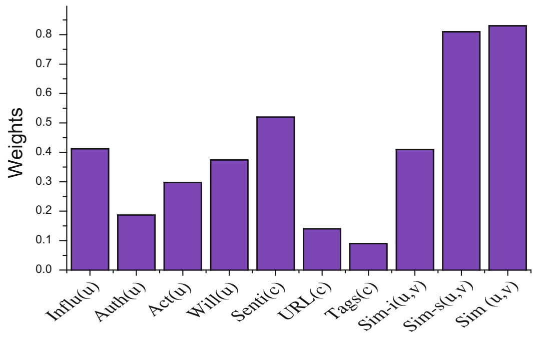


Fig. 5. Weight of each factor in information diffusion probability model

4.2.2 Experimentation Results and Conclusion

Figure 6 shows the prediction accuracy of different models on two datasets *D1*, *D2*, and the model proposed in this paper is superior to the AsIC model and the LN model at each regression point. For the results on the *D1*, *D2* dataset, all the three models are at the highest accuracy in the first regression point. With the move of the regression point, the accuracy rate decreased gradually.

Although we have made the optimization of the information cascading model, it's still a difficult problem to effectively make the prediction accurately in face of tens of thousands of information steams in the real-time social networks [18]. At present, there are some researches on the prediction of real-time network communication, and many achievements have been made in this field [19]. The future task is how to effectively predict the diffusion in a real-time social network rather than analyze based on the limit dataset.

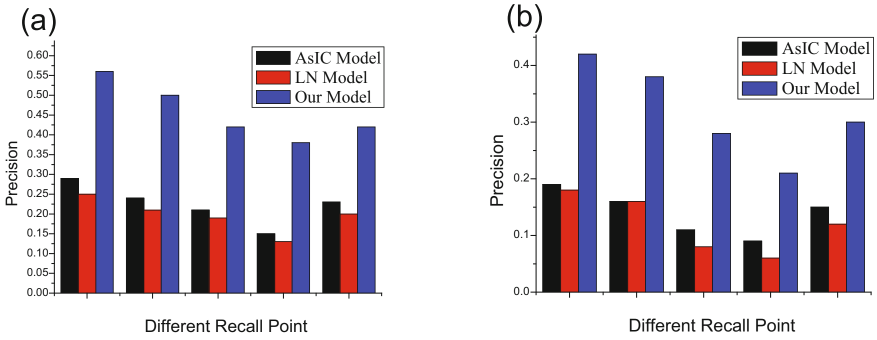


Fig. 6. Statistics of the precision values on the datasets D1 (a) and D2 (b)

References

1. Dreżewski, R., Sepielak, J., Filipkowski, W.: The application of social network analysis algorithms in a system supporting money laundering detection. Inf. Sci. **295**(295), 18–32 (2015)
2. Fogues, R., Such, J.M., Espinosa, A., et al.: Open challenges in relationship-based privacy mechanisms for social network services. Int. J. Hum. Comput. Interact. **31**(5), 350–370 (2015)
3. Qin, Y., Ma, J., Gao, S.: Efficient influence maximization under TSCM: a suitable diffusion model in online social networks. Soft Comput. **21**(4), 1–12 (2016)
4. Saito, K., Kimura, M., Ohara, K., et al.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Asian Conference on Machine Learning: Advances in Machine Learning. Springer, pp. 322–337 (2009)
5. Goyal, A., Bonchi, F., Lakshmanan, L.V.S.: Learning influence probabilities in social networks. In: International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, DBLP, pp. 241–250 (2010)
6. Cui, P., Jin, S., Yu, L., et al.: Cascading outbreak prediction in networks: a data-driven approach. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 901–909 (2013)
7. Zhou, F., Jiao, J., Lei, B.: A linear threshold-hurdle model for product adoption prediction incorporating social network effects. Inf. Sci. **307**(20), 95–109 (2015)
8. Bozorgi, A., Haghighi, H., Zahedi, M.S., et al.: INCIM: a community-based algorithm for influence maximization problem under the linear threshold model. Inf. Process. Manage. **52** (6), 1188–1199 (2016)
9. Guille, A., Hacid, H., Favre, C., et al.: Information diffusion in online social networks: a survey. ACM SIGMOD Rec. **42**(2), 17–28 (2013)
10. Baggio, S., Luisier, V., Vladescu, C.: Relationships between social networks and mental health: an exponential random graph model approach among Romanian adolescents. Swiss J. Psychol. **76**(1), 5–11 (2017)
11. Xu, L.: Constructing the affective lexicon ontology. J. China Soc. Sci. Tech. Inf. **27**(2), 180–185 (2008)
12. Rodriguez, M.G., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: International Conference on Machine Learning. arXiv, pp. 561–568 (2011)
13. Christakis, N.A., Fowler, J.H.: Social network sensors for early detection of contagious outbreaks. PLoS ONE **5**(9), e12948 (2010)
14. Qiao, Y., Van, L.B., Lelieveldt, B.P., et al.: Fast automatic step size estimation for gradient descent optimization of image registration. IEEE Trans. Med. Imaging **1**(6), 391–403 (2015)
15. Da, M.S.B.A., Anderson, C.W.: Restricted gradient-descent algorithm for value-function approximation in reinforcement learning. Artif. Intell. **172**(4–5), 454–482 (2008)
16. Lagnier, C., Denoyer, L., et al.: Predicting information diffusion in social networks using content and user's profiles. In: European Conference on Advances in Information Retrieval, pp. 74–85 (2013)
17. Wiens, T.S., Dale, B.C., Boyce, M.S., et al.: Three way k-fold cross-validation of resource selection functions. Ecol. Model. **212**(3–4), 244–255 (2008)

18. Saito, K., Ohara, K., Yamagishi, Y., et al.: Learning diffusion probability based on node attributes in social networks. In: International Symposium on Foundations of Intelligent Systems, Proceedings, ISMIS 2011, 28–30 June 2011, Warsaw, Poland, DBLP, pp. 153–162 (2011)
19. Fris, M., Nilsson, M., Sollerhed, V.: Real-time social network - exploring the design space for a multi-user real-time visualisation tool for social network analysis. Astrophys. J. Lett. **705**(1), L67–L70 (2011)

Multi-target Detection of FMCW Radar Based on Width Filtering

Mingfei Liu(✉), Yi Qu, and Yefeng Zhang

Information Engineering Department,
University of the People's Armed Police Force, Xi'an, Shaanxi, China
460166636@qq.com

Abstract. Considering the issue of how to implement multi-target detection rightly in the complex environment of LFMCW (linear frequency modulated continuous wave) radar, a framework was proposed to analyze target detection based on width information in the frequency domain. The method involves three procedures. First, CFAR (constant false alarm rate) processor was implemented in frequency spectrum of beat frequency for the echo. Besides, a clustering algorithm was introduced to obtain width and amplitude information by calculate the continuous interval width of the frequency spectrum in positive frequency domain after the CFAR, Finally, multi-target detection by remove tiny spectrum cluster through width filtrating and spectrum line association utilize amplitude and width information in frequency domain. And a computer simulation was carried out. The result showed that the framework could effectively eliminate false target caused by clutter. The method could be used in target location and tracking signal processing in continuous wave radar system.

1 Introduction

FMCW has a series of merits such as high range resolution, no ranging blind area, low probability of intercept and simple structure [1], making it applied in areas of target detection and characteristic research, such as microwave altimeter, airport surveillance system of radar, automatic driving radar, military attack and many other fields. Not only does it play an important role in those areas, but also it has important economic and military significance.

In a practical application, the key problem is that how to implement multi-target detection rightly in the complex environment, which restricts its application and popularization. As is known to all, there are two types of false FMCW radar target, one kind is produced by the clutter and the noise of radar. This part of the interference signal will be with the whole process of radar signal processing, and eventually embodies in false alarm time-varying characteristics. For example, in automotive radar: the vehicles in the adjacent lane, the barrier nearby the line lane, roadside trees and the aerial or distant buildings and so on, all of these have made interference with the system of radar. Meanwhile, the interference changes with the environment which sensors are located in. Another kind is the mismatch of the target parameter in signal processing. The number of the false targets is particularly prominent in multi-sensor multi-target case.

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_74

In order to solve the problem of ghost targets, experts, scholars and radar workers have gradually reached a consensus through a lot of experimental research: to the first kind of false alarm, through the design of the adaptive filter and constant false alarm rate (CFAR) processor, the radar system can maintain normal work in the complex environment [2]. For the second category of false alarm, by (1) designing the radar system that has the ability of angle measuring to increase the information of target azimuth, (2) or the complex emission signal of radar which is easy to produce and has strong anti-interference performance, cooperating with efficient real-time methods of signal processing and target detection to remove false alarm [3–5].

At present, most of the detection algorithms only use the information of angle and distance which are about location. Not only does the radar echo have location information, but also it often includes related information which reflects the signal strength (such as amplitude power, etc.). In the literature [6], the amplitude and phase characteristic of the multi-objective echo data is analyzed. In literature [7], spectrum area of the target in echo signal of difference frequency signal is used as the basis of object matching. There are few literatures that analyze the frequency width of target information. Even the point target has spectrum width and parameter estimation can't be directly conducted. This kind of problem will encounter in the above methods by adding the right match.

This article put forward a method to achieve target matching and clutter cancellation which is based on amplitude and width information in frequency domain. The idea is as follows: firstly, the frequency resolution of the FMCW radar echo signal has influence on spectrum width of the target echo, which should be analyzed. Secondly, echo signal processing should be conducted in such parameters of radar. Finally, the validity of the method should be verified through the simulation (Fig. 1).

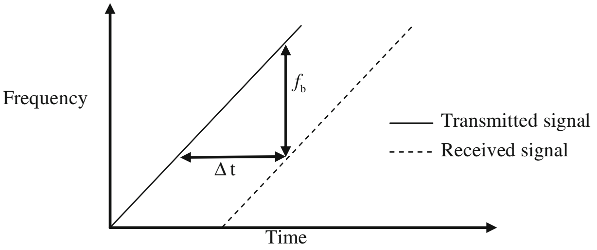


Fig. 1. The transmitting and receiving waveform of up frequency modulation period

2 Related Work

2.1 The Basic Theory of FMCW Radar

Consider that the symmetric triangular wave whose period, amplitude, carrier frequency, bandwidth is constant Tm, A, f_0, B, individually, is used for signal waveform for launch. Take radar echo of the upper frequency modulation band as an example merely.

For the CW radar, the frequency of the transmitted signal is a constant, which is the same as carrier frequency, and simply assumed that the initial phase $\varphi_{\mathrm{TX}}(0) = 0$, without loss of generality. In practice, the initial echo phase is interrelated with range, phase-frequency characteristic of RF Circuit and phase overgrow reflected in the target interface. If the initial phase remains constant all along in the transmitting period, the phase information of signal during coherent integration process will extracting the range and speed information which carried.

In the ascent section of linear frequency modulation continuous wave waveform, $T \in [0, Tm/2]$, the launched LFMCW sinusoidal signals can be expressed as:

$$\mathrm{S_{TX}} = \mathrm{A} \cdot \cos(\varphi_{\mathrm{TX}}(\mathrm{t})) = \mathrm{A} \cdot \cos(2\pi(f_0 t + \frac{B}{Tm} t^2)) \tag{1}$$

Target echo signal in this period, compared with the transmitted signal $\mathrm{S_{TX}}$, Just a amplitude attenuation and a delay in time due to dual - path difference τ,

$$\mathrm{S_{RX}} = A' \cdot \cos(\varphi_{\mathrm{RX}}(\mathrm{t})) = A' \cdot \cos(\varphi_{\mathrm{TX}}(\mathrm{t} - \tau)) \tag{2}$$

Where $\varphi_{\mathrm{RX}}(\mathrm{t})$ is the instantaneous phase of the signal that received.

So, instantaneous beat frequency S_{IF} generated by the mixer (DDC) expressed as:

$$S_{IF} = \mathrm{A}'' \cdot \cos(\varphi_{\mathrm{IF}}(\mathrm{t})) = \mathrm{A}'' \cdot \cos(\varphi_{\mathrm{TX}}(\mathrm{t}) - \varphi_{\mathrm{RX}}(\mathrm{t})) \tag{3}$$

The phase of beat frequency signal is:

$$\varphi_{\mathrm{IF}}(\mathrm{t}) = \varphi_{\mathrm{TX}}(\mathrm{t}) - \varphi_{\mathrm{TX}}(\mathrm{t} - \tau) = 2\pi(f_0\tau + \frac{2B}{Tm} t\tau - \frac{B}{Tm} t^2) \tag{4}$$

Due to electromagnetic wave propagation speed is too fast relative to the target distance, if $\tau/\mathrm{Tm} << 1$, the last item in the phase $\varphi_{\mathrm{IF}}(\mathrm{t})$ can be ignored. Replace with $\tau = 2 \cdot (\mathrm{R} - \mathrm{v} \cdot \mathrm{t})/\mathrm{c}$, then

$$\varphi_{\mathrm{IF}}(\mathrm{t}) \approx 2\pi[f_0 \frac{2R}{c} + (\frac{4B}{Tm} \cdot \frac{R}{c} - \frac{2vr}{c} f_0)\, t - \frac{4B}{Tm} \frac{vr}{c} t^2] \tag{5}$$

$$f_{\mathrm{IF}} = \frac{1}{2\pi} \frac{\partial \varphi_{\mathrm{IF}}(\mathrm{t})}{\partial t} = \frac{4B}{Tm} \frac{R}{c} - \frac{2v}{c} f_0 \tag{6}$$

First part of formula (6) associated with distance, which represents the frequency difference cause by the distance r of the position of motionless or stationary objective, while the second part correlated the speed-related, which is equivalent to the Doppler frequency shift. Therefore, the formula can be simplified as

$$f_{\mathrm{b1}} = f_{\mathrm{R}} - f_d \tag{7}$$

Where f_R and f_d indicated frequency cause range and Doppler respectively.

Through beat frequency processing, single frequency approximate rectangle modulation signal waveform which has the cycle of $Tm/2$ can be achieved. The corresponding spectrum centers is f_{b1}, and the first zero point is $2/Tm$. Therefore, the minimum frequency resolution of the system Δf is the reciprocal of system measuring time of frequency modulation (the up frequency modulation time or the lower frequency modulation time): $\Delta f = 2/Tm$.

When the target is stationary, the beat frequency signal only associated with the target distance can be expressed as:

$$f_{\mathrm{R}} = f_{\mathrm{b,stationary}} = \frac{4\mathrm{B}}{\mathrm{T_m}} \cdot \frac{\mathrm{R}}{\mathrm{c}} \tag{8}$$

For moving target, the Doppler Effect to change the size of the received echo frequency (when the Doppler is positive, the value is bigger while the value is smaller when Doppler is negative), thus, the frequency differences in the down frequency modulation band are:

$$f_{\mathrm{b2}} = f_{\mathrm{R}} + f_d \tag{10}$$

where, the Doppler frequency is:

$$f_d = 2v/\lambda_0 \tag{11}$$

2.2 Distance and Speed Estimation

Beat frequency signal can be measured through the baseband signal. Then the distance and relative speed can be calculated as follows:

$$R = \frac{\mathrm{c}(f_{\mathrm{b1}} + f_{\mathrm{b2}})T_{\mathrm{m}}}{8B} \tag{12}$$

$$v = \frac{\mathrm{c}(f_{\mathrm{b2}} - f_{\mathrm{b1}})}{4f_0} \tag{13}$$

2.3 Range and Speed Resolution

When put Δf to the expression of the difference frequency of the stationary target, namely (8), the range resolution can be obtained:

$$\varDelta R = \frac{\mathrm{c}}{4B}\varDelta f = \frac{\mathrm{c}}{2B} \tag{14}$$

Where c represents the velocity of light. The wider the bandwidth is, the larger range resolution is. Therefore, the range resolution of the narrowband radar will be limited. For example, in order to achieve distance resolution with the value of 1.5 m, FWCM needs a bandwidth of 100 MHz

When put Δf to the expression of the difference frequency of the moving target, namely (11), we can obtain that:

$$\Delta \mathrm{v} = \frac{\mathrm{c}\Delta f}{2\mathrm{f}_0} = \frac{\mathrm{c}}{f_0}\frac{1}{T_m} = \frac{\lambda_0}{T_m} \tag{15}$$

Where λ_0 is the working wavelength, f_0 is the working frequency, T_m is the waveform cycle. It is visible that when the time or frequency is high, a higher resolution of the system can be gotten.

2.4 The Influence of Radar Sampling Points on the Target Spectrum Width

The process is done in the frequency domain of beat signal in FWCM radar system. Signal sampling rate is determined by the effective width of time and the sampling rate is always corresponding to the theoretical resolution. Thus, in the FWCM radar, within the effective time width, the sampling rate of the beat signal needs to meet the Nyquist sampling theorem. After sampling, high range resolution can be obtained the sampling by fast Fourier transform. Sampling frequency is determined as follows:

$$f_s \geq 2f_{\max} \tag{16}$$

Where $f_{\max}$ is the maximum frequency of difference frequency signal, it is the result of the combined action of target distance and Doppler Effect.

When using FFT to analyze spectrum of the echo signal, the minimum resolution spectrum is directly related to sampling points N:

$$\Delta \mathrm{FFT} = f_{\max}/N \tag{17}$$

The width of the range resolution can be gotten at this time:

$$\frac{\Delta \mathrm{F}}{\Delta \mathrm{FFT}} = \frac{2N}{Tm \cdot f_{\max}} = \frac{4B \cdot \Delta R \cdot N}{c \cdot f_{\max}} \tag{18}$$

Formula (18) shows frequency domain is contain a cluster of spectral line even a point target.

Where λ_0 is the working wavelength, f_0 is the working frequency, T_m is the waveform cycle. It is visible that when the time or frequency is high, a higher resolution of the system can be gotten.

3 The Method of Width Filter and Spectrum Association

From the above analysis we can see that even the spectrum of the point target also have the phenomenon of broadening. Considering the range resolution of radar and the size of target, we can assume that the spectrum of target has a certain width.

Thus, according to the width information of the signal detection after CFAR to remove clutter signal, which is also a kind of methods of clutter suppression and can bring about benefits of subsequent parameter estimation.

For the same target, the beat signal spectrums produced by upper and negative frequency modulation period have the same amplitude and shape [7]. In practical applications, because of the interference of environmental noise, target fluctuation and so on, the echo spectrums width produced by the same target in the up/down modulation period have the largest similarity. Width filtering and spectrum association method is as follows:

Step1. Frequency domain CFAR adapt to the beat frequency spectrums.

Step2. Use clustering algorithms calculate the continuous interval width of the frequency spectrum in positive frequency domain after the CFAR. Remove false target signals by filtrating tiny spectrum with narrow width based on width information of the signals. This method also can be used as a means of recognition. And the width parameter's design should need to be determined in accordance with experimental measured data.

Step3. Find the biggest amplitude of each frequency units that escapes from width filter, record them as $(a_{i,up}, b_{j,down},)i \in (1, n), j \in (1, m)$, and identify the frequency that corresponds to the sequence number.

Step4. Soft association used in amplitude. According to the range of the beat frequency of both modulation periods, you can search some similar values in the down modulation period to match each peak frequency within the up modulation period. Peak matching thresholds can be appropriate to widen, so each frequency corresponds to more than one possible value in the frequency combination lists.

Step5. Introducing width information in secondary association procedure, obtain corresponding frequency unit finally.

The amplitude spectrum sampling sequence number produced by the ith target in the upper frequency modulation period is $\{S_{\mathrm{i}},\ \mathrm{up}(1),\ S_{\mathrm{i}},\ \mathrm{up}(2), \ldots, S_{\mathrm{i}},\ \mathrm{up}(n)\}$. The width' value of spectral line is used as reference information:

$$N_{\mathrm{i,up}} = S_{\mathrm{i}},\ \mathrm{up}(n) - S_{\mathrm{i}},\ \mathrm{up}(1) \tag{19}$$

The width of amplitude spectrum in lower frequency modulation is as follows:

$$N_{\mathrm{i,down}} = S_{\mathrm{i}},\ \mathrm{down}(n) - S_{\mathrm{i}},\ \mathrm{down}(1) \tag{20}$$

From the analysis above, $N_{i,up}$ is most similar to $N_{i,down}$, The spectrum line width matrix W of n objects is as follows:

$$W = \begin{pmatrix} N_{1,up} & N_{2,up} & \cdots & N_{n,up} \\ N_{1,down} & N_{2,down} & \cdots & N_{n,down} \end{pmatrix} \tag{21}$$

The difference matrix ΔW produced by up frequency modulation band and down frequency modulation band of each target is as follows:

$$\Delta W = \begin{pmatrix} \Delta N_{1,1} & \Delta N_{1,2,} & \cdots & \Delta N_{1,n} \\ \Delta N_{2,1} & \Delta N_{2,2} & \cdots & \Delta N_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta N_{n,1} & \Delta N_{n,2} & \cdots & \Delta N_{n,n} \end{pmatrix} \tag{22}$$

Where $\Delta N_{i,j} = N_{i,up} - N_{j,down}, i,j \in (1,n)$,

Use amplitude information to soft association then received a matching matrix

$$H = \begin{pmatrix} a_{1,1} & a_{1,2,} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \tag{23}$$

Where $a_{i,j} = 1$, if the two frequencies is matched with each other, else $a_{i,j} = 0$.

The final matching matrix can be described as $D = \Delta W \cdot H$.

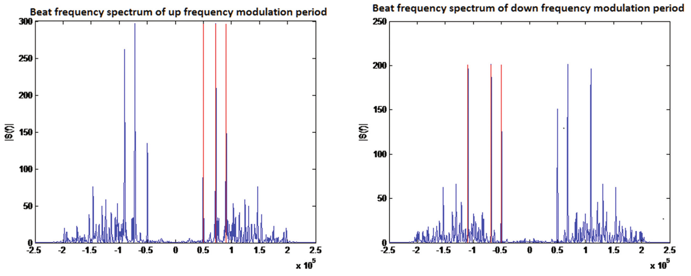
The Hungarian allocation algorithm is used to find the optimal solution at this time. $(f_{i,up}, f_{j,down},)i \in (1,n), j \in (1,m)$ So far, for echo frequency of each target in the up frequency modulation band, a corresponding frequency can be found at down frequencies, using formula (5) or (6) at this time. The range and velocity of the target can be gotten.

4 Numerical Simulation

The transmitting frequency of the signal used in the computer simulation is f_0 with the value of 24 GHz. The tuning bandwidth B is 300 MHz Besides, the symmetric periodic triangle wave Tm is 2 ms, and the points of FFT are 2048. By this time, the resolution of range and speed reaches 0.5 m and 6.25 m/s, respectively. The maximum rate of departure in the simulation is 150 km/h. The height of the radar is 1 m. Both Cartesian coordinate and polar coordinate are constructed for a point which is the projection of the position of the radar on the ground. The main lobe width of the radar used in the simulation is 110. Naturally, ground clutter is generated. The ground clutter to the noise power ratio CN is 25.

Two moving point targets are located in 35 m and 50 m, with the speed of −12 m/s and 60 m/s, respectively. Besides, a stationary point-target is set in 25 m. Figure 2(a) and (b) show the frequency spectrograms.

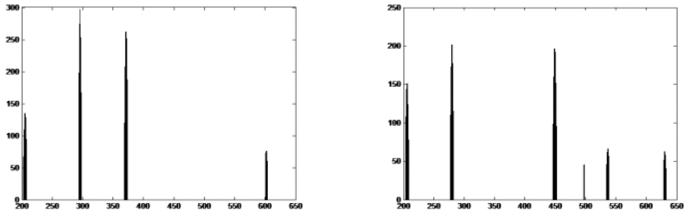
In order to detect the target in the spectrum, CA - CFAR is applied on the frequency spectrum. Obviously, clutter must exist after constant false alarm processing, particularly in the heterogeneous clutter environment. Taking factors such as antenna beam width into account, we assume the amplitude of target signal is relatively large; a fixed threshold filtering can be used to remove the small amplitude signals. The threshold is determined by average value of signal multiplied by the coefficient. In the measured



(a) (b)

Fig. 2. Beat frequency spectrum (a) up frequency modulation period, (b) down frequency modulation period

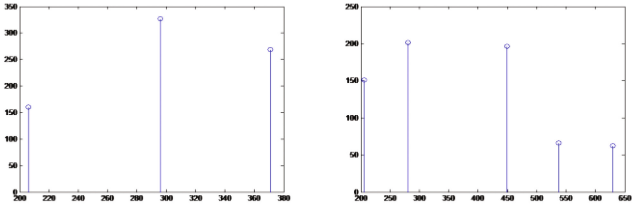
data processing, it can also be determined by the maximum value of signal multiplied by coefficient. After the comparison of threshold, frequency spectrum is shown in Fig. 3(a), (b).



(a) (b)

Fig. 3. Constant false alarm rate processing (a) up frequency modulation period, (b) down frequency modulation period

After apply CFAR in the frequency spectrum. It's conspicuous to find more than three peaks in both period of the spectrum by influence of clutter. The method of width filter and spectrum association can use in this situation, Fig. 4(a), (b) shows the biggest



(a) (b)

Fig. 4. Peaks of frequency after width filter and spectrum association (a) up frequency modulation period, (b) down frequency modulation period

Table 1. Detect target range and velocity

Metrics	Real range/m	Measured range/m	Range error/%	Real velocity/(m/s)	Measured velocity/(m/s)	Velocity error/%
Target1	35	35.0342	0.9771%	−12	−12.2070	1.725%
Target1	50	49.9878	0.0244%	60	60.2722	0.4536%
Target1	25	25.0244	0.0976%	0	0	0

spectral line after clustering algorithms. Figure 4(b) displays the effect of width filter which inhibition the spectrum line at 500 kHz. And Table 1 shows the finally result of the method.

According to the theoretical analysis to know, when two target echo amplitude difference is not big, may cause the spectrum matching error. But found in the simulation, as the two target echo amplitude of the same or similar, or cannot by the above methods for spectral matching error, only will appear in certain segment matching error, these still needs further analysis.

5 Conclusion

A method for frequency domain matching and clutter elimination is put forward in this paper, according to the spectrum characteristics of echo of the beat signal in positive and negative modulation period of FMCW radar. Both theoretical analysis and simulation experiments show that this method can effectively control the target data matching and clutter, and it is able to be applied in target location and tracking signal processing in continuous wave radar system.

References

1. Lufei, D., Fulu, G.: Principles of Radar. Xidian University Press, Xian (2002)
2. Zhangwei, D.: Research on the techniques of radar constant false alarm rate detection under complex clutter. Nanjing University of aeronautics and astronautics, Nanjing (2009)
3. Jiang, L., Xu, T., Yang, C., et al.: A method of automobile anti-collision radar to identify multi-target. Modern Radar **36**(6), 54–58 (2014)
4. Lu, X., Liu, G.: A method of improving waveform design for LFMCW radar multi-target detection. Shipboard Electron. Countermeasure **38**(2), 33–36 (2015)
5. Du, J., Song, C.: A modified millimeter wave radar multi-target detection algorithm. Commun. Technol. **48**(7), 808–813 (2015)
6. Min, T., Wang, X., Zhao, F., Xiao, S.: Modeling of multiple unresolved targets echo characteristics for LFM pulse radar. J. Syst. Simul. **25**(4), 737–741 (2013)
7. Shi, L., Zhang, L.: The signal processing based on the partner for frequency modulation continuous wave radars. J. Xidian Univ. **30**(4), 534–538 (2003)

DroidMark: A Lightweight Android Text and Space Watermark Scheme Based on Semantics of XML and DEX

Lingling Zeng[1], Wei Ren[1](✉), Min Lei[2], and Yu Yang[2]

[1] School of Computer Science, China University of Geosciences, Wuhan, China
weirencs@cug.edu.cn

[2] Information Security Center, Beijing University of Post and Telecommunications, Wuhan, China
leimin@bupt.edu.cn

Abstract. Android platform induces an open application development framework to attract more developers and promote larger market occupations at the same time. However, the open architecture also makes it easier to reverse engineering and application piracy. These result in the property loss for developers and companies, and increase the risks of mobile malicious code. Copyright protection for android application is thus of significant importance. Currently, many solutions for application copyright protection apply overload methods, assuming the availability of source code, which could be impractical for a large scale application protection. In this paper, we propose a lightweight copyright protection method for android application called DroidMark. The copyright is protected by text and space watermark based on semantics of xml and dex. Functional files are chosen as watermark carriers to increase watermark semi-fragileness and concealment. And the DroidMark can be accomplished without secret keys. Models and algorithms are proposed and analyzed all sidedly. The experiment results and analysis justified that DroidMark is secure and efficient.

1 Introduction

In recent years, copyright consciousness has been highly valued unprecedentedly. The copyright in Apps shows its value in both economy and society.

For example, "Repackage", one of the android potential safety hazard [1] for application copyright, is a technique to produce fake Android applications on the basis of the legit App. Fake Apps cause damage to the legit achievements and benefits of the developer, while reduce the user experience for customers at the same time. Moreover, fake Apps may be embedded malicious codes, which may threat to user privacy and property security. Therefore, as the most widely use mobile platform worldwide, copyright protection for large-scale Android applications attracts more and more attentions in research communities.

© Springer International Publishing AG 2018

L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_75

Till now, many schemes have been proposed in app copyright protection. However, those methods induce a large amount of computation overhead such as reverse engineering detection and code manipulation, which may not be suitable for the app protection in a large scale. Therefore, a lightweight and covert method in terms of computation cost will be more applicable. Watermarking is a technique to provide data integrity and authenticity in public or in private. For this perfect attribute, watermarking can be applied in application copyright protection.

Current methods for watermarking-based application protection can be roughly divided into three folders. First of all, put copyright information directly into the APK provides chances to adversaries to extract the watermark by keyword search, and modify and forge the information without detection. Secondly, encrypts the information with asymmetric key brings high challenge for key storage and distribution in an open environment. Lastly, the method embeds the watermark in an additional carrier file will be noticed and separated easily, resulting in the loss of the watermark. Although several watermarking-based schemes have been proposed, most of them are not suitable enough for application copyright protection.

Through the analysis above, we can conclude that the watermarking-based copyright protection should accurately tackle three aspects: good choice of carriers, message pretreatment for hiding, and easiness of embedding and extracting algorithms for requirement of lightweight and large-scale. For these regards, we proposed a novel android watermark method for the protection of Android application copyright, called DroidMark. The method is designed to be consist of DroidMark-XML and DroidMark-DEX, distinguished on the basis of the choice of the carrier files.

The contribution of this paper can be summarized as follows:

1. Secret key and encryption is avoided. The embedding and extracting process only need to scan the carrier file for characteristic strings and fields, and embed or extract watermarking information on the basis of the features and semantics of the carrier file.
2. The modification of the application is able to tracked. The watermark is semi-fragile, therefore, DroidMark can detect the modification of Apps and maintain the copyright information in Apps. Any modification of Apps, repackage, re-optimization, will be noticed.

The rest of the paper is organized as follows. Section 2 gives an overview on relevant prior work. And Sect. 3 provides the detailed description of our proposed methods and algorithms. Analyzation and evaluation for both security and performance of the scheme are provided in Sect. 4. Finally, we conclude the paper in Sect. 5.

2 Related Work

Some android copyright protection schemes are proposed recently. Sanghoon Choi et al. [2] proposed a copyright protection technology based on forensic mark, which is marked in the classes.dex file using the IMSI of the buyer to identify illegal Apps. This method aimed at personal software validation, but not simplified for the software copyright protection. Sung Ryul Kim et al. [3] proposed a hybrid copyright protection design on android applications that combines two proposed techniques: online execution class and encryption-based copyright protection. This method needs the participation of secret key, which is hard on preservation and distribution with software.

Wu Zhou et al. [8] proposed AppInk, which designs a dynamic graph based watermarking mechanism for Android Apps. This scheme involves secret key, and the key is needed before extract, therefore, the key cannot be embedded in watermark and is not suitable for negotiability. Yingjun Zhang and Kai Chen [7] proposed a picture-based watermark for Android Apps. The extraction of the scheme has to use the same sequence of events to find basic blocks in the execution path, which has the same problem with the former scheme. To solve these problems, we suggest text watermarking, which doesn't require keys in the process.

To satisfy the resistance of compilation and packaging, the watermark for Android copyright protection can based on text semantics. Some watermarks schemes depend on the semantics of the text. Mikhail J. Atallah et al. [4,5] first proposed the natural language watermarking scheme, using the syntactic structure of the text. This method preserves the inherent properties of the text while embedding. Hassan et al. [6] proposed the natural language watermarking algorithm by performing the morphosyntactic alterations to the text. Mercan et al. proposed an algorithm by using typing errors, acronyms and abbreviations like cursory text in char, emails etc. However, none of these methods may not be suitable for android applications, and the efficiency and capacity is still waiting to be improved.

3 Proposed Scheme - DroidMark

3.1 Design Goals

To insure the efficiency of the design, our scheme should achieve the following properties:

- ***Identifiability:*** It ensures that the scheme can identify the correct watermark in the carrier file. This is the basic property for extraction.
- ***Concealment:*** It guarantees that the attacker cannot distinguish the watermark information even if they obtain the carrier file. This property requires that the string added to carrier should be similar to the original content, and be as natural as possible.
- ***Transparency:*** After addition of the watermark information into the xml file, the APK should be able to operate normally, which indicates that the addition of watermark information should not be detected by operating the APP.

- ***Semi-fragile:*** It ensures the stability of the watermark in the circumstance of being recompiled and attacked. Besides, the modification or damage of the watermark can be able detected.
- ***Capacity:*** The high capacity will improve the robustness and applicability of the scheme, especially when watermarks have large volume but carriers have small volume. Achieving high capacity as possible is a necessary goal for the proposed scheme.
- ***Efficiency:*** It ensures the high performance of embedding and extracting in terms of computation overhead and timing cost.

3.2 Notations

For fast checking the short notation for Sect. 3, we list the major notations used in the remainder of the paper in Table 1.

Table 1. Notation

Symbol	Meaning
W	Array W[0,1 . . . L−1] to store watermark in a format
L	Length of the array W
Loc	Array Loc[0,1 . . . N−1] to save field location
N	Length of the array Loc
S_1, S'_1	String "</application>" and "<application> </application>"
S_2, S'_2	String "/>" and "</>"
S_3, S'_3	String "</activity>" and "<activity> android:name=Cache.i"
S_{3Name}	Content of Row With String "android:name"
$Cache$	String in the cache
APK_w, APK_o	Apk File with or without Watermark
CAR_w, CAR_o	AndroidManifest.xml file with or without Watermark
WTM_p, WTM_b	Plaintext or Binary watermark information
i, e, b	Random number, $i \in [1, 10]$, $e \in \{1, 3, 5, 7, 9\}$ and $b \in \{2, 4, 6, 8, 10\}$
$Seed$	Seed of random-number generator
T	Length of $Seed$
$H[j]$	Array to save random address

3.3 Scheme Construction

In this section, we will present our scheme construction for lightweight copyright protection for Android applications. The scheme is described as follows in algorithm form, followed by explanations:

3.3.1 Scheme Construction for DroidMark-XML

- **Embedding Procedures of DroidMark-XML**

1. Pretreat: Before embedding watermark into the APP, pretreat the watermark information and prepare the carrier file.
   a. $APK_o \rightarrow CAR_o$: get the AndroidManifest.xml file as the carrier file by decompiling the APK_o of the application to be marked.
   b. $WTM_p \rightarrow WTM_b$: turn the watermark into binary data, and save the data in $W[0, 1 \dots L-1]$. Calculate the array length as L.
2. Embed: Scan AndroidManifest.xml file from beginning to the end successively for S_1, S_2 and S_3, and embed the binary data $W[0, 1 \dots L-1]$ in turn by adding strings S_1', S_2' and S_3' respectively at the corresponding position in the following rules:
   a. f_{embed_1}: $S_1, S_2 \rightarrow CAR_o$. After find S_1 (or S_2), identify $W[j], j \in (0, L-1)$. If $W[j] = 0$, generate b, and put equivalent amount of string $S1'$ (or $S2'$) after $S1$ (or $S2$). While if $W[j] = 1$, generate e, and do the same as aforementioned.
   b. f_{embed_3}: $S_3 \rightarrow CAR_o$. At the process of scanning, $Cache$ will just record the latest string after string S_{3Name}. After find S_2 in the carrier, identify $W[j]$, and generate random number b or e. Also, add string $S3'$ after $S3$.
3. Recompile: $f_{compile_r}$. When $j = L$, which identifies the end of embedding, stop scanning and save the new AndroidManifest.xml. Recompile APK with the new file with watermark to get APK_w, in this case, the watermark will be embedded into an application.

- **Extracting Procedures of DroidMark-XML**

1. Decompile: Decompile APK to get AndroidManifest.xml file, which is the CAR_w, before extracting watermark from APK_w.
2. Extract: Scan CAR_w from beginning to the end successively for the specified strings S_1', S_2' and S_3', and extract the binary data in turn in the following rules:
   a. $f_{extract_1}$: $CAR_w \rightarrow S_1'$, S_2'. After find S_1' or S_2' in the CAR_w, count the amount of successive strings as x, an even x represents binary 0, while the odd x represents binary 1.
   b. $f_{extract_3}$: $CAR_w \rightarrow S_3'$. Set the initial value of $Cache$ as "null". At the process of scanning, compare the latest S_{3Name} with $Cache$. If S_{3Name} contains $Cache$, which means S_{3Name} is an anthropogenic watermark, get the rest part of S_{3Name} as ".i". Judge i if it's an even or an odd and get the binary data as above.
   c. Transform: Scan to the end of the CAR_w, and record the binary data successively in a new file. Finally, turn WTM_b into WTM_p. In this case, we can extract the watermark information.

3.3.2 Scheme Construction for DroidMark-DEX

- **Embedding Procedures of DroidMark-DEX**

1. Pretreat: Before embedding watermark into the APP, pretreat the watermark information and prepare the carrier file.
   a. $APK_o \rightarrow CAR_o$: Get the classes.dex file by decompressing APK_o to be marked.
   b. $WTM_p \rightarrow WTM_b$: Turn the plaintext or cryptographic text watermark into hexadecimal data.
   c. $CAR_o \rightarrow Loc$: For the present dex file, some fields keep empty or insignificance in the whole lifecycle of the application. For this reason, these fields can be used for watermarking or digital forensics. As far as we know, there's redundancy information exist between $Header\_item$ and Map_list. In this experiment, we scan for the field *unused* in Map_list. And store the corresponding location of file in the array *Loc*.
2. Embed:
   a. *Seed*: $H[j]$: generate a number as *Seed* to produce random numbers in the function $RAND(Seed)$, calculate the length of *Seed* as T. If the address is used, generate $H[j] = H[j] + 1$;
   b. $WM, H[j] \rightarrow CAR_o : WTM_b, Seed$: Embed *Seed* in the first T unit of *Loc*, and WM in the corresponding $Loc[H[j]]$.
3. Repackage: $CAR_w \rightarrow APK_w$: After finish embedding, recalculate checksum of dex file, and repackage the APK into APK_w. In this case, the watermark will be embedded into an application.

- **Extracting Procedures of DroidMark-DEX**

1. Decompress:Decompress its APK to get classes.dex file, which is CAR_w, before extracting watermark from APK_w.
2. Extract:
   a. CAR_w: $Seed, Loc$: Scan for satisfactory fields and number these locations in hexadecimal unit. Then store the corresponding location of file in the array *Loc*.
   b. *Seed*: $H[j]$; $CAR_w, H[j] \rightarrow WM : WTM_b$: Extract the first T unit of *Loc* to get *Seed*. Generate the random address using function. If the generation repeated, extract from one unit after. In this way, the WTM_b is extracted.
3. Transform: Finally, turn WTM_b into WTM_p. In this case, we can extract the watermark information.

4 Security Analysis and Performance Evaluation

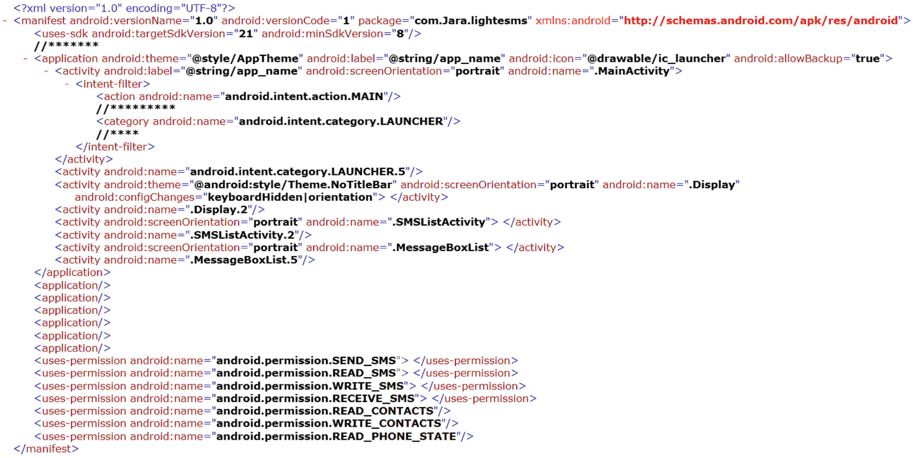
In this section, we evaluate the security and property performance of scheme in experiments. All the following experiments are based on language C and python, and tested on hardware AMD Athlon(tm) II X4 630 with Intel Pentium 2.7 GHz processor and 6 GB memory.

4.1 Security Analysis

To evaluate the security performance of the scheme, we analyze the four security properties in its design goal for DroidMark to evaluate its security performance. We analyze the security of our proposed scheme in the following two aspects under the assumption that APK is arranged:

- **Soundness and Concealment:**

DroidMark cannot be perceived by mainstream watermark detector.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <manifest android:versionName="1.0" android:versionCode="1" package="com.Jara.lightesms" xmlns:android="http://schemas.android.com/apk/res/android">
      <uses-sdk android:targetSdkVersion="21" android:minSdkVersion="8"/>
      //*******
    - <application android:theme="@style/AppTheme" android:label="@string/app_name" android:icon="@drawable/ic_launcher" android:allowBackup="true">
        - <activity android:label="@string/app_name" android:screenOrientation="portrait" android:name=".MainActivity">
            - <intent-filter>
                  <action android:name="android.intent.action.MAIN"/>
                  //*********
                  <category android:name="android.intent.category.LAUNCHER"/>
                  //****
              </intent-filter>
          </activity>
          <activity android:name="android.intent.category.LAUNCHER.5"/>
          <activity android:theme="@android:style/Theme.NoTitleBar" android:screenOrientation="portrait" android:name=".Display"
              android:configChanges="keyboardHidden|orientation"> </activity>
          <activity android:name=".Display.2"/>
          <activity android:screenOrientation="portrait" android:name=".SMSListActivity"> </activity>
          <activity android:name=".SMSListActivity.2"/>
          <activity android:screenOrientation="portrait" android:name=".MessageBoxList"> </activity>
          <activity android:name=".MessageBoxList.5"/>
      </application>
      <application/>
      <application/>
      <application/>
      <application/>
      <application/>
      <application/>
      <uses-permission android:name="android.permission.SEND_SMS"> </uses-permission>
      <uses-permission android:name="android.permission.READ_SMS"> </uses-permission>
      <uses-permission android:name="android.permission.WRITE_SMS"> </uses-permission>
      <uses-permission android:name="android.permission.RECEIVE_SMS"> </uses-permission>
      <uses-permission android:name="android.permission.READ_CONTACTS"/>
      <uses-permission android:name="android.permission.WRITE_CONTACTS"/>
      <uses-permission android:name="android.permission.READ_PHONE_STATE"/>
  </manifest>
```

Fig. 1. Effect of embedding in XML

We put an small-scale experiment on DroidMark-XML, and the effect is shown in Fig. 1. The binary data embedded is "1, 1, 0, 1, 0, 0, 1, 0", and the random number generated by the tool is "7, 9, 4, 5, 2, 2, 5, 6", which has the same oddity as the binary data. From Fig. 1 we can see that the embedding of watermark uses three methods to induces only slightly modification on the basis of original strings conform to the file semantics. In this way, the format of file embedded watermark is similar to the normal AndroidManifest.xml file. Therefore, the watermark is hard to be distinguished apart from the carrier. Watermark embedding and extracting rely on the detection and judgment of characteristic strings rather than the secret "Key". The scheme is free of insecurity in the process of "Key" distribution.

- **Identifiability:**

DroidMark can extract watermark accurately. DroidMark achieves good confusion by using the semantics of carrier files, and is secure in the mainstream watermark detector. Instead, DroidMark can extract the watermark embedded in carriers accurately.

- **Transparency:**

DroidMark has no influence on the operation of the APP. We did experiment on 100 different APKs of embedding watermark in the two carrier files. The experiment indicates the overhead on the installation and operation are the same in millisecond. Therefore, we conclude that DroidMark has no influence on APP's operation.

- **Semi-fragile:**

On the one hand, DroidMark can defend against the attack of decompilation and recompilation, on the other hand, the artificial modification will be aware of. We simulate attacks of recompiling and decompiling, turn out that the watermark in both xml and dex file are stable in the experiment. Therefore, the experiment justified that DroidMark can defend both attack. Besides, AndroidManifest.xml is a configuration file, while classes.dex is for execution, both are the essential part for an APP. If the adversary modifies the app artificially, both file will have to be modified incidentally. The destruction of the watermark will be simultaneous with the modification of carrier files. The damage of the watermark can thus both confirm the counterfeit of the APP.

4.2 Performance Analysis

The performance of DroidMark can be evaluated in two folders: efficiency of embedding and extraction, and capacity of watermark. Efficiency is related to the pretreatment of scanning for the location, and the process of embedding and extracting the watermark. Meanwhile capacity is influenced by the size and type of the carrier files, and the algorithm for watermarking as well.

4.2.1 Efficiency

During the process of embedding and extraction, the cost of computation is mainly induced by two aspects: carrier file scan, watermark implant and extraction.

In our algorithm, the length of watermark decide the time for scanning for embedding, while in the process of extraction, it's decided by the size of carrier file. Therefore, the efficiency cost grows along with the scale of the watermark and the length of carrier file. Further more, to inquire which element has more influence, we design two experiments as follows.

- **Efficiency of DroidMark-XML:**

Firstly, we embedded and extracted different watermark in the same AndroidManifest.xml file to control the environment variables of the length of the carrier file. Watermarks range from 1 to 1600 bits increasing by 100 bits, and the Fig. 2 shows that the overhead of the whole process of embedding and extracting watermarks. We can see that the cost of time is linearly other than exponentially increasing with the number of bit of watermarking, which indicates that the overhead increases along with the size of the watermark embedded is within a controllable scale. That is also the inevitable cost for embedding and extracting operations. Therefore, we can conclude that the cost of this proposed scheme is efficient and stable.

Secondly, the length of the file only has fatal effect on the scanning overhead of watermark's extraction. To explore the relationship between the scanning overhead and the length of the carrier file, we design the experiment to embed and extract the same watermark in 8 bits in a series of incremental size of carrier files. We keep the content of watermark as 8 bits to ensure the same cost of embedding and extracting, so that to achieve the influence of the scanning only. Besides, the increasing content of the carrier file is the repetition of central body in 1271 characters based on 1527 characters of the original file. We adopt this method to guarantee the validity of AndroidManifest.xml file, and the arithmetic progression is in favor of statistics of overhead as well. The result of the experiment is shown in the Fig. 3. The graph indicates that the overhead of scanning increases only less than 2 seconds in the file augment of more than 20 thousand characters, which is much less than the overhead of the process of embedding and extracting.

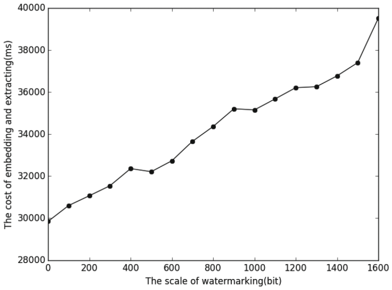


Fig. 2. Overhead in different length of watermark

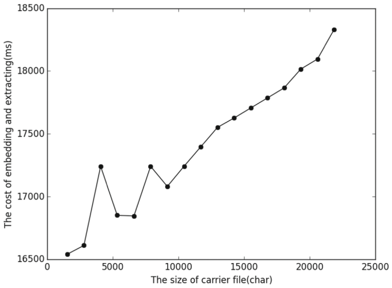


Fig. 3. Overhead in different size of file

- **Efficiency of DroidMark-DEX:**

In the process of DroidMark-Dex, the extra overhead compares to the DroidMark-XML is the confusion of embedding location, which is proved to be low-time-consuming operation. Hence the overhead of DroidMark-DEX has the similar tendency with DroidMark-XML. And for this circumstance and space limitations, the DroidMark-DEX experimental result is abridged.

Therefore, we can draw the conclusion that, the overhead mainly comes from the process of embedding and extraction, and the linearly growth with in a small scale justifies the efficiency and stability of our scheme.

4.2.2 Capacity

The capacity refers to how much bits of the original content are needed to hide one bit watermark information. Under the requirement of practicability and security, the higher capacity indicates the greater ability, applicability, and efficiency of watermark algorithm.

- **Capacity of DroidMark-XML:**

In order to detect the capacity of DroidMark-XML, we choose 100 different AndroidManifest.xml files decompiled from 100 Apps. And we calculate each characters of xml file as the x axis, while collect the totality location for watermark embedding as the y axis. The relationship between x and y is shown in Fig. 4. According to Fig. 4, we draw a baseline slopes 0.2, representing that the capacity of the AndroidManifest.xml is around 20 percent of the full text, which is pretty high compares to other android watermark algorithms. Therefore, we can conclude that the this proposed scheme is stable and high-capacity.

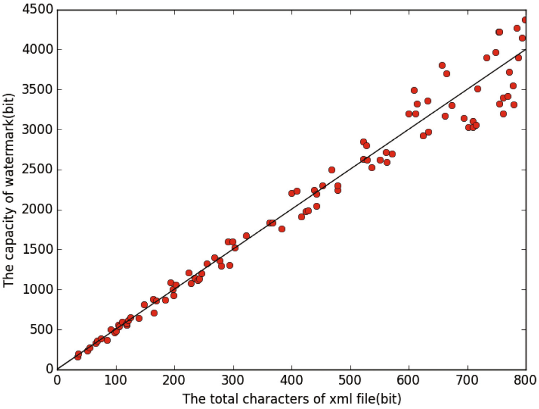


Fig. 4. Capacity of DroidMark-XML

- **Capacity of DroidMark-DEX:**

As for the capacity of DroidMark-DEX, the location in which the watermark is properly embedded is relatively fixed. Therefore, the capacity of DroidMark-DEX is fixed in a range rather than grows linearly. On account of redundancy between $Header\_item$ and Map_list in classes.dex, the available field of dex file we choose are "*unused*" fields in each "map_list_item[]" which is used to format the alignment. The amount of unused field is usually less than 20. Calculate this fields, we can conclude that the capacity of DroidMark-DEX is $20*2*8 = 320$ bits at most, which is much less than what in DroidMark-XML. But at the watermark in DroidMark-DEX has better effect of confusion and has tighter organization, which will offer better protection of watermark integrity.

5 Conclusions

In this paper, we proposed an watermark tool, called DroidMark, to protect the copyright of android applications. This method embeds the watermark in application configuration files, which achieves great concealment. At the same time, the semi-fragile property insure the authenticity for DroidMark. More over, the method achieves the disuse of secret key in the process of embedding and extracting, which caters applications property of circulation and lightweight. DroidMark is designed to achieve Android copyright protection, and can also be used in information hiding, secrete communication, provenance-based forensics, and key distribution. The watermarking algorithms and carrier selection guarantee the reliability and applicability of DroidMark. The achievements of security, high-performance and lightweight property of DroidMark is extensively analyzed and thoroughly verified by experiments.

Acknowledgements. The research was financially supported by the National Natural Science Foundation of China under Grant No.s 61310306028.

References

1. Sufatrio, Tan, D.J.J., Chua, T.-W., Thing, V.L.L.: Securing android: a survey, taxonomy, and challenges. ACM Comput. Surv. **47**, 58–102 (2015)
2. Choi, S., Jang, J., Jae, E.: Android applications copyright protection technology based on forensic mark. In: Proceedings of the 2012 ACM Research in Applied Computation Symposium (RACS 2012), pp. 338–339 (2012)
3. Kim, S.R., Kim, J.H., Kim, H.S.: A hybrid design of online execution class and encryption-based copyright protection for android apps. In: Proceedings of the 2012 ACM Research in Applied Computation Symposium (RACS 2012), pp. 342–343 (2012)
4. Atallah, M.J., McDonough, C., Nirenburg, S., Raskin, V.: Natural language processing for information assurance and security: an overview and implementations. In: Proceedings of the 2000 Workshop on New Security Paradigms (NSPW 2000), pp. 51–65 (2000)
5. Atallah, M.J., Raskin, V., Crogan, M., Hempelmann, C., Kerschbaum, F., Mohamed, D., Naik, S.: Natural language watermarking: design, analysis, and a proof-of-concept implementation. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, pp. 185–200. Springer, Heidelberg (2001). doi:10.1007/3-540-45496-9_14
6. Meral, H.M., et al.: Natural language watermarking via morphosyntactic alterations. Comput. Speech Lang. **23**, 107–125 (2009)
7. Zhang, Y., Chen, K.: AppMark: a picture-based watermark for android apps. In: Eighth International Conference on Software Security and Reliability (SERE 2014), pp. 58–67 (2014)
8. Zhou, W., Zhang, X., Jiang, X.: AppInk: watermarking android apps for repackaging deterrence. In: Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security (ASIA CCS 2013), pp. 1–12 (2013)

Research on CSER Rumor Spreading Model in Online Social Network

An-Hui Li[1(✉)], Jing Wang[1,2], and Ya-Qi Wang[1]

[1] Department of Electronic Technology,
Engineering College of the Chinese Armed Police Force,
Xi'an 710086, Shaanxi, China
anhui9307@163.com

[2] College of Primary Command, Rocket Force University of Engineering,
Xi'an 710025, China

Abstract. In the field of rumor spreading, kill rumor is a very important concept. In the previous study of rumor spreading, only considering the rumor from the outside of the network, while ignoring the node itself has the discriminability. In this paper, a new CSER rumor propagation model is proposed and a kind of node with the ability of killing rumor is introduced in the model, that node is called rumor-killer. By means of complex network theory and mean-field method, we establish the differential equations of propagation dynamics. In this paper, the model is used to study the process of rumor spreading both in homogeneous network and heterogeneous network. Through theoretical analysis and experimental simulation, it is found that the node with the ability of killing rumor can mitigate the spreading rumor in the network. This conclusion provides us a new way to control the rumor spreading from the inside of network.

1 Introduction

The emergence of the online social network not only greatly enhanced the link between people, but also provided a way for the wide spread of rumors. In the year of 2011, after the incident of the Japan Fukushima nuclear leak [1], because of one message "iodized salt can prevent radiation", the public crazily to buy iodized salt, and iodized salt was sold out, which caused a great panic, it seriously disrupted the normal market order and brought a great loss to the whole society. In order to effectively deal with the adverse effects of rumors such as "iodized salt can prevent radiation", many scholars at home and abroad are devoted to research rumor spreading mechanism, hoping to reveal the inherent laws of rumor spreading so that can timely control the spread of rumors.

This work is supported by the National Natural Science Foundation of China (Grant No. 61402531), Shaanxi Provincial Natural Science Basic Research Project (Grant No. 2014JQ8358, 2015JQ6231, 2014JQ8307), the China Postdoctoral Science Foundation(Grant No.2015M582910).

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7_76

There are many similarities between the rumor spreading and the virus spreading, so we can learn from the virus spreading mechanism [2] to study the rumor spreading. Based on complex network theory and mean field method, we proposed rumor spreading model and explored the inherent law and mechanism of rumor spreading. The earliest rumor spreading model is the D-K model [3], which was proposed by Daley and Kendall in 1965. It divided the population into three categories: those who not heard rumors, those who are spreading the rumors, those who heard rumors but do not spread it, and the rumors spread through the mutual contact between individuals. Potts model [4] is a rumor spreading model proposed from a physical point of view, the model mainly from the perspective of semantic change to do quantitative research of rumor spreading process. In the book “Complex System Simulation and Application”, Xuan Hui-Yu and Zhang Fa [5] made a complete introduction to the cellular automaton-based rumor spreading model. Cellular automata as a discrete mathematical model, which can better show how rumors spreading among individuals. By taking into account some factors such as belief, forgetting probability, etc. [6], we can express the process of rumor spreading more really reflect. The above mentioned models only considering the spreading process itself, but do not take into account the differences in the nodes and the impact of network topology on rumor spreading. Moreover, rumor spreading is also a social discipline, not just from a mathematical point of view to carry out research. So, the Complex network theory [7] appeared. Zanette [8, 9] and others used the complex network theory to study the rumor spreading. They firstly found the law that rumor spreading has a critical value in the small-world network [10]. Moreno [11] established a rumor-spreading model on scale-free networks, and introduced mean-field equations for homogeneous networks. He concluded that there was no critical value for rumor spreading in homogeneous networks.

In all the above studies on the rumor spreading model, the main consideration is that the nodes passively accept the rumors, not considering the entire network which include a small part of the node is active when the rumors began to spread in the whole network, it will interfere with the normal spreading of rumors. Therefore, we introduce a new type of node that called rumor-killers. When the rumors begin to spread throughout the network, the rumor-killers will play their effectiveness and begin to kill rumors, after contacting the rumor-killers, the spreader will be a certain probability into a rumor-killer or rational. In real life, there are often some wise people or functional departments, In the early days when rumors began to spread, they will take appropriate measures to prevent its continued spreading. Our rumor spreading model that introduced the rumor-killer nods, it will be more realistic.

2 CSER Rumor Spreading Model

This paper proposes a CSER rumor spreading model, which is based on the CSR model, taking into account the fact that some users in social organizations or organizations have similar functions to the police, and can control rumor spreading after rumors began to spread. Compared with the traditional rumor transmission model, the main advantages are as follows:

(1) The forgetting mechanism is considered in the traditional rumor-spreading model, and the rumor does not generally appear when propagated on micro-messages because rumors have been transmitted over the Internet, not by word of mouth, it is not relevant to the memory capacity.

(2) The traditional rumor spreading model is derived from the epidemic model, but these spreading models are often offline. In offline network, each node's neighbor nodes is limited. That is to say, everyone comes into contact with few people than online social network, and these models do not take into account the existence of rumor-killers, only taking into account the loss of interest or passive conversion. In the rapid development of the network today, the rate of rumor spreading is clearly faster, so we introduce the rumor-killer in the network, it is more consistent with today's actual situation.

Based on the above description, in the CSER model, we divide the population into four categories: the credulous C, the spreader S, the rumor-killer E, and the rational R. The rumor spreading rules in the model are as follows:

i. When a credulous contacts with the spreader, the credulous is transformed into a spreader with a certain probability λ, it is possible that the credulous is not interested in the rumor but is directly converted to the rational, the Probability is β.

ii. When a spreader contacts a rumor-killer in the process of transmission, it is probable that the spreader will be converted to a rumor-killer with the probability α.

iii. When a spreader communicates with other spreaders, rumor-killers, or rational, they will be translated to the rational with the probability σ.

iv. Spreaders in the dissemination process may lose the interest and translate into rational, the probability is δ.

The Fig. 1 shows the state transition diagram of rumor spreading, the total number of users in the network is N, At the moment t, the density of the credulous, the spreader, the rumor-killer, the rational is $C(t)$, $S(t)$, $E(t)$ and $R(t)$, respectively. The equation of $C(t) + S(t) + E(t) + R(t) = 1$ is also satisfied. The dynamic equations of the CSER model are as follows:

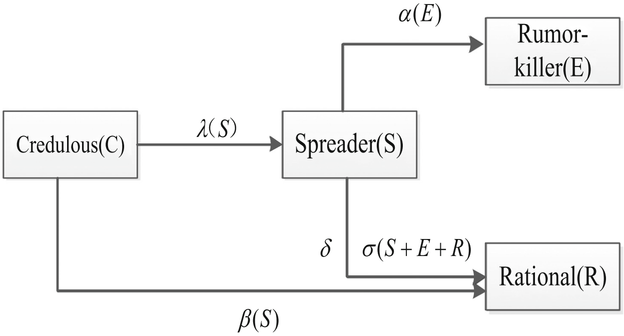


Fig. 1. CSER rumor spreading model state transition diagram

$$\frac{dC(t)}{dt} = -(\lambda + \beta)\langle k\rangle C(t)S(t) \tag{1}$$

$$\frac{dS(t)}{dt} = \lambda\langle k\rangle C(t)S(t) - \delta S(t) - (\alpha + \sigma)\langle k\rangle S(t)E(t) \tag{2}$$

$$\frac{dE(t)}{dt} = \alpha\langle k\rangle S(t)E(t) \tag{3}$$

$$\frac{dR(t)}{dt} = \beta\langle k\rangle C(t)S(t) + \delta S(t) + \sigma\langle k\rangle S(t)[E(t) + S(t) + R(t)] \tag{4}$$

The $\langle k\rangle$ is the average degree of the whole network node, the four differential equations express that the changing rate of credulous, spreaders, rumor-killer, rational nodes with the time.

3 Analysis

In the whole process of rumor spreading, the number of spreaders is increased to a peak and then began to decrease until to zero, the entire network also reached a steady state at this time. Suppose the final number of rumors is R, $R = \mathit{final}\{R(t)\} = \lim_{t\to\infty} R(t) = R(\infty)$. We can use the value of R to measure the impact of the spread of rumors, for example, if the $R = 0.9$, 90% of the nodes in the entire network have heard rumors. At the beginning of the rumor spreading, suppose there is one spreader and one rumor-killer, the initial conditions are as follows: $C(0) = \frac{N-2}{N} \approx 1$, $S(0) = \frac{1}{N} \approx 0$, $E(0) = \frac{1}{N} \approx 0$, $R(0) = 0$.

Dividing Eq. (4) by Eq. (1), we can get the Eq. (5)

$$\frac{dR(t)}{dC(t)} = \frac{\beta\langle k\rangle C(t)R(t) + \sigma\langle k\rangle S(t)[C(t) + R(t) + S(t)] + \delta S(t)}{-(\lambda + \beta)\langle k\rangle C(t)S(t)} \tag{5}$$

Equation (5) is simplified to obtain the Eq. (6)

$$dR(t) = \frac{\sigma - \beta}{\lambda + \beta} d(t) - \left(\frac{\sigma}{\lambda + \beta} + \frac{\delta}{(\lambda + \beta)\langle k\rangle}\right)\frac{dC(t)}{C(t)} \tag{6}$$

The Eq. (6), respectively, on both sides and the derivative to $R(t)$ and $C(t)$, Combining the initial conditions $C(0) = 1$, $R(0) = 0$, $C(\infty) = 1 - R(\infty) = 1 - R$, we can get Eq. (7)

$$-(\lambda + \sigma)R = (\sigma + \frac{\delta}{\langle k\rangle})\ln(1 - R) \tag{7}$$

Equation (7) for constant deformation, we can get the following transcendental equation:

$$R = 1 - e^{-\varepsilon R} \tag{8}$$

Including

$$\varepsilon = \frac{(\lambda + \sigma)\langle k\rangle}{\sigma\langle k\rangle + \delta} \tag{9}$$

Equations (8) and (9) are the transcendental equations in the CSER rumor spreading model.

Conclusion when $\varepsilon > 1$,the equation $x = 1 - e^{-\varepsilon x}$ has two solutions, the one is a zero solution, the other one is a non-zero solution R, which can meet the condition $0 < R < 1$.

Prove. We can easily get $x = 0$ is a solution of the equation $x = 1 - e^{-\varepsilon x}$. Suppose the $y = x - 1 + e^{-\varepsilon x}$,and find the first derivative and the second derivative of x, so we can get $y' = 1 - \varepsilon e^{-\varepsilon x}$ and $y'' = \varepsilon^2 e^{-\varepsilon x} > 0$. By the second derivative can be obtained y is a concave function. And $y'(0) = 1 - \varepsilon < 0$, $y(1) = e^{-\varepsilon x} > 0$,the equation has a solution in the interval from 0 to 1. So the conclusion proved to be completed.

When meet the condition $\varepsilon > 1$,It has been proved that the equation has a nontrivial solution in addition to the zero solution, so it can be concluded that the CSER rumor spreading model has no spreading threshold in the homogeneous network, which is consistent with the conclusion in reference [10].

4 Experimental Simulation

4.1 Experimental Environment

In this paper, all experiments are implemented in the Intel Core i5-2450 M 2.50 GHz frequency, Windows 8.1 operating system, MATLAB R2014a platform.

4.2 Homogeneous Network

In the simulation experiment of homogeneous network, we first generate a small-world network with 2000 nodes and average node degree of 7. Then, we conduct CSR rumor spreading model and CSER rumor spreading model on this small-world network.

Figure 2 shows the variation of the nodes in the CSR model with time. From Fig. 2 we can see that there was only rational in the network when steady state is reached, which is similar to the classical virus spreading SIR model.

Figure 3 shows the variation of the nodes in the CSER model with time. We can see that the whole network reaches equilibrium at the time $t = 25$. After the introduction of the rumor-killer node in the model, as rumors begin to spread in the network, the number of rumor-killers begin to increase, it plays a certain inhibitory effect to the spreaders and reduces the spreading range of rumors, so in the CSER model simulation

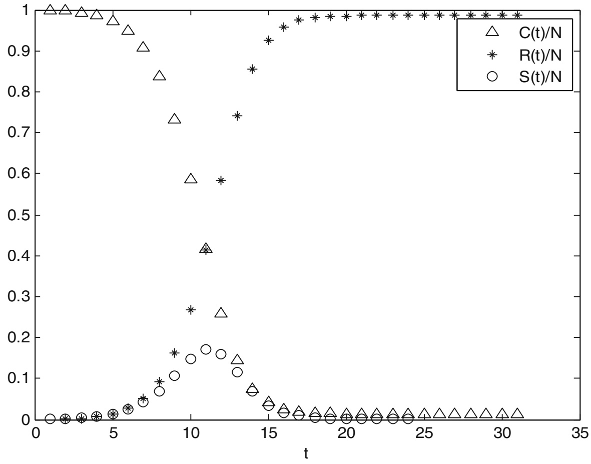


Fig. 2. CSR model rumor spreading simulation results

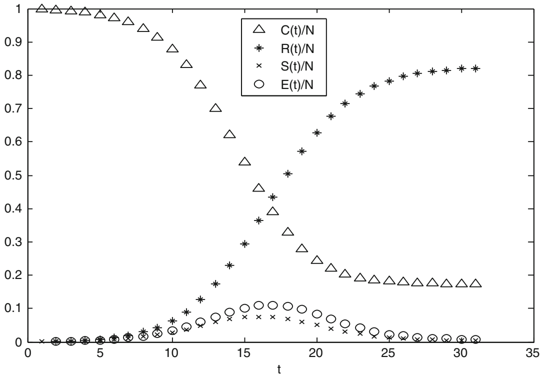


Fig. 3. CSER model rumor spreading simulation results

experiments, the number of rational nodes are relatively less than CSR model, it verified the reliability and correctness of the proposed model.

Figure 4 shows the rumors spreading process in the CSR and CSER models with the effective spreading rate λ changed. It can be seen from Fig. 4 that as the effective spreading rate λ increases, the rumor spreading range increases, but when the value of λ are the same, the values of $R(\infty)$ in the CSER spreading model are relatively smaller than in the CSR model. After the introduction of rumor-killers, it is not only closer to reality, but also inhibit the spread of rumors. It provides a new way of control the rumor spreading.

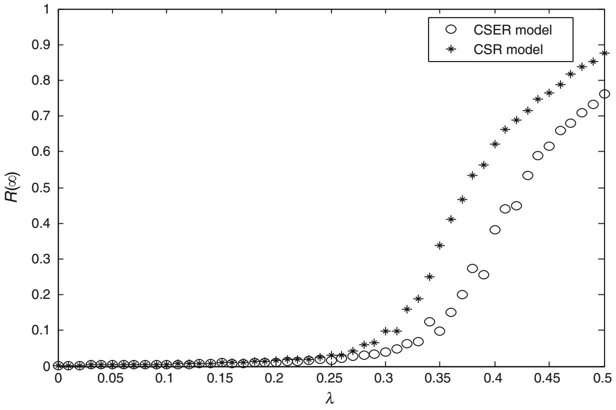


Fig. 4. Rumor spreading in the homogeneous network with the effective spreading rate λ

4.3 Heterogeneous Network

In this simulation, we use the BA scale-free network to compare the CSR and CSER models in heterogeneous networks. The network contains 2000 nodes and is subject to exponential power-law distribution, which has the exponential index is $\rho = 2$. Figure 5 shows the out-of-scale distribution of the scale-free networks used.

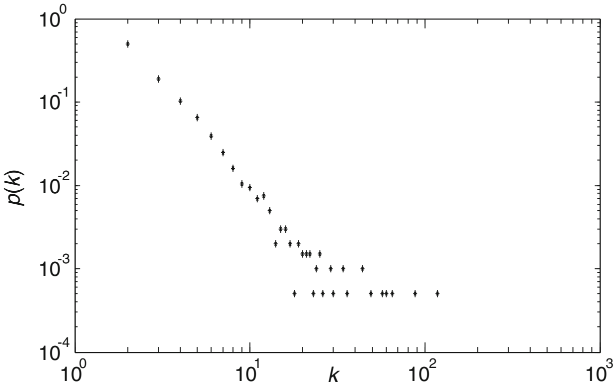


Fig. 5. The degree distribution of BA-scale network

Similar to the homogeneous network, in the heterogeneous network, we also draw the changing pattern of the rumor spreading with the effective spreading rate under CSR and CSER models. The simulation value is selected as $\alpha = 0.2$, $\beta = 0.4$, $\delta = 0.4$, $\sigma = 0.1$. From Fig. 6, it can be seen that the value of $R(\infty)$ varies with the effective spreading rate λ. We can see the same result as in the non-homogeneous network. That is to say, the proposed CSER can reduce the range of rumors spreading.

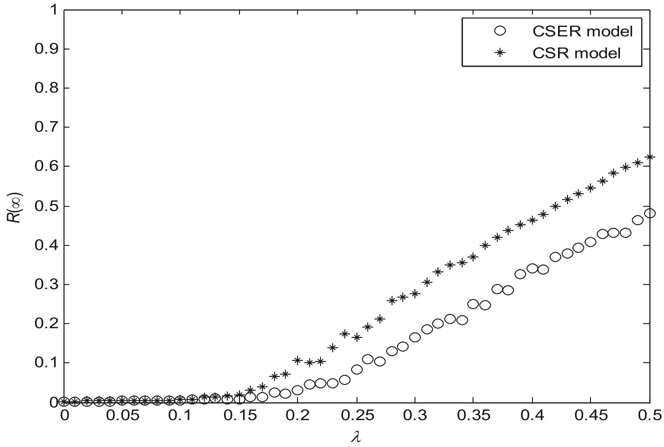


Fig. 6. Rumor spreading in the heterogeneous network with the effective spreading rate λ

5 Conclusions

In this paper, we consider the rumor spreading model of social network after adding rumor-killer. Based on the analysis of information transmission characteristics of online social network, a new CSER model is proposed based on CSR model, and the CSER model is verified on homogeneous network and heterogeneous network. It was proved that adding the rumor-killer is better to depict the characteristics of the online social network and provide a new feasible scheme for us to control rumor within the network. However, the real online social networks have both Homogeneous and Heterogeneous network characteristics, in order to better adapt to the reality, the next step in the study should be further strengthened.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 61402531), Shaanxi Provincial Natural Science Basic Research Project (Grant No. 2014JQ8358, 2015JQ6231, 2014JQ8307).

References

1. Wang, Z.T., Chai, G.H.: Japan Fukushima Nuclear Accident. Atomic Energy Press (2014)
2. Zhou, T., Fu, Z.Q., Niu, Y.W., Wang, D., Zeng, Y., Wang, B.H.: A survey of propagation dynamics on complex networks. Prog. Nat. Sci. **15**(5), 513–518 (2005)
3. Maki, D.P., Thompson, M.: Mathematical Models. Methods and Applications. Springer, Singapore (2015)
4. Shao, C.G.: The timid rumors spread the rumors of the Potts model. (Doctoral dissertation, Huazhong University of Science and Technology) (2003)
5. Xuan, H.Y.: Complex System Simulation and Application. Tsinghua University Press, Beijing (2008)

6. Zhang, F., Si, G.Y., Luo, P.: A survey for rumor propagation models. Complex Syst. Complexity Sci. **6**(4), 1–11 (2009)
7. Wang, X.F., Li, X., Chen, G.R.: Complex Network Theory and Its Application. Qing Hua University Publication, Beijing (2006)
8. Zanette, D.H.: Criticality of rumor propagation on small-world networks. arXiv preprint cond-mat/0109049 (2001)
9. Zanette, D.H.: Dynamics of rumor propagation on small-world networks. Phys. Rev. E **65** (4), 041908 (2002)
10. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)
11. Moreno, Y., Nekovee, M., Pacheco, A.F.: Dynamics of rumor spreading in complex networks. Phys. Rev. E **69**(6), 066130 (2004)

Author Index

© Springer International Publishing AG 2018
L. Barolli et al. (eds.), *Advances in Internetworking, Data & Web Technologies*,
Lecture Notes on Data Engineering and Communications Technologies 6,
DOI 10.1007/978-3-319-59463-7