Jeffrey W. Tweedale
Lakhmi C. Jain (Eds.)

Communications in Computer and Information Science 246

Advanced Techniques for Knowledge Engineering and Innovative Applications

16th International Conference, KES 2012
San Sebastian, Spain, September 2012
Revised Selected Papers

Springer

Communications in Computer and Information Science 246

Jeffrey W. Tweedale Lakhmi C. Jain (Eds.)

Advanced Techniques for Knowledge Engineering and Innovative Applications

16th International Conference, KES 2012
San Sebastian, Spain, September 10-12, 2012
Revised Selected Papers

Volume Editors

Jeffrey W. Tweedale
Defence Science and Technology Organisation
Edinburgh, SA 5111, Australia
E-mail: jeffrey.tweedale@dsto.defence.gov.au

and affiliated as an adjunct with the
University of South Australia
School of Engineering
Adelaide, SA 5095, Australia
E-mail: jeffrey.tweedale@unisa.edu.au

Lakhmi C. Jain
University of Canberra
Faculty of Education, Science, Technology
and Mathematics
Canberra, ACT 2601, Australia
E-mail: lakhmi.jain@canberra.edu.au

ISSN 1865-0929 e-ISSN 1865-0937
ISBN 978-3-642-42016-0 e-ISBN 978-3-642-42017-7
DOI 10.1007/978-3-642-42017-7
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013954675

CR Subject Classification (1998): H.2.8, I.2, I.4, I.5, H.2, H.3, H.4

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

This book contains a selection of the best papers presented during the invited sessions of the 16th International Conference on (KES). There were over 254 papers submitted to these sessions and only 34 authors were invited to extend their papers into full chapters. Following a second review process, 21 papers were selected for inclusion in this book.

KES International was born from a community of like-minded researchers promoting excellence and innovation in knowledge management techniques. As a result of the collective action of these researchers seeking formal feedback through peer review, the annual broad-spectrum KES intelligent systems conference series was initiated. For over 15 years, KES international events have served as a platform for sharing the latest developments in the pursuit of knowlege using intelligent systems. The 16th Annual KES conference was held in the beautiful city of San Sebastian, in the north of Spain (see http://kes2012.kesinternational.org/index.php). At the request of the committee, the editors invited a limited number of authors to submit extended papers. A subset of this competitive field of submissions were selected in order to share the latest achievements in this domain. As such, the contents of this book represent the best contributions from the "invited sessions" received and presented at the conference by leading international experts. The quality of these contributions clearly shows that knowledge engineering is more than just a trendy topic; it is a continuous living and evolving set of technologies aimed at improving the design and understanding of systems and their relations with humans.

Knowledge engineering relies on the exploitation of AI techniques to employ human-like intelligence in machine systems so as to solve specific problems. Researchers continue to improve existing techniques within the domain. This evolution in information processing has become a pervasive phenomenon within the community. Mobile computing continues to promote the ubiquitous access of information resources. Technolgy is providing increased processing capabilities to hand-held devices, forcing more innovative access techniques onto existing intelligent systems. Society is beginning to demand everyday applications that provide convenient access to the wealth of information-processing systems serving the public. To achieve this, we must take advantage of the most recent research in information technologies. The major research threads have manifested in semantics, artificial intelligence, and knowledge engineering that support domain-specific applications. Intelligent systems are becoming ubiquitous in a wide range of situations. These include facets of simple everyday actions on mobile devices to more advanced enterprise-level applications in logistics systems and in the medical domain. Society benefits daily, through digital news, socialization of relations, and enhancements derived from expert decision making in knowledge-based systems.

Dynamic ontologies play a major role in the development of distributed knowledge engineering, especially in the Semantic Web. This influences the design of modern decision-support systems. They are used for the specification of natural language semantics, information modeling and retrieval in querying systems, geographical information systems, and medical information systems—the list is growing continuously. Ontologies allow for easy modeling of heterogeneous information, flexible reasoning for the derivation of consequents or the search of query answers, specification of a priori knowledge, and increasing accumulation of new facts and relations (e.g., reflexive ontologies). Therefore, they are becoming key components of adaptable information-processing systems. Classic problems such as ontology matching or instantiation have new and more complex formulations and solutions, involving a mixture of underlying technologies, from traditional logic artificial neural networks and fuzzy logic. A selection of these highly active research topics, techniques, and applications in computational intelligence are provided in this book.

This publication would not have been possible without the suport of the KES International Program Committee and the conference chairs. The editors wish to express their gratitude for the support and dedication they provided in hosting the 16th Annual KES Conference. They are the academic backbone supporting activity.

As editors, we are proud to present a number of hand-selected papers that have been extended. These contributions span several theoretical and research conceptualizations, complete with real-world applications. We must thank a large number of people that have contributed to the success of this endeavor as reviewers of one or more submisisons. They include:

J. Abe
B. Ayerdi
D. Barbucha
F. Bellas
R. Chaves
A. Consoli
P. Cutler
R. Duro
M. Favorskaya
W. Filipowicz
Y. Fujita
M. Grana
L. Jain
P. Jedrzejowicz
R. Kountchev
N. Martin
M. Moreno
K. Nakamatsu
H. Prado
E. Ratajczak-Ropel
E. Sanchez
C. Sanin
M. Sato-Ilic
A. Savio
F. Segovia
C. Sioutis
A. Skakovski
E. Szczerbicki
M. Takahashi
M. Termenon
C. Toro
J. Tweedale
M. Veganzones

A special thanks is extended to 16th annual conference chairs for hosting the conference in San Sebastian, Spain, during September 10-12, 2012. We also wish to acknowledge the support of the Basque Government, Vicomtech-IK4, and the University of the Basque Country for contributing to the success of this meeting. They include: Manuel Graña, Jorge Posada, Carlos Toro and Robert J. Howlett.

May 2013

Jeffrey W. Tweedale
Lakhmi C. Jain

Acronyms

k-NN	k-Nearest Neighbor
3G	3^{rd} Generation of mobile telecommunications technology
A-Team	Asynchronous Team
ACE	The Ace Orb
ADF	Australian Defence Force
ADIIB	Australian Defence Force ISR Integration Backbone
ADL	Activities of Daily Living
ADO	Australian Defence Organisation
AD	Alzheimer's Disease
AEW&C	Airborne Early Warning and Control
AHI	Answers that Have Integrity
AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
ANN	Artificial Neural Network
API	Application Programming Interface
ASAP	As Soon As Possible
ASCEL	Airborne Systems Connectivity Environment Laboratory
ATU	Answers That are Uncertain
BDC	Bootstrapped Dendritic Classifiers
BUT	Break-Up Time
CAAT	Computer-Assisted Auditing Technique
CAD	Computer Aided Diagnosis
CA	Cellular Automata
CBT	Computer-Based Test
CCIS	Communications in Computer and Information Science
CDR	Clinical Dementia Rating
CDSA	Connecting Soldiers to Digital Applications
CFP	Call-for-Proposal
CIOG	Chief Information Officer Group
CORBA	Common Object request Broker Architecture
CPD	Collaborative Product Design
CQA	Consistent Query Answering
CSP	Constraint Satisfaction Problem
CTAN	Comprehensive TEX Archive Network
CT	Computed Tomography
DARPA	Defense Advanced Research Projects Agency
DBA	Data Base Agent
DBMS	Data-Base Maangement System
DC	Dendritic Classifiers

DDS	Data Distribution Service
DD	Double-Dipping
DES	Differential Equations Solver
DoD	Department of Defence
DSTO	Defence Science and Technology Organisation
DTI	Diffusion Tensor Imaging
ECA	Evolving Cellular Automata
ECG	Eelectrocardiogram
EIS	Enterprise Information System
ELM	Extreme Learning Machines
EMG	Electromyogram
EV	Evaluating Variable
FIPA	Foundation for Intelligent Physical Agents
FIS	Fuzzy Inference System
FL	Fuzzy Logic
FN	False Negatives
FOCAL	Future Operations Centre Analysis Laboratory
FPGA	Field Programmable Gate Array
FP	False Positives
FTTH	Fiber-To-The-Home
GBM	Graph-Based Matcher
GICS	Global Industry Classification Standard
GM	Gray Matter
GPRS	General Packet Radio Service
GPS	Global Positioning System
GPU	Graphic processing unit
GSEG	Gradient SEGmentation
Hb	Hemoblobin
HIS	Hyperspectral image
HRMF	Hidden Markov Random Fields
HR	Heart Rate
HSV	Hue, Saturation, Value
ICBM	International Consortium for Brain Mapping
ICT	Information and Communication Technology
IIA	Institute of Internal Auditors
INRIA	Institut pour la Recherche en Informatique et Automatique
IPA	Information Technology Promotion Agency
ISR	Intelligence, Surveillance and Reconnaissance
ITSS	Information Technology Skill Standards
IT	Information Technology
JABAT	JADE based A-Team
JADE	Java Agent Development Framework
KM	Knowledge Modelling
LA	Logical Agent
LBP	Local Binary Pattern

LEP	Local Edge Pattern
LOO-CV	Leave One Out - Cross Validation
LTE	Long-Term Evolution
LVQ	Learning Vector Quantization
MAS	Multi-Agent System
METI	Ministry of Economy, Trade and Industry
ML	Machine Learning
MMSE	Mini-Mental State Examination
MOPET	Mobile Personal Trainer
MP-RAGE	Magnetization-Prepared Rapid Gradient Echo
MRI	Magnetic Resonance Imaging
MSM	Matcher for Semantic Matching
MST	Mission System Testbed
MVNO	Mobile Virtual Network Operator
MVS	Model Validation Service
NCW	Network Centric Warfare
NIRS	Near-Infrared Spectroscopy
NN	Neural Network
NP-hard	Non-deterministic Polynomial hard
OAEI	Ontology Alignment Evaluation Initiative
OASIS	Open Access Series of Imaging Studies
OA	Overall Accuracy
OECD	Organization for Economic Co-operation and Development
OMG	Object Management Group
ORB	Object Request Broker
OWL	Ontology Web Language
PCA	Principal Component Analysis
PCA-SIFT	Principal Component Analysis – Shift Invariant Feature Transformation
PC	Principal Component
PDE	Product Design Engineering
PDF	Portable Document Format
PIP	Personal Injury Protection
PR	Pattern Recognition
PSPLIB	Project Scheduling Problems LIBrary
QoS	Quality of Service
RCPSP	Resource-Constrained Project Scheduling Problem
RGB	Red, Green, Blue
RMT	Random Matrix Theory
RMT-PCA	Random Matrix Theory – Principal Component Analysis
ROI	Region Of Interest
RT-CORBA	Real Time - CORBA
RVGA	Real Valued Genetic Algorithm
S2PL	Strict Two-Phase Locking

SBM	String-Based Matcher
SCS	Speed Control Service
SDK	Software Development Kit
SGS	Serial Generation Scheme
SIFT	Shift Invariant Feature Transformation
SI	International System
SIS	System Identification Service
SLFN	Single Layer Feedforward Network
SME	Subject Matter Expert
SNLDC	Single Neuron Lattice Model with Dendrite Computation
SOA	Service Oriented Architecture
SPECT	Single-Photon Emission Computed Tomography
S&T	Science and Technology
SVM	Support Vector Machines
TA-Teams	Team of A-Teams
TA	TutorAgent
TAO	Adaptive Communication Environment
TCU	Tribunal de Contas da Uniao
TN	True Negatives
TOEIC®	Test of English for International Communication
TPO	Training Protocol Optimizer
TP	True Positives
UniSA	University of South Australia
UTIC	Uncertainty-tolerant Integrity Checking
VBM	Voxel Based Morphometry
W3C	World Wide Web Consortium
WBT	Web-Based Test
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WIRE	Wedgetail Integration and Research Environment
XMPP	Extensible Messaging and Presence Protocol

Table of Contents

Part I: Data Mining

Part II: Classifiers

Part I

Data Mining

Brain Activity Measurement for the Scores of On-line English Grammar Tests with White and Blue Backgrounds

Atsuko K. Yamazaki[1], Kaoru Eto[2], Akane Nakabayashi[1], and Hitomi Shimada[1]

[1] College of Engineering, Shibaura Institute of Technology, 307 Fukasaku, Minuma-ku, Saitamashi, Saitama, Japan
atsuko,f09070,f09050@sic.shibaura-it.ac.jp
[2] Faculty of Engineering, Nippon Institute of Technology, 4-1 Gakuendai, Miyashiro-cho, Minami Saitama-gun, Saitama, Japan
eto@nit.ac.jp

Abstract. Yamazaki' study and a study by Yamazaki and Eto indicated that a combination of black text and a background color with high luminance and high brightness was not considered preferable for Web-based tests (WBTs). In this study, the authors conducted an experiment to examine if the scores of on-line English tests differed depending on the characteristics of a background color. By using near-infrared spectroscopy, relative changes in Hemoblobin (Hb) concentrations in the brains of test takers were observed to see how background colors can affect the functions of their brains. Twenty four subjects in their twenties took web-based English grammar and non-linguistic tests with white and blue background colors with black text. The average scores of the linguistic and non-linguistic tests for the blue background were higher than those for the white background among the subjects. In particular, a significant difference in subjects' performance was found between the white and blue backgrounds for the non-linguistic task. Two dimensional images of the Hb concentration changes obtained in the experiment showed that areas in the brain related to the frontal eye field were observed to be more active while the subjects were taking the tests with a white background. These results indicate that the combination of black text and white background encourages a test-taker to concentrate more on visual input from the screen rather than the test questions. They also suggest that the combination may not be the best choice for a background color of a WBT to assess test-takers' linguistic performance even though a white background is commonly used for WBTs.

Keywords: Background color, Web-based test, Test performance, Brain Functions, Near-Infrared Spectroscopy, Brodmann Areas.

1 Introduction

The selection of background colors of web pages has been evaluated in many studies in order to improve the usability of websites [1,2,3,4]. Many studies on

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 3–15, 2013.

colors used in web design have pointed out that the combination of background and text colors is important for performing tasks on Web pages in terms of readability. From their experimental results, Hall and Hanna pointed out that greater contrast ratios between the background and font colors improved the readability of a Web page. Their results, however, showed that the combination of background and text colors did not significantly affect users' retention [1]. On the other hand, other studies found that a luminance contrast ratio between the text and background colors of a Web page can affect its readability [2,3,4]. According to a study of background colors conducted by Mehta and Zhu [5], red is beneficial for some kinds of mental processing while blue is better for others tasks. In their experiments, the results suggested that red backgrounds were better for tasks that require an attention to detail, and blue enhanced performance on approach-based and exploratory tasks. Their experiments showed that the subjects performed better on a word-recall task and a proof-reading task when the screen background was red, while the subjects came up with better quality and more creative ideas for things to do with a brick when the screen was blue, rather than red.

Computer-Based Tests (CBTs) and web-based tests Web-Based Tests (WBTs) have been widely used to assess human knowledge and understanding. This media takes advantage of the characteristics that provide cost efficiency and immediate feedback on test takers' performance. In particular, web-based language testing to assess English proficiency, such as University of Cambridge ESOL Examinations[1] and Test of English for International Communication (TOEIC®)[2], has been utilized by many schools and companies to assess the English proficiency of students or employees. In spite of the increasing popularity and influence of CBTs and WBTs, not enough attention has been given to the visual designs of the interface in relation with test takers' performance. In particular, not much research has been done to investigate whether the background colors of CBTs or WBTs affect test takers' scores. Many English CBTs and WBTs use black text against a white background, being resembled to conventional paper-based tests. This text-background combination has rarely been evaluated in terms of test takers' performance.

2 Previous Studies

This section describes two previous studies of the background color effects on the scores of a WBT and a CBT in relation with this study. One is a study by Yamazaki, which evaluated the effects of eight background colors on the performance of WBT takers [6]. The other is a preliminary experiment by Yamazaki and Eto, which investigated if the background colors of a CBT can affect the brain activity of a test taker.

[1] See at `http://theenglishacademy.ie/cambridge-exams`
[2] See at `http://www.ets.org/toeic`

2.1 Test Taker's Performance Experiment

Yamazaki previously examined whether the background colors of a Web-based test has significant influence on the scores of test takers [6]. Eight combinations of black text and a background color were chosen to see if test takers' scores differed, depending on the characteristics of a background color. The characteristics of the background colors used in the study are listed in Table 1. The table summarizes their hexadecimal color codes, brightness and color differences between the text and background, and luminance ratios between the text and the two background colors. Formulas suggested by the World Wide Web Consortium[3] were used to calculate these values. The maximum values of brightness, color difference and the luminance ratio are 255, 765 and 1 respectively. Two hundred and forty six Japanese college students participated in the experiment and they were divided into eight groups. A WBT with 40 questions that resembled grammar questions in the TOEIC® Test was used and the subjects answered the same questions displayed on a background of eight different colors.

The results of the experiment in Yamazaki's studies demonstrate that the background color of a computer-based test can affect the performance of a test-taker. Table 2 shows the average test scores of the subjects in a background color groups. The results suggest that blue colors may be better for the background color of a WBT when question sentences are displayed in black. White, yellow and light yellow have similar characteristics when the text color is black, as shown in Table 1. They have higher brightness differences and luminance ratios to black. Since the subjects who took the test with these background colors tended to scored low in Yamazaki's experiments as shown in Table 2, a combination of black text and a background color with high luminance and brightness is considered not preferable for WBTs [6].

Table 1. Brightness difference, color difference and luminance ratio between each background color and black text color

Background color	Hexadecimal color code	Brightness difference	Color difference	Luminance ratio	Luminance
Light Blue	#C0C0FF	199	639	12.225	0.56126
Blue	#0000FF	29	255	2.44	0.00722
Pink	#FFC0C0	211	639	13.553	0.62765
Light Green	#C0FFC0	229	639	18.306	0.86532
Red	#FF0000	76	255	5.252	0.2126
Light Yellow	#FFFFC0	284	702	20.317	0.96586
White	#FFFFFF	255	765	21	1
Yellow	#FFFF00	226	510	19.55	0.927

The results also indicate that the background color of a WBT can make a difference to the concentration level of a test taker, and that primary colors with

[3] See `http://www.w3.org/TR/AERT`

Table 2. Average test scores, concentration and tiredness levels indicated by the subjects: primary colors with low luminance ratios tend to cause test takers to lose their concentration [6]

Background color	# of subjects	Average test score	"Able to concentrate" very well and well (%)	"Felt very tired" and "Felt tired" (%)
Light Blue	30	129.00	56.81	56.82
Blue	28	125.71	32.14	50.00
Pink	37	124.86	37.93	53.44
Light Green	36	123.33	31.15	54.09
Red	25	122.00	20.83	70.84
Light Yellow	34	121.41	42.50	64.11
White	31	120.97	41.94	61.29
Yellow	23	109.13	36.79	60.87

low luminance ratios, such as blue and red, cause test takers to lose their concentration. In addition, the percentages of the subjects who answered that they had felt tired were a little higher for the high-luminance background colors, such as yellow and white, than for the blue and light blue backgrounds, whose luminance rations are relatively low. On the other hand, the results from the study suggest that the color preference of the user is not an importance factor in choosing the best background color for a CBT, although Hall and Hanna demonstrated that preferred color would lead to higher rating of behavior intention [1]. The results of concentration and tiredness levels indicated by the subjects are also summarized in Table 2.

In Yamazaki's experiments, the levels of fatigue and difficulty that test takers reported after taking the tests did not correlate with their test performance. Also, the blue colors resulted in the two highest test score averages in the study, but their color characteristics are significantly different in terms of luminance and brightness [6]. Therefore, physiological factors or neurological factors associated with color characteristics were suspected to explain the differences among the test score averages for these background colors.

2.2 Preliminary Observation of CBT Test Taker's Brain Functions

Yamazaki and Eto conducted a preliminary experiment to see how a background color can affect the brain functions of CBT takers by recording relative changes in Hemoblobin (Hb) concentrations of the brains by using Near-Infrared Spectroscopy (NIRS) with 16 channel points [7]. The channels covered the frontal area of prefrontal cortex. In the experiment, seven male subjects in their twenties took computer-based English grammar tests with different background colors with black text. Among the background colors, average scores for the tests taken with the light blue and blue backgrounds were significantly higher than that for the white background. This result coincided with the finding of Yamazaki's study

wherein the blue backgrounds resulted in higher average scores than those with the white background, even though the grammar tests were different from ones used in Yamazaki's study [6].

The total Hb concentration changes in subjects' brain recorded during the experiment were mapped onto two-dimensional images of the brain. The images resulting from the subjects taking the CBT with the blue background showed that the anterior part of the frontopolar region in the frontal cortex, which roughly corresponds to Brodmann Areas 10 (BA 10), tended to experience higher levels of change in Hb concentration. These images suggested that the frontal part of the prefrontal cortex had been more highly activated than other brain areas while the subjects were taking the test with the blue background. On the other hand, high levels of change in Hb concentrations were observed in broader regions of the frontal cortex including Brodmann Areas 8 (BA 8) and 46 (BA 46), while the test takers were responding to the test with white background.

BA 10 is known to be one of the areas associated with cognitive tasks, and present studies suggest that this part of the brain involves strategic processing for memory retrieval and executive function [8,9]. Regions in BA 8 have been found to play an important role in the control of eye movements [10]. The images obtained in the previous CBT study indicate that the brain parts associated with memory and cognitive tasks were more highly activated while the subjects were taking the tests with blue backgrounds. By contrast, brain areas related to eye movements exhibited higher levels of change in Hb concentrations when the subjects were taking the test with the white background. These results suggest that reading black text on white background may encourage a CBT test-taker to concentrate more on dealing with visual input from the screen rather than with the syntax or semantics features of English exam questions. However, the activation of brain areas associated with linguistic tasks, such as Broca's and Wernicke's areas, was not observed due to the limitation of the number of NIRS channels.

3 Purpose of This Study

In this study, the authors investigated whether a background color can affect the linguistic functions of WBT test taker's brain and the results obtained in Yamazaki's studies can be explained in association with the activation of linguistic regions in the brain. We examined the activities of WBT test takers' brains while they were taking a Web-based English test with white and blue screen background colors. In order to identify which part of WBT test taker's brain is activated while he is taking a WBT with one of the background colors, we measured relative changes in Hb concentrations in the brain of the test taker by using a NIRS system with more channels than the one used for the previous preliminary study [7], in order to observe the activation of brain areas associated with linguistic tasks.

4 Experiment

This section describes two WBTs developed for this study, the experimental method and the subjects who participated in the experiment. The experiment was conducted to identify if there is a difference in subjects' WBT scores in relation with the backgrounds. In addition, the brain activity of each subject was observed in the experiment while the subject was taking the tests.

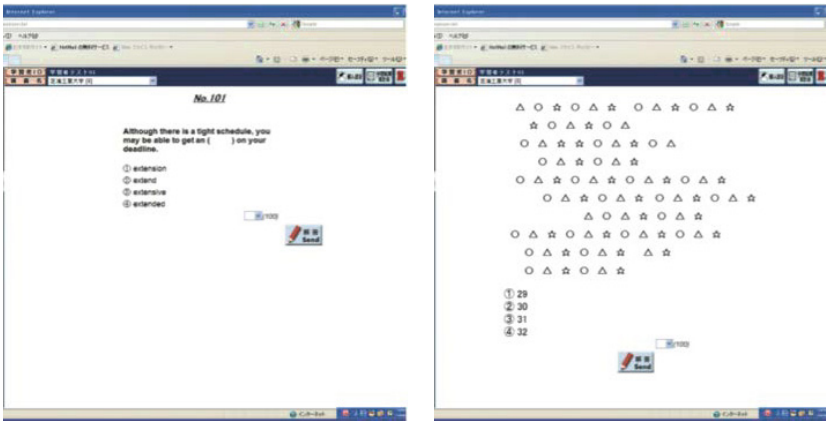
4.1 Web-Based Test

In this study, the authors conducted an experiment with two sets of Web-based tests, which consisted of English grammar questions and non-linguistic questions as rest tasks. White background color was used for one set and blue background color was chosen for the other set. The text color of all questions was black for both sets. Both test sets had 15 English grammar questions that were very similar to questions in Part V of a standardized English examination called the TOEIC® Test. The TOEIC® Test is a multiple choice exam that consists of two parts with 100 questions each: the Listening Section and the Reading Section. The Reading Section is divided into three parts: Incomplete Sentences (Part V of the test), Text Completion (Part VI of the test) and Reading Comprehension (Part VII). The Incomplete Sentences part has sentence completion questions and each question contains a sentence with a blank and four choices for the blank. The web-based English tests constructed for the experiment had 15 incomplete-sentence questions of the same type and a similar difficulty level as Part V of the TOEIC® Test. The tests were presented to the test takers and a response for each question was indicated by selecting one of the numbers 1 through 4 for the answer choices with the pull-down menu. An example of the test pages with the white background is shown in Figure 1.

In both sets of the WBTs, a non-linguistic task was presented at the beginning of the test and after every three English grammar questions as a rest task. All the rest task pages were designed in the same way in which circles, stars and triangles were randomly drawn on each page. A test taker was directed to count the number of circles on the page and select a choice corresponding to the number of circles from the pull down menu. Every rest task page had a different arrangement of circles and the numbers of circles on a rest task page varied from 29 and 32. An example of the circle-counting task pages with the white background is shown in Figure 1.

4.2 Method

A total of 24 subjects participated in this experiment and took the WBTs with the two background colors. The subjects were university students and graduate school students in their twenties, and their first language was Japanese. Seventeen of them were male (70.8%) and seven of them were female (29.2%). They were all right-handed and none of them was reported to have a color vision deficiency at the time of the experiment. They had all taken a paper-based TOEIC®



(a) English grammar question (b) Circle-counting test

Fig. 1. Example pages of the web-based English and circle-counting tests used in the experiment with the white background

Bridge Test or TOEIC® Test prior to the experiment. Their English proficiency levels ranged from pre-intermediate to upper-intermediate (their TOEIC® Test score ranged from 370 to 695). The subjects were given instructions for taking the WBT prior to undertaking the test. All the subjects took the tests with black characters and background in blue and white. Test scores obtained from the subjects were analyzed to see if there were any differences between the test sets with the two background colors.

In the experiment, we used a near-infrared spectroscopy system developed by Hitachi Medical Corporation (EGT 4000) to observe and record relative changes in hemoglobin concentration ratios in the brain of every subject, while the subject was taking the English tests and performing the rest tasks. Figure 2 shows a subject wearing a headgear with 24 probes and another subject taking the English WBT with the white background while wearing a headgear for applying NIRS probes to his forehead. The headgear included 52 optical source-detector channels to monitor relative changes in hemoglobin concentration in subject's brain. The channels covered the frontal area of prefrontal cortex and Broca's Area, which roughly corresponds to Brodmann Areas (BA) 8, 9, 10, 44, 45 and 46. In order to assess the activation of the brain functions associated with these areas, the blood hemoglobin concentrations of each subject were observed from the beginning of the tests until the subject complete the last grammar question, and relative changes in oxy-hemoglobin, deoxy-hemoglobin and total hemoglobin (oxy-Hb, deoxy-Hb, total Hb) concentrations from the 52 channel points were simultaneously measured and recorded for each subject.

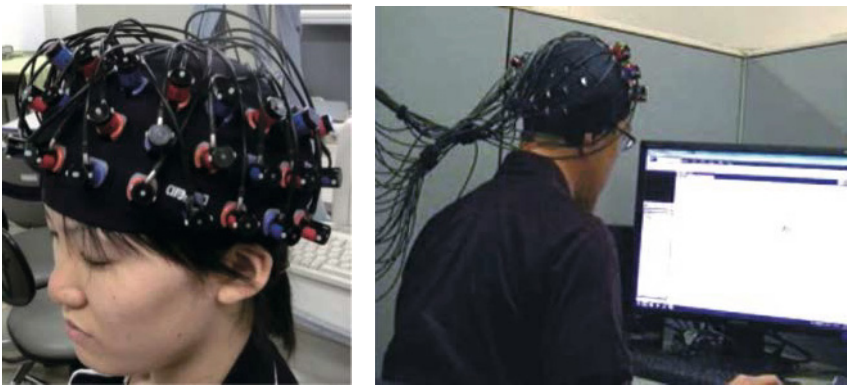


Fig. 2. Photographs of a subject wearing NIRS probes (right) and another subject taking the web-based English test with optical topography probes applied to his forehead (left)

5 Results

The authors obtained the scores of the subjects for the English grammar and circle-counting tests to examine if there was a difference in their test performance in relation to background colors. In addition, the brain activity of each subject was measured by means of NIRS while the subject was taking the tests with blue and white backgrounds. In this section, the test results and optical topography images of subjects brain activity are shown, and the analyses of the test scores and NIRS measurements are presented.

5.1 Test Taker's Performance Results

The average percentages of English and circle-counting questions answered correctly for each background color were calculated. The average percentage for the English grammar test taken with the blue background was significantly higher than that for the white background. The average percentage of correct answers for the non-linguistic tasks with the blue background was also higher than that for the white background. These results coincided with the findings of Yamazaki's study [6] and the study by Yamazaki and Eto [7], wherein the blue backgrounds resulted in higher average scores than those with the white background. The average percentages of correct answers for both the English and circle-counting questions are also summarized in Table 3. A t-test on the number of correct answers by all subjects showed that there was a significant difference between the scores for the white and blue backgrounds at $p < 0.001$ level for the circle-counting tasks. Another t-test also found a difference in the scores of English grammar tests between the two backgrounds at $p < 0.05$ level. The results of the t-tests are listed in Table 4.

The data was further analyzed to see if there is a difference between male and female subjects in terms of their test performance. For the English tests, the

Table 3. Average percentages of correct answers for the web-based English tests and the circle-counting tasks with the white and blue background colors

Background color	# of subjects	Average percentage of English questions answered correctly	Average percentage of circle-counting questions answered correctly
Blue	24	59.17%	95.00%
White	24	49.72%	78.33%

Table 4. Results of T-tests on the numbers of correct answers for the web-based English tests and the circle-counting tasks between the white and blue background colors. ($\alpha = 0.05$)

	English tests (full score = 15)	Circle-counting (full score = 5)
White background Average score	7.46	3.92
Blue background Average score	8.88	4.75
The number of subjects	24	24
t	−2.41	−3.97
df	45	33
p	0.020	1.84E-4

average test scores of both female and male subjects were higher for the blue background. On average, the female subjects performed better than the male subjects for both the English grammar tests and the circle-counting tests. The average score of the female subjects for the circle-counting tests was the same for the blue and white backgrounds. On the other hand, the scores obtained from the male subjects demonstrated that they had performed much better when they took the English and non-linguistic tests on the blue background, compared to when they took the tests with white background. The test performance results by sex are shown in Table 5.

Table 5. Average percentages of correct answers for the web-based English tests and the circle-counting tasks with the white and blue background colors by subject's sex

Sex	Background color	# of subjects	Average percentage English questions answered correctly	Average percentage circle-counting answered correctly
Male	Blue	17	58.43%	94.11%
	White		47.84%	58.43%
Female	Blue	7	60.95%	94.14%
	White		54.29%	94.14%

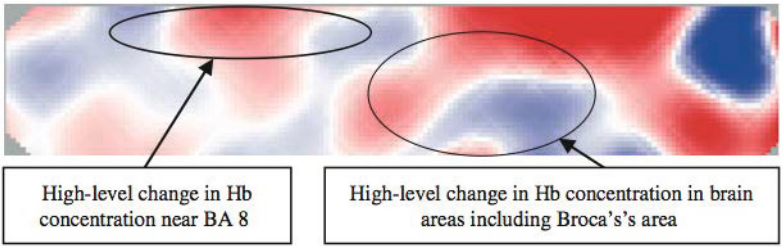


Fig. 3-a White background

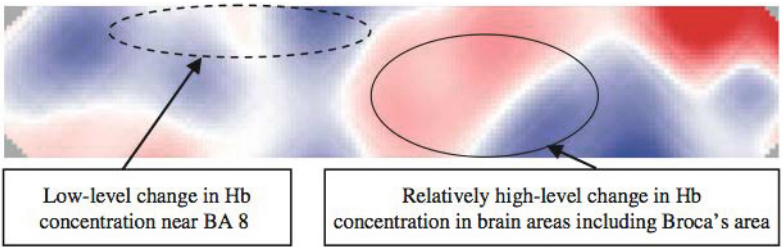


Fig. 3-b Blue background

Fig. 3. NIRS topography images showing relative changes in total Hb concentrations in the brain of Subject No.7 while taking the Web-based English test with white (Fig. 3-a) and blue (Fig. 3-b) backgrounds

5.2 Measurement Results of WBT Test Taker's Brain Functions

The total Hb concentration changes in subjects' brain recorded during the experiment were mapped onto optical topography images of the brain. The images were also analyzed in relation to the functions of brain regions. Figures 3 and 4 show the examples of Hb concentration images created from the data measured by NIRS while the subjects were taking the WBT tests with the blue and white backgrounds. The images in Figures 3 and 4 show Hb concentration changes in the brain of a subject with lower English proficiency (Subject No.7). Figures 3-a and 3-b are the topography images of the NIRS data taken while the subject was taking the Web-based English test with white and blue backgrounds respectively. Figures 4-a and 4-b show the topography images of Hb concentration changes in the brain of the same subject while the subject was counting the number of circles on white and blue backgrounds respectively.

As Figures 3-b and 3-b demonstrate, higher levels of change in Hb concentrations were observed in brain areas including Broca's area, whose activation relates to syntactic and semantic processing [8], while the subject was solving English grammar questions with both blue and white background colors. The comparison between the two figures showed that the broader regions of the frontal cortex had higher levels of Hb concentration changes while the subject was responding to the test with a white background. These activated regions include BA8 and the function of BA8 is known to be associated with eye

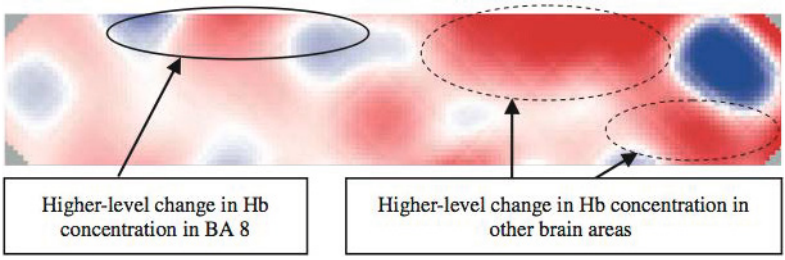


Fig. 4-a White background

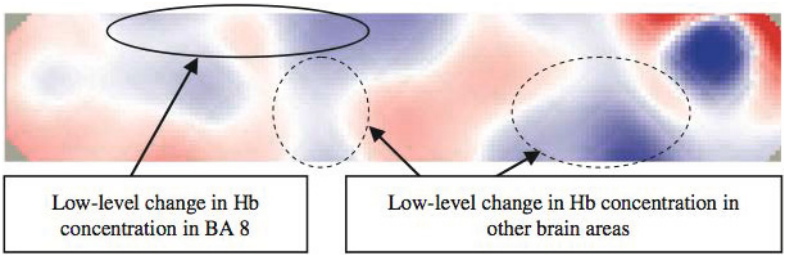


Fig. 4-b Blue background

Fig. 4. NIRS topography images showing relative changes in total Hb concentrations in the brain of a subject while taking circle-counting tests with white (Fig. 4-a) and blue (Fig. 4-b) backgrounds

movements. On the other hand, the images resulting from the subject taking the WBT with the blue background showed that the area near BA8 tended to experience lower levels of change in Hb concentration. These images suggested that the frontal eye field had been more highly activated than other brain areas while the subjects were taking the test with the white background. The same tendency was observed in Hb concentration changes in the brain of the same subject while the subject was performing the circle-counting tasks as seen in the two Hb concentration images of Figure 4. In addition, topographical images of brain activities obtained from other subjects, including ones with higher English proficiency, also showed a similar tendency for regions near BA8.

6 Discussion and Conclusion

In this study, the authors examined if the performance of WBT test takers differs when an English grammar test and a non-linguistic test are given on a white background and a blue background. The results of the experiment showed that the subjects had performed significantly better when WBT questions written in black color were given on a blue screen background than when they answered the questions on a white background. This result coincides with findings in the previous studies [6,7]. In addition, by using near-infrared spectroscopy, relative changes of blood hemoglobin concentrations in the brain were observed to see

if the brain function activities of WBT test takers can be affected by the background colors of the tests, in particular the areas associated with eye movements and linguistic tasks.

The results of the experiment in this study demonstrate that the background color of a WBT can affect test taker's performance and hemoglobin concentrations in the brain of a test taker. They suggest that blue colors may be better than white for the background of a WBT when question sentences are displayed using black text in terms of activating brain functions. The results also suggest that the background color can make a difference in brain activities related to linguistic tasks, indicating that a white background provokes states that require more eye movements and cause test takers to divert concentration away from language processing. The NIRS images of relative Hb concentration changes in subjects' brains also indicate that their brain areas related to eye movements tended to be more active while performing tasks on a computer screen with the white background than while doing the same task with a blue background. However, a recent study of BA8 has shown that the activation of the brain area is also elicited with decision uncertainty [11]. Therefore, the activations of BA8 observed in this study need to be carefully analyzed in terms of the brain functions in relation to specific tasks.

In this study, the authors could not analyze if there is a significant difference between white and blue backgrounds in terms of Hb concentration changes in Broca's area of the brain, which is associated with grammar and vocabulary processing [8]. We did not observe blood hemoglobin concentrations in the brain for Web-based English tests with a light blue background, which marked the high average score for Web-based and computer-based English grammar tests in the previous studies [6,7]. In the next study, we plan to conduct experiments to see if Hb concentrations in Broca's Area can be affected by differences in the case that the background colors of WBTs are white, blue and light blue. In addition, it is important to increase the number of subjects in the next study in order to obtain a more general sampling. In particular, the number of female subjects should be increased to statistically confirm if there are differences between male and female subjects. It is also necessary to see differences in brain activities according to English proficiency levels since brain activities are known to vary among individuals and also to be associated with the level of language proficiency [12].

Acknowledgement. The authors would like to thank Dr. Eiji Kamioka, Dr. Ryota Horie and Dr. Kenji Muto at Shibaura Institute of Technology for their assistance and advice for this study. We also would like to extend our gratitude to Chieru Co., Ltd. for its assistance in developing a set of WBTs with different background colors and to the students at Shibaura Institute of Technology and Tokyo University of Marine Science and Technology who participated in this study. This work was supported by JSPS KAKENHI Grant Number 23501171.

References

1. Hall, R., Hanna, P.: The Impact of Web Page Text-Background Color Combinations on Readability, Retention, Aesthetics, and Behavioral Intention Citation. Behaviour & Information Technology 23(3), 183–195 (2004)
2. Lin, C.: Effects of contrast ratio and text color on visual performance with TFL-LCD. International Journal of Industrial Ergonomics 31(2), 65–72 (2003)
3. Notomi, K., Hiramatsu, A., Saito, K., Saito, M.: A fundamental study on visibility of background and character colors at the web browsing. Biomedical Soft Computing and Human Sciences 9(1), 17–25 (2003)
4. Nishiuchi, N., Yamanaka, K., Beppu, K.: A study of visibility evaluation for the combination of character color and background color on a web page. In: The International Conference on Secure System Integration and Reliability Improvement, pp. 191–192 (July 2008)
5. Mehta, R., Zhu, R.: Blue or Red? Exploring the Effect of Color on Cognitive Task Performances. Science 323(5918), 1226–1229 (2009)
6. Yamazaki, A.K.: An Analysis of background-color effects on the scores of a computer-based English test. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part II. LNCS, vol. 6277, pp. 630–636. Springer, Heidelberg (2010)
7. Yamazaki, A.K., Eto, K.: A Preliminary Examination of Background-Color Effects on the Scores of Computer-Based English Grammar Tests Using Near-Infrared Spectroscopy. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part III. LNCS, vol. 6883, pp. 31–39. Springer, Heidelberg (2011)
8. Sakai, K.: Language acquisition and brain development. Science 310(5749), 815–819 (2005)
9. Duncan, J., Phillips, L., McLeod, P. (eds.): Measuring the Mind: Speed, Control, and Age. Oxford University Press (2005)
10. Hu, X.S., Hong, K.S., Ge, S.S., Jeong, M.Y.: Kalman estimator- and general linear model-based on-line brain activation mapping by near-infrared spectroscopy. BioMedical Engineering OnLine 9(82) (2010)
11. Volz, K.G., Schubotz, R.I., von Cramon, D.Y.: Variants of uncertainty in decision-making and their neural correlates. Brain Research Bulletin 67(5), 403–412 (2005)
12. Sakai, K.L., Nauchi, A., Tatsuno, Y., Hirano, K., Muraish, Y., Kimura, M., Bostwick, M., Yusa, N.: Distinct roles of left inferior frontal regions that explain individual differences in second language acquisition. Human Brain Mapping 30, 2440–2452 (2009)

Loss Aversion Behavior Utterances Extraction in Internet with Expected Utility

Suzuki Nobuo[1], Fujita Yoshikatsu[2], and Tsuda Kazuhiko[3]

[1] KDDI Corporation, Iidabashi 3-10-10, Chiyoda, Tokyo 102-8460, Japan
nu-suzuki@kddi.com
[2] Digital and Network Technology Development Center, Panasonic Corporation, Saedo-cho 600, Tsuzuki-ku, Yokohama City 224-8539, Japan
fujita.yoshikatsu@jp.panasonic.com
[3] Graduate School of Buisiness Sciences, University of Tsukuba, Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
tsuda@gssm.otuka.tsukuba.ac.jp

Abstract. Recent research advances in new on knowledge of the human behavior has been stimulated by mining sensor data and huge text on the Internet. Behavioral modification research is ongoing in a social science research area. These make users to change their behavior and realize the better society which is represented by prohibition of smoking and a route guidance of a GPS navigation system. This research aims to construct the model of the behavioral modification using information technology. This paper examines the technique of extracting the knowledge about the behavioral modification from the online forums on the Internet in which user's problems and opinions often appear. Specifically, the extraction of the utterances expressing the loss aversion is carried out from a series of online forum text sentences. The loss aversion means that men have a strong tendency to select to avoid a loss rather than get a profit. Therefore, it is possible to build a system which presents users a series of actions of maximizing a profit by investigating what kind of loss aversion actions are performed. This paper shows examples of the loss aversion utterances in online forum text sentences and tries to classify loss aversion utterances on the basis of the expected utility index that is computed only from the text sentences.

Keywords: Loss Aversion, Expected Utility, Online Forums, Information Extraction.

1 Introduction

Many studies of the human behavior understanding are forwarding with sensors and huge data on the Internet in recent years [1,2]. On the other hand, the behavioral modification research is ongoing. These make users to change their behavior and realize more acceptable approach to certain activities represent prohibition smoking or the driving with mobile phones and car navigation system. This study aims at construct a model of behavioral modification using information technology. It examines the technique of extracting the knowledge about behavioral modification from the online forums on the Internet in which user's problems and opinions often appear. Behavioral Economics is

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 16–24, 2013.

known as an area of research regarding to human's behavioral modification [3]. Heuristics, Discount Utility Theory, Expected Utility Theory and Prospect Theory are studied in Behavioral Economics. Expected Utility Theory and Prospect Theory in this area often express the tendency of the human selecting behavior under the uncertain situation and show the view of the risk when avoiding a loss. Behavioral economics is one of the economics aiming at studying that a financial expert like typical economics is not premised, but the experiment by an actual human being and its observation are important, and what will be happened when man chooses and acts something. Heuristics refers to experience-based techniques for problem solving, learning, and discovery. This technique is used within decision-making processes and can suufer Bias, due to divergent priority collect and evaluate data. For example, people have a tendency to choose information that they remembers first. Discount Utility Theory means that making the present discount rate into 20% does not change both of value when giving the same satisfaction with present 120 dollars and 100 dollars at one year after. Expected Utility Theory shows that the expected value calculated from probability and the choice actually chosen including risks are not in agreement. Prospect Theory means that people feel the increase in the amount of a loss is earlier than the increase in the amount of money by a profit. For example, they show that the human has a tendency to choose actions avoiding a loss rather than obtaining a profit [4]. Therefore, the construction of the behavioral modification model based on the loss aversion actions can be expected by collecting the loss aversion utterances and analyzing characters for them. For example, we can consider a system that predicts human's loss aversion actions and presents choices of the right profit maximization. The term utterance means sentences which one person speaks in an online forum.

This paper proposes the technique of automatically extracting the utterances expressing a loss aversion from a series of online forum text sentences from the online forums the problem appear from users and that opinions in order to collect the utterance sentences showing such a loss aversion. First, this paper shows examples of the loss aversion utterances in online forum text sentences, next explains the Expected Utility Score which is an index of the expected utility computed with text sentences. Then, it describes the technique of extracting the loss aversion utterances on the basis of this score, and the contents of the evaluation experiment. It also shows that this method can classify the threads that include loss aversion utterances by 89.4%, and extract the loss aversion utterances by 68.6% performance as a result of the evaluation.

2 Related Researches

First, researches that extract information from user utterances on Internet sites such as Web sites. Researches of classification utterances in those works are automatically pros and cons decision of opinions [5], detection of improper utterances in Web sites [6] and evaluation utterances of conferences [7]. One of such researches is IBM WebFountain which is most biggest works and analyzes 100 TB text data using 1 million servers [8]. Syntactic Analysis and machine learning technologies are used to applied in such research areas. Researches that extract behaivors from such text information on the Internet also are going on these days.

The example of extracting such human behaviors from text information is Ono's work [9]. He considered an extraction failure stories method from text information. It constructed the stories with actions, results, and reasons. They proposed how to decide each part. In particular, indefinite causes are limitted using an original score which used n-gram in the judgment of causes. N-gram is a set of the fixed number of the row of the character started from a certain character string. The character of the original character string can be acquired by analyzing kinds and frequencies of the set. It got good precision that was 0.679. Although this failure story is different from our loss aversion, it is a method that extracts human behavior from text information. Next, the method uses the connecting probability to extract the loss aversion utterances. Ishizaka et al. also sdudied ectracting text expressions with these connecting probabilities [10]. Ishizaka assumed that it is easy to connect with failure expressions so that connection probability is high, and controlled the reliability of extraction by setting up a threshold value. This research also improves extraction accuracy by incorporating the connecting probability value which it is easy to connect with loss aversion expressions. It also extracts the utterances which maximize an utility by using the expected utility theory in order to extract loss evasion utterances. Akiba et al. studied to use such expected utility in text processing [11]. They have proposed a method of choosing a suitable reply group by comparing the expected value of a utility function when a certain respondent group is extracted in order to obtain common responses from online forums.

Syntax Analysis and Machine Learning are required prior learning works and much time for it as known in those reserch works. This research uses the characteristics of actions efficiently using connection probabilities based on n-gram technology when it extracts utterances show loss aversion actions.

3 The Loss Aversion Utterances in Online Forums

In recent years, the online forums are frequently used as means to solve various problems mutually. Such sites are used widely from dealing with the problems of individuals such as 'Yahoo! Wisdom' to problems specialized in the specific company such as 'Q and A plus'. Online forums express explicitly a dissatisfaction and a demand, because they has the form that a questioner presents a question and a respondent replies to the question explicitly. Therefore, many loss aversion utterances we try to extract will be appeared in those sites. Table 1 shows the actual examples of utterances which have a loss aversion. The respondent in the 1st utterance recommends an action that inserts MicroSD card again in order to avoid the loss which the user has never used it. In the 2nd utterance, in order to avoid a loss that the power supply of a mobile phone will fall, the respondent recommends the action which reduces the capacity of a data folder. Such utterances are called 'Loss Aversion Utterances' in this chapter. The definition of a loss aversion utterance is as follows.

1. Loss is included clearly: For example, the 2nd utterance in Table 1 obviously shows the loss that a screen becomes black and returns to the beginning screen. Although the loss of behavioral economics treats an economical loss, the loss by this research points out a phenomenon of the disadvantageous for a questioner or a speaker.

2. Avoiding a loss policy or candidates of them are included: For example, the 2nd utterance of Table 1 shows that a phenomenon is improved by lowering the capacity of a data folder.

Table 1. The examples of the loss aversion utterances in online forums

Questions	Answers	Explanation
'A microSD card access error' is displayed when listening to music by LISMO or while tampering with the data folder. What is this caused by?	When data has broken, you need to put it again, since it isnt restorable.	In order to avoid the loss that it cannot use again, it puts in again.
My screen becomes black suddenly and key operation completely becomes impossible in these days. After a while, it returns to the same screen as the time of starting. How are other W41CA users?	If your data folder is full, such a phenomenon may occur. You need to reduce a data folder capacity at most 70%.	In order to avoid the loss of power supply, the capacity of a data folder is reduced.

4 Extraction of the Loss Aversion Utterances by Using Expected Utility

This clause explains the method of separation and extraction of the loss aversion utterances from other utterances shown in the preceding clause. First, it focuses a related structure about the question and response of the loss aversion utterances, and proposes about the index which shows them by definition. Next, it describes the extraction method of them by using that index.

4.1 The Trouble Sentence Probability and the Expected Value by Using N-gram

Since a loss aversion utterance is accompanied by a decision-making situation in many cases, it basically appears in a reply sentence instead of a questioning one. A certain issue is highlighted in a question statement, and a loss aversion expression tends to appear in the subject of the solution to the problem. This chapter calls these statements in which the issue is included 'Trouble Sentence'. Therefore, this technique computes the probability which shows the trouble sentences at first. Specifically, it prepares the corpus to learn the trouble sentences and the word sequence dictionary expressing troubles. Table 2 shows the examples of the word sequences expressing the troubles. This dictionary is called 'Trouble Sentence Dictionary'. This dictionary has the characteristic expressions extracted from the loss aversion utterances previously collected. Next, it

constructs an n-gram model by referring to connecting morphemes with the morphemes of the words registered into the word sequence dictionary. A morpheme is the smallest semantic unit in a language. Morphological analysis is able to process segmenting a sentence into a row of morphemes [12]. This n-gram model is called 'Trouble Sentence N-gram Model' and use bi-gram at this time. The connection intends for two morphemes connected before and after the word sequence registered into the trouble sentence dictionary. This trouble sentence n-gram model can calculate the connecting probability for each morphological sequence. This probability is called 'Trouble Sentence Probability'. Here, the average of the trouble sentence probabilities is taken if two or more trouble sentences exist.

Table 2. Trouble sentence dictionary

Japanese Trouble Sentense	
Nakunari/Masu Wakari/Mase/N Deki/Mase/N Te/Shimai/Mashi/Ta Te/Tamari/Mase/N ...	Note: '/' is a separator of morphemes.

Next, this method computes the expected value with the word sequences expressing the uncertainty which characteristically appears in the loss aversion utterances of the reply sentence. For example, it uses the negative expressions such as 'It is dangerous' and 'It costs a lot', and the frequency of the word sequence showing the uncertainty such as 'Darou' and 'Youda' in Japanese as the expected value. Table 3 shows the example of the negative expressions and the uncertainty expressions. Such expressions express how uncertain expectation of the loss aversion for the trouble. We call a set of such word sequences 'Uncertainty Dictionary'.

Table 3. Uncertainty Dictionary

Japanese Uncertainty Word Sequence	
O/Susume/Shi/Masu Kanou/Sei/Ga/Ari/Masu Kiken/Desu Kosuto/Ga/Kakaru Sake/Taka/Ta ...	Note: '/' is a separator of morphemes.

4.2 Extracting the Loss Aversion Utterances

The expected utility score S is defined by Eq. 1 using the expected value E and the trouble sentence probability P mentioned above. This method extracts the threads including these utterances that the expected utility score S exceeds the threshold C as the loss aversion utterances. C is a value computed experimentally by an experiment.

$$S(x_i) = \sum E(w_{ij}) \frac{\sum P(t_{ik})}{N} > C \tag{1}$$

For instance, when there are five utterances in thread A, the utterances from x_1 to x_5 are obtained. When ten word sequences are contained in this utterance x_1, $w_{1,1}$ to $w_{1,10}$ show the word sequence. The expected values $E(w_{1,1})$ to $E(w_{1,10})$ are calculated using this each word sequence, and it calculates total value as $\sum E(w_{ij})$. Next, when utterance x1 includes three trouble sentences, it calculates the trouble probabilities as $P(t_{1,1})$ to $P(t_{1,3})$, and averages $\sum P(t_{i,k})$ by number of trouble sentences $N = 3$.

5 Evaluation Experiment

Figure 1 shows the procedures of this evaluation experiment. First, the text information of questions and responses are collected from online forums, and loss aversion utterances are manually extracted as a learning. The extracted utterances are analyzed by a morphological analysis, then the trouble sentence dictionary and the uncertainty dictionary are created. Next, the text information of questions and responses are similarly acquired from online forum for an evaluation test. Loss aversion utterances are extracted by the morphological analysis of these data and calculating the expected utility scores. At the same time, the loss aversion utterances are judged manually and it compares with the previous loss evasion utterances. The extraction of the loss aversion utterances using the technique described above was experimented. The first 1,014 utterances were collected which included the loss aversion utterances from the online forums for a learning. The uncertainty word dictionary and the trouble sentence n-gram model were created based on this data. ChaSen was used as a morphological analysis tool [13]. As a result, the connecting probability in the trouble sentences were calculated as shown in Table 4, and the word sequences registered into the uncertainty dictionary accumulated 45 pieces of knowledge.

Next, 272 threads (3,193 sentences) were collected which were not considered the existence of the loss aversion utterances from the online forums for the evaluation without the learning purpose. The expected value and the trouble sentence probability for every thread were computed with this data. The threshold value C of the expected utility score that judges whether it is a loss aversion utterance was set to 0.3. The performance was evaluated by comparing with the loss aversion utterances extracted by this method and by human hands. This method finally decided that 35 threads corresponded to the loss aversion utterances among 274 threads, and 239 threads didn't. According to this result, the classification accuracy of the loss aversion utterances is 89.4%. The extracting accuracy of the loss aversion utterances also was 68.6% by the Eq. 2, the recall was 96.0% by the Eq. 3, and F-measure was 0.80 by the Eq. 4. Here, Precision shows the rate correctly judged in the utterances judged by this system to be Loss Aversion

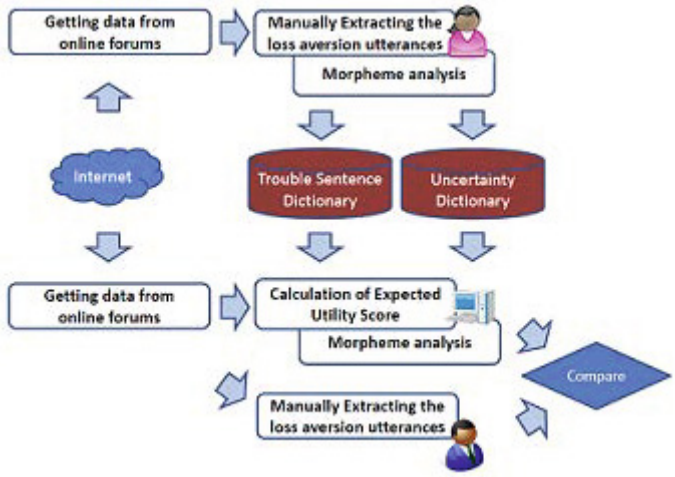


Fig. 1. The Procedures of The Evaluation Experiment

Table 4. The connecting probability in the evaluated data set

Japanese word sequences	The connecting probability
Deki+Nakunari/Masu	0.954
Nakunari/nasu+Ne	0.090
Ha+Wakari/Mase/N	0.384
Wakari/Mase/N+Ga	0.384
Kitai+Deki/Mase/N	0.028
...	...

Note: '/' is a separator of morphemes and '+' means connecting with some words.

Utterance, and it serves as an accuracy rate. Recall shows the rate judged correctly with this method in the utterances manually judged to Loss Aversion Utterance. F-measure is a harmonic average of precision and recall, and if F value is high, it means good performance [14].

$$Precision = \frac{Number\ of\ Correct\ Threads}{Number\ of\ threads\ that\ this\ method\ decided\ to\ Loss\ Aversion\ Utterances} = 68.6\% \quad (2)$$

$$Recall = \frac{Number\ of\ Correct\ Threads}{Number\ of\ all\ threads\ that\ Loss\ Aversion\ Utterances\ were\ decided\ by\ human\ hands} = 96.0\% \quad (3)$$

$$F-measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = 0.80 \quad (4)$$

Figure 2 also shows the accuracy by changing the threshold value C, recall, and F-measure. We can realize $C = 0.6$ is an appropriate condition, because it is the best balance between the accuracy, the recall, and the F-measure according to this graph.

Next, the incorrect utterances are focused by this method in the utterances extracted as the loss aversion utterances by human hands. Table 5 shows the classification of 15

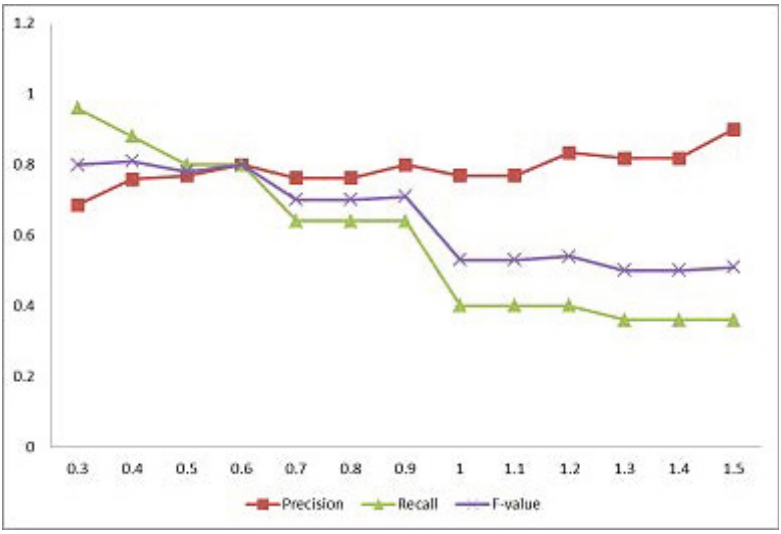


Fig. 2. The performance change according to the threshold

incorrect threads. The cause of most inaccurate solutions is that there is no solution avoiding a loss. Second most cause are not solutions and explanations of the cause of the problem. The fewest cause was shortage of the trouble sentence dictionary. The shortage of a trouble sentence dictionary can be improved easily by substantial dictionary in the future.

Table 5. The classification of the incorrect threads

Classification	Percentage
There is no solution for avoiding the loss.	66.7%
It explains the cause instead of the solution.	22.2%
The trouble sentence dictionary does not have appropriate sentences.	11.1%

6 Conclusion

This chapter proposes the method of extracting the utterances which shows loss aversion from the online forums to examine a human behavioral modification system. This method uses situations that there are the trouble sentences when the loss aversion utterances are appeared. It also use the expected utility based on an uncertain situation in the reply sentences. The good performance is observed that it classified the threads included the loss aversion utterances by 89.4% and extracted such utterances by 68.6% according to the result of the experiment. The extraction accuracy will be raised by increasing the learning data in our future work. Furthermore, the loss aversion utterances

were collected by using this method, and examine the behavioral modification system which can obtain the maximum profit.

References

1. Mizuno, M.: Preference Formation and Behavioral Inducement. In: The 25th Annual Conference of the Japanese Society for Artificial Intelligence, pp. 1–4 (2011)
2. Miltenberger, R.: Behavior Modification: Principles and Procedures. Wadsworth Pub. (2011)
3. Thaler, R.: The Winner's Curse: Paradoxes and Anomalies of Economic Life. Princeton University Press (1994)
4. Nick, W.: An Introduction to Behavioral Economics. Palgrave Macmillan (2007)
5. Matt, T., Bo, P., Lillian, L.: Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In: Proceedings of EMNLP, pp. 327–335 (2006)
6. Ichifuji, Y., Konno, S., Sone, H.: A Method to Monitor a BBS Using Feature Extraction of Text Data. In: Shimojo, S., Ichii, S., Ling, T.-W., Song, K.-H. (eds.) HSI 2005. LNCS, vol. 3597, pp. 349–352. Springer, Heidelberg (2005)
7. Dustin, H., Mari, O., Elizabeth, S.: Detection of agreement vs. disagreement in meetings: training with unlabeled data. In: Proceedings of the Conference of the NAACL on Human Language Technology, pp. 34–36 (2003)
8. Yi, J., Niblack, W.: Sentiment Mining in WebFountain. In: Proceedings of the 21st International Conference on Data Engineering, pp. 1073–1083 (2005)
9. Hiroki, O., Akira, U.: Automatic extraction of failure stories from Weblogs. In: The 26th Annual Conference of the Japanese Society for Artificial Intelligence, 4I1-R-9-4, pp. 1–4 (2012)
10. Tatsuya, I., Kazuhide, Y.: Abuse expressions extraction for 2 channel. In: The 16th Annual Conference of the Association for Natural Language Processing, pp. 178–181 (2010)
11. Tomoyoshi, A., Atsushi, F., Katsunobu, I.: Question Answering using Common Sense Knowledge latent in Corpora and Utility Maximization Principle. Study group of Natural language processing, pp. 131–138 (2004)
12. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson (2010)
13. Matsumoto, H.: A Morpheme Analysis System 'ChaSen'. Information Processing 41(11) (2000)
14. Ian, H., Eibe, F., Mark, A.: Data Mining: Practical Machine Learning Tools and Techniques, 3rd edn. The Morgan Kaufmann Series in Data Management Systems (2011)

Extracting Market Trends from the Cross Correlation between Stock Time Series

Mieko Tanaka-Yamawaki[1], X. Yang, T. Kido, and A. Yamamoto

[1] Department of Information and Knowledge Engineering
Graduate School of Engineering Tottori University
mieko@ike.tottori-u.ac.jp

Abstract. In this paper, the RMT-PCA is applied on daily-close stock prices of American Stocks in NYSE for 16 years from 1994 to 2009 to show the effectiveness and consistency of this method by analyzing the whole data of 16 years at once, as well as analyzing the cut data in various lengths between 2-8 years. The extracted trends are consistent to the actual history of the markets. The authors further analyze the intra-day stock prices of Tokyo Stock Market for 12 quarters extending from 2007 to 2009 and attempted to answer to the two remaining question of the RMT-PCA. The first issue is the number of principal components to examine, and the second issue is the number of eminent elements to examine out of the total N components of the chosen eigenvectors. While the second issue is still open, the authors have found for the first issue that only the second largest principal component is sufficient to examine, based on the comparison of this scenario and the use of the largest ten principal components. This paper argues on this point that the positive elements, and the negative elements, of the eigenvector components individually form collective modes of industrial sectors in the second eigenvector u_2, and those collective modes reveal themselves as trendy sectors of the market in that time period. The authors also discuss on the problem of setting the effective border between the noise and signals considering the artificial correlation created in the process of taking log-returns in analyzing the price time series.

Keywords: Quarterly trends in the stock market, RMT-PCA, Cross correlation matrix, Eigenvector components.

1 Introduction

Recently, there have been wide interests on the use of the RMT (Random Matrix Theory) in many fields of sciences [1–10]. In particular, the use of asymptotic formula of the eigenvalue spectrum of cross correlation matrix between independent time series of random numbers [11, 12], as a reference to the corresponding spectrum derived from a set of different stock price times series in order to extract principal components effectively in a simple way [13–16], has attracted much attention in the community of econo-physics [17, 18]. The main advantage of this method as a principal component analysis is its simplicity. While the standard PCA (Principal Component Analysis) gives out a way to find the largest PC (Principal Component) and subtract this component from the entire data, and apply the same procedure recursively on the remaining

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 25–38, 2013.

data one by one, the RMT-PCA (RMT-based principal component analysis) can process all the "non-random" components at once by subtracting the RMT formula from the eigenvalue spectrum of a cross correlation matrix. Plerau, et al. [13] was one of the first attempts to apply this technique on stock price time series. By using the daily close stock prices of NYSE/S&P500, they successfully extracted eminent stocks out of massive data of price time series.

However, this method suffers from two difficulties. One is the restriction on the data structure. The entire set of N×T data are needed for analysis, due to the fact that the basic quantity is the cross correlation matrix whose elements are the equal-time inner products between a pair of stocks.

Another difficulty is the restriction of the parameter size. Since the RMT formula is derived in the limit of N and T being infinity, a special care is needed to keep the ranges of the parameters in which the RMT formula is valid.

By using machine-generated random numbers, such as rand(), etc., the authors have tested the validity of the RMT formula in various range of N and T, and have clarified that $N \geqslant 300$, is the safe range unless T is not too close to N, and the validity decreases for smaller N, and the borderline is around $50 < N < 100$. Since the size of stocks dealt in the major markets exceeds 400, the applicability of RMT formula is justified.

Due to the restriction of the methodology to prepare the length of the time series, T, larger than the dimension of the correlation matrix, N, all the data extending to several years had to be combined into a single correlation matrix in [13–16], in which daily-close prices were used. Thus it was difficult to pin-point a short term trend or to compare trends of different time periods.

By employing intra-day (tick-wise) data containing all the transactions made every day, it has become possible to analyze one-year data to compare the result of different years. The authors carried out the same line of study used in [13, 14] by setting up the algorithm of RMT-PCA to be applied on intra-day equal-time price correlations. Based on this approach, the papers [19, 20] have shown that this handy methodology works well to extract the trend change of 4 year interval, from 1994 to 2002.

The authors have applied the same algorithm to a wider set of stock price data including daily-close prices of American stocks in the database of S&P500 for 16 years from 1994 to 2009, by cutting the 16 years into 2, 4 and 8 pieces and check the consistency and effectiveness of the proposed methodology in various data lengths. In this paper, the RMT-PCA is applied to a tickwise price data of Tokyo Stock Market from 2007 to 2009, in order to study quarterly trends of the market and attempt to clarify the remaining two technical problems of this algorithm.

There are still some technical problems remaining in the application of the RMT-PCA. One is the number of principal component to be analyzed. It is well known that the first principal component corresponding to the largest eigenvalue of the cross correlation matrix does not give out much information on the trendy sectors, since this mode is almost parallel to the major index of the market made of large-sized popular stocks thus extremely stable [13]. The next largest mode represented by the second eigenvector, u_2 is the major source of information the trendy sectors of that period can be extract from it. Based on the condition $\lambda_i > \lambda_+$, there are 11 to 20 principal components extracted from each data set. Whether u_2 is sufficient for redthe purpose that is to

determine the trendy sectors, or some of the remaining states are to be considered is the focus of question.

Another problem is how many elements are to be picked up in order to identify the trendy sectors from the total N dimensional eigenvector, such as u_2. In the previous work of the authors, a fixed number (say 5 or 10) of the largest elements are chosen from each of the positive and negative elements. This point is examined by comparing the use of the fixed number of elements and the fixed accumulative rate.

This paper is organized as follows. After introduction, the methodology of RMT-PCA is summarized in Section 2. The result of daily-close prices of American stocks in the database of S&P500 for 16 years from 1994 to 2009 is shown in Section 3. The result of tickwise price data of Tokyo Stock Market from 2007 to 2009 is given in Section 4, in order to study quarterly trends of the market and attempt to clarify the remaining two technical problems of this algorithm. Then Section 5 is devoted to discuss remaining problems of this methodology.

2 Eigenvalue Problem of Correlation Matrix for Stock Prices

The methodology of the RMT-PCA is outlined as folows. The first step is to prepare the price time series into an N×(T+1) matrix named S, whose i-th row contains the price time series of length T+1. This matrix S is converted into a matrix of log-return as follows

$$r(t) = \log(S(t+\Delta t)) - \log(S(t)) \tag{1}$$

Each string of time series is normalized by

$$x_i(t) = \frac{r_i(t) - \langle r_i \rangle}{\sigma_i} \ (i = 1, \ldots, N) \tag{2}$$

The correlation $C_{i,j}$ between two stocks, i and j, can be written as the inner product of the two log-profit time series, $x_i(t)$ and $x_j(t)$,

$$C_{i,j} = \frac{1}{T} \sum_{t=1}^{T} x_i(t) x_j(t) \tag{3}$$

Here the suffix i indicates the time series on the i-th member of the total N stocks.

The correlations defined in Eq.(3) makes a symmetric ($C_{i,j}$ = $C_{j,i}$), square matrixwhose diagonal elements are all equal to one ($C_{i,i}$) and off-diagonal elements are in general smaller than one ($|C_{i,j}| \leq 1$). As is well known, a real symmetric matrix C can be diagonalized by a similarity transformation $V^{-1}CV$ by an orthogonal matrix V satisfying $V^t = V^{-1}$, each column of which consists of the eigenvectors of C. Such that

$$C_{i,j} = \lambda_k V_k \ (k = 1, \ldots, N) \tag{4}$$

where the coefficient λ_k is the k-th eigenvalue and is the k-th eigenvector.

According to the random matrix theory (RMT, hereafter), the eigenvalue distribution spectrum of C made of random time series is given by the following formula [8, 9]

$$P_{RMT}(\lambda) = \frac{Q}{2\pi}\frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \text{ where } \lambda_\pm = (1 \pm Q^{-\frac{1}{2}})^2 \tag{5}$$

in the limit of $N \to \infty$, $T \to \infty$, $Q = T/N =$ const where T is the length of the time series and N is the total number of independent time series (i.e. the number of stocks considered). This means that the eigenvalues of correlation matrix C between N normalized time series of length T distribute in the following range.

$$\lambda_- < \lambda < \lambda_+ \tag{6}$$

The criterion of the RMT-PCA propoed in this paper is to identify the principal components if the eigenvalues are larger than the upper bound given by the RMT.

$$\lambda_+ < \lambda \tag{7}$$

However, the authors have proved based on extensive numerical analysis using the pseudo random generators that a process of taking the log-return in Eq.(1) adds extra randomness to the data [21–24]. This percolation always occurs and the maximum front of the continuum spectrum extends to about 20% larger than the upper limit λ_+ of RMT. This fact suggests that the upper limit λ_+ is not appropriate to separate the signal from the noise due to the percolation of the random spectrum over λ_+ but an effective upper bound λ_{eff}=1.2λ_+. Thus a new criterion is introduced for choosing the principal components

$$1.2\lambda_+ = \lambda_{eff} < \lambda \tag{8}$$

instead of Eq.(7), as illustrated in Fig. 1 above.

3 Trendy Industrial Sectors form the Daily-close Stock Prices

A rectangular matrix of $S_{i,k}$ is constructed by normalizing the N stock returns of the length where i=1, ... ,N represents the stock symbol and k=1, ... ,T represents the traded time of the stocks. The i-th row of this price matrix corresponds to the price time series of the i-th stock symbol, and the k-th column corresponds to the prices of N stocks at the time k. The algorithm to extravt significant principal components is summaried in Fig. 2.

However, a detailed analysis of the eigenvector components has shown that the random components do not necessarily reside below the upper limit of RMT, λ_+, but percolate beyond the RMT due to extra randomness added in the process of computing the log-return in Eq.(1). Based on extensive numerical analysis, this percolation always occurs and the maximum front of the continuum spectrum extends to about 20% larger

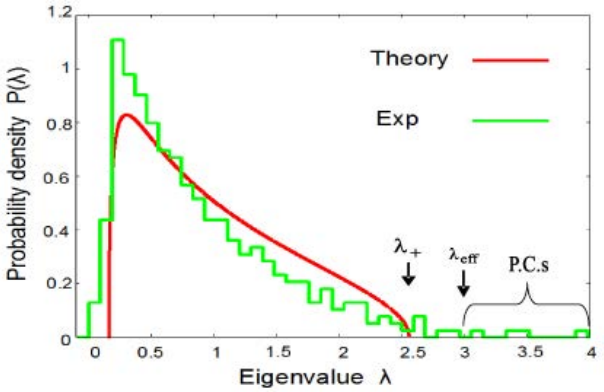


Fig. 1. The states corresponding to the eigenvalues satisfying $\lambda > 1.2\lambda_+ = \lambda_{eff}$ are identified as the principal components by the RMT-PCA

Algorithm of RMT-PCM :

(1) Select N stock symbols for which the traded price exist for all t=1,...,T, corresponding to all the working days of that term.

(2) Compute logreturn r(t) for the selected N stocks. Normalize the time series to have mean=0, variance=0, for each stock symbol, i=1,..., N.

(3) Compute the cross correlation matrix C and obtain eigenvalues and eigenvectors.

(4) Select eigenvalues larger than λ_+, the upper limit of the RMT spectrum, and $\lambda_\pm = (1 \pm Q^{-1/2})^2$ $P_{RMT}(\lambda) = \frac{Q}{2\pi\lambda}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}$ and identify those eigenstates as the principal components.'

(5) Sort the eigenvector components corresponding to the eigenvalues identified in the step (4) above, in the descending order and identify the business sectors of the the largest 20 components. If those 20 components belong to any particular sector, that is the leading sector in that term.

Fig. 2. The algorithm to extract the significant principal components in RMT-PCA

than the upper limit λ_+ of RMT. This fact suggests that the upper limit λ_+ is not appropriate to separate the signal from the noise due to the percolation of the random spectrum over λ_+ but an effective upper bound λ_{eff} = 1.2 λ_+ about 20% larger than the upper limit λ_+ of RMT. Then λ_+ in the step (4) of the RMT-PCA algorithm in Fig. 2 is to be replaced by λ_{eff}.

According to the step (4) in the RMT-PCA algorithm in Fig. 2 and, those 14 eigenstates are the principal components, based on. The authors find the business sectors of the companies of 20 largest components in the corresponding eigenvectors. If those components are concentrated in any particular business sector, that sector is defined as the trend makers during that time period. It can be proved mathematically that the eigenvector of the largest eigenvalue is consist of components of the same sign, and the corresponding sectors are not concentrated to a particular sector but distributed to any sectors, because the largest principal component show the global feature of the market thus corresponds to its representative index, such as S&P500, in this case of dealing with American stocks. The eigenvectors of the other eigenvalues have components of

Table 1. Results for 16, 8, 4 year data (Eigenvalues larger than $2\lambda_+$ are highlighted in bold-Itaric)

	94-09	94-01	02-09	94-97	98-01	02-05	06-09
N	373	373	464	373	419	464	468
T	3961	2015	1946	1010	1002	1006	936
Q	10.6	5.40	4.19	2.71	2.17	2.17	2
λ_+	1.7	2.1	2.2	2.6	2.8	2.8	2.9
λ_1	***74***	***41***	***150***	***37.2***	***53***	***116***	***200***
λ_2	***11***	***13***	***15***	***8.7***	***19***	***14***	***18***
λ_3	***8.8***	***8.8***	***12***	***5.8***	***13***	***13***	***14***
λ_4	***7.7***	***6.9***	***11***	4.6	***9.2***	***9.1***	***8.9***
λ_5	***5.1***	***4.8***	***6.5***	3.3	***6.6***	***6.3***	5.3
λ_6	***4.3***	***4.2***	***5.1***	3.2	***5.8***	5.3	5.0
λ_7	3.3	3.5	3.8	2.8	4.7	4.8	4.4
λ_8	2.9	3.1	3.4	2.6	4.2	4.6	3.5
λ_9	2.5	2.7	3.3	2.4	3.8	4.0	3.2
λ_{10}	2.4	2.2	2.8	2.4	3.8	4.0	3.2
λ_{11}	2.0	2.2	2.4	2.3	2.8	2.9	2.7
λ_{12}	1.9	2.1	2.3	2.3	2.7	2.9	2.5

both signs. It has been known that the positive components and the negative components belong to the two separate business sectors, if they are strongly concentrated to particular sectors. Summing up those knowledge the authors have, the 2nd principal component reflects the trend of the time period of the data if any concentration of the sectors are observed.

The sectors are classified according to GICS (Global Industry Classification Standard) coding system, that classifies the business sectors of stocks into 10 categories. The authors denote them by a single capital letter, A-J as follows.

A: Energy, B: Materials, C: Industrials, D: Service, E: Consumer Products, F: Health Care, G: Financials, H: Information Technology, I: Telecommunication, and J: Utility.

If taking λ_{eff} instead of λ_+, as it has been explained in the last paragraph of Section 3, then there are 10 eigenstates corresponding to the eigenvalues $\lambda_1 = 74.3, \ldots, \lambda_{10} = 2.41$, actual number of the principal components is less than 14. However, the concentration of business sectors in the eigenvector components occurs only for the 4-5 largest eigenvalues and quickly becomes blur for smaller eigenvalues. Based on this observation, the authors might increase λ_{eff} to the range of $\lambda_{eff} = 2\lambda$, 100% larger than the theoretical criterion. In any case, the difference is irrelevant as long as only several principal components are taken. There are 8 bars corresponding to $v_2(+), v_2(-), v_3(+), v_3(-), v_4(+), v_4(-), v_5(+), v_5(-)$, where $v_k(+)/v_k(-)$ indicates the positive-sign part/negative-sign part of the vector of k-th principal component, by partitions corresponding to 10 sectors of A-J, and the corresponding eigenvalues and the sign of the components below each bar.

It can be observed from the graphs in Fig. 3 that the sector H (InfoTech) dominates the (+) components of v_2 and the sector J (Utility) dominates the (-) components of v_2.

The result of 8 years data, 1994-2001 and 2002-2009 are shown in Fig. 4, the left figure of which shows the dominance of J (Utility) and H (InfoTech) during the term

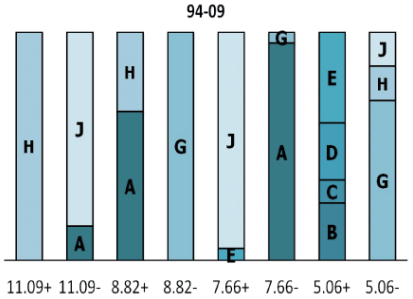


Fig. 3. Trends of 16 years from 1994 to 2009 are shown. The sector H (Information Technology) and J (Utility) are the most eminent sectors in this period.

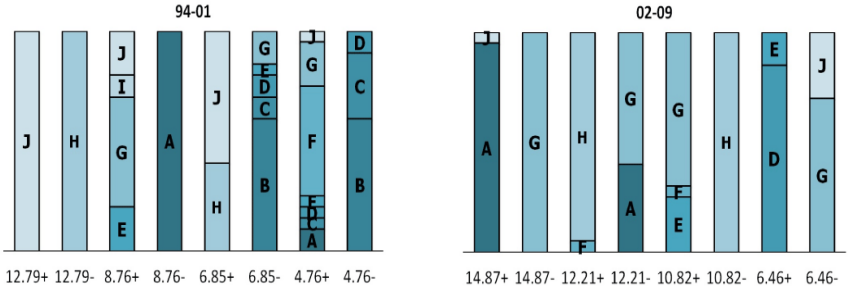


Fig. 4. Trends of 8 years, 1994-2001 (left) and 2002-2009 (right). In 1994-2001, the sector J (Utility) and H (Information Technology) dominance, but in 2002-2009, A (Energy) and G (Financial) dominance the market.

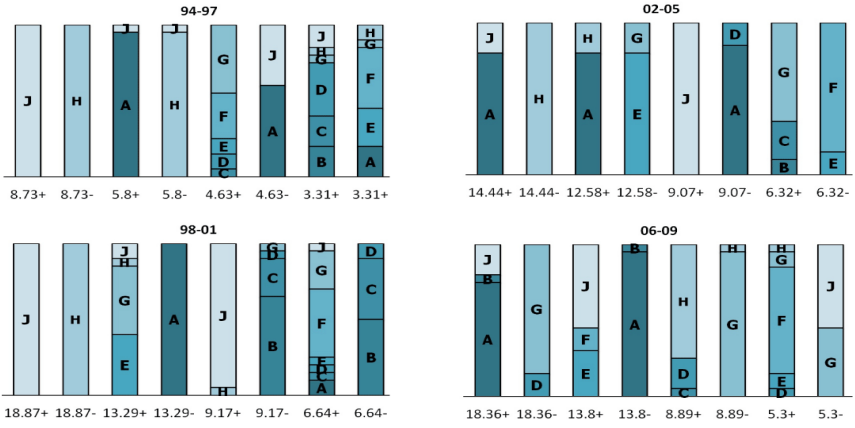


Fig. 5. Trends of 4 years each are shown. Both in 1994-1997 and 1998-2001, J (Utility) and H (Information Technology) dominance, while A (Energy) and H (Information Technology)dominance in 2002-2005 and A (Energy) and G (Financial) dominance in 2006-2009.

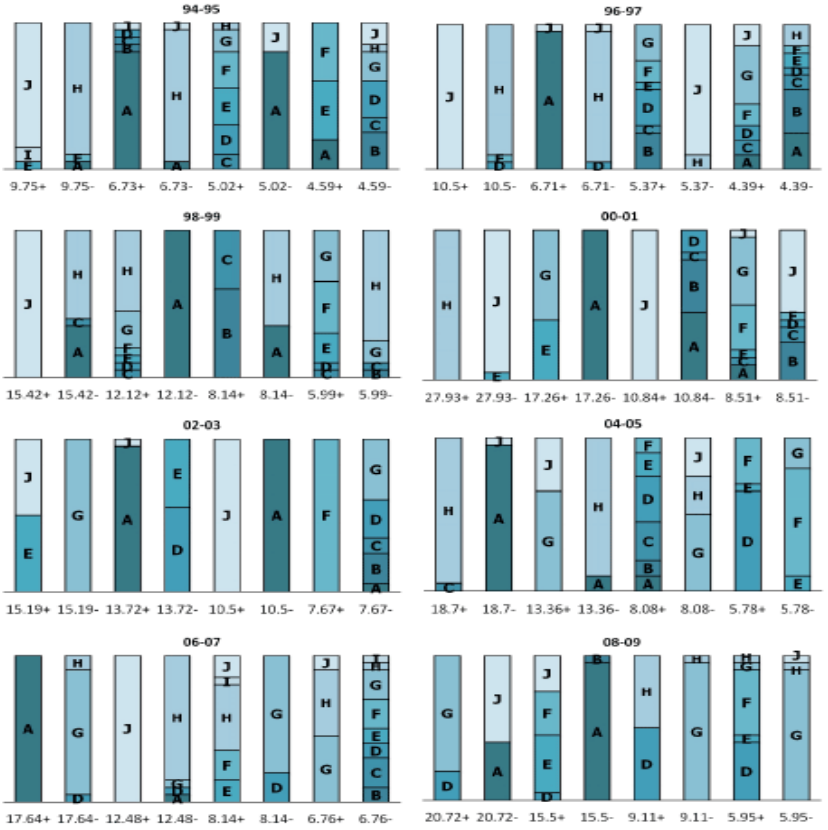


Fig. 6. Trends of 2 years each are shown.The trend change can be observed from J(Utility) and H (Information Technology) dominance towards A (Energy) and G (Financial) dominance.

1994-2001, and the right figure shows the dominance of A (Energy) and G (Financials) during the term 2002-2009. This means the active sector has changed from J (Utility) and H (InfoTech) to A (Energy) and G (Financials) at the turn of the century.

The results of 4 year data, 1994-1997, 1998-2001, 2002-2005, and 2006-2009 are in Fig. 5, showing the dominance of J (Utility) and H (InfoTech) both in 1994-1997 and 1998-2001, the dominance of A (Energy) and H (InfoTech) in 2002-2005, and A (Energy) and G (Financials) dominance in 2006-2009. The corresponding result of 2 year data is shown in Fig. 6. No clear structure is seen after 2002, except weak dominance of G (Financials) and A (Energy).

The authors have pointed out that the trend of each time period can be successfully depicted by the concentrated business sectors in the positive components and the negative components of the eigenvector corresponding to the 2nd principal components. Although the condition $\lambda > \lambda_+$ dramatically reduces the number of principal components compared to the conventional method of PCA. Moreover, the method proposed in this paper is considerably simple with much shorter in process to extract principal components, which is a great advantage in the case of analyzing the stock market.

Table 2. Parameters and the resulting numbers of PCs for the 12 quarters)

YEAR	Quarter	T	N	Q=T/N	λ_+	$\lambda > \lambda_+$	λ_{eff}	$\lambda > \lambda_{eff}$
	I	642	486	1.32	3.50	13	4.20	9
2007	II	681	486	1.40	3.40	22	4.08	13
	III	681	489	1.39	3.41	18	4.10	12
	IV	675	492	1.37	3.44	14	4.12	8
	I	642	488	1.32	3.50	11	4.20	7
2008	II	681	491	1.39	3.42	14	4.10	9
	III	692	492	1.41	3.40	15	4.08	11
	IV	664	487	1.36	3.45	13	4.14	10
	I	642	490	1.31	3.51	11	4.21	7
2009	II	659	486	1.36	3.46	13	4.15	10
	III	681	485	1.40	3.40	13	4.08	7
	IV	670	483	1.39	3.42	15	4.10	10
2007	all	2682	477	5.62	2.02	20	2.43	13
2008	all	2682	480	5.59	2.03	19	2.43	13
2009	all	2655	476	5.58	2.03	16	2.43	10

The conventional PCA can extract the largest principal component and subtract this element from the entire data, and apply the same procedure recursively on the remaining data one by one. This kind of method requires a lot of computational time and is not suitable for analyzing a system of the large dimension, such as a set of stocks in the market. Another method of PCA uses the eigenvalues of the correlation matrix of times series, which pick up the components whose eigenvalues are larger than one, or the accumulated sum of eigenvalues exceeds 80 percent of the total sum, etc. Neither one is suitable for analyzing the stocks in the market, since the number of principal components thus obtained usually exceeds 100 for N=400-500, while the RMT- PCA has derived the number of principal components in the range of 5-13 in Section 4 in this paper. This point is illustrated in Fig. 9.

4 Trendy Industrial Sectors form the Tickwise Stock Prices

The original tick-wise stock prices are converted to 30 minutes data by selecting the stocks which have at least one trade in the range of each 30 minutes period. For example, the first quarter of the year 2007, from January to March, 2007 had N=486 stocks satisfied this condition and the length of time series of this period was T=642. The numbers of principal components thus computed are listed in the rightmost column of Table 1. Although there are 7-13 principal components whose eigenvalues λ larger than λ_{eff}, for each set of quarterly (or yearly) data, firstly, focus on the second largest eigenvalue λ_2 and its eigenvector u_2, and ignore the rest. Then comparing the above result to the corresponding results of considering all the first ten eigenvectors, in order to show the superiority of the information from u_2, over the noisy results of using other eigenvectors.

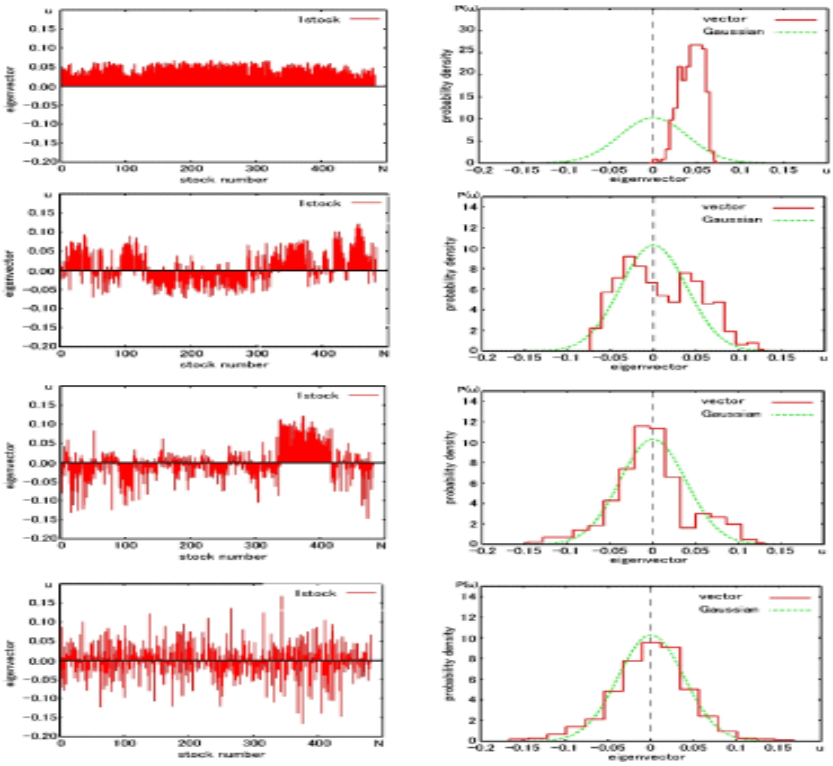


Fig. 7. Eigenvector components (left) and the histograms (right) of u_1 - u_4

Table 3. The industrial sectors represented by the code numbers

ID:Sector	ID:Sector	ID:Sector
13:Fishery/Agri/Forestry	15:Mineral Mining	17:Construction
20:Food	30:Fiber/Paper	40:Chemistry/Medicine
50:Resources/Material	60:Machine/Elec.Machinery	70:Automobile/Trans.apparatus
80:Commerce	83:Finance/Insurance	88:Real Estate
90:Transport/Telecom	95:Electric/GasPowerSupply	96:Service

First of all, the largest principal component corresponding to the largest eigenvalue λ_1 and its eigenvector u_1 are unfortunately not suitable for extracting trendy sectors. The components of u_1 are almost equally sized around the average value $0.05 = 1/\sqrt{500}$ and do not have any distinguished components, as shown in the first row in Fig. 4. This fact is in common to most markets and is often referred to the 'market mode'. It is known that this component is strongly correlated to the index consist of dominant and stable stocks such as so called blue-chip stocks.

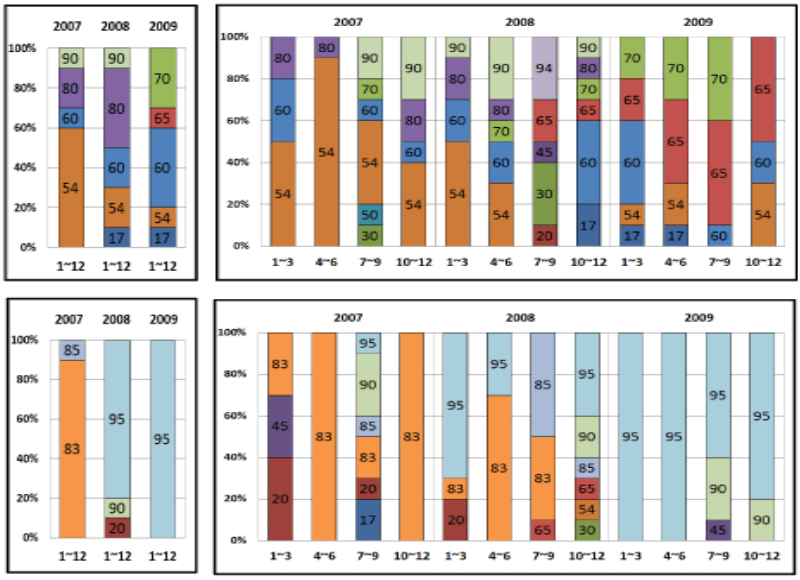


Fig. 8. Trendy sectors in the positive and the negative sectors extracted as collective modes of u_2, obtained from 30 minutes price time series in 2007-2009. The numbers in each bar are the codes of sectors shown in Table 3.

The authors thus focus on the second largest eigenvalue λ_2 and its eigenvector u_2, which exhibits a certain trend that changes from time to time. Moreover, as shown in the second row of Fig. 4, the positive components aggregate to form a certain collective mode by their internal attractive force, and the negative components do the same, so that both + and - components individually form their own collective modes, which represent temporary trend of active sectors in the market.

The third largest eigenvalue λ_3 and its eigenvector u_3 exhibit the similar feature as the second component in a vague manner, and the fourth eigenvalue λ_4 and its eigenvector u_4 do not show any clear feature and behave more like Gaussian, as shown in the third and the fourth raw of Fig. 4. For the fifth or further eigenstates, the sizes of the N components behave more random and the corresponding histograms reach the Gaussian.

Comparing the first four eigenvectors, it is clear that the second eigenvector u_2 exhibits the existence of two collective modes in the positive and the negative sides in a most clear sense. On the other hand, the components of the first eigenvector u1 are distributed evenly and do not show a sign of aggregation. The components of the fourth or higher eigenvectors are highly random and the distribution is close to Gaussian. The third eigenvector seems transient in between.

Based on the above observation, it can be concluded that only u_2 shows a clear sign of the collective modes that make the trendy industrial sectors of each period of time.

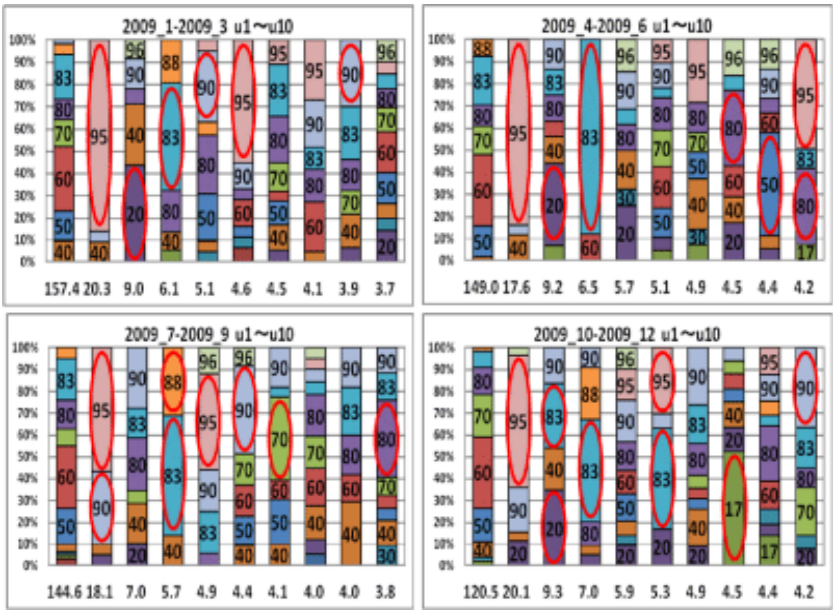


Fig. 9. Noisy result of the top ten eigenstates, u_1-u_{10}, in 2009 are shown for the sake of comparison. The industrial sectors are partitioned in each bar showing 10 eigenstates, ordered from the leftmost bar to the rightmost bar in each figure, with the corresponding eigenvalue below each bar. The first quarter (Jan.-Mar.) to the last quarter (Oct. -Dec.) are shown in four figures from the top to the bottom.

5 Trendy Industrial Sectors Extracted from the Collective Modes in the Second Eigenvector u_2

The trendy industrial sectors are identified as the sectors that the distinguishably large elements of the chosen eigenvectors belong to. However, how many of such elements to be taken is not given in any sense. In this paper, the authors followed on this point two different scenarios. One scenario is to take a fixed number of elements from + and - elements each. Fig. 8 shows the result of this scenario for u_2 only. The numbers inside the graphs show the industrial sector according to the codes defined in Table 3. Another scenario is to use the accumulation of large elements in the descending order of the sizes up to 20% of the total amount. Fig. 8 shows the result of this scenario for all the largest ten eigenstates.

6 Conclusion and Discussion

In this paper, it has shown the result of applying the RMT-PCA on 30 minutes traded prices of 4 quarters each in three years from 2007 to 2009 of Tokyo Market, and compared the result on daily data in sixteen years from 1994 to 2009 of S&P.

By analyzing the size distribution of the N components of the first four eigenvectors, the authors found that only the second eigenvector u_2 has a useful feature for the

sake of extracting the trendy industrial sectors from their collective modes, formed independently in the positive parts and the negative pas of the components. This is the conclusion that have reached as to the first of the two unresolved technical problem in the practical application of the RMT-PCA.

As to the second of the unresolved technical problem, the authors have simply compared Fig. 8 where the fixed number of positive and negative elements are selected in the descending order, and Fig. 9 where the accumulation of large elements in the descending order of the sizes up to 20% of the total amount. It is observed that the extracted sectors shown by circles in Fig. 9 coincide with the result of Fig. 8, and no more useful information is offered by Fig. 9 other than noisy details. Thus the authors conclude that the use of u_2 is sufficient to extract trendy sectors, while the number of large elements to consider is inconclusive.

Finally, the authors discuss on the consistency of our result to the actual historical incidence. Both Fig. 8 and Fig. 9 indicate the change of trendy sectors from 83 (banks) in 2007 to 95 (power supply) in 2009. Also a disappearance of major sectors in the third quarter of 2007 and the fourth quarter of 2008 represent the extremely confusing market conditions caused by the sub-prime loan problem in August 2007 and the bankruptcy of Lehman Brothers in October 2008.

References

1. Mehta, M.L.: Random matrices, 3rd edn. Academic Press (2004)
2. Edelman, A., Rao, N.R.: Random matrix theory. Acta Numerica, 1–65 (2005)
3. Zhidong, B., Silverstein, J.: Spectral analysis of large dimensional Random Matrices. Springer (2010)
4. Tao, T., Vu, V.: Random matrices: universality of ESD and the circular law (with appendix by M. Krishnapur). Annals of Probability 38(5), 2023–2065 (2010)
5. Beenakker, C.W.J.: Random-matrix theory of quantum transport. Reviews of Modern Physics 69, 731–808 (1997)
6. Kendrick, D.: Stochastic control for economic models. McGraw-Hill (1981)
7. Bahcall, S.R.: Random matrix model for superconductors in a magnetic field. Physical Review Letters 77, 5276–5279 (1976)
8. Franchini, F., Kravtsov, V.E.: Horizon in random matrix theory, the Hawking radiation, and flow of cold atoms. Physical Review Letters 103, 166401 (2009)
9. Peyrache, A., et al.: Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution. Journal of Computational Neurosience 29 (2009)
10. Sánchez, D., Büttiker, M.: Magnetic-field asymmetry of nonlinear mesoscopic transport. Physical Review Letters 93, 106802 (2004)
11. Marcenko, V.A., Pastur, L.A.: Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik 1, 457–483 (1994)
12. Sengupta, A.M., Mitra, P.P.: Distribution of singular values for some random matrices. Physical Review E 60, 3389–3392 (1999)
13. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E.: Random matrix approach to cross correlation in financial data. Physical Review E 65, 066126 (2002)
14. Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Stanley, H.E.: Scaling behaviour in the growth of companies. Physical Review Letters 83, 1471–1474 (1999)
15. Laloux, L., Cizeaux, P., Bouchaud, J.-P., Potters, M.: Noise dressing of financial correlation matrix. Physical Review Letters 83, 1467–1470 (1999)

16. Bouchaud, J.-P., Potters, M.: Theory of Financial Risks. Cambridge Univ. Press (2000)
17. Mantegna, R.N., Stanley, H.E.: An Introduction to econophysics: correlations and complexity in finance. Cambridge University Press (2000)
18. Iyetomi, H., et al.: Fluctuation-dissipation theory of input-output interindustrial relations. Physical Review E 83, 016103 (2011)
19. Tanaka-Yamawaki, M.: Extracting principal components from pseudo-random data by using random matrix theory. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part III. LNCS, vol. 6278, pp. 602–611. Springer, Heidelberg (2010)
20. Tanaka-Yamawaki, M.: Cross correlation of intra-day stock prices in comparison to random matrix theory. Intelligent Information Management (2011)
21. Yang, X., Itoi, R., Tanaka-Yamawaki, M.: Testing randomness by means of RMT formula. In: Watada, J., Phillips-Wren, G., Jain, L.C., Howlett, R.J. (eds.) Intelligent Decision Technologies. SIST, vol. 10, pp. 589–596. Springer, Heidelberg (2011)
22. Yang, X., Tanaka-Yamawaki, M.: Testing randomness by means of Random Matrix Theory. In: 2011 Kyoto Workshop on NOLTA, p. 1 (2011)
23. Yang, X., Itoi, R., Tanaka-Yamawaki, M.: Testing randomness by means of Random Matrix Theory. Progress of Theoretical Physics Supplement (194), 73–83 (2012)
24. Tanaka-Yamawaki, M., Yang, X., Itoi, R.: Moment approach for quantitative evaluation of randomness based on RMT formula. In: Watada, J., Watanabe, T., Phillips-Wren, G., Howlett, R.J., Jain, L.C. (eds.) Intelligent Decision Technologies, Vol. 2. SIST, vol. 16, pp. 423–432. Springer, Heidelberg (2012)

A Framework for a Posteriori Method of Transitional Analysis to Diffuse ICT Services Based on Text Mining

Motoi Iwashita

Chiba Institute of Technology, Chiba 275-0016 Japan
iwashita.motoi@it-chiba.ac.jp

Abstract. It is essential to know how customers perceive their wants and needs and what they require in the diffusion of Information and Communication Technology (ICT) services. Customer perceptions/requirements of ICT services have continuously been changing and have depended on new developments in technology and improvements to user-operability. Therefore, it is necessary to sense customer wants and needs by continuously monitoring their behaviours could be used in marketing.

This chapter first proposes the framework of a posteriori method of transitional analysis to solve these issues, which is used to analyze blogs and news events on ICT services as time-series data based on a text mining technique. Therefore, a person in enterprise can detect changes in customer perceived requirements through the co-occurrences and transition rates of terms by taking into consideration the network or its ICT services. The proposed method was applied to Worldwide Interoperability for Microwave Access (WiMAX) as a high-speed wireless access services by collecting and analyzing blog data. And obtained practical results show adequacy in the right of the fact that the service has been evolved.

Keywords: Semantic content, Co-occurrence, Time-series analysis, Data mining, Diffusion mechanism, ICT services, Text mining.

1 Introduction

Fiber-To-The-Home (FTTH) has been provided since 2002 in Japan. This provides domestic access to ultra-high-speed broadband access infrastructure. The coverage of FTTH for telephone customers widely adopted by 84% of the population in 2009 [1]. However, the percentage of FTTH customers was initially low at about 35%. Long-Term Evolution (LTE)[1] for high-speed wireless broadband access has only been provided since December 2010. The LTE coverage rate (expressed as a percentage) for the population was about 25% and the customer rate was about 4% at the end of the 2011 fiscal year. Another kind of high-speed wireless access, [2], has also been provided in metropolitan areas since 2009 and the coverage rate for the population in government-decreed cities with populations greater than 500 000 has been set about 95% and there have been about one million customers. This has a great effect on enterprise management when the use is very low, because of the high cost of maintaining these facilities.

[1] See http://www.3gpp.org/Technologies/Keywords-Acronyms/LTE

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 39–55, 2013.

Therefore, the use of facilities is an issue of high priority for providers/municipalities, and it is necessary to strategically install these economically.

A highly efficient strategic framework to diffuse Information and Communication Technology (ICT) services should be constructed, as shown in Fig. 1. It can be seen that service diffusion involves two types of activities, which are "sales promotion" and "service optimization". The plan for sales promotion includes advertisements, price reduction campaigns, and ample supply of commemorative goods. The two main features of sales promotion are that it requires no investment in facilities but does require an understanding of its effect on customers. The plan for service optimization includes providing value-added services, improving operability, and expanding the coverage area. The two main features of service optimization are investment in facilities and an understanding of customers' requirements such as bandwidth expansion, easy access anytime and anywhere, etc.

The effect of service diffusion for sales promotion is only to attract potential demand that has already been planned, and not to attract new business. It is important to identify customer perceptions and requirements for ICT services to optimize them when generating new types of customers from the viewpoint of service diffusion. That is to say, the kinds of factors have a great effect on the behaviour of service choices. Take Worldwide Interoperability for Microwave Access (WiMAX) as an example. If customers feel the WiMAX coverage area is too limited, an effective solution would be for the provider to expand coverage areas by installing more base stations. If customers often complain about communication speeds, these need to be improved. A provider can find suitable solutions, when it understands customer perceptions/requirements. However, customer perceptions and requirements for ICT services are continuously changing and depend on new technologies and service operability. For example, customers initially complain about complex connection settings, and then about simple security settings against application downloads. Therefore, it is necessary to uncover customer perceptions and requirements by monitoring changes in their behaviours on a daily, monthly, and yearly basis. Providers can take quick action in advance and gain a high level of customer satisfaction, if they can effectively comprehend customer perceptions/requirements.

This chapter proposes a framework for service diffusion and a method of identifying customer perceptions/requirements as a new means of marketing. The basic idea is based on time-series data that are able to be used to detect changes in customer perceptions/requirements. Section 2 describes related issues concerned with the detection of perceptions/requirements and time-series analysis. Section 3 is devoted to the characteristics of customer perceptions/requirements for ICT services. The categorization of data and the framework for the method of analysis are introduced in Sections 4 and 5. The experimental results and other considerations are discussed in Section 6. The conclusion and future research are followed.

2 Related Work

This chapter especially focuses on WiMAX as an ICT services. WiMAX network topology is explained at first. Mobile gadget such as smartphone, ultra mobile PC can access the nearest base station by air in each region, as shown in Fig. 2. The base station transmits signals to provider' server through carrier's network. Then, the users can receive

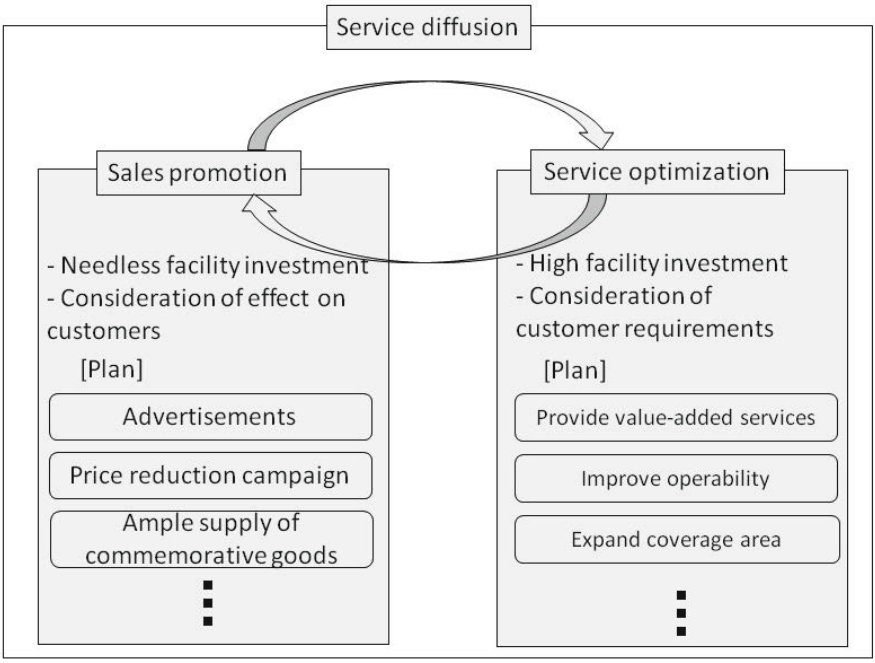


Fig. 1. Service diffusion framework

wireless high-speed internet access services. If the numbers of areas which have a base station increases, the customers can access high-speed internet over a wider region.

Text-mining, such as that in morphological analyses, syntax analyses, or co-occurrence relations, is an effective technique [3–7]. This can be applied to analyses of customer questionnaires in product development, word searches at portal sites, analyses of term frequency in Web logs (blogs) or customer-generated media, classification by keywords in news articles, and evaluation indexes of a company's image. Morphological analysis is mainly applied in these areas to survey trends by analyzing the frequency of terms in selected texts. A keyword is extracted as the topic of a sentence in terms of the features of the network structure [8–12]. Clustering and co-occurrence related methods have been proposed to classify keywords and relate them to synonymous terms, different words having the same meaning, and synonyms, which have similar meanings [13, 14]. Clustering methods based on supervised learning have also been proposed [15–18]. They have mainly been applied to searching the abstracts of research papers and the automated scoring of descriptive answers, and they are effective for searching for similarities between texts based on given trend terms/information.

Customer opinions about text mining in ICT services are strongly related to service conditions such as technologies and service operability at that point. Therefore, although it is insufficient to understand trends in keywords or to find similar content, it is essential to understand the meanings of sentences and extract customer requirements. The present author has proposed a method of classifying and construct-

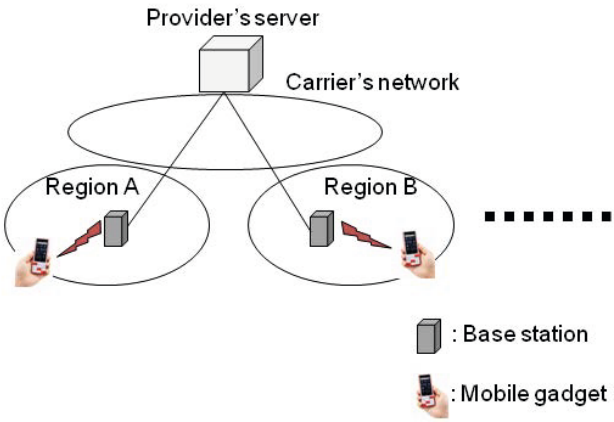


Fig. 2. WiMAX networks topology

ing the meaning of sentences as customer enquiries about telecommunication services [19] that can be used to extract large volumes of customer requirements. The main limitation of this method is in detecting small volumes of customer requirements but this will be addressed in the future. This is because the method does not continuously analyze data. That is to say, not only understanding semantic content, but also detecting changes in customer perceptions/requirements at that point is necessary to take strategic action quickly. Therefore, it is necessary to detect changes in customer behaviour with time-series analysis. Although a system that can display views of time-series data has been developed, there have previously been no methods of analysis. Moreover, such time-series analyses would be done at individual enterprises by experts without scientific or systematic methods. If such methods of analysis could be established, sudden incapacity in telecommunication networks caused by unanticipated increases in traffic could be expected and prevented. Therefore, effective method of detecting changes in customer perceptions/requirements is constructed with making a correlation between network service components and provider's events in this chapter.

3 Features of Blog Textual Data in ICT Services

Blog textual data would be effective for time-series analysis to understand customer perceptions/requirements. Although questionnaires administered to customers or closed data in enterprises are the most effective kinds of information, only specific enterprises can use them. Therefore, not only specific but also general enterprises with related services/networks (software/hardware vendors) can use blogs as open data. The blog data is easily collected by crawler which is a freeware crawling web sites automatically. The specified blogs are collected including past ones (e.g. several years ago) after the keywords are input into crawler. This chapter uses WiMAX as a keyword because it is available and popular these days.

The three main advantages of using blogs are below:

- It is easy to use them collect customer thoughts and comments posted on the Internet.
  "One of the best ways for mobile networking is to use UQ-WiMAX or eMobile. Although WiMAX is superior in speed, its wireless area coverage is limited, so I'm worried. This is because the concept underlying WiMAX is that it is primarily a wireless LAN for the last mile..."
- We can obtain customers' thoughts and comments after thoughtful consideration through blogs, which is different to Twitter's real-time logs.
- We are able to obtain large volumes of data for extended periods.

However, blogs have two main disadvantages.

- It is difficult to decide whether the topics in blogs fit our intentions or not even if we collect blogs by narrowing them down with keywords. This is because there are several sentences in a blog, and the topics they contain are generally long.
  "Name: Mr. WiMAX, posting date: 8/28/2011, Sun., 07:04:53. My franchise stores cover a large area..."
- There are also data that lack credibility.

Therefore, the three main problems to solve are as follows:

1. How to extract words related to a given keyword in several long sentences.
   "Radio waves are of the most concern to WiMAX. It was hard to establish connections in doorways even in Tokyo two years ago."
   The above sentence means that WiMAX connections were hard to establish two years ago, and it does not describe recent situations.
2. How to filter the possibility that the same message has been transmitted by a specific person.
3. How often to collect blog information (every month, quarterly, yearly, or when a new service is provided or a new product is released).

The relationship between term frequencies and their rankings for blogs is plotted in Figs. 3 and 4. The terms (nouns) were induced by morphological analysis and counted, and those that appeared from the first to the 200th ranking are shown for one month of data at the four points in Fig. 3.

The graph indicates almost a 1/n feature, so they seem to follow a power law [20, 21]. A power law implies that there are many kinds of terms in the textual data and means that they could contain many types of customer interests. The gradient of the results becomes more gradual over time. This implies that customers experience a greater variety of perceptions and requirements. Although there are few selected terms that have a relation to "WiMAX", the same behaviour can be seen in the graph in Fig. 4. However, high-frequency terms that reflect customer interests are limited for WiMAX services for all periods.

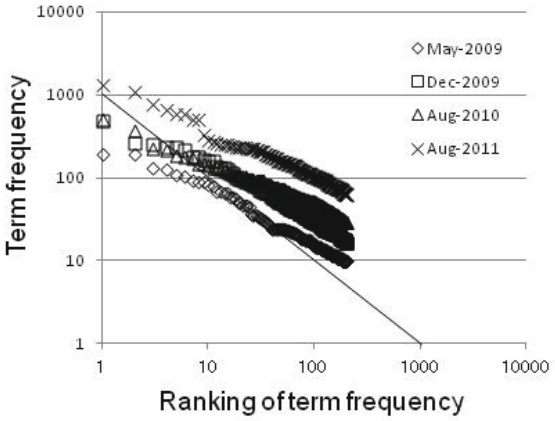


Fig. 3. Relationship between term frequency and rankings: All terms

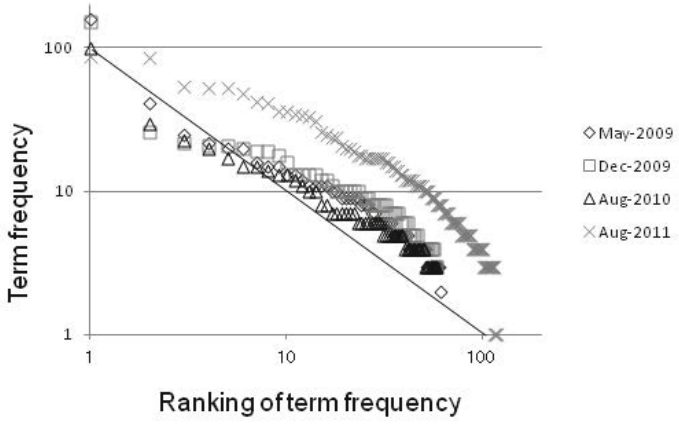


Fig. 4. Relationship between term frequency and rankings: WiMAX related

4 Data Categorization

Since blogs contain long sentences, it is difficult to apply text-mining techniques directly to raw textual data for semantic or structural classification. Moreover, it is difficult to find which factors in an information network are essential because an information network consists of several components. Taking the example phrase, "We need to spend a lot of time to access WiMAX". Such phrase is always induced by the cause of networks/services, so the provider should quickly take measure against network components/services, otherwise increasing customer churn. However, it is hard to find the cause as to whether this is caused by the performance of gadgets, the provider's server, or the network. Therefore, it is needed that modifications to distinguish the features of WiMAX services. WiMAX services are generally provided by an end-to-end network consisting of a mobile gadget, access point, carrier network, and the provider's

server, as outlined in Fig. 5. There are clearly customer thoughts/comments on all elements of the network. It can be predicted that equipment, such as mobile gadgets, is strongly related to easy operation/long battery life. Therefore, a semantic representation can be constructed by designating the network factor as one event (Category A) and customer thoughts/comments as another event (Category B). If the changes in customer perceptions/requirements are analyzed, transitional events such as the start of commercial services, promotion campaigns, and service menu upgrades are necessary for Category C. This is because customer perceptions/requirements generally depend greatly on such transitional events. Therefore, it is necessary for Categories B and C to have a relationship for time-series analysis.

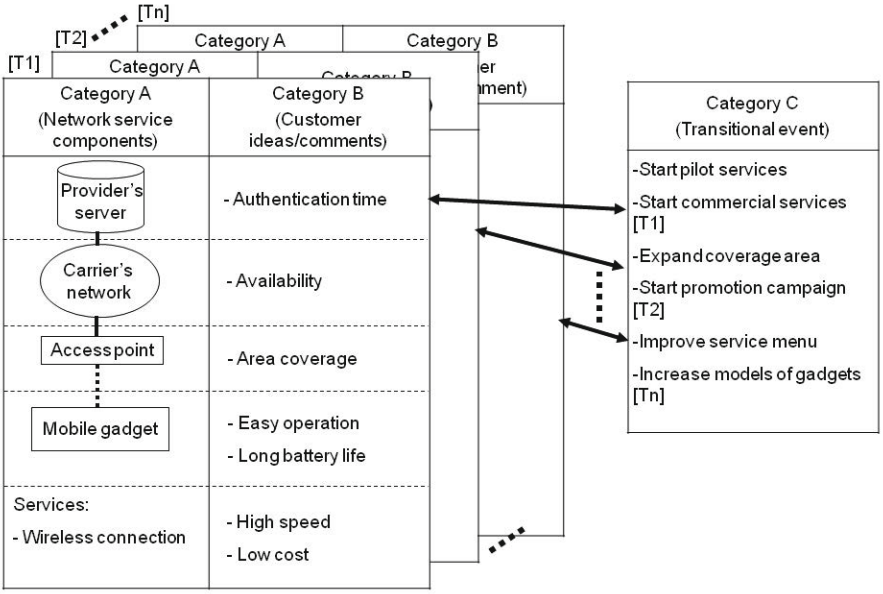


Fig. 5. Categorization of WiMAX features

5 A Posteriori Method of Transitional Analysis

This section explains a posteriori method. A framework of the method is described in 5.1. The rest of the three subsections are for the method detail. Term classification as preprocessing customer perceptions/requirements and topics extraction with news events are explained in 5.2 and 5.3, respectively. 5.4 is devoted to transitional analysis.

5.1 Framework of Method

The framework for time-series analysis to find customer perceptions/requirements is outlined in Fig. 6. Many blogs are input. Morphological analysis and co-occurrence

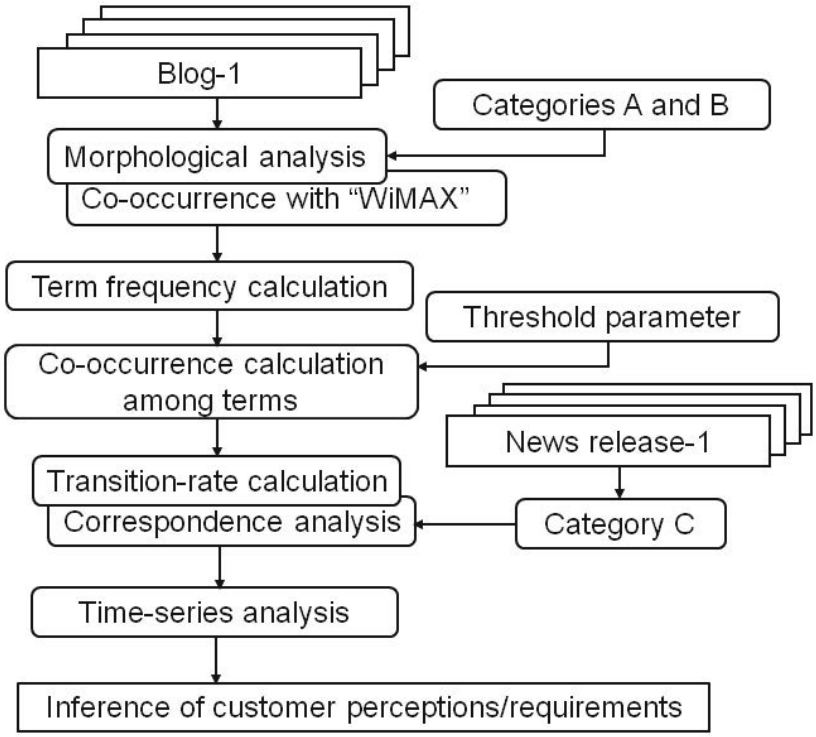


Fig. 6. Framework for method of time-series analysis

with "WiMAX" are effective preprocessing steps for classifying this input in terms of categories. The requirements to classify blogs are based on the ability to cover all textual data. It is wanted that this classification to be determined from the viewpoints of term frequency and calculation of the co-occurrence rate. Term frequency can tell us what kinds of blogs often appear according to categories A and B. The co-occurrence rate can be used for the terms we obtain to determine whether they are related to a specific topic or not.

Then the relationship between terms and events are individually found. Therefore, news releases are one of the inputs for category C. Correspondence analysis is applied to the relationship with terms and category C. Then, the calculation of the transition rate tells us the relationship between terms so that we can extract phrases with meaning. Finally, we construct a transitional graph from the results by using transition-rate calculations and correspondence analysis, and find trends in customer perceptions/requirements.

5.2 Term Classification with Relation to Specific Topic

It is important in semantic analysis to carry out classification by preprocessing customer perceptions/requirements. The outputs at this level are sets of textual data classified by

the relationship between terms and categories. The classification rules for procedures 1 and 2 follow. They are also shown graphically in Figs. 7 and 8.

Procedure 1: Term extraction by morphological analysis and co-occurrence (Fig. 7)

1. First, textual data are classified by morphological analysis.
2. The T-value of each term is calculated and terms are selected if their T-value is greater than 1.5.
3. Terms are selected that have a high co-occurrence rate for "WiMAX".

Since blogs contain several long sentences, terms that have a high co-occurrence rate for WiMAX are selected to check whether the sentences descriptions are concerned with WiMAX or not. Before the co-occurrence rate is calculated, the T-value for each term is checked to determine whether it is greater than 1.5. Here, the T-value means whether the term has a relation to a specific topic/keyword, and is calculated in Eq. 1, where X is the real value of the co-occurrence rate and μ is the expected value of the co-occurrence rate.

$$T - value = (X - \mu)/\sqrt{X}\,, \tag{1}$$

The threshold, "1.5", was determined from an empirical study in the statistical field (as a value of more than two was reliable, we extracted more terms than those in that case).

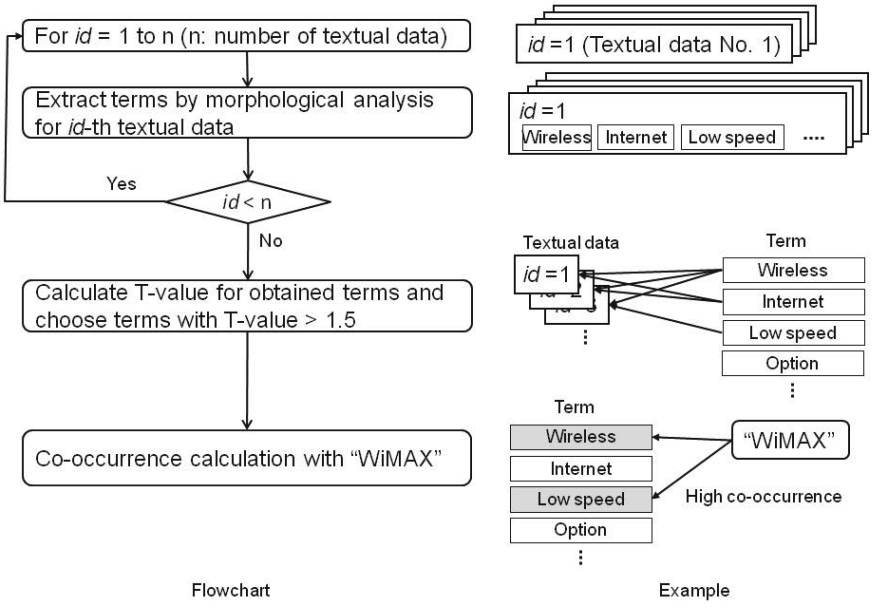


Fig. 7. Term extraction with relation specific topic

Procedure 2: Classification by type of category (Fig. 8):

1. The textual data are classified in terms derived from Procedure 1. Category A indicates network service components, such as mobile gadgets, access points, or wireless modems, while Category B indicates customer thoughts/comments, such as long battery life, area coverage, or high speed. Classification is permitted to contain several terms for one item of textual data.
2. Similar terms are grouped as a representative term.

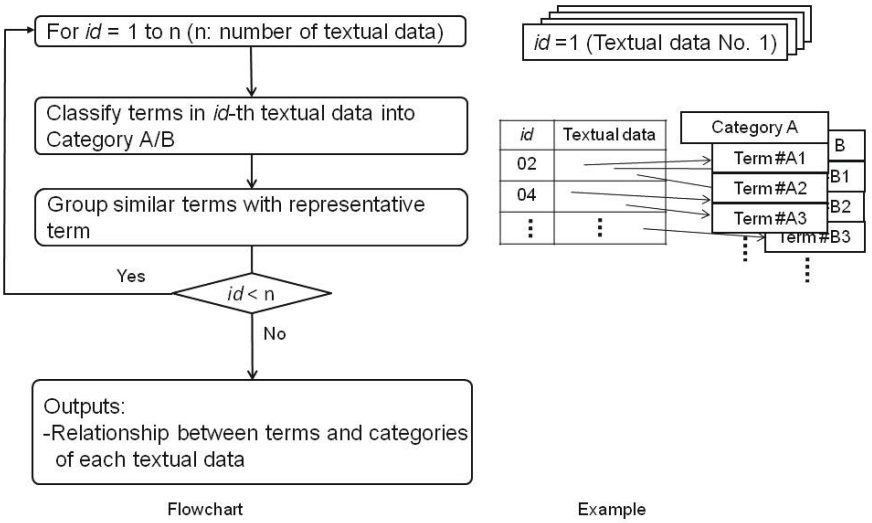


Fig. 8. Classification by category

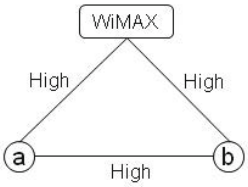
5.3 Correspondence Analysis and Calculation of Transition Rate

Related terms, i.e., a pair of terms, for WiMAX are extracted by using transition-rate calculation. Although it is necessary to construct a pair of terms precisely with a combination of all terms, the number of terms is excessive and inefficient. Therefore, the related terms with news events are firstly extracted by carrying out correspondence analysis.

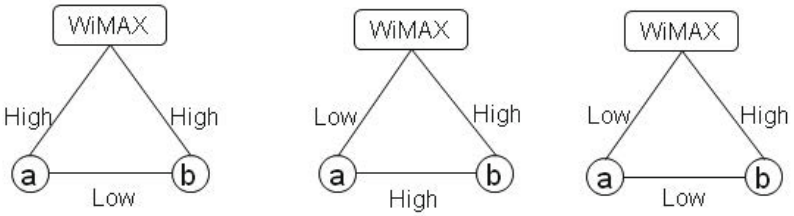
Procedure 3: Extraction of topics

1. A table is constructed where terms related to WiMAX and their co-occurrence rates are in each cell as time-series data.
2. Correspondence analysis extracts terms as topics.
3. Transition-rate calculation is done for the pair of topics shown in Fig. 9.

The aim of the procedure for transition-rate calculation is to automatically reconstruct meaningful sentences. When there are three terms including WiMAX, all co-occurrence rates are calculated. Both “a” and “b” are estimated to be terms related to



(a) "a" and "b" are estimated to be terms related to WiMAX



(b) If there is term with low co-occurrence rate, pair of "a" and "b" is not estimated to be terms related to WiMAX

Fig. 9. Concept underlying transition-rate calculation

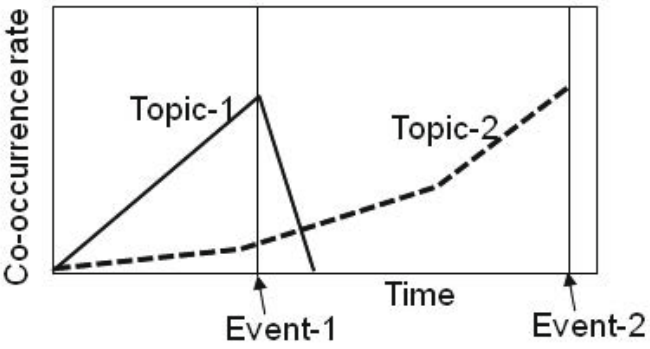
WiMAX, when all pairs of co-occurrence rates are high, as shown in Fig. 9 (a). If there is at least a pair with a low co-occurrence rate, a pair of terms, "a" and "b", is not supposed to be related to WiMAX, as seen in Fig. 9 (b). This transition concept extends to a four or more term relationship so that it is possible to reconstruct meaningful long sentences.

5.4 Transitional Analysis

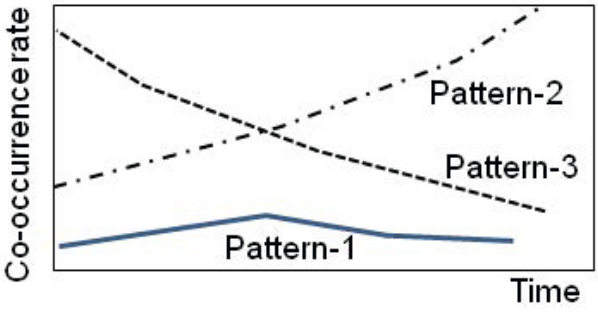
The concept underlying time-series analysis and the results are presented in Fig. 10. There are two types of trends. The first is for topics concerned with news events, and the second is for continuously appearing terms. The trend analysis of topics concerned with events is first plotted in Fig. 10 (a). The co-occurrence rate of topics concerned with events is obviously the highest when corresponding events occur. We can also see the behaviour of the co-occurrence rate around that time. When the co-occurrence rate increases with corresponding events and then decreases (Topic-1), the situation implies a topic is improved. When that rate increases gradually with corresponding events without decreasing behaviour (Topic-2), the topic becomes popular or interesting, or worsens the situation depending on good or poor reputations.

The transitional change of a pair of terms that is not a topic but appears continuously is plotted in Fig. 10 (b). Each term is located on the graph at a point in the relationship over time. Then, the terms are characterized from the viewpoint of the co-occurrence rate at all four points. When the term appears several times and has almost the same co-occurrence rate at each time (Pattern-1), it is not affected very much by changes in

events. That is to say, customers feel the same all the time. When the co-occurrence curve gradually increases (Pattern-2), changes in events such as increased demand, increased coverage area, and new technological developments would have a continuous effect and become interesting. When the co-occurrence curve gradually decreases (Pattern-3), changes in events have less effect and become uninteresting.



(a) Trend analysis of topics concerning events



(b) Trend analysis of continuously appearing terms

Fig. 10. Analysis of time-series data

6 Results from Evaluation

6.1 Assumptions

This chapter targeted at WiMAX as an example to evaluate the proposed method, which is a wireless broadband access service that started in 2009 in Japan. The plans for service evolution and the results obtained from a company that provided WiMAX were investigated. Blog data was collected during two in 2009 (May and December), one

in 2010 (August), another in 2011 (August). This was analyzed to determine changes in customer perceptions/requirements. The terms were classified into network services/components and customer thoughts/comments.

6.2 Term Extraction with Relation to WiMAX

First, terms are checked with a T-value, then, selected terms are sorted in descending order of their rate of co-occurrence, as shown in Fig. 11. The features of extracted terms are different each time. Next, the terms are classified according to categories A and B (indicated in a term cell); this means terms not belonging to either category are eliminated at this step manually.

Term	T-value	Rate
Try	2.121825	0.833333
Vendor (A)	1.89356	0.8
All	2.829697	0.75
Customer (A)	2.102785	0.714286
Make	2.102785	0.714286
UQ (A)	11.86485	0.686695
Brand (B)	1.609145	0.6
Special topic (B)	1.609145	0.6
Relationship (B)	1.609145	0.6
Theme	2.758737	0.529412
Communications (A)	4.091326	0.5
Sensitivity (B)	2.045663	0.5
Option (B)	1.584564	0.5
Ward	1.584564	0.5
Built-in (B)	3.286724	0.481481
⋮	⋮	⋮

(a) May 2009

Term	T-value	Rate
DIS	2.984053	0.909091
Electrical power (B)	2.967851	0.833333
UQ-WiMAX (A)	2.098587	0.833333
Title	3.378187	0.8125
Half price (B)	3.242243	0.8
SS	1.87191	0.8
On (B)	1.613726	0.75
Inclusion	1.613726	0.75
Customer (A)	1.846292	0.666667
UQ (A)	11.38162	0.618474
BIC	2.592937	0.615385
Lend	2.413367	0.583333
Air (B)	1.820674	0.571429
Let's	2.029847	0.555556
UD	2.219403	0.545455
⋮	⋮	⋮

(b) December 2009

Fig. 11. Term extraction with relation to WiMAX

6.3 Correspondence Analysis and Transition-Rate Calculation

At first, the news is morphologically analyzed to find the relationship between the obtained terms and news events, and news events are selected each time. Then, correspondence analysis is applied to the terms and events ("time" has been used in this chapter), as shown in Fig. 12.

The terms obtained with strong relationships to events are called "topics". A pair of topics for WiMAX is extracted with transition-rate calculation according to the concept in Fig. 9. The pairs of topics listed in Table 1 are selected each time.

The topics that satisfied threshold α are selected, which were very difficult to achieve. Consequently, it is found that the best way of doing this was first to set a high value and then gradually decrease it by checking the appearance of pairs of terms.

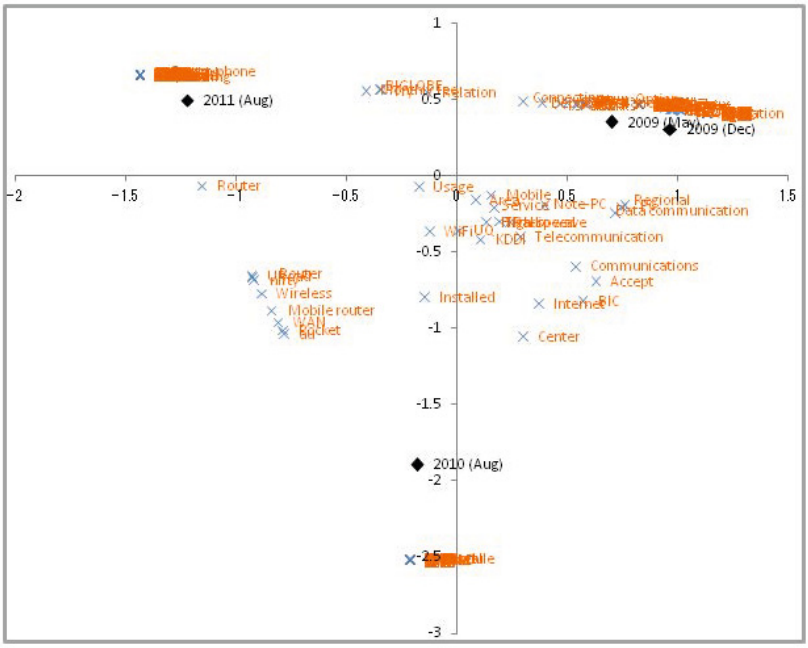


Fig. 12. Correspondence analysis

Table 1. News events and most related topic

	2009 (May)	2009 (Dec)	2010 (Aug)	2011 (Aug)
-News event	*-Service start*	*-Area expansion in high gear* *-Half price campaign for mobile terminals*	*-International services*	*-Line-up of new routers* *-MVNO introduction* *-Connection with WiFi*
Topics	Monitor and service (0.01)	Area and map (0.73)	WORLD and use (0.01)	Routers and comparisons (0.04)
	Provision and data communication (0.03)	Area and expansion (0.68)	-	Charges and comparisons (0.09)
	-	Area and readiness (0.29)	-	EVO and tethering (0.09)
	-	For X days and trial (0.39)	-	EVO and smart-phones (0.04)
	-	-	-	EVO and switching (0.16)
	-	-	-	HTC and tethering (0.05)
	-	-	-	HTC and smart-phones (0.04)
	-	-	-	MVNO and comparisons (0.07)
	-	-	-	MVNO and charges (0.03)
	-	-	-	Xi and high speed (0.06)
	-	-	-	Tethering and equipping (0.05)
	-	-	-	High and speed (0.22)
	-	-	-	Advantages and charges (0.06)
	-	-	-	Comfort and connections (0.01)

6.4 Transition Analysis

Figure 13 plots the transition in the co-occurrence rate from time-series analysis for the most related and selected pairs of terms (as topics) in Table 1. The topics behave differently. For example, although the pair of "area" and "expansion" has a high co-occurrence rate in 2009 (December), it decreases after that. "Area" and "readiness" behave the same way. One can understand such kinds of behaviours from Fig. 13.

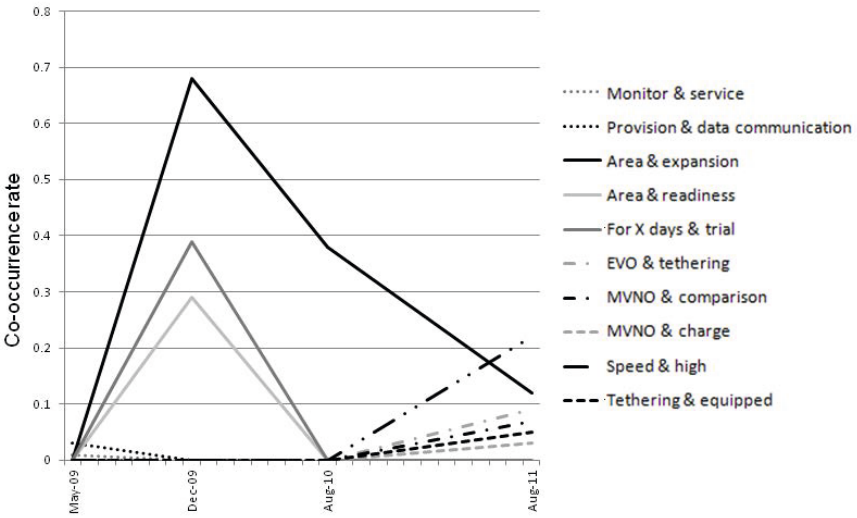


Fig. 13. Time-series data for topics

The reduced co-occurrence rate with "area" and "expansion" (or "area" and "readiness") can be explained by the effect of area expansion by the provider by combining terms with the news events summarized in Table 1. However, "speed" and "high" increased late in 2011 (August). This is because service features were clarified by service penetration. Moreover, the Mobile Virtual Network Operator (MVNO) and "comparison", and "MVNO" and "charge", had the same behaviours. This is because customers tried to compare services after several competitors appeared on the MVNO. Topics containing "tethering" also recently appeared. That situation was related to WiFi connections as news events.

Figure 14 plots the results for transition for pairs of terms that appeared continuously during certain periods. The trends could be analyzed by using patterns 1, 2, and 3, as seen in Fig. 10. The later behaviours of "high speed" and "mobile", and "high speed" and "communication" increased, which meant many customers recognized WiMAX to be a useful high speed mobile communication service. "mobile" and "PC", and "mobile" and "service" decreased, which meant mobile communications became popular and were broadly penetrating services. "UQ" and "service" always appeared because UQ-communication was the provider's name and only WiMAX provider during those periods.

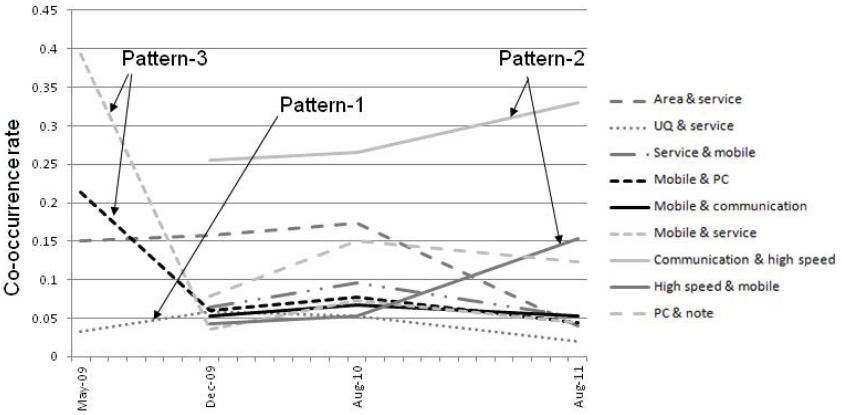


Fig. 14. Time-series data for pairs of continuously appearing terms

7 Conclusion

It is essential to identify customer perceptions and requirements for service diffusion. Customer perceptions/requirements for ICT services are continuously changing depending on new technologies and service operability. Therefore, it is necessary to assess customer perceptions and requirements by monitoring changes in their behaviours. Providers can take quick action in advance and gain high levels of customer satisfaction if they can identify sufficient customer perceptions/requirements. This chapter proposed a framework for service diffusion and a method of time-series analysis to solve these issues by uncovering customer perceptions/requirements in a new approach to marketing. The method worked on the basis of blogs and news events that acted as time-series data. The two types of trends are analyzed, one is for topics concerned with news events, and the other is for continuously appearing terms. Therefore, time-series changes in customer perceptions/requirements could be detected through the co-occurrence of texts by taking into account the structure of network services, and through the transition rate of relationships between words. The proposed method was applied to WiMAX by collecting and analyzing blog data. And obtained practical results show adequacy in the right of the fact that the service has been evolved.

Not only specific but also general enterprises with related services/networks (e.g. software/hardware vendors) benefit by applying this method, because blog data can be easily obtained.

It is intended that further studies will be done on methods of reading signs of customer perceptions/requirements in advance by analyzing textual data so that service providers can react quickly to changing customer perceptions/requirements.

References

1. The Ministry of Information and Communications: 2009 White Paper Information and Communications in Japan (2009)

2. Harte, L.: Introduction to 802.16 WiMAX. NTT Publishing, Tokyo (2007)
3. Srivastava, A., Sahami, M.: Text Mining: Classification, Clustering, and Applications. Chapman and Hall/CRC (2009)
4. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2007)
5. Sato, S., Fukuda, K., Sugawara, S., Kurihara, S.: On the relationship between word bursts in document streams and clusters in lexical co-occurrence networks. IPSJ 48(SIG14), 69–81 (2007)
6. Sullivan, D.: Document Warehousing and Text Mining. John Wiley (2001)
7. Toda, H., Kataoka, R., Kitagawa, H.: Clustering news articles using named entities. IPSJ, SIG TR, 2005-DBS-137, pp. 175–181 (2005)
8. Cutting, D., Kager, D., Tuky, J.: Scatter/gather: A cluster-based approach to browsing large document collection. In: Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 318–329 (1992)
9. Ho, X., Ding, C., Zha, H., Simon, H.: Automatic topic identification using webpage clustering. In: Proc. 2001 IEEE Int. Conf. on Data Mining, pp. 195–202 (2001)
10. Leuski, A.: Evaluating document clustering for interactive information retrieval. In: Proc. 2001 ACM Int. Conf. on Information and Knowledge Management, pp. 33–40 (2001)
11. Matsuo, Y., Ohsawa, Y., Ishizuka, M.: A Document as a small word. In: Terano, T., Nishida, T., Namatame, A., Tsumoto, S., Ohsawa, Y., Washio, T. (eds.) JSAI-WS 2001. LNCS (LNAI), vol. 2253, pp. 444–448. Springer, Heidelberg (2001)
12. Ohsawa, Y., Benson, N.E., Yachida, H.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. In: Proc. IEEE Forum on Research and Technology Advances in Digital Libraries, pp. 12–18 (1998)
13. Naganuma, K., Isonishi, T., Aikawa, T.: DIAMining: Text Mining Solution for Customer Relationship Management. Mitsubishi Technical Report 79(4), 259–262 (2005)
14. Rodoriguezd, M., Gomez-Iliadlgo, J., Diaz-Agudo, B.: Using wordnet to complement training information in text categorization. In: Proc. Recent Advances in Natural Language Proceedings, pp. 12–18 (1998)
15. Akiba, Y., Tanaka, T., Suyama, T., Nagata, M.: Grading Examinee's Answer Sentences by Verifying Syntactic and Semantic Compatibility. IPSJ SIG TR, 2006-NL-174(6), pp. 31-35 (2006)
16. Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., Harris, M.D.: Automated scoring using a hybrid feature identification technique. In: Proc. of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics, ACL-COLING 1998, pp. 206–210 (1998)
17. Takahashi, S., Takahashi, S., Yasuda, N., Takahata, N., Ishikawa, T.: A Meaningful Keywords Extracting System based on A Sentence-Semantic Analysis Method. IPSJ, AI TR, 90-8, pp. 65–72 (1992)
18. Taira, H., Mukouchi, T., Haruno, M.: Text Categorization Using Support Vector Machine. IPSJ, NL TR, 128-24, pp. 173–180 (1998)
19. Iwashita, M., Shimogawa, S., Nishimatsu, K.: Semantic analysis and classification method for customer enquiries in telecommunication services. Engineering Applications of Artificial Intelligence 24(8), 1521–1531 (2011)
20. Newman, M.: Power laws, pareto distributions and zipf's law. Contemporary Physics 46, 323–351 (2005)
21. Clauset, A., Shalizi, C.R., Newman, M.: Power-law distributions in empirical data. SIAM Rev. 51(4), 661–703 (2009)

Text-Shared Collaboration in Second Language Using Groupware for an Idea Generation

Takaya Yuizono and Zeying Yu

Japan Advanced Institute of Science and Technology,
1-1 Asahidai Nomi, Ishikawa 923-1292, Japan
{yuizono,yu}@jaist.ac.jp

Abstract. In the age of globalization, a second language can play an important role in collaboration between different countries. Text-based communication is slower than verbal communication, although its speed allows sufficient time to think which leads to an increased collaboration. Some effects of a second language using a groupware on the distributed and cooperative KJ method (the DC-KJ method) were investigated. The DC-KJ method is an arranged creative task referring to the KJ method developed by Jiro Kawakita well known in Japan. Post-it notes are used for writing and sharing ideas in the KJ method. The DC-KJ method consists of three steps: generating ideas by brainstorming, grouping ideas by concept formation, and framing a concluding statement from the previous steps. Thirty Chinese students took part in the experiments to investigate the effects. A group of three students carried out the collaboration task twice; one case used the Japanese language as a second language and the other case used the Chinese language as a native language. Those results were compared in terms of quantity and quality by means of a log data analysis, a questionnaire survey, and a writing satisfaction valuation as the final result. The results showed that (1) Chinese people using the Japanese language produced similar Chinese language usage result quantities and quality and (2) the ability to think 84 percent of opinions in the Japanese language were utilized to obtain those results. These results show a potential of text-shared collaboration with the groupware by using a second language.

Keywords: Groupware, Second Language, Creative Task, the KJ method, Chinese Participants, Japanese Language.

1 Introduction

Innovation of computer networks and air transport has increased the opportunity for people to collaborate with foreign people in using a second language. During human communication research, voice was better than text at accomplishing a problem-solving task by transmitting assigned information [2] and this result was suited to multimedia communication in similar group tasks, but not to creative tasks [3, 16]. It is not clear whether groupware in using the second language helps or disturbs a creative collaboration, while groupware has been expected to promote the creative collaboration. Therefore, it is aimed to understand the effects of groupware in using the second language to reveal the impact of groupware in the age of globalization.

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 56–69, 2013.

The use of Post-it notes (sticky notes) is commonly used during brainstorming sessions. Groupware research to assist intellectual production within text-shared collaboration used Post-it notes has been done in distributed computing environment. The KJ method with stickies (Post-It notes) is well known in Japan [7–9, 18] and is often applied to collaboration in Japanese organizations, and is partially known as the Affinity Diagram for a tool for a quality control [12] or a contextual design method [1, 4]. GUNGEN groupware supports the distributed and cooperative KJ method (the DC-KJ method), which is an arranged version of the original KJ method for groupware technology [14, 28]. The video and voice tools in GUNGEN, which are some of its multimedia conferencing tools, affected the text-chat communication used to determine participant state, but did not affect various other factors governing the output performance of the DC-KJ method, such as the number of opinions, groups, and characters of conclusion sentences [16]. This result shows some potential of the DC-KJ method as text-shared collaboration to help a reasonable collaboration in not rich communication, for example usage of second language.

In addition, the effects of groupware that supports the DC-KJ method were investigated by using the mother tongue only [14, 16]. It is necessary to investigate the effects of a second language because it could limit intellectual ability or greater understanding of the groupware technology in the globalization era. Recently, high economic progress in China has resulted in more frequent exchange with Japan, and then many Chinese have visited Japan as students. In this research, the effects of a second language on groupware were investigated using Chinese students.

In this research, related works are described in Section 2 and the experimental method used to investigate the effects of a second language are described in Section 3. The results of the experiments are shown and discussed in Section 4, while the conclusion to this research is presented in Section 5.

2 Related Work

Groupware is software technology to support a group sharing common task with computer network, and is expected to support cross-cultural communication in the age of globalization. In 2.1, groupware for an idea generation to support the KJ method is introduced. In 2.2, researches of cross cultural collaboration are overviewed in the view of two approaches.

2.1 Groupware for an Idea Generation and the KJ Method

Groupware is commonly used to promote inclusive dynamic environment that encourages spontaneous as creative generation of ideas. The tools used for supporting idea generation using the KJ method [7–9] has been researched since the 1990's in Japan. The groupware research realized visual editors for the KJ method, such as the KJ-Editor [17] and D-Abductor [11]. Another research supports a distributed form of group work based on the KJ method. GUNGEN (groupware for a new idea generation support system) [14, 28] and KUSANAGI [29, 30] was developed for supporting the DC-KJ method. These research studies aimed to develop and understand idea generation technology.

The DC-KJ method was adapted to the groupware work from the original KJ method by Munemori [14]. This task has three steps: generating ideas by brainstorming, grouping ideas by concept formation similarities, and framing a concluding statement from the previous steps. The creative problem solving was required to harmony of divergent thought to make many ideas and convergent thought to focus good ideas. The DC-KJ method included the two thoughts; brainstorming requires divergent thought and both grouping and writing require convergent thought [22]. The DC-KJ method was very similar to a task supported by Cognoter system [21] that was a part of the Colab project to demonstrate the potential of groupware technology for office workers. It means that the Cognoter task was very similar to the KJ method but the system did not clarify the performance of new idea generation. While GUNGEN was utilized in order to understand its performance by comparing with that in paper and pencil, and how the multimedia communication (video and text) affect the DC-KJ method [14, 16].

The effects of communication mode and distributed environments on the groupware for an idea generation are revealed in the above research studies. This research will focus on the effects of a second language for greater understanding of creative collaboration in the age of globalization.

2.2 Cross-Cultural Collaboration and Two Approaches

In 1990, Ishii pointed out that cross-cultural collaboration was an important issue in the groupware technology [6]. In the 21st century, the cross-cultural collaboration becomes more important in advanced globalization. In recent, the researches for the cross-cultural collaboration are divided into two approaches: focusing on language barriers and focusing on cultural differences.

The representative approach to focus on the language barriers is consideration of machine translation (MT). Ishida proposed and opened the Language Grid services that help building language resources by the Internet community [5]. The MT services have been applied to studies of cross-cultural collaboration. Yamashita found some difficulties to use MT in text chat communication and in problem solving task [27]. Wang found that machine MT helped non-native English speakers produce ideas in the brainstorming session but that both native and non-native English speakers viewed MT-mediated messages as less comprehensible than English messages [26]. Veinott found that video images were helpful to understand each other's sate for second language speaking pairs but not for native speaking pairs in negotiation tasks [23]. On the other hand, Munemori developed Pictograph Chat Communicator to support a statement with only pictograms as common language [13]. The system could support ice breaking between people from different countries but not consider creative tasks, problem solving tasks, and negotiation tasks.

The approach to focus on the cultural differences tended to consider the distinction between Western individualistic, low-context cultures and Eastern collectivistic, high-context cultures. Surveys of instant messaging found that multi-party chat, audio-video chat and emoticons were much more popular in Asia than in North America [10]. Setlock compared communication and performance of low-context American dyads, high- context Chinese dyads, and mixed American-Chinese dyads on a negotiation task [19, 20]. The negotiation task was done under two possible media conditions: audio

conferencing or video conferencing. The results showed no effects of culture or medium on conversational efficiency although they found communication styles in cultural differences, although Chinese participants used Chinese language or English language. On the other hand, Wang used the brainstorming task to understand the cross cultural collaboration [24–26]. He found that text-only medium reduced cultural differences in talkativeness. And he found that Chinese and Americans were conversationally more similar in Chinese-majority groups than American-majority groups, when they used rich media (video-enhanced chat room). The pictures were helpful to convert cultural diversity into a creative outcome.

The effects of MT and cultural differences to cross cultural collaboration are revealed on the problem solving task, the negotiation task, and the brainstorming task. This research will show the effects of the groupware in using a second language to the DC-KJ method that includes both divergent fought and convergent thought.

3 The Experimental Method

This section explains laboratory experiments to investigate the effects of second language using groupware for an idea generation. In 3.1, a procedure of the experiments is described. And then, in 3.2, a collaboration task for an idea generation and a case of the collaboration task with groupware KUSANAGI are introduced.

3.1 Procedure

In this research, participants carried out the DC-KJ method as text-shared collaboration. The details of this task with KUSANAGI [29, 30] groupware are described in next section. The participants in the experiments were all Chinese and masters students from Japan Advanced Science Institute of Technology, and they executed the cooperative task twice. They answered the questionnaire after each task. In addition, they were asked what language was used for each opinion and were required to translate their data in Chinese to data in Japanese.

There were thirty participants organized into ten groups of three participants. The discussion theme was entitled, “School environment for international students” labeled Theme 1 (T1) or “Sightseeing in Japan” labeled Theme 2 (T2), based on each participant’s interest. Each group performed a task twice and considered the two themes. The order of experiments in Japanese or Chinese and the two themes were alternated in order to counterbalance the tasks.

The experiment was conducted in a distributed environment in which they could not see each other’s faces. The environmental setup is shown in Figure 1. The communication media was text-chat communication only for the purpose of avoiding usage of the Chinese language when the Japanese language was used to accomplish the task.

3.2 Collaboration Task with KUSANAGI Groupware

KUSANAGI groupware was used for this experiments because it was available and supports the DC-KJ method, was utilized as groupware software for the task. KUSANAGI

Fig. 1. A shot of an experimental setup

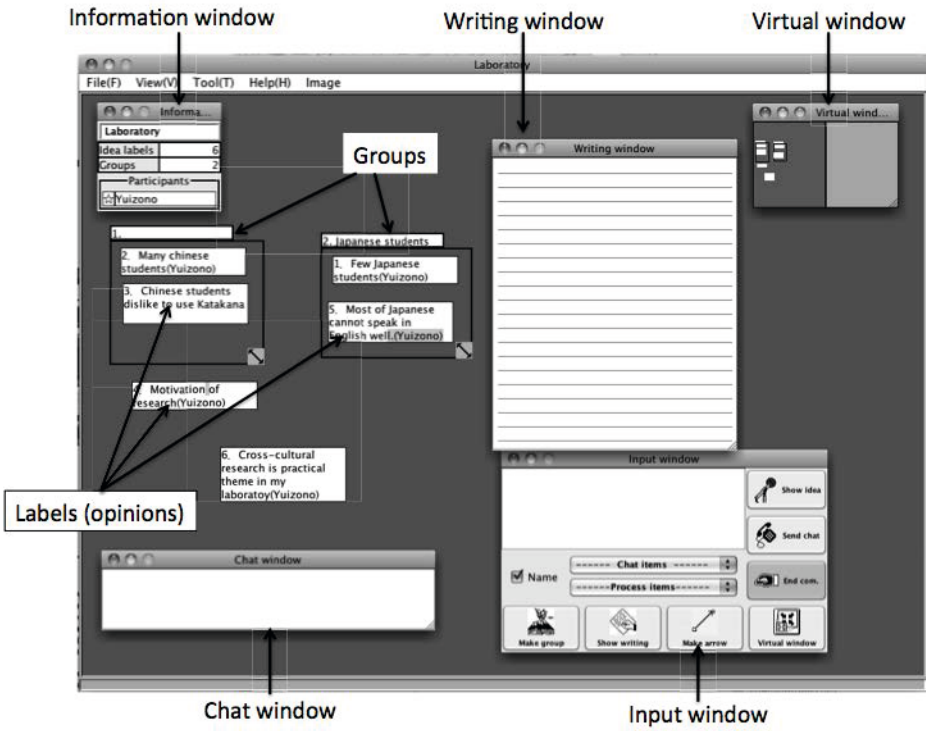


Fig. 2. Multi-user interfaces of KUSANAGI groupware

was developed with the Java Programming Language, which has multi-language (Japanese, Chinese, Korean) support with UNICODE characters. KUSANAGI supports the concept of showing an opinion as a label and grouping labels and naming each group on a shared screen. It also supports the framing of concluding statements on the shared window for writing. The multi-user interfaces of KUSANAGI are shown in Figure 2.

The operation of KUSANAGI through the DC-KJ method is described as follows. The first step was inputting idea opinions. Each participant freely input user's idea

opinions in the input window, shown on the lower right in Figure 2, and clicked the ‘Show idea’ button in the input window, and then the user could put the idea opinion as a label on the shared screen. The all idea opinions from them were shown in Figure 3. In the case, they showed 103 labels (idea opinions) before the next step. The case by them was output under the theme “Sightseeing in Japan.” The labels in the case were looked like chat conversations and sometimes participant commented on a previous idea opinion with new opinion. In the all experiments under this research, all groups were encouraged to show idea opinions as possible as they think. So, the limitation of experimental time was not set through the all three steps.

Fig. 3. A result of first step: inputting idea opinions using Japanese as the second language

The next step was grouping idea opinions. The result of grouping all ideas from the previous step was shown in Figure 4. The participants used twice screens with the virtual window interface and built eleven groups. The titles of groups were “6. Amusement park,” “3. Ocean’s bounty – attractive Ryukyu,” “9. Colored leaves,” “4. Trip and life philosophy,” “7. Desire for a trip in Japan,” “8. Snow country – desire for a trip in Hokkaido,” “8. Winter and a hot spring,” “11. Problem presentation – travel destinations,” “2. Budget consideration,” “10. About Ishikawa,” and “5. Gifu Prefecture and Lake Biwa.”

Each participant looked over the idea opinions on the shared screen and gathered associable idea opinions to make a group. If the participant wanted to add an idea opinion, the adding was allowed. In this case, five idea opinions were added in this grouping step. It is considered that the grouping should not be classification by known criteria,

Fig. 4. A result of second step: grouping idea opinions using Japanese as the second language

because idea generation of group is encouraged to find a new concept formation. In the KJ method, one idea opinion can become a group, if it can represent a specific concept in the all idea opinions. A user pushed the 'Make group' button and then the user put a group frame on the shared screen. If some labels were put into the group, the group could move with the labels and a concrete group name was made and represented on the contents of labels within the group.

The final step was writing the conclusion. The conclusion text from the previous two steps in the case was shown in Figure 5. The number of characters was 797 and the value of result was 4.1. The value of each conclusion text was evaluated by same three persons, not participants in the experiments, with the evaluation method for the result sentence of the KJ method proposed by Munemori [15]. A participant wrote the theme result in the writing window. In this step, each group name was encouraged to use as a keyword for writing. In this case, the text contained the all titles of groups. The native speaker in Japanese could understand its content but detected the text was written by non-native speaker in Japanese because there were strange usages on adposition and conjunction in the Japanese syntax.

日本の観光といえば、やはり留学生から最初に出されたのは北海道と沖縄でしょう、美しい景色や食文化が絶大の人気を誇る。日本は経済的発展しているとともに、伝統や自然も守りつつ、観光の非常にいい選択肢である。日本にいる留学生達も留学の便利を乗って、いろいろ自然の恵みと出会いことができる。日本人は些細ところまで人の気持ちを配慮して、どこに行っても不便が感じられない。
ただ、旅行といったら、食の豊かさはまだまだ満足でき無いところがいっぱいがある。種類豊富な料理もこれから欠かせないので、伝統を守りながら、日本の食はイノベーションをしなければならない。大都市は世界は大体同じと言い過ぎないから、出来るだけ京都あるいは金沢のような日本らしいさをたっぷり表現できるようなところが行ったほうが日本人の心の優しいさが感じられる。遊園地と言えば、富士急、USJ、ディズニーランドなどたくさんある。ただ待ち時間と刺激を受けやすいから、遊園地に苦手する人もいる。
私たちは都会の中華街、白川郷、琵琶湖や温泉のような地方の自然風景などをわたり、さまざまな場所へいきたいが、現実に旅行はやはり金銭につながる贅沢な活動、それでも私たちは旅行に対する憧れは衰えない、旅行を通じて美しい自然と暖かい人たちに出会えるからだ。が、石川県内で日帰り旅行で紅葉を鑑賞するなども良さそう。
いま、中国の富裕層の増加に伴う中国人の旅行者の急激的な日本へ、経済的には日本に対してとてもメリットだが、中国人の留学生としては、もっと中国人の観光客が日本の文化やマナーを尊重してもらいたいのである。短いに滞在により、日本人の方に迷惑かけるの避けるべきだし、そのうちに中国人として残ったの印象も考えなければならない。政治は敏感ですけど、友好交流はお互い良いところが見なければならない。
ですので、両国の文化交流や友好発展のために、これからは順調に祈っている。

Fig. 5. A result of last step: writing the conclusion using Japanese as the second language

4 Results and Discussion

As described in the previous section, the thirty Chinese students took part in the experiments to investigate the effects of second language using the groupware for an idea generation. The results are shown in 4.1 and discussed in 4.2.

4.1 Results

The results of each experiment using Japanese as the second language are shown in Table 1. These result data consist of the number of ideas, the number of groups and time required for grouping, the characters in a concluding statement and time required

Table 1. Results of each experiment using Japanese as the second language

		Inputting idea opinions		Making groups		Writing the conclusion		Total	
	Theme	Idea opinions	Time (min.)	Groups	Time (min.)	Written characters	Time (min.)	Value of results	Time (min.)
Ex-A	T2	132	70	3	18	353	29	2.0	116
Ex-B	T2	131	89	11	29	503	20	3.0	139
Ex-C	T2	108	72	11	35	797	28	4.1	136
Ex-D	T2	182	59	10	28	374	18	2.5	104
Ex-E	T2	100	85	8	29	816	37	3.9	151
Ex-F	T1	70	67	7	10	373	13	3.6	90
Ex-G	T1	122	56	8	18	431	37	2.3	111
Ex-H	T1	84	67	4	13	1271	61	4.6	140
Ex-I	T1	103	69	8	23	611	40	3.7	132
Ex-J	T1	108	82	5	22	582	19	3.9	124
Average		114.0	71.5	7.5	22.6	611.1	30.1	3.4	124.2

Fig. 6. A screenshot of KUSANAGI using Chinese as the native language

for writing, and the result values. A sample screen of a result obtained from the collaboration task using the Japanese language were shown in Figures 3, 4 and 5, performed by a group Ex-C in Table 1.

The results of each experiment using Chinese as the native language are shown in Table 2. These result data consist of the number of ideas, the number of groups and time required for grouping, the characters in a concluding statement and time required

Table 2. Results of each experiment using Chinese as the native language

		Inputting idea opinions		Making groups		Writing the conclusion		Total	
	Theme	Idea opinions	Time (min.)	Groups	Time (min.)	Written characters	Time (min.)	Value of results	Time (min.)
Ex-A	T1	137	56	3	21	274	18	1.9	95
Ex-B	T1	138	66	7	26	560	13	4.3	105
Ex-C	T1	151	80	9	27	659	17	3.6	124
Ex-D	T1	214	60	9	17	280	16	1.7	94
Ex-E	T1	126	66	6	17	940	21	4.3	103
Ex-F	T2	92	70	5	15	349	18	3	103
Ex-G	T2	147	66	13	20	400	22	4.2	108
Ex-H	T2	102	75	6	27	584	21	4.1	123
Ex-I	T2	122	61	6	32	437	33	2.5	126
Ex-J	T2	152	87	13	43	612	41	3.3	171
Average		138.1	68.8	7.7	24.5	509.5	22.0	3.3	115.2

Table 3. Comparison of the text-shared collaboration by Chinese using Japanese and Chinese

	Using Japanese (Second language)	Using Chinese (Native language)
Idea opinions	114.0	138.1
Time for idea opinion inputting (min.)	71.4	68.8
Groups	7.5	7.7
Time for grouping (min.)	22.6	24.5
Characters in conclusion sentences	611.1	509.5
Time for writing (min.)	30.1	22.0
Value of results	3.4	3.3
Total time (min.)	124.1	115.3
N	10	10

for writing, and the result values. A sample screen of a result obtained from the collaboration task using the Chinese language is shown in Figure 6.

These results are compared in Table 3. There were no differences in text-shared collaboration by Chinese students between the cases using Japanese as a second language and those using Chinese as the native language in a statistical Mann-Whitney U test analysis. There were no differences in the resulting values between the two conditions. Therefore, the content and performance by Chinese students in Japanese was not inferior to those in Chinese in terms of both quality and quantity.

After the group work, the participants answered the five-scale questionnaires about their interest in the theme, collaboration, and satisfaction rate with the result obtained in each experiment. The results are shown in Table 4, and the overall score is approximate to 4, which is more than a neutral score, which is ranked at 3. There were no differences between the cases using the Japanese language and those using the Chinese language in a statistical Mann-Whitney U test analysis, the same as in Table 3.

The participants felt that both they and their partners were interested in the theme and they were friendly with the partner, had good collaboration (inputting opinions,

Table 4. Comparison of the text-shared collaboration by Chinese using Japanese and Chinese

Questions	Using Japanese (Second language)	Using Chinese (Native language)
Participant's interest in the theme	4.2	3.9
Ease of expressing opinions	4.3	4.1
Success of conference	4.4	4.2
Success of inputting opinions	4.1	4.1
Success of grouping	4.1	4.2
Success of writing	4.0	4.0
Satisfaction with the result	4.1	4.2

grouping, and writing), and were satisfied with the results. In short, it seemed that participants had less trouble using the DC-KJ method.

4.2 Using Japanese as a Second Language

The participants were asked which language was used for thinking of each opinion, in Japanese or in Chinese, in the collaboration using the Japanese language. After the collaboration task, each participant reviewed all of own opinions on a hardcopy of a screen shot like that in Figure 4. A participant looked at the own opinion and then the person marked 'J' from a thought in the Japanese language, or marked 'C' from a thought in the Chinese language. The results are shown in Table 5. According to the participants, Chinese native speakers thought the 84 percent of all opinions in the Japanese language. The percentage of opinions thought in Japanese was checked for correlation with the written characters and the result values. The correlation value between the percentage and the characters was -0.48, and the correlation value between the percentage and the result values was -0.17. Therefore, the rate of thinking in Japanese did not directly influence the collaboration result. It is assumed that the ability to show the 84 percent of the opinions in the second language could lead to appropriate results in text-shared collaboration.

In 1990, Ishii expressed the opinion that e-mail was a good communication tool for non-native speakers because it allowed thinking [6]. Although text-based communication is slow in comparison to voice communication, text-based communication is recorded and readable on the shared screen. This property of text-shared collaboration with groupware technology is helpful for a second language collaborator.

The most researches [19, 23–27] described in the section 2.2 treated English as the second language, the common language and the native language. In the navigation task and the problem-solving task, the output in the second language as English has no overcoming that in the native language or some demerits [19, 20, 23, 27]. If the usage of Japanese as second language by Chinese is similar to that of English as second language by Chinese, the results in the DC-KJ method were applicable but the experiments in English as second language remain as future work.

The case of brainstorming was most relational case to this research, and the case with MT were helpful [26], so the use with MT in the DC-KJ method has potential to help the first step. But the effects with MT in the grouping and writing steps were not clear

Table 5. Opinions by thinking in Japanese

	Idea opinions	Thinking in Japanese		Thinking in Chinese		Written characters	Result values
Ex-A	132	102	77%	30	23%	353	2.0
Ex-B	70	57	81%	13	19%	503	3.0
Ex-C	131	123	94%	8	6%	797	4.1
Ex-D	122	114	93%	8	7%	374	2.5
Ex-E	108	80	74%	28	26%	816	3.9
Ex-F	84	76	90%	8	10%	373	3.6
Ex-G	182	168	92%	14	8%	431	2.3
Ex-H	103	69	67%	34	33%	1271	4.6
Ex-I	100	100	100%	0	0%	611	3.7
Ex-J	108	78	72%	30	28%	582	3.9
Average	114.0	96.7	84.0%	22.6	16.0%	611.1	3.4

from previous and this researches, so the DC-KJ method with MT remains as future work, too.

5 Concluding Remarks

The effects of using a second language with the DC-KJ method were investigated. Ten groups of three Chinese students carried out text-shared collaboration as remote intellectual collaboration using KUSANAGI groupware. The groups collaborated twice, using the Japanese language and the Chinese language, alternately.

The results showed that (1) Chinese people using the Japanese language in text-based group work with the groupware produced results of similar quantities and quality as in the usage of the Chinese language and (2) the ability to think 84 percent of opinions in the Japanese language as a second language could be required to obtain those results. These results mean that a groupware to support text-based collaboration is a potential tool to share and reflect ideas by a collaborator in using a second language.

In the future, experiments using Japanese and Chinese pairs will be carried out to consider the cultural effects, and not the linguistic effects. In addition, experiments using English language as a second language and those of the DC-KJ method with MT will be expected to explore advanced potential of the groupware to support text-based collaboration.

Acknowledgement. This research was partially supported by Japan Society for the Promotion of Science (JSPS) and the Grant-in-Aid for Scientific Research 24500143, 2012.

References

[1] Beyer, H., Holtzblatt, K.: Contextual design: defining customer-centered systems. Morgan Kaufmann Pub. (1998)

[2] Chapanis, A.: Interactive Human Communication. Scientific American 232, 36–42 (1975)
[3] Gale, S.: Human Aspects of Interactive Multimedia Communication. Interacting with Computers 2(2), 175–189 (1990)
[4] Harboe, G., Minke, J., Ilea, I., Huang, E.M.: Computer support for collaborative data analysis: augmenting paper affinity diagrams. In: Proceedings of CSCW 2012, pp. 1179–1182 (2012)
[5] Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration. In: Proceedings of SAINT 2006, pp. 96–100 (2006)
[6] Ishii, H.: Cross-Cultural Communication and Computer-supported cooperative work. Whole Earth Review, 48–49 (Winter 1990)
[7] Kawakita, J.: Idea Generation Method. Chuokoron-sha, Tokyo (1967) (in Japanese)
[8] Kawakita, J.: KJ Method. Chuokoron-sha, Tokyo (1986) (in Japanese)
[9] Kawakita, J.: The original KJ method. Kawakita Research Institute (1991)
[10] Kayan, S., Fussell, S.R., Setlock, L.D.: Cultural differences in the use of instant messaging in Asia and North America. In: Proceedings of CSCW 2006, pp. 525–528 (2006)
[11] Misue, K., Nitta, K., Sugiyama, K., Koshiba, T., Inder, R.: Enhancing D-ABDUCTOR towards a Diagrammatic User Interface Platform. In: Proceedings of KES 1998, pp. 359–368 (1998)
[12] Mizuno, S. (ed.): Management For Quality Improvement - The 7 New QC Tools. Productivity Press (1988)
[13] Munemori, J., Fukuda, T., Mohd Yatid, M.B., Nishide, T., Itou, J.: Pictograph Chat Communicator III: A Chat System That Embodies Cross-Cultural Communication. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part III. LNCS, vol. 6278, pp. 473–482. Springer, Heidelberg (2010)
[14] Munemori, J., Nagasawa, Y.: GUNGEN: Groupware for a New Idea Generation Support System. Inf. and Soft. Technology 38(3), 213–220 (1996)
[15] Munemori, J., Yagishita, K., Sudo, M.: Evaluation of an idea generation method and its supporting groupware. In: Proceedings of KES 1999, pp. 54–57 (1999)
[16] Munemori, J., Yuizono, T., Nagasawa, Y.: Effects of Multimedia Communication on GUNGEN (Groupware for an Idea Generation Support System). In: Proceedings of 1999 IEEE Inter-national Conference on Systems, Man, and Cybernetics, vol. II, pp. 196–201 (1999)
[17] Ohiwa, H., Takeda, N., Kawai, K., Shimomi, A.: KJ editor: A Card-Handling Tool for Creative Work Support. Knowledge-Based Systems 10, 43–50 (1997)
[18] Scupin, R.: The KJ Method: A Technique for Analyzing Data Derived from Japanese Ethnology. Human Organization 56(2), 233–237 (1997)
[19] Setlock, L.D., Quinones, P.A., Fussell, S.R.: Does culture interact with media richness? The effects of audio vs. video conferencing on Chinese and American dyads. In: Proceedings of HICSS 2007 (2007)
[20] Setlock, L.D., Fussell, S.R.: Culture or fluency? Unpacking interactions between culture and communication medium. In: Proceedings of CHI 2011, pp. 1137–1140 (2011)
[21] Stefik, M., Foster, G., Bobrow, D.G., Kahn, K., Lanning, S., Suchman, L.: Beyond The Chalkboard: Computer Support for Collaboration and Problem Solving in Meetings. Communications of ACM 30(1), 32–47 (1987)
[22] Treffinger, D.J., Isaksen, S.G., Stead-Dorval, K.B.: Creative Problem Solving, 4th edn. Prufrock Press Inc. (2006)
[23] Veinott, E., Olson, J., Olson, G., Fu, X.: Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other. In: Proceedings of CHI 1999, pp. 302–309 (1999)
[24] Wang, H.-C., Fussell, S.R., Setlock, L.D.: Cultural difference and adaptation of communication styles in computer-mediated group brainstorming. In: Proceedings of CHI 2009, pp. 669–678 (2009)

[25] Wang, H.-C., Fussell, S.R., Cosley, D.: From diversity to creativity: Stimulating group brainstorming with cultural differences and conversationally-retrieved pictures. In: Proceedings of CSCW 2011, pp. 265–274 (2011)
[26] Wang, H.-C., Fussell, S.R., Cosley, D.: Machine Translation vs. Common Language: Effects on Idea Exchange in Cross-Lingual Groups. In: Proceedings of CSCW 2011, pp. 265–274 (2013)
[27] Yamashita, N., Ishida, T.: Effects of Machine Translation on Collaborative Work. In: Proceedings of CSCW 2006, pp. 515–524 (2006)
[28] Yuizono, T., Munemori, J., Nagasawa, Y.: GUNGEN: groupware for a new idea generation consistent support system. In: Proceedings of APCHI 1998, pp. 357–362. IEEE Press (1998)
[29] Yuizono, T., Kayano, A., Shigenobu, T., Yoshino, T., Munemori, J.: Groupware for a New Idea Generation with the Semantic Chat Conversation Data. In: Khosla, R., Howlett, R.J., Jain, L.C. (eds.) KES 2005. LNCS (LNAI), vol. 3681, pp. 1044–1050. Springer, Heidelberg (2005)
[30] Yuizono, T., Jin, Z.: The Effects of Individual Differences in Two Persons on the Distributed and Cooperative KJ Method in an Anonymous Environment. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part III. LNCS, vol. 6278, pp. 464–472. Springer, Heidelberg (2010)

Capturing and Scaling Up Concurrent Transactions in Uncertain Databases

Alfredo Cuzzocrea[1], Hendrik Decker[2], and Francesc D. Muñoz-Escoí[2]

[1] ICAR-CNR and University of Calabria, I-87036 Cosenza, Italy
cuzzocrea@si.deis.unical.it
[2] Instituto Tecnológico de Informática, UPV, E-46022 Valencia, Spain
{hendrik,fmunyoz}@iti.upv.es

Abstract. This chapter provides *a framework for capturing and scaling up concurrent transactions in uncertain databases*. Models and methods proposed in the context of this framework for *managing data uncertainty* are innovative as previous studies have not considered the specific case of *concurrent transactions*, which may worsen the uncertainty of database management activities beyond the simplest case of *isolated transactions*. Indeed, as this chapter demonstrates, *inconsistency tolerance of integrity management*, *constraint checking* and *repairing* easily scale up to concurrent transactions in a natural way, and query answers in concurrent transactions over uncertain data remain *certain* in the presence of uncertainty. This analytical contribution is enriched by means of a *reference architecture* for uncertain database management under concurrent transactions that strictly adheres to models and methods that are the main contributions of this research.

Keywords: Uncertain Databases, Concurrent Transactions, Inconsistence Tolerance of Integrity Management.

1 Introduction

Uncertainty in databases is closely related to inconsistency and a lack of integrity. The validity of answers in inconsistent databases obviously is uncertain. Uncertainty of data can be modeled by integrity constraints that act as guards against undesirable properties of uncertainty. Thus, each constraint violation corresponds to some uncertainty in the database, no matter if the constraint models a regular integrity assertion or some specific uncertainty condition. This chapter addresses the mentioned relations between uncertainty and inconsistency.

For instance, the denial $\leftarrow$ *item(x, y), y* $< 75\%$ constrains entries *x* in the *item* table to have a probability (certainty) *y* of at least 75%. In a similar manner, the constraint $I = \leftarrow$ *uncertain(x)*, where uncertain is defined by the database clause *uncertain(x)* $\leftarrow$ *email(x, from(y)),* $\sim$*authenticated(y)*, bans each email message *x* that is uncertain because its sender *y* has not been authenticated. Likewise, uncertain could be defined, for instance, by *uncertain(x)* $\leftarrow$ *item(x, null)*, indicating an uncertainty about each item x the attribute of which has a null value.

An advantage of representing uncertainty by constraints is that the evolution of uncertainty across updates can then be monitored by inconsistency-tolerant methods for

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 70–85, 2013.

integrity checking, and uncertainty can then be eliminated by integrity repairing. For instance, each update U that tries to insert an email by a non-authenticated sender will be rejected by each method that checks U for integrity preservation, since U would violate I, in the preceding example. Ditto, stored email entries with unauthenticated senders or items with unknown attributes can be eliminated by repairing the violations of I in the database.

Conventional approaches to integrity management unrealistically require total constraint satisfaction before an update is checked and after a repair is done. However, methods for checking or repairing integrity or uncertainty must be inconsistency-tolerant as soon as data that violate some constraint are admitted to persist across updates. Decker and Martinenghi [22] have shown that the total consistency requirement can be waived without further ado for most (though not all) known methods. Thus, they can be soundly applied in databases with persistent constraint violations, for example with extant inconsistency and uncertainty.

Rather than pretending that consistent databases certainlt remain consistent across updates (as conventional methods do), inconsistency-tolerant methods just assure that inconsistency, that is, uncertainty is not increased, neither by updates nor by repairs. Such increase or decrease is determined by violation measures [20] (also called 'inconsistency metrics' [19]). Some of these measures also serve to provide answers that have integrity in the presence of uncertainty, by adopting an inconsistency-tolerant approach named *answers that have integrity* (Answers that Have Integrity (AHI)) [18].

Inconsistency tolerance also enables uncertainty management for concurrent transactions. For making any guarantees of integrity preservation across concurrent transactions, the usual requirement is that each transaction maps each consistent state to a consistent successor state. Unfortunately, that excludes any prediction for what is going to happen in the presence of constraint violations, that is, of uncertainty. However, it will become obvious that the inconsistency tolerance of integrity management easily scales up to concurrent transactions, and concurrent query answering with AHI remains certain in the presence of uncertainty.

After some preliminaries in Section 2, inconsistency-tolerant integrity management (checking, repairing and query answering) is recapitulated in Section 3. In Section 4, an example of how to manage uncertainty expressed by constraints is going to be elaborated. Section 5 outlines how inconsistency-tolerant constraint management scales up to database systems with concurrent transactions. In Section 6, a reference architecture for uncertain database management under concurrent transactions is provided, in which our approach can be embeded. In Section 7, related work is addressed. Section 8 contains concluding remarks and provides a glance of future research directions.

2 Formal Terminology and Definitions

In this paper, terminology and formalisms are those that are common for *datalog* [1]. Also, some familiarity with transaction concurrency control [6] is assuned. Throughout the chapter, symbols like *D, I, IC, U* are used for representing a database, an integrity constraint (in short, constraint), a finite set of constraints (also called *integrity theory*) and, resp., an update. The result of executing an update U on D is denoted by D^U, and the truth value of a sentence or a set of sentences S in D is denoted by $D(S)$.

Constraints often are asserted as denials, that is, clauses with empty head of the form $\leftarrow B$, where the body B is a conjunction of literals that state what should not be *true* in any state of the database. For each constraint I that expresses what should be *true*, a denial form of I can be obtained by re-writing $\leftarrow\sim I$ in clausal form [17]. Instead of leaving the head of denial constraints empty, a predicate that expresses some lack of consistency may be used in the head. For instance, the clause $uncertain \leftarrow B$ explicitly states an uncertainty that is associated to each instance of B that is *true* in the database.

3 Management of Uncertainty that Tolerates Inconsistency

As argued in Section 1, violations of constraints, that is, the inconsistency of given database states with their associated integrity theory, reflect uncertainty. Each update may violate or repair constraints, and thus increase or decrease the amount of uncertainty. Hence, checking updates for such increases, and decreasing uncertainty by reparing violated constraints, are essential for uncertainty management. Also mechanisms for providing answers that are certain in uncertain databases are needed. In Section 3.1, 3.2, 3.3, measure-based inconsistency-tolerance for integrity checking [19] is recapitulated and extended, as well as repairing [20] and, resp., query answering [18], all in terms of uncertainty.

3.1 Measure-Based Uncertainty-Tolerant Integrity Checking

The integrity constraints of a database are meant to be checked upon each update, which usually is committed only if it does not violate any constraint. Since total integrity is rarely achieved, and in particular not in databases where uncertainty is modeled by constraints, integrity checking methods that are able to tolerate uncertainty are needed.

Decker and Martinenghi [22] have formalized and discussed inconsistency-tolerant integrity checking. In particular, it has been shown that many (but not all) existing integrity checking methods tolerate inconsistency and thus uncertainty, although most of them have been designed to be applied only if all constraints are totally satisfied before any update is checked. As shown by Decker [19], integrity checking can be described by 'violation measures' [20], which are a form of inconsistency measures [27]. Such measures, called 'uncertainty measures' below, size the amount of violated constraints in pairs (*D, IC*). Thus, an update can be accepted if it does not increase the measured amount of constraint violations.

Definition 1. *A tuple $(\mu, \preccurlyeq)$ is caled an uncertainty measure (in short, a measure) if μ maps pairs (D, IC) to some metric space $(\mathbb{M}, \preccurlyeq)$ where $\preccurlyeq$ is a partial order, that is, a binary relation on $\mathbb{M}$ that is antisymmetric, reflexive and transitive. For two elements E, E' in $\mathbb{M}$, let $E \prec E'$ denote that $E \preccurlyeq E'$ and $E \neq E'$.*

Various axiomatic properties of uncertainty measures that go beyond Definition 1 have already been proposed [19,20,25,27]. Here, such axioms are not dealt with, since the large variety of conceivable inconsistency measures has been found to be "too elusive to be captured by a single definition" [25]. Moreover, several properties that are

standard in measurement theory [4] and that are postulated also for inconsistency measures in [25,27] do not hold for uncertainty measures, due to the non-monotonicity of database negation, as shown in [20].

Definition 2, below, captures each integrity checking method $\mathscr{M}$ (in short, method) as an I/O function that maps updates to $\{ok, ko\}$. The output *ok* means that the checked update is acceptable, and *ko* that it may not be acceptable. For deciding to *ok* or *ko* an update, $\mathscr{M}$ uses an uncertainty measure. That definition also captures what it means that a method is uncertainty-tolerant. In the remainder, UTIC is used as an acronym of *uncertainty-tolerant integrity checking*.

Definition 2. *An* integrity checking method $\mathscr{M}$ *maps triples (D,IC,U) to* $\{ok, ko\}$. *Each such method* $\mathscr{M}$ *is called sound (resp., complete) for uncertainty-tolerant integrity checking if there is an uncertainty measure* $(\mu, \preccurlyeq)$ *such that, for each (D,IC,U), (1) (resp., (2)) holds.*

$$\mathscr{M}(D,IC,U) = ok \Rightarrow \mu(D^U,IC) \preccurlyeq \mu(D, IC) \tag{1}$$

$$\mu(D^U,IC) \preccurlyeq \mu(D,IC) \Rightarrow \mathscr{M}(D,IC,U) = ok \tag{2}$$

If $\mathscr{M}$ is sound, it is also called a μ*-based Uncertainty-tolerant Integrity Checking (UTIC) method*.

The only real difference between conventional integrity checking and UTIC is that the former additionally requires total integrity before the update, that is, that $D(IC) = true$ in the premise of Definition 2. The range of the measure μ is used by conventional methods as the binary metric space, or $(\{true, false\}, \preccurlyeq)$, where $\mu(D,IC) = true$ means that *IC* is satisfied in *D*, given $\mu(D,IC) = false$ and that it is violated. Here, $true \prec false$, since, in each consistent pair (*D,IC*), there is a zero amount of uncertainty, which is of course less than the amount of uncertainty of each inconsistent pair (*D,IC*).

More differentiated uncertainty measures are given, for example, by comparing or counting the sets of instances of violated constraints, or the sets of 'causes' of inconsistencies. Such causes are characterized more precisely in 3.3.1, as the data whose presence or absence in the database is responsible for integrity violations [18,19]. Other violation measures are addressed by Decker [19].

In fact, many conventional methods can be turned into measure-based uncertainty-tolerant ones, simply by waiving the premise $D(IC) = true$ while comparing violations in (*D,IC*) and (D^U, *IC*) [19]. If there are more violations in (D^U,*IC*) than in (*D,IC*), they output *ko*; otherwise, they may output *ok*. According to Decker and Martinenghi [22], the acceptance of *U* by an uncertainty-tolerant method guarantees that *U* does not increase the set of violated instances of constraints.

More generally, the following result states that uncertainty can be monitored and its increase across updates can be prevented by each UTIC method, in as far as uncertainty is modeled in the syntax of integrity constraints.

Theorem 1. Let *D* be a database and *IC* an integrity theory that models uncertainty in *D*. Then, the increase of uncertainty in *D* by any update *U* can be prevented by checking *U* with any sound UTIC method.

3.2 Uncertainty-tolerant Integrity-preserving Repairs

In essence, repairs consist of updates that eliminate constraint violations [30]. However, hidden or unknown violations may be missed when trying to repair a database. Moreover, as known from repairing by triggers [10], updates that eliminate some violation may inadvertedly violate some other constraint. Hence, uncertainty-tolerant repairs are called for. Below, the definition of partial and total repairs in [22] is recapitulated. Such repairs are uncertainty-tolerant since some violations may persist after partial repairs. But they may not preserve integrity.

Definition 3. *For a triple (D,IC,U), let S be a subset of IC such that D(S) = false. An update U is called a repair of S in D if $D^U(S)$ = true. If $D^U(IC)$ = false, U is also called a partial repair of IC in D. Otherwise, if $D^U(IC)$ = true, U is called a total repair of IC in D.*

Example 1. Let $D = \{p(1,2,3), p(2,2,3), p(3,2,3), q(1,3), q(3,2), q(3,3)\}$ and $IC = \{\leftarrow p(x,y,z) \wedge \sim q(x,z), \leftarrow q(x,x)\}$. Clearly, both constraints are violated. $U = \{$delete $q(3,3)\}$ is a repair of $\{\leftarrow q(3,3)\}$ in D and a partial repair of IC. It tolerates the uncertainty reflected by the violation of $\leftarrow p(2,2,3) \wedge \sim q(2,3)$ in D^U. However, U also causes the violation of $\leftarrow p(3,2,3) \wedge \sim q(3,3)$ in D^U. Thus, the partial repair U' = {*delete* $q(3,3)$, *delete* $p(3,2,3)$} is needed to eliminate the violation of $\leftarrow q(3,3)$ in D without causing any other violation.

As illustrated in Example 1, there is a need to check if a given update or partial repair is integrity-preserving, that is, does not increase the amount of uncertainty. This problem is a generalization of what is known as repair checking [2]. The problem can be solved by UTIC, as stated in Theorem 2.

Theorem 2. Let $(\mu, \preccurlyeq)$ be an uncertainty measure, $\mathscr{M}$ a UTIC method based on $(\mu, \preccurlyeq)$, and U a partial repair of IC in D, for a tuple (D,IC). U preserves integrity wrt. μ, that is, $\mu(D^U,IC) \preccurlyeq \mu(D,IC)$, if $\mathscr{M}(D,IC,U) = ok$.

For computing partial repairs, any off-the-shelve view update method can be used, as follows. Let $S = \{\leftarrow B_1, \ \ldots \ , \leftarrow B_n\}$ be a subset of constraints to be repaired in a database D. Candidate updates for satisfying the view update request can be obtained by running the view update request *delete violated* in the database $D \cup \{violated \leftarrow B_i \mid 0 \leq i \leq n\}$. For deciding if a candidate update U preserves integrity, U can be checked by UTIC, according to Theorem 2.

3.3 Certain Answers in Uncertain Databases

Violations of constraints that model uncertainty may impair the integrity of query answering, since the same data that cause the violations may also cause the computed answers. Hence, there is a need of an approach to provide answers that either have integrity and thus are certain, while tolerating some uncertainty. An approach to provide answers that are certain in uncertain databases is outlined in 3.3.1, and generalized in 3.3.2 to provide answers that tolerate uncertainty.

3.3.1 Answers that are Certain

Consistent Query Answering (CQA) provides answers that are correct in each minimal total repair of *IC* in *D* [3]. CQA uses semantic query optimization [11] which in turn uses integrity constraints for query answering. A similar approach is to abduce consistent hypothetical answers, together with a set of hypothetical updates that can be interpreted as integrity-preserving repairs [24].

An approach to provide AHI and thus certainty has been proposed by Decker [18]. AHI determines two sets of data: the causes by which an answer is deduced, and the causes that lead to constraint violations. More precisely, for databases *D* and queries without negation in the body of clauses, causes are minimal subsets of ground instances of clauses in *D* by which positive answers or violations are deduced. For clauses with negation in the body and negative answers, also minimal subsets of ground instances of the only-if halves of the if-and-only-if completions of predicates in *D* [12] take part in causes.

An answer then is defined to have integrity if it has a cause that does not intersect with any of the causes of constraint violations, that is, if it is deducible from data that are independent of those that violate constraints. Definition 4 below is a compact version of the definition of AHI in terms of certainty. Decker [18,19,20] provides precise definitions of causes and details of computing AHI.

Definition 4. *Let θ be an answer to a query $\leftarrow B$ in (D, IC), where θ is either a substitution such that $D(\forall(B\theta)) = true$, or $\theta = no$, in which case $D(\leftarrow B) = true$. a) Let B_θ stand for $\forall(B\theta)$ if θ is a substitution, or for $\leftarrow B$ if $\theta = no$. b) θ is certain in (D,IC) if there is a cause C of B_θ in D, such that $C \cap C_{IC} = \emptyset$, where C_{IC} is the union of all causes of constraint violations in (D,IC).*

3.3.2 Answers that Tolerate Uncertainty

AHI is closely related to UTIC, since some convenient violation measures are defined by causes: cause-based methods accept an update *U* only if *U* does not increase the number or the set of causes of constraint violations [19]. Similar to UTIC, AHI is uncertainty-tolerant since it provides correct results in the presence of constraint violations. However, each answer accepted by AHI is independent of any inconsistent parts of the database, while UTIC may admit updates that violate constraints. For instance, *U* in Example 1 causes the violation of a constraint while eliminating some other violation. Now, suppose *U* is checked by some UTIC method based on a violation measure that assigns a greater weight to the eliminated violation than to the newly caused one. Thus, *U* can be *ok*-ed, since it decreases the measured amount of inconsistency.

In this sense, AHI is going to be relaxed to Answers That are Uncertain (ATU): answers that tolerate uncertainty. ATU sancions answers that are acceptable despite some amount of uncertainty involved in their derivation.

To quantify that amount, some 'tolerance measure' is needed. Unlike uncertainty measures which size the amount of uncertainty in all of (*D,IC*), tolerance measures only size the amount of uncertainty involved in the derivation of given answers or violations.

Definition 5. *(ATU)*
a) For answers θ to queries $\leftarrow B$ in (D,IC), a tolerance measure maps tripletes (D,IC, $B\theta$) to ($\mathbb{M}$, $\preceq$), where $\mathbb{M}$ is a metric space partially ordered by $\preceq$. b) Let th be a threshold value in $\mathbb{M}$ up to which uncertainty is tolerable. Then, an answer θ to some query $\leftarrow B$ in (D,IC) is said to tolerate uncertainty up to th if τ(D,IC,$B\theta$) $\preceq$ th.

A first, coarse tolerance measure τ could be to count the elements of $C\theta \cap C_{IC}$ where $C\theta$ is the union of all causes of Bθ and C_{IC} is as in Definition 4. Or, taking application semantics into account, a specific weight may be assigned to each element of each cause, similar to the tuple ranking in [5]. Then, τ can be defined by adding up the weights of elements in C$\theta \cap C_{IC}$. Another possibility is to define τ in terms of application-specific weights that are assigned to each ground instance *I'* of each *I* in *IC*. Then, τ would sum up the weights of those *I'* that have a cause *C'* such that $C\theta \cap C' \neq \emptyset$.

For example $\tau(D, IC, B_\theta) = |\, C\theta \cap C_{IC}\,|$counts elements in C$\theta \cap C_{IC}$, where $|\cdot|$ is the cardinality operator. Or, $\tau(D, IC, B_\theta) = \sum\{\omega(c) \mid c \in C_\theta \cap C_{IC}\}$ adds up the weights of elements in $C_\theta \cap C_{IC}$, where ω is a weight function.

4 Uncertainty Management in Databases with Concurrent Transactions – An Example

In this section, the management of uncertainty by inconsistency-tolerant integrity management is described, and also some more conventional alternatives are discussed. In particular, uncertainty-tolerant integrity management is going to be compared with brute-force constraint evaluation, conventional integrity checking that is not uncertainty-tolerant, total repairing, and CQA, in 4.1 – 4.6.

The predicates and their attributes below are open to interpretation. By assigning convenient meanings to predicates, it can be interpreted as a model of uncertainty in a decision support systems for, for example, stock trading, or controlling operational hazards in a complex machine.

Let *D* be a database with the following definitions of view predicates *ul*, *um*, *uh* that model uncertainty of low, medium and, respectively, high degree:

$ul(x) \leftarrow p(x,x)$

$um(y) \leftarrow q(x, y), \sim p(y, x);$ $\quad um(y) \leftarrow p(x,y), q(y, z), \sim p(y, z), \sim q(z, x)$

$uh(z) \leftarrow p(0, y), q(y, z), z > th$

where *th* is a threshold value greater or equal 0. Now, let uncertainty be denied by the following integrity theory: $IC = \{\leftarrow ul(x), \leftarrow um(x), \leftarrow uh(x)\}$.

Note that *IC* is satisfiable, for example, by $D = \{p(1, 2), p(2, 1), q(2, 1)\}$. Hence, the extensions of *p* and *q* in *D* be populated with the following facts.

$p(0,0),\ p(0,1),\ p(0,2),\ p(0,3),\ \ldots,\ p(0, 10000000),$
$p(1,2),\ p(2,4),\ p(3,6),\ p(4,8),\ \ldots,\ p(5000000,\ 10000000),$
$q(0,0),\ q(1,0),\ q(3,0), q(5,0), q(7,0),\ \ldots,\ q(9999999, 0)$

It is easy to verify that the low-uncertainty denial $\leftarrow ul(\mathrm{x})$ is the only constraint that is violated in D, and that this violation is caused by $p(0,0) \in D$. Next, consider the update results of $U = insert\ q(0, 9999999)$.

4.1 Brute-Force Uncertainty Management

For the purpose of comparison, the general cost of a brute-force evaluation of IC in D^U is analyzed in this subsection. Evaluating $\leftarrow ul(\mathrm{x})$ involves a full scan of p. The process of evaluating $\leftarrow um(\mathrm{x})$ involves access to the whole extension of q, a join of p with q, and possibly many lookup transaction for p and q in order to test the negative literals. Evaluating $\leftarrow uh(\mathrm{x})$ involves a join of p with q plus the evaluation of possibly many ground instances of $z > th$.

For large extensions of p and q, brute-force evaluation of IC clearly may last too long, in particular for safety-critical uncertainty monitoring in real time. In Section 4.2, the reader is going to see that it is far less costly to use an UTIC method that simplifies the evaluation of constraints by confining its focus on the data that are relevant for the update.

4.2 Uncertainty Management by UTIC

First of all, note that the use of customary methods that require the satisfaction of IC in D is not feasible in our example, since $D(IC) = false$. Thus, conventional integrity checking has to resort on brute-force constraint evaluation. The reader is going to see in this subsection that checking U by an UTIC method is much less expensive than brute-force evaluation.

At update time, the following simplifications of medium and high uncertainty constraints are obtained from U. (Here, no low uncertainty is caused by U, since $q(0, 9999999)$ does not match $p(x, x)$.) The simplifications displayed below are obtained at hardly any cost, by simple pattern matching of U with pre-simplified constraints that can be compiled at constraint specification time.

$\leftarrow \sim p(9999999, 0)$
$\leftarrow p(x, 0), \sim p(0, 9999999), \sim q(9999999, x)$
$\leftarrow p(0, 0), 9999999 > th$

By a simple lookup of $p(9999999, 0)$ for evaluating the first of the three denials, it is inferred that $\leftarrow um$ is violated. Now that a medium uncertainty has been spotted, there is no need to check the other two simplifications. Yet, let us do that, for later comparison in Section 4.3.

Evaluation of the second simplification from left to right essentially equals the cost of computing the answer $x = 0$ to the query $\leftarrow p(x,0)$ and successfully looking up $q(9999999,0)$. Hence, the second denial is *true*. Thus, there is no further medium uncertainty. Clearly, the third simplification is violated if $9999999 > th$ holds, since $p(0,0)$ is *true*, then the possibilitiy of U may cause high uncertainty.

Summarizing this subsection, it can be observed that the cost of validating U using the UTIC method according to Theorem 1 essentially consists in a simple access to the

p relation. Only one additional look-up is needed for evaluating all constraints. Thus, apart from a significant cost reduction, UTIC prevents medium and high uncertainty constraint violations that would be caused by U if it were not rejected.

4.3 Methods That are Uncertainty-Intolerant

UTIC is sound. In general, however, methods that are uncertainty-intolerant, that is, not uncertainty-tolerant (such as in Gupta et al. [26] and Lee [28]), are unsound. That is shown below.

Obviously, p is not affected by U. Thus, $D(ul(x)) = D^U(ul(x))$. Each integrity checking method that is uncertainty-intolerant assumes $D(IC) = true$. Thus, the method in [26] concludes that the unfolding $\leftarrow p(x, x)$ of $\leftarrow ul(x)$ is satisfied in D and D^U. Thus, it infers that also $\leftarrow p(0, 0)$, $9999999 > th$ (the third of the simplifications in 4.2) is satisfied in D^U. However, that is wrong if $9999999 > th$ holds. Hence, uncertainty-intolerant integrity checking may wrongly infer that the high uncertainty constraint $\leftarrow uh(z)$ cannot be violated in D^U.

4.4 Uncertainty Management by Repairing (D, IC)

Conventional integrity checking requires $D(IC) = true$. To comply with that, all violations in (D,IC) must be repaired before each update. However, such repairs can be exceedingly costly, as argued below. Especially when the identification of all violations in (D,IC) may be prohibitively costly at update time. But there is only a single low uncertainty constraint violation in our example: $p(0,0)$ is the only cause of the violation $\leftarrow ul(0)$ in D. Thus, to begin with repairing D means to request $U = delete\ p(0,0)$, and to execute U if it preserves all constraints, according to Theorem 2.

To check U for integrity preservation means to evaluate the simplifications $\leftarrow q(0, 0)$ and $\leftarrow p(x,0)$, $q(0, 0)$, $\sim q(0,x)$, then the two resolvents of $\sim p(0,0)$ and the clauses defining um, since U affects no other constraints. The second one is satisfied in D^U, since there is no fact matching $p(x, 0)$ in D^U. However, the first one is violated, since $D^U(q(0,0)) = true$. Hence, also $q(0, 0)$ must be deleted. That deletion affects the clause $um(y) \leftarrow p(x, y), q(y, z), \sim p(y, z), \sim q(z, x)$ and yields the simplification $\leftarrow p(0, y), q(y, 0), \sim p(y, 0)$.

As is easily seen, this simplification is violated by each pair of facts of the form $p(0,o), q(o,0)$ in D, where o is an odd number in $[1, 9999999]$. Thus, deleting $q(0, 0)$ for repairing the violation caused by deleting $p(0, 0)$ causes the violation of each instance of the form $\leftarrow um(o)$, for each odd number o in $[1, 9999999]$.

Hence, repairing each of these instances would mean to request the deletion of many rows of p or q. those deletions shall not be tracked down any further, since it should be clear already that repairing D is complex and tends to be significantly more costly than UTIC. Another advantage of UTIC: since inconsistency can be temporarily tolerated, UTIC-based repairs do not have to be done at update time. Rather, they can be done off-line, at any convenient point of time.

4.5 Uncertainty Management by Repairing (D^U, IC)

Similar to repairing (*D, IC*), repairing (D^U, *IC*) also is more expensive than to tolerate extant constraint violations until they can be repaired at some more convenient time. That can be illustrated by the three violations in D^U, as identified in 4.1 and 4.2: the low uncertainty that already exists in *D*, the medium and high uncertainties caused by *U*. To repair them obviously is even more intricate than to only repair the first of them, as tracked in 4.4.

Moreover, for uncertainty management in safety-critical applications, it is not a good idea to simply accept an update without checking for potential violations of constraints, and to attempt repairs only after the update is committed, since repairing takes time, during which an updated but unchecked state may contain possibly very dangerous uncertainty.

4.6 AHI and ATU for Uncertainty Management

Checking and repairing uncertainty constraints involves their evaluation, by querying them. As already mentioned in 3.3.1, CQA is an approach to cope with constraint violations for query evaluation. However, the evaluation of constraints or simplifications thereof by CQA is unprofitable, since consistent query answers are defined to be those that are *true* in each minimally repaired database. Thus, for each queried denial constraint *I*, CQA will by definition return the empty answer, which indicates the satisfaction of *I*. Thus, answers to queried constraints computed by CQA have no meaningful interpretation.

For example, CQA computes the empty answer to the query $\leftarrow ul(x)$ and to $\leftarrow uh(z)$, for any extension of *p* and *q*. However, the only reasonable answers to $\leftarrow ul(x)$ and $\leftarrow uh(z)$ in *D* are $x = 0$ and, resp., $x = 9999999$, if $9999999 > th$. These answers correctly indicate low and high uncertainty in *D* and, resp., D^U.

For computing correct answers to queries (rather than to denials representing constraints), AHI and ATU are viable alternatives to CQA. Decker [18] has sketched a comparison that turned out to be advantageous for AHI. ATU goes beyond CQA and AHI by providing reasonable answers even if these answers depend on uncertain data that violate constraints, as already seen in 3.3.2.

5 Scaling Up Uncertainty Management to Concurrency

The number of concurrently issued transactions increases with the number of online users. Up to this point in the text, only serial executions of transactions have been considered. Such executions have many transactions wait for others to complete. Thus, the serialization of transactions severely limits the scalability of applications. Hence, to achieve high scalability, transactions should be executed concurrently, without compromising integrity, that is, without increasing uncertainty.

Standard concurrency theory guarantees the preservation of integrity only if each transaction, when executed in isolation, translates a consistent state into a consistent successor state. More precisely, a standard result of concurrency theory says that, in a

history H of concurrently executed transactions $T_1, \ldots, T_n$, each T_i preserves integrity if it preserves integrity when executed non-concurrently and H is serializable, that is, the effects of the transactions in H are equivalent to the effects of a serial execution of $\{T_1, \ldots, T_n\}$. This can be captured using Eq. 1:

$$\textit{isolated integrity} + \textit{serializability} \Rightarrow \textit{concurrent integrity}\ (*) \qquad (1)$$

If uncertainty corresponds to integrity violation, and each transaction is supposed to operate on a consistent input state, then (*) does not guarante that concurrently executed transactions on uncertain data would keep uncertainty at bay, even if they would not increase uncertaity when executed in isolation and the history of their execution was serializable.

Fortunately, this approach and the results provided in Section 3 can be easily scaled up using concurrent transactions, as shown for inconsistency-tolerant integrity checking by Decker and Muñoz-Escoí [23]. This methodology is based on a mesure that compares sets of violated instances of constraints before and after a transaction.

Theorem 3 adapts the result by Decker and Muñoz-Escoí [23] to measure-based UTIC in general. It asserts that a transaction T in a history H of concurrently executing transactions the do not increase uncertainty if H is serializable. T preserves this integrity whenever it is executed in isolation. On one hand, this research weakens their theorey by assuming a Strict Two-Phase Locking (S2PL), rather than abstracting away from any implementation of serializability [23]. On the other hand, a generalized theorem is provided by using the arbitrary uncertainty measure μ, instead of the inconsistency measure mentioned above. A full-fledged generalization that would not assume any particular realization of serializability is possible using Decker and Muñoz-Escoí [23], but would be out of proportion in this chapter.

Theorem 3. Let H be a S2PL history, μ an uncertainty measure and T a transaction in H that uses a μ-based UTIC method for checking the integrity preservation of its write operations. Further, let D be the committed state at which T begins in H, and D^T the committed state at which T ends in H. Then, $\mu(D,IC) \preccurlyeq \mu(D^T,IC)$.

The essential difference between (*) and Theorem 3 is that the latter is uncertainty-tolerant, the former is not. Thus, as opposed to (*), Theorem 3 identifies useful sufficient conditions for integrity preservation in the presence of uncertain data. Another important difference is that the guarantees of integrity preservation that (*) can make for T require the integrity preservation of all other transactions that may happen to be executed concurrently with T. As opposed to that, Theorem 3 does away with the standard premise of (*) that all transactions in H must preserve integrity in isolation; only T itself is required to have that property. Thus, the guarantees that Theorem 3 can make for individual transactions T are much better than those of (*).

To outline a proof of Theorem 3, the cases that T either terminates by aborting or by committing its write operations are distinguished. If T aborts, then Theorem 3 holds vacuously, since, by definition, no aborted transaction could have any effect whatsoever on any committed state. So, it can be supposed that T commits. Let $\mathscr{M}$ be the μ-based method used by T. Since T commits, it follows that $\mathscr{M}(D,IC,WT) = \textit{true}$, where WT

is the write set of T, that is, $D^T = D^{WT}$, since otherwise, the writes of T would violate integrity and thus T would abort. Since H is S2PL, it follows that there is an equivalent serialization H' of H that preserves the order of commited states in H. Thus, D and D^T are also the committed states at beginning and end of T in H'. Hence, Theorem 3 follows from $\mathscr{M}(D,IC,WT) = true$ and Definition 2 since H' is serial, that is, non-concurrent. D It follows from Theorem 2 that, similar to UTIC, also integrity repairing scales up to S2PL concurrency if realized as described in 3.2, that is, if UTIC is used to check candidate repairs for integrity preservation.

Also AHI and ATU as defined in 3.3 scale up to concurrency, which can be seen as follows. Concurrent query answering is realized by read-only transactions. In S2PL histories, such transactions always read from committed states that are identical to states in equivalent serial histories, as described in the proof of Theorem 3. Hence, each answer can be checked for certainty or for being within in the confines of tolerable uncertainty as described in 3.3.1 or, resp., 3.3.2.

6 A Reference Architecture for Uncertain Database Management under Concurrent Transactions

Figure 1 depicts a modular reference architecture for uncertain database management under concurrent transactions, integrated with a classical Data-Base Maangement System (DBMS) architecture. It illustrates the main result of our research.

The modules shown in Figure 1 are described as follows, bottom to top:

Physical Layer: Physical Layer: This is the basic layer of classical DBMS, where data are stored according to some *storage scheme* (for example, based on fixed record length).

Access Methods: This layer contains the collection of access methods needed to retrieve data from the physical layer as a reaction to the execution of standard SQL statements that occur in the relational layer.

Relational Layer: This is the relational layer of classical DBMS, where tuples are processed according to the relational data model.

Uncertainty Detection Layer: This is the layer where data uncertainty is detected, according to a given uncertain data model (for example, *probabilistic uncertainty model*).

Uncertainty Management Layer: This is the layer where data uncertainty is managed by using the elements (denoted by *u_mi*(.,.,.)) of an integrated approach to uncertainty management, such as a combination of UTIC and ATU, as proposed in Section 3 and Section 4.

Relaxed Concurrency Protocol Layer: In this layer, concurrency is handled in an inconsistency-tolerant manner, as proposed in Section 3 and Section 4, such that sequentializable histories of concurrent transactions can proceed without rollbacks that would be due to extant uncertainty.

Concurrent Transaction Layer: This is the layer where concurrent transactions (modeled in terms of *directed graphs*) occur, for instance in the application scenario of an e-commerce web database system.

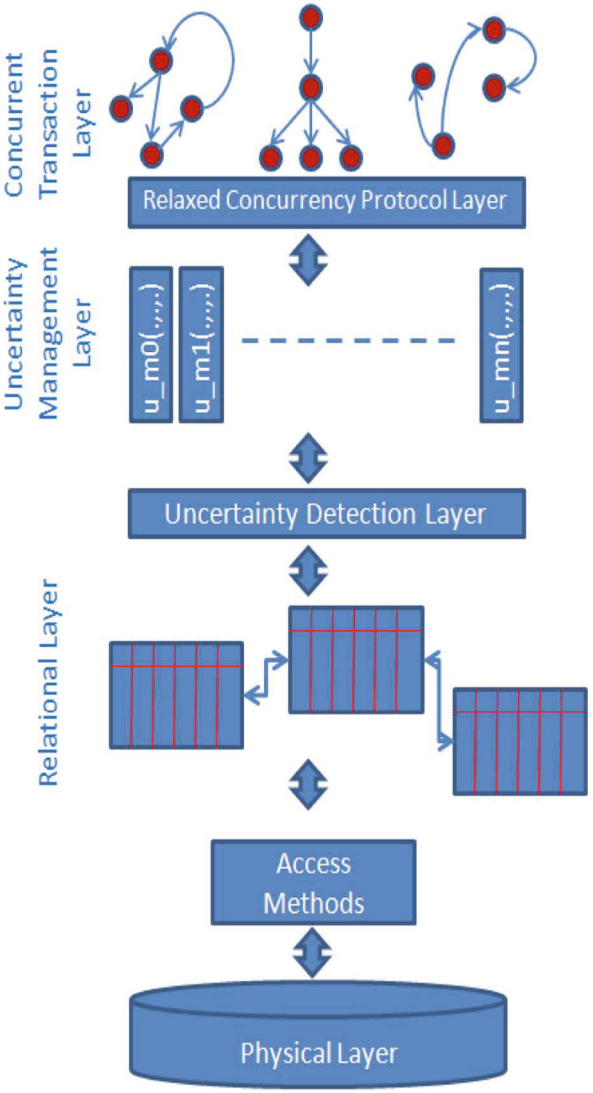


Fig. 1. Reference Architecture for Uncertain Database Management Under Concurrent Transactions

7 Related Work

An early, not yet measure-based attempt to conceptualize some of the material in 3.1 has been made by Decker and Martinenghi [21]. Apart from that, it seems that integrity maintenance and query answering in the presence of uncertain data never have been approached in a uniform way, as in this chapter. That is surprising since integrity, uncertainty

Semantic similarities and differences between uncertainty and the lack of integrity are observed in Motro and Smets [29]. In that book, largely diverse proposals to handle data that suffer from uncertainty are discussed. In particular, approaches such as probabilistic and fuzzy set modeling, exception handling, repairing and para-consistent reasoning are discussed. However, no particular approach to integrity maintenance (checking or repairing) is considered. Also, no attention is paid to concurrency.

Several paraconsistent logics that tolerate inconsistency and thus uncertainty of data have been proposed, for example, in Bertossi et al. and Carnielli et al. [7,9]. Each of them departs from classical first-order logic, by adopting some annotated, probabilistic, modal or multi-valued logic, or by replacing standard axioms and inference rules with non-standard axiomatizations. As opposed to that, UTIC fully conforms with standard datalog and does not need any extension of classical logic.

Work concerned with semantic inconsistencies in databases is also going on in the field of measuring inconsistency [27]. However, the violation measures on which UTIC is based have been conceived to work well also in databases with non-monotonic negation, whereas the inconsistency measures in the literature do not scale up to non-monotonicity, as argued by Decker [20].

8 Conclusions and Future Work

In this paper, an extension of recently developed concepts of logical inconsistency tolerance has been applied to problems of managing uncertainty in databases. It has been shown that the uncertainty of stored data can be modeled by integrity constraints and maintained by uncertainty-tolerant integrity management technology. In particular, updates can be monitored by UTIC, such that they do not increase uncertainty, and extant uncertainty can be partially repaired while tolerating remaining uncertainty. Also, a concept of how databases can provide reasonable answers in the presence of uncertainty has been developed. Moreover, the paper has exposed that uncertainty tolerance is necessary and sufficient for scaling up uncertainty management in databases to concurrent transactions. This result is significant since concurrency is a common and indeed indispensable feature of customary database management systems.

As illustrated in Section 4, the use of uncertainty-tolerant methods is essential, since wrong, possibly fatal conclusions can be inferred from deficient data by using a method that is uncertainty-intolerant. Decker and Martinenghi [22] and Decker [19] have featured a lot of UTIC methods, including some intolerant ones.

In ongoing research, the authors are elaborating a generalization of the results in Section 5 to arbitrary serializable histories. Also, the authors are working on a further scale-up of their results to replicated databases and recoverable histories. Moreover, further applications of inconsistency-tolerant uncertainty management are envisaged in the

fields of OLAP, data mining, and data stream query processing, in order to complement previous work by Cuzzocrea [8,13,14,15,16].

Acknowledgement. The second and the third author have been supported by FEDER and the Spanish grants TIN2009-14460-C03, TIN2010-17139.

References

1. Abiteboul, S., Hull, R., Vianu, V.: Foundations of Databases. Addison-Wesley (1995)
2. Afrati, F., Kolaitis, P.: Repair checking in inconsistent databases: algorithms and complexity. In: Proc. 12th ICDT, pp. 31–41. ACM Press (2009)
3. Arenas, M., Bertossi, L.E., Chomicki, J.: Consistent query answers in inconsistent databases. In: Proceedings of PODS, pp. 68–79. ACM Press (1999)
4. Bauer, H.: Maß- und Integrationstheorie, 2nd edn. De Gruyter (1992)
5. Berlin, J., Motro, A.: TupleRank: Ranking discovered content in virtual databases. In: Etzion, O., Kuflik, T., Motro, A. (eds.) NGITS 2006. LNCS, vol. 4032, pp. 13–25. Springer, Heidelberg (2006)
6. Bernstein, P.A., Hadzilacos, V., Goodman, N.: Concurrency Control and Recovery in Database Systems. Addison-Wesley (1987)
7. Bertossi, L., Hunter, A., Schaub, T. (eds.): Inconsistency Tolerance. LNCS, vol. 3300. Springer, Heidelberg (2005)
8. Budhia, B.P., Cuzzocrea, A., Leung, C.K.: Vertical frequent pattern mining from uncertain data. In: KES, pp. 1273–1282 (2012)
9. Carnielli, W., Coniglio, M., D'Ottaviano, I. (eds.): The Many Sides of Logic. Studies in Logic, vol. 21. College Publications, London (2009)
10. Ceri, S., Cochrane, R., Widom, J.: Practical applications of triggers and constraints: Success and lingering issues (10-year award). In: Abbadi, A.E., Brodie, M.L., Chakravarthy, S., Dayal, U., Kamel, N., Schlageter, G., Whang, K.Y. (eds.) Proceedings of 26th International Conference on Very Large Data Bases, VLDB 2000, Cairo, Egypt, September 10-14, pp. 254–262. Morgan Kaufmann (2000)
11. Chakravarthy, U.S., Grant, J., Minker, J.: Logic-based approach to semantic query optimization. ACM Trans. on Database Syst. (TODS) 15(2), 162–207 (1990)
12. Clark, K.: Negation as failure. In: Gallaire, H., Minker, J. (eds.) Logic and Data Bases, pp. 293–322. Plenum Press (1978)
13. Cuzzocrea, A.: Olap over uncertain and imprecise data: Fundamental issues and novel research perspectives. In: Proc. 21st DEXA Workshop, pp. 331–336. IEEE CSP (2001)
14. Cuzzocrea, A.: Retrieving accurate estimates to OLAP queries over uncertain and imprecise multidimensional data streams. In: Bayard Cushing, J., French, J., Bowers, S. (eds.) SSDBM 2011. LNCS, vol. 6809, pp. 575–576. Springer, Heidelberg (2011)
15. Cuzzocrea, A., Decker, H.: Non-linear data stream compression: Foundations and theoretical results. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part III. LNCS, vol. 7208, pp. 622–634. Springer, Heidelberg (2012)
16. Cuzzocrea, A., Gunopulos, D.: Efficiently computing and querying multidimensional OLAP data cubes over probabilistic relational data. In: Catania, B., Ivanović, M., Thalheim, B. (eds.) ADBIS 2010. LNCS, vol. 6295, pp. 132–148. Springer, Heidelberg (2010)
17. Decker, H.: The range form of databases and queries or: How to avoid floundering. In: Proc. 5th ÖGAI. Informatik-Fachberichte, vol. 208, pp. 114–123. Springer (1989)
18. Decker, H.: Answers that have integrity. In: Schewe, K.-D. (ed.) SDKB 2010. LNCS, vol. 6834, pp. 54–72. Springer, Heidelberg (2011)

19. Decker, H.: Inconsistency-tolerant integrity checking based on inconsistency metrics. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS, vol. 6882, pp. 548–558. Springer, Heidelberg (2011)
20. Decker, H.: Measure-based inconsistency-tolerant maintenance of database integrity. In: Schewe, K.-D., Thalheim, B. (eds.) SDKB 2013. LNCS, vol. 7693, pp. 149–173. Springer, Heidelberg (2013)
21. Decker, H., Martinenghi, D.: Integrity checking for uncertain data. In: Proc. 2nd TDM Workshop on Uncertainty in Databases. CTIT Workshop Proceedings Series, vol. WP06-01, pp. 41–48. Univ. Twente, The Netherlands (2006)
22. Decker, H., Martinenghi, D.: Inconsistency-tolerant integrity checking. Transactions on Knowledge and Data Engineering 23(2), 218–234 (2011)
23. Decker, H., Muñoz-Escoí, F.D.: Revisiting and improving a result on integrity preservation by concurrent transactions. In: Meersman, R., Dillon, T., Herrero, P. (eds.) OTM 2010. LNCS, vol. 6428, pp. 297–306. Springer, Heidelberg (2010)
24. Fung, T.H., Kowalski, R.: The iff proof procedure for abductive logic programming. J. Logic Programming 33(2), 151–165 (1997)
25. Grant, J., Hunter, A.: Measuring the good and the bad in inconsistent information. In: Proc. 22nd IJCAI, pp. 2632–2637 (2011)
26. Gupta, A., Sagiv, Y., Ullman, J.D., Widom, J.: Constraint checking with partial information. In: Proceedings of PODS 1994, pp. 45–55. ACM Press (1994)
27. Hunter, A., Konieczny, S.: Approaches to measuring inconsistent information. In: Bertossi, L., Hunter, A., Schaub, T. (eds.) Inconsistency Tolerance. LNCS, vol. 3300, pp. 191–236. Springer, Heidelberg (2005)
28. Lee, S.Y., Ling, T.W.: Further improvements on integrity constraint checking for stratifiable deductive databases. In: VLDB 1996, pp. 495–505. Kaufmann (1996)
29. Motro, A., Smets, P.: Uncertainty Management in Information Systems: From Needs to Solutions. Kluwer (1996)
30. Wijsen, J.: Database repairing using updates. Transaction on Database Systems 30(3), 722–768 (2005)

Part II

Classifiers

Collusion and Corruption Risk Analysis Using Naïve Bayes Classifiers

Remis Balaniuk[1], Pierre Bessiere[2], Emmanuel Mazer[3], and Paulo Cobbe[4]

[1] MGCTI - Catholic University of Brasilia, SGAN 916, Módulo B, Asa Norte, cep:70790-160, Brasília DF and Tribunal de Contas da União, Setor de Administracão Federal Sul, SAFS quadra 4, lote 1, cep:70042-900 Brasília - DF, Brazil
remis@ucb.br
[2] LPPA - Collège de France, 11 place Marcelin Berthelot, 75231 Paris cedex05 France
pierre.bessiere@college-de-france.fr
[3] CNRS, E-Motion, LIG - INRIA, 655 avenue de l'Europe, 38334 Montbonnot, France
[4] Information Technology Department, UniCEUB College, SEPN 707/907 Asa Norte, cep:70790-075 Brasília - DF, Brazil

Abstract. Fighting corruption connected with public procurement and governmental agencies requires a strong and effective audit function. The scale and the complexity of the roles to be considered can prevent the use of most audit methods and technologies so successful on the corporate world. The aim of this chapter is to propose a data mining method, based on naïve Bayes classifiers, to support a generic risk assessment process for audit planning. The method can sort auditable units by total risk score, fostering dedicated audit coverage to high-risk areas. Audit organizations can transition from a reactive response to a proactive approach to identify and correct issues that may be indicative of fraud, waste or abuse. Extensive databases containing records from government operations can be combined to auditors knowledge, fraud profiles, impact factors or any other relevant metric in order to rank an audit universe.

Keywords: probabilistic classifiers, data mining, public sector corruption, risk analysis.

1 Introduction

Computer technology provides auditors a large set of techniques to examin the automated business environment. Analytical techniques have become not only more powerful but also widely used by auditors. Computer-assisted tools and auditing techniques have become standard practice in corporate internal auditing.

Audit software applications permit auditors to obtain a quick overview of the business operations and drill down into the details of specific areas of interest. Audit software can also highlight individual transactions that contain characteristics often associated with fraudulent activity. A 100% verification can identify suspect situations such as the existence of duplicate transactions, missing transactions, anomalies. More sophisticated analysis can be achieved by recalculating relevant ratios and figures by comparing data from different locations to assist in identifying possible discrepancies.

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 89–100, 2013.

Auditors usually look for patterns that indicate fraudulent activity. Data patterns such as negative entries in inventory received fields, voided transactions followed by "No Sale," or a high percentage of returned items may indicate fraudulent activity in a corporation. Auditors can use any data patterns to develop a "fraud profile" early in their review of operations. The patterns can function as auditor-specified criteria and transactions fitting the fraud profile can trigger auditor reviews [1].

Nevertheless, when it comes to government auditing, the scale and the complexity of the roles to be considered can prevent the use of most methods and technologies so successful on the corporate world.

As stated by the US Comptroller General in [2], the major roles and responsibilities of governmental auditing teams include:

- Combating corruptions;
- Assuring accountability;
- Enhancing economy, efficiency, and effectiveness;
- Increasing insight; and
- Facilitating foresight.

Corruption can exist within government or on the part of contractors and others who conduct business with government. Therefore, fighting corruption requires a strong and effective audit function. For that to occur, government audit agencies must be assured free access to process, routines and records of government organizations.

Government auditing can also add value by analyzing the efficiency and effectiveness of government organizations, programs and resources. Insights into the operations of the organizations can help government audit agencies to evaluate the extent to which organizational goals are being met, the cost-effectiveness of program performance or whether programs duplicate, overlap or conflict with one another.

Compared to corporate internal auditing, government auditing acts on a much larger universe. These activities range from auditing specific operations or contracts to evaluating program effectiveness and performance.

As indicated by a survey by the Global Audit Information Network [3], the top challenge facing all government audit organization is adequate audit staffing. The large audit universe, the diversity and complexity of the topics being covered makes it impossible for audit agencies to perform a 100% verification on all government operations or entities. The same survey also indicated the ability to plan based on risk as being part of the top ten challenges imposed to all government audit organizations.

Traditional methods for audit planning are usually based on management requests, auditors experience or expertise or simply on statues or regulations. These methods can work fine inside a corporation but are much less effective at the government level. Risk-based audit planning, on the other hand, can result in disciplined analytical approaches to evaluate the audit universe, highlights potential risks that might otherwise be unknown, fosters dedicated audit coverage to high-risk areas, and allocates resources where pay-back is greatest [4].

A risk assessment process for audit planning proposed by the Institute of Internal Auditorss (IIAs) [4] is based on five steps:

- Define the audit universe;
- Identify and weight risk factors;
- Establish a mechanism and score risk factors for auditable units;
- Sort the auditable units by total risk score; and
- Develop the audit plan based on the ranked audit universe.

The aim of this chapter is to propose a method, based on a probabilistic classifier, to support this generic risk assessment process. Our method can score auditable units using a formally defined mathematical framework. Extensive databases containing records from government operations can be combined to auditors knowledge, fraud profiles, impact factors or any other relevant metric in order to rank an audit universe.

2 Related Work

A number of areas can be discussed in relation to previous effort in this domain and two appropriate topics are provided. Therefore this chapter includes a brief discusion on fraud dectection using data mining 2.1 and naive Bayes classifiers 2.2 to highlight any associated research.

2.1 Fraud Detection and Risk Analysis Using Data Mining

Data mining has become an increasingly popular tool used by business, organizations and governments for aiding audit professionals in risk analysis and fraud detection. Several scholarly works have been written on application techniques, particularly for such businesses as financial and security exchange institutions, telecommunications and insurance companies, whom, along with their clients, incur incalculable financial losses due to fraud every year around the world [5,6].

When properly applied, data mining techniques are able to identify trends that indicate suspicious or fraudulent activities, casting light on transactions hidden among the crowd. By reducing the universe of transactions or activities to a smaller subset, data mining allows decision makers to concentrate their efforts on higher risk transactions, and act to mitigate the repercussion of fraudulent activity in affected organizations [5,7,8].

Risk prediction is an important contribution of data mining to a decision making process. By quantifying the possibility that a given event may occur in the future, it provides the decision maker with a basis for comparing alternative courses of action under uncertainty [9]. Despite the fact that it is impossible to ascertain with complete certainty events which have yet to take place, risk analysis allows decision makers to define possible outcomes and assess the risks associated with each, and make a decision based on these possible alternatives. This assessment does not shield the decision maker from negative outcomes, but should ensure that positive outcomes are reached more often than not [10].

2.2 Naïve Bayes Classifiers and Fraud Detection

Naïve Bayes is a mining technique not commonly associated with fraud detection in the scientific literature. In their review of the academic literature on data mining applications in financial fraud detection, Ngai et al. [5] indicate very few studies that applied naïve Bayes algorithms for this purpose.

That trend contrasts with Viaene, Dering and Dedene [11] findings, who present a successful application of naïve Bayes algorithm to PIP claims data for the State of Massachusetts. Their findings suggest that this algorithm can contribute to the implementation of efficient fraud detection systems for insurance claim evaluation support.

Viaene et al. [12] also indicates that naïve Bayes algorithms showed comparative predictive performance to more complex and computationally demanding algorithms, such as Bayesinan Learning Multilayer Perceptron, least-squares support vector machine and tree-augmented naïve Bayes classification, in a benchmark study of algorithm performance for insurance fraud detection.

The emphasis of research of complex unsupervised algorithms presents to Phua et al. [7] a problem for the future. The authors suggest that in order for fraud detection to be successfully implemented in real time applications, less complex algorithms, such as naïve Bayes, have to be considered as the only viable options.

3 Description of the Method

The method proposed in this chapter performs risk evaluation based on classification rules. Rules are built by exploring large databases collected in daily activity of government agencies. Typically, the auditable universe is large and we want to classify auditable units to one of two groups: high risk units and low risk units. Moreover, these units should be sorted with respect to their risk.

3.1 Naïve Bayes Classifiers

The probability model for a classifier is a conditional model as shown in Eq. 1.

$$P(C|H_1,...,H_n) \tag{1}$$

The model is defined over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables H_1 through H_n. Using the very well known Bayes theorem it is possible to write (see Eq. 2):

$$P(C|H_1,...,H_n) = \frac{P(H_1,...,H_n|C)P(C)}{P(H_1,...,H_n)} \tag{2}$$

Because the denominator does not depend on C we are usually interested only in the numerator of the right side fraction. The values of the features are also given and consequently the denominator is constant. The numerator is equivalent to the joint probability model (see Eq. 3):

$$P(C,H_1,...,H_n) \tag{3}$$

The problem is that if the number of features is large or when a feature can take on a large number of values, the computation of such a model can be infeasible. The "naïve" conditional independence assumption assumes that each feature is conditionally independent of every other feature (see Eq. 4):

$$P(H_i|CH_j) = P(H_i|C) \tag{4}$$

This strong assumption can be unrealistic in most cases, but empirical studies related to fraud detection show that most frequently the method presents good performance [11]. Under these independence assumptions the conditional distribution over the class variable can be expressed as Eq. 5, where Z is a constant if the values of the feature variables are known. $P(C)$ is called the class prior and $P(H_i|C)$ are the independent probability distributions.

$$P(C|H_1,...,H_n) = \frac{1}{Z}P(C)\prod_{i=1}^{n}P(H_i|C) \tag{5}$$

By using this naïve Bayesian algorithm it is then possible to obtain a probability distribution of objects belonging into classes. Threshold rules can be used to decide when a probability is strong enough to assign an object into a group. Depending on these rules it happens that an object is not assigned to any group or to more than one group, like in fuzzy logic. Naïve Bayes also naturally deals with missing values, what is difficult to achieve using other methods like decision trees or neural networks. Resulting models are self explainable, unlike other methods like neural networks. Further information about Bayesian Classifiers can be is provided by phua et al. [7].

3.2 Adapted Naïve Bayes Classifier

A major difficulty imposed on risk evaluation by government auditing agencies is the lack of consistent fraud databases. Most methods used in Artificial Intelligence are based on supervised learning, where a set of examples are used to define an inference model. Neural networks and decision-trees are examples of supervised learning methods.

Detected fraud cases are usually reported in unstructured documents. These are typically disconnected, identified and described in very specific contexts and are not statistically relevant considering the number of variables and states in which a systemic fraud could be described.

This lack of information compromises the typical use of Bayesian classifiers. Conditional probabilities for the output classes cannot be established without a consistent base of tagged examples distributed on the input space.

In order to overcome this handicap the standard naïve Bayes classification approach was adapted making two assumptions:

- Statistically, the high risk group is much smaller than the low risk group: fraud is an exception (typically less than 1% of the auditing units).
- High risk units can be described by rules: auditing experts are able to describe fraud profiles from their field expertise.

From these assumptions and having a non labeled extensive database describing our auditing universe, our modeling problem is decomposed in two different problems:

- Find the conditional probabilities for the low risk group (Eq 6): from the first assumption it is possible to consider that the fraud cases inside the large database are not statistically relevant, so to estimate this conditional probability all the data contained in the database will be used as if there were no fraudulent units there.

$$P(H_i|C = lowrisk) = P(H_i) \tag{6}$$

- Find the conditional probabilities for the high risk group: the audit expert directly defines the shape of a probability distribution based on his field expertise.

3.3 The Risk Assessment Process

Distinct risk assessment processes are defined for distinct sets of auditable entities. The first step in implementing a successful risk based audit process is to clearly define and understand a chosen audit universe. The adoption of Computer-Assisted Auditing Techniques (CAATs) by government audit agencies enables the compilation of timely, reliable, and meaningful data that can be used for planned audits on an ongoing basis. Based on compiled data, once CAATs are implemented, auditors or auditing systems can routinely review significant financial and non-financial processes or identify potential audit issues. Once an audit issue is identified the corresponding audit universe will be defined as the set of entities to be classified in order to reflect the risk associated to that issue. An audit issue can be related to a business process as government purchasing / procurement, general ledger, treasury, payroll, accounts payable, inventory and fixed assets. Broader issues can be related to government programs. More specific issues can be related to public employees or public contractors. Correspondently, the audit universe can be a set o purchases, contracts, assets, programs, public employees or contractors. Once the audit universe is defined the second step is to identify its risk factors. The risk factors of an audit universe are related to the audit issue and the business rules associated to the entities set. The risk factors will define the conditional features of the probabilistic classifier, so it is desirable to consider the mutual conditional independence while choosing risk factors. There is no need to define weight factors while using probabilistic classifiers once the nature of the feature variables define a common framework to all measures. To illustrate how the choice of the risk factors occurs, consider as audit issue the fraud detection on public purchasing/procurement. Public procurement comprises government purchasing of goods and services required for State activities, the basic purpose of which is to secure best value for public expenditures. In both developed and developing economics, however, the efficient functioning of public procurement may be distorted by the problems of collusion or corruption or both [13].

Considering fraud on public procurement our audit issue, our audit universe will be the set of public purchases, preferably over a large period of time (five to ten years) and a large scope of purchasing organizations, contractors and purchased goods and

services. The risk assessment would be ideally performed based on a detailed database describing all relevant aspects of each purchase.

One obvious first risk factor associated to public purchases is the space left for competition between potential sellers. The corresponding high risk rule would be: *less competition = higher risk*. To compute the competition level within a procurement process one could use a heuristic model based on its characteristics. The existence or not of an open call for bids or tenders, the type of procurement and the number of registered bidders can be the input for this model. The heuristics can be based on business rules: restricted or invited tenders are more suitable for fraud; electronic procurement auctions (e-procurement, e-reverse auctioning, e-tendering), on the other hand, tend to be more competitive and less suitable for fraud; the competition of a bid is proportional to the number of registered bidders.

Other risk factors to consider could be:

- Value: *more money = higher risk*;
- Bid protests: *more protests = higher risk*;
- Bid winner profile: *no qualified winner = higher risk*; and
- Bid winner previous sanctions: *sanctioned winner = higher risk*.

Some risk factors can require the use of external data, not directly related to the audit universe. The winner previous sanctions, for instance, would require a sanctions database. This a common issue when adopting CAATs. Analytical auditing databases require information covering a broad scope of related themes.

Factors associated with different concerns than the risk itself can be added to the model:

- Political and economic impact: *expenditure contributes to the achievement of policy goals = higher economic impact*;
- Social impact: *public health spend = higher social impact*; and
- Timing and effectiveness: *older purchases = less auditing effectiveness*.

The only requirements to include a new factor to the model are the possibility to estimate a corresponding numerical value for each entity belonging to the audit universe and the existence of a rule for the high risk/impact/effectiveness group. Once the risk factors are identified and computed for the whole audit universe, the next step will be to run the naïve Bayes algorithm and compute the probability distribution of entities belonging into classes. Because there are only two classes (high and low risk) two values are obtained for each entity: $P(C = highrisk|H_1, ..., H_n)$ and $P(C = lowrisk|H_1, ..., H_n)$. If a choice of subsets of risky and non-risky entities is required a threshold rule must be defined in order to decide when a probability is strong enough to assign an entity into a class. Moreover, the probability $P(C = highrisk|H_1, ..., H_n)$ can be directly used to sort the auditable units by total risk score. The ranked audit universe can then be used to develop an audit plan.

As in any knowledge discovery process, the auditing results should be feedbacked to refine the assumptions upon which the whole model was constructed.

4 Experiments

This approach was tested by the Tribunal de Contas da Uniao (TCU), the Brazilian Court of Audit. The TCU have been using CAATs in the last five years intensively, and has gathered extensive information about the Brazilian public sector. It receives information on an ongoing basis from IT systems of major Brazilian government agencies and all relevant data is assembled into a large data warehouse. This test was done as part of a research project involving the TCU, the Institut pour la Recherche en Informatique et Automatique (INRIA) and the company Probayes.ProBayes developed the ProBT engine [14], a powerful tool that facilitates the creation of Bayesian models. A number of risk assessment models were built in order to analyze major audit issues like high risk private contractors, collusion between private contractors and corruption between public bodies and private contractors. The audit issues, the audit universes and the risk factors were designed by TCU auditors. In the following we detail one of these models.

4.1 Corruption in Public Procurements

As pointed out by the Organization for Economic Co-operation and Development (OECD) Global Forum on Competition Debate on Collusion and Corruption in Public Procurement in October 2010 [13]:

> "Corruption occurs where public officials use public powers for personal gain, for example, by accepting a bribe in exchange for granting a tender. While usually occurring during the procurement process, instances of post-award corruption also arise. Corruption constitutes a vertical relationship between the public official concerned, acting as buyer in the transaction, and one or more bidders, acting as sellers in this instance."

The proposed naïve Bayes approach was used to assess risk of corruption between a public body and private companies awarded with its public contracts. The audit universe was the set of pairs of public and private parties of all public contracts signed by the Brazilian federal administration between 1997 and 2011, totalizing 795,954 pairs. The risk factors chosen by the auditors were:

- Competition: purchases awarded to restricted or invited tenders without a open call bid were considered risky;
- Post-award renegotiations of values rising the initial awarded purchase value raise corruption risk;
- Post-award renegotiations of values reducing the initial awarded purchase value reduce corruption risk; and
- The existence of links between the parties was considered risky: a possible link would be public employees from the public body or their relatives which are or were partners or employees of the private company.

The risk factors were then transformed in numerical features. To reduce the computation effort the features were discretized and their values bounded. The following features were computed for each pair of (public body, private company) from the audit universe:

NISLICIT: number of purchases awarded without an open call bid.
- The final feature value was the natural logarithm of the count of purchase awards (values between 0 and 9)

TISLICIT: total amount of the purchases awarded without an open call bid.
- The final feature value was the logarithm base 10 of the total amount of purchases (values between 0 and 5)

NREFOR: the number of post-award purchase value increases.
- The final feature value was the natural logarithm of the count of purchase value increases (values between 0 and 7)

TREFOR: total amount of the purchases value increases.
- The final feature value was the logarithm base 10 of the total amount of purchase value increases (values between 0 and 6)

NANULA: the number of post-award purchase value reductions.
- The final feature value was the natural logarithm of the count of purchase value reductions (values between 0 and 7)

TANULA: total amount of the purchases value reductions.
- The final feature value was the logarithm base 10 of the total amount of purchase value reductions (values between 0 and 6)

TSESO: indicates the existence or not of links between the public body and the private company.
- The final feature value was 1 for linked pairs and 0 otherwise.

4.2 Results

The conditional probabilities for the seven features are illustrated on Figures 1- 7. Conditional probabilities for the low risk class, obtained from the purchase database are displayed on the left side of the Figures. Conditional probabilities for the high risk class directly assigned by the audit experts, are displayed on the right.

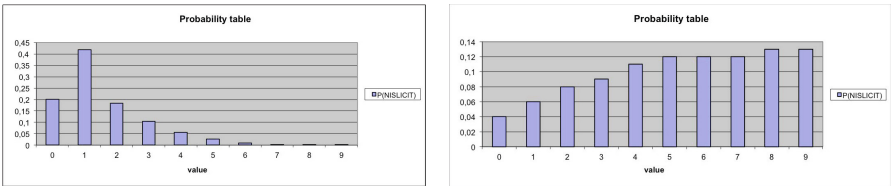


Fig. 1. Left side: $P(NISLICIT|C = lowrisk)$ and right side: $P(NISLICIT|C = highrisk)$

The previous class was assigned an empirical value $P(C) = 0.01$ based on the auditors expectation of corruption inside the audit set. The computation of $P(C = highrisk|NISLICIT, TISLICIT, NREFOR, TREFOR, NANULA, TANULA, TSESO)$ for the set of 795,954 pairs of public and private parties indicated 2,560 pairs with probability higher than 99% (0.99).

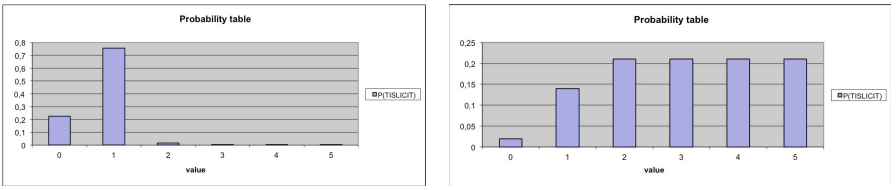


Fig. 2. Left side: $P(TISLICIT|C = lowrisk)$ and right side: $P(TISLICIT|C = highrisk)$

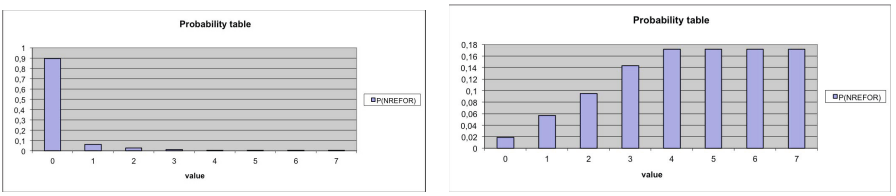


Fig. 3. Left side: $P(NREFOR|C = lowrisk)$ and right side: $P(NREFOR|C = highrisk)$

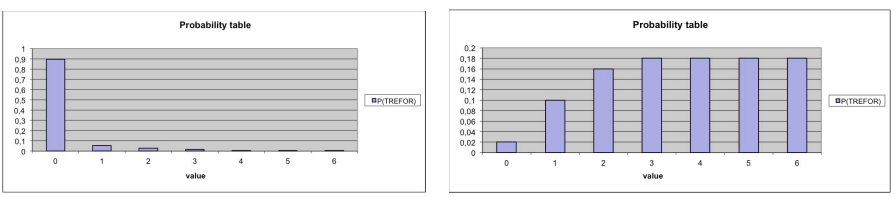


Fig. 4. Left side: $P(TREFOR|C = lowrisk)$ and right side: $P(TREFOR|C = highrisk)$

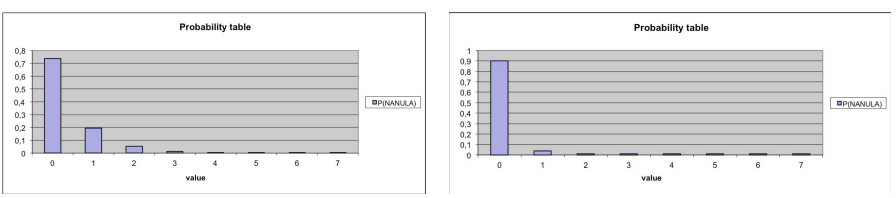


Fig. 5. Left side: $P(NANULA|C = lowrisk)$ and right side: $P(NANULA|C = highrisk)$

This result included parties with obvious links, like private not-for-profit foundations for which the Brazilian public procurement law gives special treatment. Moreover, a number of high risk pairs were known by the auditors from previous investigations where corruption was effectively found. The overall feeling from all auditors to which the high risk list was presented was that the result was very reasonable. Future audit plans will possibly be based on our risk rankings.

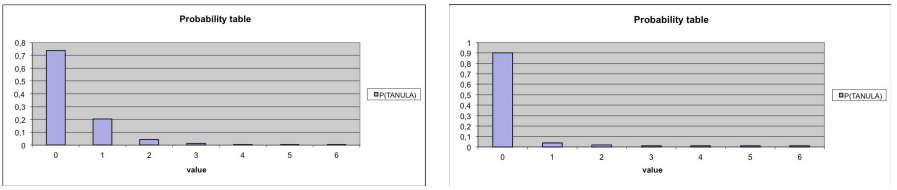


Fig. 6. Left side: $P(TANULA|C = lowrisk)$ and right side: $P(TANULA|C = highrisk)$

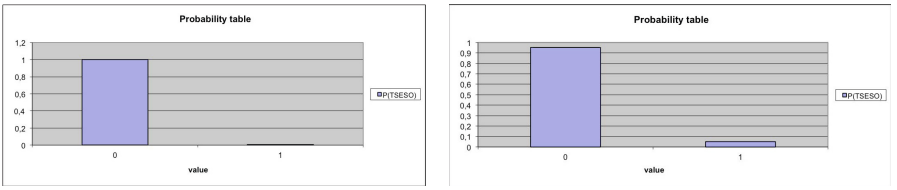


Fig. 7. Left side: $P(TSESO|C = lowrisk)$ and right side: $P(TSESO|C = highrisk)$

5 Conclusion

With a growing number of governmental agencies transitioning into unified and consolidated data information platforms, access to information and uniformity are increasing. To deal with all this information government audit organizations are increasingly adopting CAATs in order to routinely review significant financial and non-financial processes and identify potential audit issues.

The steps taken to accumulate and identify significant data elements allow audit organizations to use that data during planned and unplanned audits. The perpetual development of data repositories and implementation of ongoing monitoring can also contribute to the development of periodic government-wide risk assessment.

In this chapter we proposed a risk assessment method, based on naïve Bayes classifiers, that can be used by government audit organizations. The proposed method is suitable to the typical risk assessment process for audit planning, as formalized by the IIAs.

The main advantages of the proposed method are:

- Integration of auditors knowledge to large data repositories in order to analyze audit issues;
- Integration of quantitative risk factors to qualitative aspects to compose probabilistic features;
- Natural framework to deal with missing data, data in different scales and from different sources; and
- Low computational complexity.

Upon implementation, this semi-automated risk assessment procedure can help audit organizations transition from a reactive response to a proactive approach to identify and correct issues that may be indicative of fraud, waste or abuse.

References

1. Coderre, D.: Auditing - Computer-Assisted Techniques for Fraud Detection. The CPA Journal (August 1999)
2. United States Government Accountability Office (GAO): Government Auditing Standards (The Yellow Book) (August 2011)
3. Vito, K.W.: SPHR, CCP, Auditing Employee Hiring and Staffing. The IIA Research Foundation (June 2011)
4. The Institute of Internal Auditors: International Professional Practices Framework (IPPF). The IIA Research Foundation (2009)
5. Ngai, E.W.T., Yong Hu, Y.H., Wong, Y.C., Sun, X.: The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 559–569 (2011)
6. Panigrahi, S., Kundu, A., Sural, S., Majumdar, A.K.: Credit card fraud detection: A fusion approach using Dempster Shafer theory and Bayesian learning. Information Fusion 10(4), 354–363 (2009)
7. Phua, C., Lee, V., Smith, K., Gayler, R.: A comprehensive survey of data mining-based fraud detection research. Artificial Intelligence Review, 1–14 (2005)
8. Hormazi, A.M., Giles, S.: Data Mining: A Competitive Weapon for Banking and Retail Industries. Information Systems Management 21(2), 62–71 (2004)
9. Nilsen, T., Aven, T.: Models and model uncertainty in the context of risk analysis. Reliability Engineering and System Safety 79, 309–331 (2003)
10. Nilsen, T.: Foundations of Risk Analysis: A Knowledge and Decision-Oriented Perspective. John Wiley and Sons Ltd., West Sussex (2003)
11. Viaene, S., Derrig, R.A., Dedene, G.: A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering 16(5), 612–620 (2004)
12. Viaene, S., Derrig, R.A., Baesens, B., Dedene, G.: A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. J. Risk and Insurance 69(3), 373–421 (2002)
13. Organisation for Economic Co-operation and Development: Roundtable on Collusion and Corruption in Public Procurement (October 2010)
14. Mekhnacha, K., Ahuactzin, J.M., Bessiere, P., Mazer, E., Smail, L.: Exact and approximate inference in ProBT. Revue d'Intelligence Artificielle 21(3), 295–332 (2007)
15. Zhang, H.: The Optimality of Naive Bayes. American Association for Artificial Intelligence (2004)

Impact of Circularity Analysis on Classification Results: A Case Study in the Detection of Cocaine Addiction Using Structural MRI

Maite Termenon[1], Elsa Fernández[1], Manuel Graña[1], Alfonso Barrós-Loscertales[2], Juan C. Bustamante[2], and César Ávila[2]

[1] Grupo de Inteligencia Computacional (GIC), UPV/EHU, Spain
www.ehu.es/ccwintco
[2] Dpto. Psicología Básica, Clínica y Psicobiología, Universitat Jaume I, Castellón de la Plana, Spain
maite.termenon, manuel.grana@ehu.es

Abstract. Due to the high dimensionality of the neuroimaging data, it is common to select a subset of relevant information from the whole dataset. The inclusion of information of the complete dataset during that selection of a subset, can drive to some bias in the results, often leading to optimistic conclusions. In this study, the differences in results obtained performing an experiment free of circularity and repeating the process including a circularity effect (Double-Dipping (DD)) are shown. Discriminant features (based on voxel's intensity values) are obtained from structural Magnetic Resonance Imaging (MRI) to train and test classifiers that are able to discriminate cocaine dependent patients from healthy subjects. Feature selection is done by computing Pearson's correlation between voxel values across subjects with the subject class as control variable. As classifiers, several machine learning techniques are used: k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Extreme Learning Machines (ELM) and Learning Vector Quantization (LVQ). Feature selection process with DD obtains, in general, higher accuracy, sensitivity and specificity values.

Keywords: MRI, Circularity, Machine Learning, Pattern Recognition, Classification, Double Dipping, Computer Aided Diagnosis, SVM, ELM.

1 Introduction

Computer Aided Diagnosis (CAD) [8,14] tools are acquiring more relevance in the neuroscience community in order to improve the prediction accuracy of the neuropsychological assessments performed by expert clinicians. They apply Machine Learning (ML) and Pattern Recognition (PR) techniques to find the localization in the brain of effects correlated with neurodegenerative disorders, trying to measure the damage and the stage of the disorder. These techniques are applied to several medical imaging modalities such as Magnetic Resonance Imaging (MRI) [7,19,18], Diffusion Tensor Imaging (DTI) [3,22], Computed Tomography (CT) or Single-Photon Emission Computed Tomography (SPECT)

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 101–114, 2013.

[13] (among them) to obtain significative statistical differences that discriminate between diseases or different stages of a disease.

Due to the high dimensionality of the medical imaging data, it is a common practice to select a subset of the whole dataset to, later on, build a classifier based on that selected subset. Multiple selection processes can be applied and there are several ways to perform that selection. Common methodological practices may introduce an optimistic bias in the analysis due to Double-Dipping (DD) or analysis circularity [11,24]. This problem occurs when the same data are used for selection and selective analysis, so that statistical results are not independent of the selection criteria under the null hypothesis [11]. Frequently, results reflect the data, but they are biased. Such biases could be introduced when data are analyzed to extract a subset and then, the subset is analyzed again to obtain results. Classifier validation standard methodology consists in the separation of independent training and testing datasets. Training dataset is used to build the classifier; testing dataset to provide generalization estimates. Circular analysis may be introduced when the selection/extraction of the relevant features is performed using the whole dataset.

As case study, structural MRI database of cocaine dependent patients and healthy subjects is used. This database was provided by the Departamento de Psicología Básica, Clínica y Psicobiología of Univesitat Jaume I (UJI), in Castellón, Spain. Cocaine is one of the most consumed illegal drugs and its chronic abuse may cause consequences such as ischemic, hemorrhagic strokes, depression and neuropsychological abnormalities [5]. Studies have shown that selected regions in the brain of cocaine consumers show functional, neurochemical and structural abnormalities. These can be used to identify the differences between the brains of cocaine consumers and non-consumers [6]. Studies found structural differences in several brain regions that include: parahippocampus, posterior cingulate, amygdala, insula, striatum and cerebellum [2,6,12,21].

In this chapter, authors focus on a specific feature selection process to extract significant voxels from structural MRI volumes. The signal intensity on these voxels is used to discriminate between patients and healthy subjects. Feature selection process is done computing Pearson's correlation between voxel values across subjects with the subject class as control variable. This process is performed twice: one avoiding circularity (correct process), and the other one including circularity. The main contribution of this chapter is the direct assessment of the impact of circularity on the classification results. Both sets of selected features (with and without circularity) are tested building four different classifiers: k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Extreme Learning Machines (ELM) and Learning Vector Quantization (LVQ).

In Section 2, experimental database and preprocessing steps are described. Next, Section 3 details the procedure to select relevant features and how to avoid DD in this experiment. Section 4 shows the experimental results and finally, in Section 5, conclusion obtained from results are established.

2 Database and Preprocessing Steps

Database consists of 98 brain structural MRI divided in two groups: 50 controls (age = 33.58 ± 8.45) and 48 patients (age = 35.10 ± 7.02). Patients with cocaine addiction were recruited from the Addiction Treatment Service of San Agustín in Castellón, Spain. The inclusion criteria for cocaine addictions was based on the DSM-IV criteria. Control subjects were required to have no diagnosis of substance abuse or dependence. The exclusion criteria for all the participants included neurological illness, prior head trauma, positive HIV status, diabetes, Hepatitis C, or other medical illness and psychiatric disorders. Cocaine consumption was assessed with an urine toxicology test, which ensured a minimum period of abstinence of two to four days prior to MRI data acquisition. Groups were matched on the basis of age and level of education. All the participants were right-handed according to the Edinburgh Handedness Inventory [15]. They all signed an informed consent prior to participating in this study. Images were acquired on a 1.5T Siemens Avanto (Erlangen, Germany) with a standard quadrature head coil. A high resolution 3D T1-weighted gradient echo pulse sequence was acquired (TE=4.9 ms; TR=11 ms; FOV=24 cm; matrix=$256 \times 224 \times 176$; voxel size=$1 \times 1 \times 1$).

Images were processed using Statistical Parametric Mapping (SPM8) running on Matlab®. Preprocessing consisted of reorientation, tissue segmentation, bias correction and spatial normalization into a unified model [1]. Steps followed during image processing are shown in Figure 1. Parameter settings were: warp frequency cutoff to 25 mm, warping regularization light to (0.001), a thorough clean up of segmentations and a $1.5{-}mm^3$ voxel size resolution for normalization. Iteratively weighted Hidden Markov Random Fields (HRMF) were applied to improve the accuracy of tissue segmentation by removing isolated voxels which were unlikely to be a member of a certain tissue class and closing hole in the clusters of connected voxels of a certain class, resulting in a higher signal to noise ratio of the final tissue probability maps.

Each subject's brain was normalized to the tissue probability maps provided by the International Consortium for Brain Mapping (ICBM)[1]. The transformations consists of a first linear registration and a second nonlinear shape transformation. Segmented Gray Matter (GM) images were modulated to restore tissue volume changes after spatial normalization. Modulation was performed by multiplying the voxel intensities by the Jacobian determinant of the spatial transformation matrix derived from normalization. The modulated GM segmented images were corrected for nonlinear warping only[2], making correcting for total intracranial volume of the individual unnecessary [20]. Modulation involves scaling by the amount of contraction, so that the total amount of GM in the modulated image remains the same as it would be in the original one.

[1] http://www.loni.ucla.edu/Atlases/

[2] http://dbm.neuro.uni-jena.de/vbm/segmentation/modulation/

3 Feature Selection and Classification

Once all images are processed, only GM information is used to extract significant features to classify. Feature selection was performed computing a voxel-wise Pearson's correlation [17] with the indicator variable specifying the subject class label.

Computing the empirical distribution of the correlation coefficients, the voxels with highest absolute correlation belonging to a certain percentile of this distribution are selected. As shown in Eq. 1, Pearson's correlation at the j-th voxel site is computed as follows:

$$r_{\mathbf{v}_j,\mathbf{y}} = \frac{n\sum_i v_{ij}\, y_i - \sum_i v_{ij} \sum_i y_i}{\sqrt{n\sum_i v_{ij}^2 - (\sum_i v_{ij})^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}} \tag{1}$$

where v_{ij} is the value of the j-th voxel site in the i-th MRI volume and y_i is the class label value of that i-th volume.

3.1 Circularity Analysis (Double Dipping)

Applying Pearson's correlation, the subset of most correlated voxels is selected and used to discriminate between healthy or addict subjects. DD is committed as follows: if correlations are computed with the whole data set, information of the true classification of the test subjects is introduced into the training procedure. To avoid DD, voxel selection procedure must be performed only on the train data. In cross-validation studies, avoiding DD implies performing as many feature selection processes as cross-validation folds.

3.2 Leave One Out - Cross Validation

To split the data, we use Leave One Out - Cross Validation (LOO-CV) technique. This technique is commonly used when the number of samples is not big enough comparing to their dimensionality. It uses $N-1$ samples for training and only one to test the built classifier. Classification is repeated N times, leaving out a different sample each iteration, where N is the number of samples in the data.

3.3 Experimental Design

To avoid circularity, correlation is computed using LOO-CV, it is leaving one subject out and repeating the process for all the subjects. In this case, correlation across volumes is calculated 98 times, leaving a different subject out at each iteration. Feature selection pipelines for both approaches are shown in Figure 2 (free of circularity) and Figure 3 (including circularity).

For this experiment, four different percentiles are applied, obtaining different vectors to classify. Percentiles and their corresponding number of features are shown in Table 1. Features are tested with four different classifiers, k-NN, SVM, ELM and LVQ.

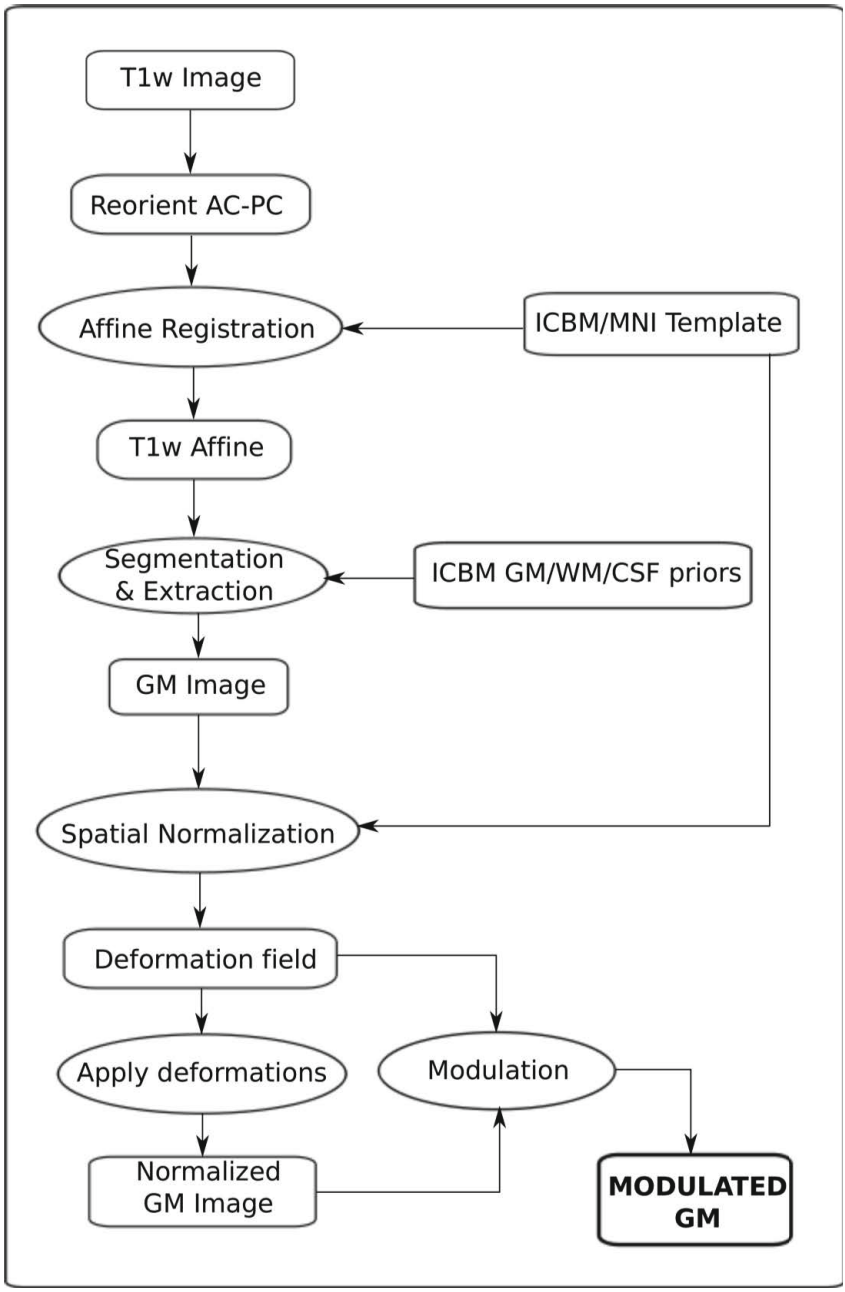


Fig. 1. Image preprocessing pipeline

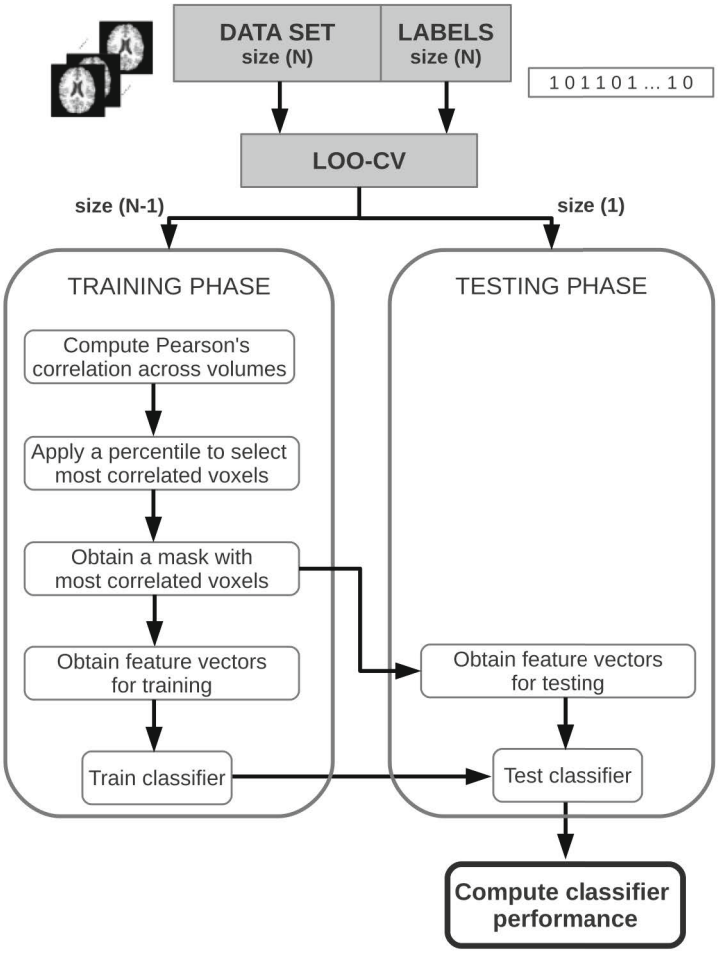


Fig. 2. Feature Selection Pipeline Free of Circularity in the Experimental Design

Table 1. Percentiles applied after computing Pearson's correlation and their corresponding number of features

Percentiles	**# Features**
99.50	2629
99.90	526
99.95	263
99.99	53

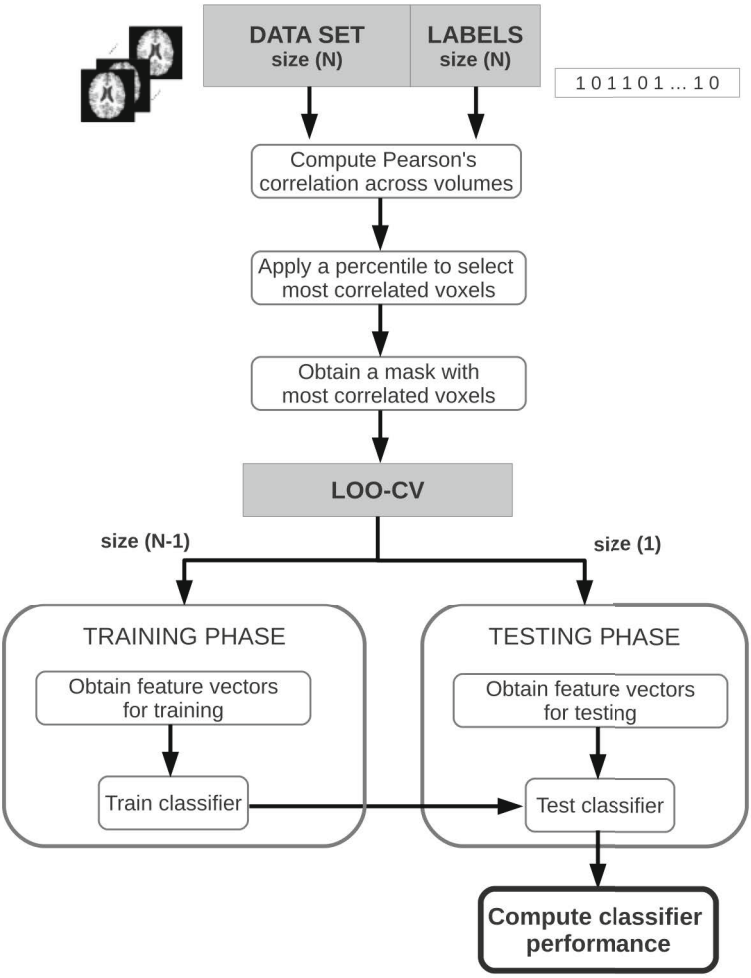


Fig. 3. Feature Selection Pipeline including Circularity in the Experimental Design

3.4 *k*-Nearest Neighbor

The k-NN algorithm [4] is one of the simplest machine learning algorithms. It is very suitable when there is little or no prior knowledge about the distribution of the data. k-Nearest neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. An object is classified by a majority vote of its neighbors (it is, the closest training examples in the feature space), with the object being assigned to the class most common among its k nearest neighbors. When $k = 1$, the object is assigned to the same class of its nearest neighbor, 1-Nearest Neighbor (1-NN).

k-NN is a type of instance-based learning where the function is only approximated locally and all computation is deferred until classification. Its decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points.

In the case of $k = 1$, a set of N pairs, $(\mathbf{x}_1, t_1) \dots (\mathbf{x}_n, t_n)$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdot\cdot\cdot, x_{in}]^T \in \mathbb{R}^n$ is the feature vector and $t_i \in [0, 1]$ is the class in a two class problem. The $\mathbf{x}_i$ take values in a metric space X upon which it is defined a metric d. A new pair $(\mathbf{x}, t)$ is given, where t is needed to be estimated utilizing the information contained in the previous set of N pairs. As stated in Eq. 2, $\mathbf{x}_j \varepsilon \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ is called a nearest neighbor of $\mathbf{x}$ if the distance between them is minimum:

$$min\, d(\mathbf{x}_i, \mathbf{x}) = d(\mathbf{x}_j, \mathbf{x}), \quad where\ i = 1, 2, \dots, n \tag{2}$$

The nearest neighbor rule decides that $\mathbf{x}$ belongs to the same category of $\mathbf{x}_j$, it is, t_j considering only the class of its nearest neighbor. The $n - 1$ remaining t_i are ignored.

3.5 Support Vector Machines

SVM [23] approach is a pattern recognition technique based on the statistical learning theory. SVMs are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. The classification approach used to solve the optimization problem is shown by Eqs. 3 and 4:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i, \tag{3}$$

subject to

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq (1 - \xi_i),\ \xi_i \geq 0,\ i = 1, 2, \dots, n. \tag{4}$$

The minimization problem is solved via its dual optimization problem Eqs. 5 and 6:

$$\min_{\alpha} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{Q}\boldsymbol{\alpha} - \mathbf{e}^T\boldsymbol{\alpha}, \tag{5}$$

subject to

$$\mathbf{y}^T\boldsymbol{\alpha} = 0,\ 0 \leq \alpha_i \leq C,\ i = 1, \ldots, l. \tag{6}$$

Where $\mathbf{e}$ is a vector of all ones, $C > 0$ is the upper bound on the error and $\mathbf{Q}$ is an $l \times l$ positive semidefinite matrix. The $\mathbf{Q}$ elements are based on a kernel function, $K(\mathbf{x}_i, \mathbf{x}_j)$, that describes the behavior of the support vectors. The chosen kernel function results in different kinds of SVM with different performance levels, and the choice of the appropriate kernel for a specific application is a difficult task. In this study, a linear kernel with $C = 1$ defined as Eq. 7 is used:

$$K(\mathbf{x}_i, \mathbf{x}_j) = 1 + \mathbf{x}_i^T\mathbf{x}_j. \tag{7}$$

This kernel shows good performance for linearly separable data.

3.6 Extreme Learning Machines

ELM algorithm was originally proposed in [9] and it makes use of the Single Layer Feedforward Network (SLFN). This method is based on the Moore-Penrose [16] generalized inverse and provides the minimum least-squares solution of general linear systems. The main concept behind the ELM lies in the random choice of the SLFN hidden layer weights and biases. For M arbitrary distinct samples $(\mathbf{x}_i, t_i)$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]^T \in \mathbb{R}^n$ and $t_i = [t_{i1}, t_{i2}, \cdots, t_{im}]^T \in \mathbb{R}^m$, a SLFN with N hidden neurons and activation function $g(x)$ is mathematically modeled as:

$$\sum_{i=1}^{N} \beta_i \cdot g(\mathbf{w}_i \cdot \mathbf{x}_j + b_i) = t_j,\ j \in [1, M], \tag{8}$$

where $\mathbf{w}_i = [w_{i1}, w_{i2}, \cdots, w_{iN}]^T$ is the weight vector connecting the i-th hidden neuron and the input neurons, $\beta_i = [\beta_{i1}, \beta_{i2}, \cdots, \beta_{iM}]^T$ is the weight vector connecting the i-th hidden neuron and the output neurons, and b_i is the threshold of the i-th hidden neuron. The expression $\mathbf{w}_i \cdot \mathbf{x}_j$ denotes the inner product of $\mathbf{w}_i$ and $\mathbf{x}_j$. Eq. 8 can be written as Eq. 9:

$$\mathbf{H}\beta = \mathbf{T}, \tag{9}$$

with $\mathbf{H}$

$$\mathbf{H} = \begin{pmatrix} g(\mathbf{w}_1\mathbf{x}_1 + b_1) & \ldots & g(\mathbf{w}_N\mathbf{x}_1 + b_N) \\ \vdots & \ddots & \vdots \\ g(\mathbf{w}_1\mathbf{x}_M + b_1) & \cdots & g(\mathbf{w}_N\mathbf{x}_M + b_N) \end{pmatrix},$$

where $\beta = (\beta_1 \ldots \beta_N)^T$ is the weight matrix and $\mathbf{T} = (t_1 \ldots t_M)^T$ is the target matrix. Training of SLFN is accomplished computing the least-squares

solution β of the linear system $\mathbf{H}\beta = \mathbf{T}$, given by $\hat{\beta} = \mathbf{H}^{\dagger}\mathbf{T}$, where $\mathbf{H}^{\dagger}$ is the Moore-Penrose inverse of $\mathbf{H}$.

For this experiment, ELM was tested with sigmoid activation function and four different number of hidden nodes: 100, 500, 1000 and 1500. In next section, only the best results were reported, with a 1500 hidden nodes. Due to the randomly initialization of ELM, experiments were repeated 50 times and average results were computed.

3.7 Learning Vector Quantization

LVQ, as introduced by Kohonen [10], represents every class $c \in \{-1, 1\}$ by a set $W(c) = \{\mathbf{w}_i \in \mathbb{R}^n; i = 1, \ldots, N_c\}$ of weight vectors (prototypes) which tessellate the input feature space. If we consider W as the union of all prototypes, regardless of class, and we denote c_i the class the weight vector $\mathbf{w}_i \in W$ is associated with, the decision rule that classifies a feature vector $\mathbf{x}$ is as shown in Eq. 3.7:

$$c(\mathbf{x}) = c_{i^*}$$

where

$$i^* = \arg\min_i \{\|\mathbf{x} - \mathbf{w}_i\|\}.$$

The training algorithm of LVQ tries to minimize the classification error on the given training set, i.e., $E = \sum_j (y_j - c(\mathbf{x}_j))^2$, modifying the weight vectors each time a feature vector is introduced. The heuristic weight updating rule is shown in Eq. 10:

$$\Delta\mathbf{w}_{i^*} = \begin{cases} \epsilon \cdot (\mathbf{x}_j - \mathbf{w}_{i^*}) & \text{if } c_{i^*} = y_j \\ -\epsilon \cdot (\mathbf{x}_j - \mathbf{w}_{i^*}) & \text{otherwise} \end{cases}, \tag{10}$$

that is, the input's closest weight is adapted either toward the input if their classes match, or away from it if not. This rule is highly unstable, therefore, the practical approach consists on performing an initial clustering of each class data samples to obtain an initial weight configuration using Eq. 10 to perform the fine tuning of the classification boundaries. This equation corresponds to a LVQ1 approach. The LVQ2 approach involves determining the two input vector's closest weights. They are moved toward or away the input according to the matching of their classes.

For this experiment, LVQ1 with 10 hidden nodes was tested.

4 Results

Average scores of the performance measures obtained from leave one out cross-validation steps for the values of correlation distribution percentiles are shown in Table 2 and 3 without and with circularity, respectively. It can be easily appreciated the increase in all performance measures obtained when DD is incurred. The quantitative performance measures are defined in Eq. 11, 12 and 13:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

assuming as the null hypothesis that the subject is a cocaine dependent patient, therefore, in these expressions, True Positives (TP) are the number of diseased patient volumes correctly classified; True Negatives (TN) are the number of control volumes correctly classified; False Positives (FP) are the number of control volumes classified as diseased patients and finally, False Negatives (FN) are the number of diseased patient volumes classified as control subjects.

Table 2. SVM classification results. Results obtained ensuring no circularity in the feature extraction procedure.

	Linear SVM			1**-NN**		
	Acc	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
99.50	98.98	100.00	97.92	93.88	89.58	98.00
99.90	98.98	100.00	97.92	94.90	89.58	100.00
99.95	97.96	100.00	95.83	95.88	93.62	98.00
99.99	91.84	96.00	87.50	85.71	79.12	92.00

	ELM (1500 nodes)			**LVQ1 (10 nodes)**		
	Acc	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
99.50	95.88	98.36	93.29	94.90	93.75	96.00
99.90	95.90	98.96	92.71	94.90	91.67	98.00
99.95	98.49	97.60	99.45	98.97	100.00	98.00
99.99	88.65	94.76	82.29	84.69	79.17	90.00

Table 3. SVM classification results. Results when there is circularity in the feature extraction procedure.

	Linear SVM			1**-NN**		
	Acc	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
99.50	100.00	100.00	100.00	94.90	93.75	96.00
99.90	100.00	100.00	100.00	97.96	95.83	100.00
99.95	98.00	98.00	100.00	96.94	95.83	98.00
99.99	93.88	92.00	95.83	91.84	91.67	92.00

Table 3. *(Continued)*

	ELM (1500 nodes)			LVQ1 (10 nodes)		
	Acc	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
99.50	98.94	99.04	98.83	96.94	97.92	96.00
99.90	99.02	99.04	99.00	96.94	100.00	94.00
99.95	98.02	96.60	99.50	98.98	100.00	98.00
99.99	91.06	90.68	91.46	87.76	87.50	88.00

In this kind of studies, it is not only important to be able to classify patients and controls. It is also important to indicate where features we are using to discriminate are located in the brain. For this purpose, the AtlasQuery tool of FSL[3] is used. Most significant voxels for the percentile 99.50% are shown in red on MNI template in Figure 4. Discriminant information was mainly found in frontal pole, insula, cerebellum, striatum and superior frontal and precentral gyrus. These areas are also reflected in the literature.

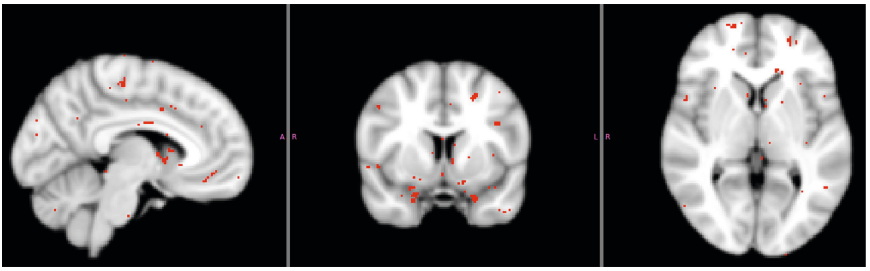


Fig. 4. Location of the features selected applying a percentile = 99.50% after computing Pearson's correlation across volumes. Discriminant features were mainly found in frontal pole, insula, cerebellum, striatum and superior frontal and precentral gyrus.

5 Conclusion

In this chapter, it is presented a comparison between a process with circularity and a process free of circularity. Database consist of 98 subjects, 50 healthy subjects and 48 cocaine dependent patients. Data were preprocessed to ensure the correspondence between voxel sites and anatomical features across all subjects. To select the most significant features, we compute Pearson's correlation using subject's class label as indicator variable. This is the critical step where processes with circularity and no circularity split. Once relevant features are selected, they are used to discriminate between controls and patients using four different classifiers: k-NN, SVM, ELM and LVQ.

[3] `http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlasquery`

Results are very good in both cases, but classification results are higher for a process with DD, as expected. The difference between classification results become higher when the dimensionality of the feature vectors is reduced (it is when percentiles are higher) except for the percentile 99.95% of LVQ, where the scores are very similar for both groups of features.

These differences, despite of not being too high, are due to introducing information from the testing dataset into the feature selection process used to select the features to classify with. It is recommended to take into account this kind of analysis bias when designing neuroimaging analysis experiments.

References

1. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage 26(3), 839–851 (2005)
2. Barrós-Loscertales, A., Garavan, H., Bustamante, J.C., Ventura-Campos, N., Llopis, J.J., Belloch, V., Parcet, M.A., Ávila, C.: Reduced striatal volume in cocaine-dependent patients. NeuroImage 56(3), 1021–1026 (2011)
3. Besga, A., Termenon, M., Graña, M., Echeveste, J., Pérez, J.M., Gonzalez-Pinto, A.: Discovering Alzheimer's Disease and Bipolar Disorder White Matter Effects building Computer Aided Diagnostic Systems on Brain Diffusion Tensor Imaging Features. Neuroscience Letters 520(1), 71–76 (2012)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
5. Ernst, T., Chang, L., Oropilla, G., Gustavson, A., Speck, O.: Cerebral perfusion abnormalities in abstinent cocaine abusers: A perfusion MRI and SPECT study. Psychiatry Research: Neuroimaging 99(2), 63–74 (2000)
6. Franklin, T.R., Acton, P.D., Maldjian, J.A., Gray, J.D., Croft, J.R., Dackis, C.A., O'Brien, C.P., Childress, A.R.: Decreased gray matter concentration in the insular, orbitofrontal, cingulate, and temporal cortices of cocaine patients. Biological Psychiatry 51(2), 134–142 (2002)
7. García-Sebastián, M., Savio, A., Graña, M., Villanúa, J.: On the Use of Morphometry Based Features for Alzheimer's Disease Detection on MRI. In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) IWANN 2009, Part I. LNCS, vol. 5517, pp. 957–964. Springer, Heidelberg (2009)
8. Graña, M., Termenon, M., Savio, A., Gonzalez-Pinto, A., Echeveste, J., Pérez, J.M., Besga, A.: Computer Aided Diagnosis system for Alzheimer disease using brain Diffusion Tensor Imaging features selected by Pearson's correlation. Neuroscience Letters 502(3), 225–229 (2011)
9. Huang, G.-B., Zhu, Q.-Y., Siew, C.-K.: Extreme learning machine: Theory and applications. Neurocomputing 70(1-3), 489–501 (2006)
10. Kohonen, T.: Learning vector quantization. In: The Handbook of Brain Theory and Neural Networks, pp. 537–540. MIT Press (1998)
11. Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I.: Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience 12(5), 535–540 (2009)
12. Lim, K.O., Wozniak, J.R., Mueller, B.A., Franc, D.T., Specker, S.M., Rodriguez, C.P., Silverman, A.B., Rotrosen, J.P.: Brain macrostructural and microstructural abnormalities in cocaine dependence. Drug and Alcohol Dependence 92(1-3), 164–172 (2008)

13. López, M.M., Ramírez, J., Górriz, J., Álvarez, I., Salas-Gonzalez, D., Segovia, F., Chaves, R.: SVM-based CAD system for early detection of the Alzheimer's Disease using kernel PCA and LDA. Neuroscience Letters 464(3), 233–238 (2009)
14. Martínez-Murcia, F., Górriz, J., Ramírez, J., Puntonet, C., Salas-González, D.: Computer Aided Diagnosis tool for Alzheimer's Disease based on Mann-Whitney-Wilcoxon U-Test. Expert Systems with Applications 39(10), 9676–9685 (2012)
15. Oldfield, R.C.: The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia 9(1), 97–113 (1971)
16. Penrose, R.: A generalized inverse for matrices. Mathematical Proceedings of the Cambridge Philosophical Society 51(03), 406–413 (1955)
17. Rodgers, J.L., Nicewander, W.A.: Thirteen Ways to Look at the Correlation Coefficient. The American Statistician 42(1), 59–66 (1988)
18. Savio, A., Charpentier, J., Termenon, M., Shinn, A.K., Graña, M.: Neural classifiers for schizophrenia diagnostic support on diffusion imaging data. Neural Network World 20, 935–949 (2010)
19. Savio, A., García-Sebastián, M.T., Chyzhyk, D., Hernandez, C., Graña, M., Sistiaga, A., López de Munain, A.L., Villanúa, J.: Neurocognitive disorder detection based on feature vectors extracted from VBM analysis of structural MRI. Computers in Biology and Medicine 41(8), 600–610 (2011)
20. Scorzin, J.E., Kaaden, S., Quesada, C.M., Müller, C., Fimmers, R., Urbach, H., Schramm, J.: Volume determination of amygdala and hippocampus at 1.5 and 3.0T MRI in temporal lobe epilepsy. Epilepsy Research 82(1), 29–37 (2008)
21. Termenon, M., Graña, M., Barrós-Loscertales, A., Bustamante, J.C., Ávila, C.: Cocaine Dependent Classification Using Brain Magnetic Resonance Imaging. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part II. LNCS, vol. 7209, pp. 448–454. Springer, Heidelberg (2012)
22. Termenon, M., Graña, M., Besga, A., Echeveste, J., Gonzalez-Pinto, A.: Lattice Independent Component Analysis feature selection on Diffusion Weighted Imaging for Alzheimer's Disease Classification. Neurocomputing
23. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience (1998)
24. Vul, E., Pashler, H.: Voodoo and circularity errors. NeuroImage 62(2), 945–948 (2012)

An Evolved Cellular Automata Based Approach to Hyperspectral Image Processing

B. Priego, D. Souto, F. Bellas, F. López-Peña, and R.J. Duro

Integrated Group for Engineering Research, University of A Coruña,
Mendizabal s/n 15403 Ferrol, A Coruña, Spain
{blanca.priego,dsouto,fran,flop,richard}@udc.es

Abstract. This chapter addresses the problem of processing Hyperspectral images (HI) in real time. It is a relevant problem as working with these types of images usually involves processing very large quantities of data due to their high spectral and spatial resolutions. To achieve real time performance requires resorting to extremely distributed architectures, such as Graphic processing units (GPUs). Most of the algorithms that have been developed for Hyperspectral image (HIS) processing are currently sequential and too cumbersome to achieve the necessary parallelism. A very promising approach for solving this problem is using Cellular Automata (CA) based algorithms as this is an intrinsically distributed computational paradigm that would adapt very snugly to these high performance architectures. The main problem of CAs is that it is necessary to endow them with rule sets which, through iterative cycles of interactions among cells, provide the final desired result. Determining these rule sets is non-trivial. In fact it is a highly difficult task, especially in complex processing operations. Here, the objective is to highlight how this problem can be solved through the use of evolutionary techniques. The proposed algorithm has been named Evolving Cellular Automata (ECA) and it has been tested in two standard operations within hyperspectral imaging: edge detection and segmentation. ECA is applied over synthetic and real HISs and the results are compared to those of well-established techniques, confirming the successful response of this approach and its potential for execution in GPU type architectures. This allows their execution in real time even when the size of the images (both spectrally and spatially) or the speed at which they are taken is large.

Keywords: Hyperspectral imaging, Evolution, Cellular Automata, Segmantation.

1 Introduction

Lately, the use of hyperspectral images is becoming more common as a very important source of remote sensing and industrial processing information due to the large amount of information they are capable of providing (imaging data with a very high level of spectral detail) as compared to more standard imaging techniques. Whereas regular images represent the colour of each pixel using three spectral bands (RGB)[1],

[1] Corresponding to red, green and blue bands respectively.

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 115–135, 2013.

HISs provide, for the same pixel, hundreds of bands covering the spectrum from the visible to the longwave infrared (400-2500 nm). This means that any HIS provides a very detailed description of the spectral signature of each pixel in the image, and consequently, facilitates the detection and identification of individual materials or classes. Moreover, these very large amounts of data, through the use of new algorithms and image processing techniques, facilitate the segmentation and classification of different elements within the scenes. Typically, a HIS may consist of two spatial dimensions of anywhere from 250 pixels to a couple of thousand pixels and a spectral dimension of up to a thousand bands per pixel. Therefore, a HIS is a 3D matrix of values that is usually referred to as a hyperspectral cube.

The first applications using hyperspectrometers were in the field of remote sensing and the images were obtained from satellites or high flying planes [1]. As a consequence, most of the analysis methods developed were aimed at providing the ratio of endmembers present in every pixel when analysing different types of covers. In fact, in all of these applications the emphasis was placed on the processing of the spectral characteristics found in the different pixels while the spatial or morphological features were not taken into account. Currently, the field of application of this technology is changing due to the popularization of imaging sensors and the advances in digital photography and video capture technology, which are leading to new implementations of smaller designs and platforms. These technological improvements have opened up new areas, especially, in ground based applications [2, 3], where the images are taken close enough to the subject to obtain a relatively detailed view. Examples are inspection tasks such as medical imaging [4, 5], quality control in processing plants [6, 7], surveillance tasks, etcetera)

As mentioned above, the main characteristic of HISs is the large data sets that make up each image due to the spectral detail they include for each pixel. The size of these data sets becomes even larger when high spatial resolution is involved, which is often the case. This represents a severe processing problem that requires the development of specialized techniques and the use of very high levels of parallelism, especially when real time processing is desired. Thus, it is the case that one would like to have an algorithm that iteratively performs simple operations over pixels and their immediate surroundings so that the current state of the art distributed architectures, such as those based on Field Programmable Gate Arrays (FPGAs) [8] or GPUs [9], can be easily applied. This is not the usual case for most algorithms developed in this field; since their algorithmic structure makes them rather cumbersome when run using this type of architectures and ad-hoc very specialized implementations have to be developed in order to benefit from their characteristics.

The rest of the chapter is structured as follows. The next section is devoted to a brief review of the background on edge detection and segmentation of high dimensional images. After this section 2 provides a brief introduction to cellular automata and the particular type we consider here, the ECA. Section 4 is devoted to the evolution of cellular automata for the detection of edges. The problem of segmenting HISs is considered in Section 5. Finally, some conclusions and indications for future research are presented in Section 6.

2 Background

Our current line of work is concerned with the development of algorithms for efficiently performing different operations over high dimensional images. In particular, the work presented here is related to two basic operations: edge detection and segmentation in HISs. These operations are two of the most basic operations when spatially extended elements need to be obtained from the images and often require the combined use of both the spectral and the spatial information present, leading to what is usually called spatial-spectral processing.

In the area of edge detection no standardized algorithms exist for high dimensional images, as in the case, for example, of grey scale images, where these processes are well studied producing good results [10]. Some work has been carried out for hyperspectral and multispectral images using non statistical approaches, such as extensions of the Sobel or Prewitt operators, but they do not really address how to exploit the spectral detail provided, as these techniques are usually applied to each band and the results summed or aggregated through an OR operator [11].

In terms of the segmentation of HISs, there also exist different approaches which range from simple extensions of those considered in regular image processing such as watershed algorithms [12] to other more advanced techniques that try to take into account the information provided by the spectral detail of this type of images, such as in the work by [13–15], or even resort to lattice computing techniques as in [16]. Most of these approaches require complex processing stages that are very hard to implement in limited computing resources or highly parallel systems such as GPUs and are, consequently, inappropriate when contemplating real time processing.

Another, more promising approach would be to choose a computing or processing paradigm that is really adapted to this type of extremely distributed computing architectures and develop new algorithms that conform to this paradigm and that carry out the functions sought. This is the approach that has been followed here. This work addresses the problem of creating new algorithms for performing the operations indicated above over HISs using Cellular Automata (CA) based computing structures.

The application of cellular automata to solve a particular problem requires determining what each cell within it must do so that the whole system through their recurrent interactions solves the task in hand. That is, it is necessary to determine what rules implemented in each cell will produce, as the CA iterates, the desired global behaviour for the CA based system. This is the so called inverse problem and it is very difficult to solve. To attempt to solve the inverse problem, different approaches have been developed by different authors within the CA community (see, for example, [17]). However, the most popular approach has been using evolutionary techniques in order to evolve the set of rules [18–21].

Regarding HISs, CAs are currently being used by different authors in the development of new techniques in order to improve processing efficiency. For instance in [22] the authors propose an implementation of CAs for edge detection. However, the CAs they propose have been hand created ad-hoc and, even though they do perform quite well, they are still cumbersome involving two processing stages. Others have presented implementations to address segmentation tasks (see [23–25]). Again, the CAs they propose have been hand created and, even though some do consider multidimensional

images, they are still far from the dimensions of HISs and are usually projected onto a lower dimension during the segmentation process.

Here, the evolution of the CAs rule set will be considered in order to address the edge detection and the segmentation problems over HISs. To this end a particular type of CA is contemplated: the nine-cell neighbourhood CA (it is usually called a Moore neighbourhood). This is a system that, with appropriate rules, has also been shown to be capable of universal computation [26]. The results obtained will be compared to more traditional techniques.

3 Cellular Automata

Cellular automata are a biologically inspired decentralized computing paradigm first proposed by Von Neumann and Ulam [27]. They are usually presented as a lattice of CA cells that can only communicate with their immediate neighbours. Each cell is basically an automaton (usually a finite state automaton) and the state or value of a cell each instant of time is determined by the previous states or values of neighbourhood cells around it as well as its own value. A cell can take any value from a finite set or a continuous interval and the value it takes is determined by the application of a set of transition rules. Thus, it is by adequately choosing these rules and by iterating the state transition process in time that the full computational capabilities of the whole system are achieved. Though CA construction is simple and the computations each cell carries out are quite elementary and easily implementable in the computing architectures mentioned above, the resulting system may present a complex emergent self-organizing behaviour that needs to be harnessed. In fact, Von Neumann himself demonstrated that CAs can be used to make a universal computing machine.

This work considers the Moore type CA with a neighbourhood of one. As shown in Fig. 1., it is made up of nine cells. The value or state of the central cell is updated according to a set of rules. These rules depend on the values or states of the eight neighbouring cells as well as that of the cell itself.

A relationship may be established between a HIS and a 2D cellular automata such that each cell of this lattice corresponds to a pixel of the HIS. This means that each cell will be a vector which contains the spectrum values of its corresponding pixel. Therefore, in each step, each cell of the CA acts according to a set of predefined rules over its state (si) or, in image terminology, each pixel of the image is modified by the CA through a set of rules that affect its spectrum. These CAs make use of the information obtained from the eight pixels surrounding each pixel.

4 Evolving Cellular Automata for Detecting Edges

In this work a series of methods are considered to identify edges and for the unsupervised segmentation of HISs through evolved cellular automata (ECA). The difficulty when developing algorithms for these types of images is the large amount of data that must be processed. The methods developed to attempt to solve this problem are based on the use of Evolutionary Algorithms (EA) to evolve the rule sets for the cellular automata. In particular, this section is concerned with the problem of detecting edges in

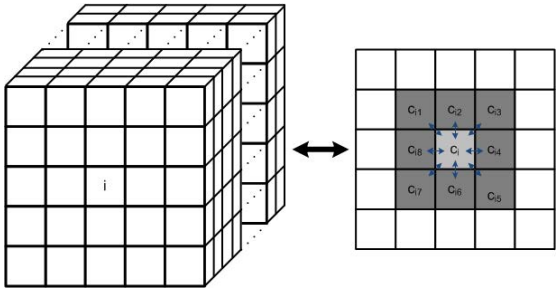


Fig. 1. Association between hyperspectral cube and CA. The CA operates over each pixel i depending on integrated measures related to some metrics with respect to the state of its eight neighbouring pixels

HISs using ECA. In this case, the function the ECA needs to perform is a projection or dimensionality reduction operation whereby each HIS, consisting of hundreds of bands per pixel, is projected onto a 2D binary image corresponding to a black and white representation of edges (white implies edge and black non edge). Thus, a CA needs to be defined and the appropriate rule set obtained so that when it is applied over HISs this projection is the desired one.

Subsection 4.1 describes how the ECA algorithm is used in the edge-detection method developed as well as the specific CA used and how it is encoded. Subsection 4.2 deals with the Evolutionary process carried out to produce the rules of the ECA. Finally, Subsection 4.3 details the experiments carried out and the results obtained.

4.1 Definition and Encoding of the CA

In order to define the ECA, apart from the fact stated above that it is a Moore type CA, it is first necessary to determine what inputs are going to be used so that it can perform its processing through the application of its rule set. Taking the direct values of the spectra of the neighbouring pixels when processing a given pixel (which would be the typical approach in the case of grey level images), would imply a huge number of inputs[2] making the rule set very difficult to obtain. To reduce the inputs, it was decided to provide a set of integrated measures related to the distances to the neighbours' states. Despite the fact that these measures reduce the number of inputs, they still contain sufficient information to process the image (as the results will show).

Thus, as a first step in a two-step process, the spectral vector s_i of each pixel i of matrix S is associated to the spectral vector of its neighbour through a distance metric. In this case, the distance metric chosen is the spectral angle α_{ij}. It is defined by Eq. (1).

$$\alpha_{ij} = \frac{2}{\pi} arc\cos\left(\frac{\sum \mathbf{s_{ij}} \mathbf{s_i}}{\sqrt{\sum \mathbf{s_{ij}^2}} \sqrt{\sum \mathbf{s_i^2}}}\right), \text{for } \mathrm{j} = 1,2,...,8 \tag{1}$$

[2] For each pixel, the inputs are eight neighbours times the number of bands in each spectrum, which can be over one hundred.

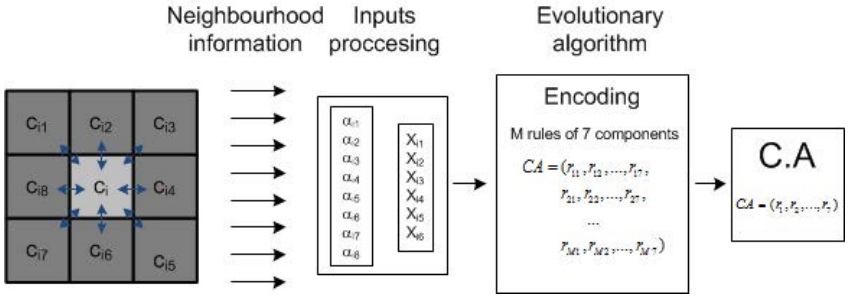


Fig. 2. Schematic of the algorithm used to obtain the cellular automata rules

This parameter provides an indication of the difference between spectra and thus takes into account the spectral information present. It is a very interesting parameter as it can be calculated in the same way regardless of the dimensionality of the spectrum. The spectral information is, admittedly, very summarized, but it can be taken as a valid first approximation to considering the effect of having spectral information present in the pixels. Thus, for each pixel i the eight spectral angles associated to the eight pixels that surround it can be calculated. These 8 spectral angles can be grouped into an Angle vector (Eq. (2)):

$$A_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{i8}) \tag{2}$$

Consequently, through this first step, a large reduction of the number of inputs the CA needs has been achieved. Initially, each pixel was represented as its complete spectral vector (usually more than 100 values), and now, the pixel contains a vector with only eight components.

With the objective of simplifying the rules that need to be applied by the CA and thus facilitate their evolution, it would be convenient to provide information that is as meaningful as possible to the CA. Consequently, in a second step, the angle vector is transformed into a new vector made up of six components $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$. These components represent aggregated descriptors derived from spectral angles of the angle vector and, consequently, depend on the features of the 8 neighbouring pixels too. They are calculated using the Eq. (3).

$$\begin{aligned}
x_{i1} &= mean\,(A_i) \\
x_{i2} &= -(\alpha_{i1} + 2 \cdot \alpha_{i2} + \alpha_{i3}) + \alpha_{i6} + 2 \cdot \alpha_{i7} + \alpha_{i8} \\
x_{i3} &= -(\alpha_{i1} + 2 \cdot \alpha_{i4} + \alpha_{i6}) + \alpha_{i3} + 2 \cdot \alpha_{i5} + \alpha_{i8} \\
x_{i4} &= -(\alpha_{i2} + 2 \cdot \alpha_{i3} + \alpha_{i5}) + \alpha_{i4} + 2 \cdot \alpha_{i6} + \alpha_{i7} \\
x_{i5} &= -(\alpha_{i4} + 2 \cdot \alpha_{i1} + \alpha_{i2}) + \alpha_{i7} + 2 \cdot \alpha_{i8} + \alpha_{i5} \\
x_{i6} &= std\,(A_i)
\end{aligned} \tag{3}$$

where:

x_{i1}	the mean spectral angle of the pixels surrounding pixel i
x_{i2}	vertical spectral gradient
x_{i3}	horizontal spectral gradient

x_{i4}, x_{i5} diagonal spectral gradients
x_{i5} is the number of pixels of region i
x_{i6} standard deviation of the spectral angles surrounding pixel i

And where all of these values are normalized in the $[0:1]$ range, and are commonly used in edge detection systems.

The output of the automaton is determined by the rules governing the CA and it can be white (meaning edge) or black (meaning non edge). In this case the CA is controlled by a set of M rules containing 6+1 components each (the values of the six inputs and the corresponding output value), where M is a configurable parameter that must be specified by the user and that is fixed for each evolution. Thus, the CA can be encoded as a vector of 7M floating point values between 0 and 1. This vector is shown in Eq. (4).

$$CA = (r_{11}, r_{12}, ..., r_{17}, r_{21}, r_{22}, ..., r_{27}, ..., r_{M1}, r_{M2}, ..., r_{M7}) \tag{4}$$

The first six components of the rules are related to the information on the six descriptor parameters and the seventh component of the each rule s (r_{s7}) is the value which decides whether a pixel is an edge or belongs to the background. To encode this decision the binarization given by Eq. (5) has been established:

$$Transition \begin{cases} \text{edgeif} & r_{i7} > 0.5 \\ \text{backgroundif} & r_{i7} <= 0.5 \end{cases} \tag{5}$$

Having reached this point, it is necessary to introduce a screening method to choose from the M rules the one that must be applied to a pixel at a given point of the process. The aim is to choose the output of the rule that is closest or most similar to the inputs. To achieve this, after computing the descriptor vector for pixel i $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$ the Euclidean distance between the descriptor vector and the first six components of each rule k $(r_{k1}, r_{k2}, r_{k3}, r_{k4}, r_{k5}, r_{k6}, r_{k7})$ is calculated. The rule that results in the minimum Euclidean distance is selected and the corresponding pixel of the output image takes the output value of the rule.

This process takes place for all the pixels in the HIS, producing a regular 2D black and white image where the white pixels represent edges and the black pixels non-edges. Obviously, for this to work appropriately and produce a real edge detection process, the rules that govern the CA must be created appropriately. As indicated above, and in order to prevent tedious and error prone hand design processes for the rules, in this work these rules will be evolved.

4.2 Evolutionary Process

As in any other evolutionary process, the object to be evolved must be somehow encoded into a representation that the evolutionary algorithm used can handle. In this case, the object of evolution are the rules and, as indicated in the previous section, these rule sets are given by a vector of dimension 7M whose values are floating point numbers in the [0, 1] interval. Consequently, a direct representation can be used where the genes directly correspond to the seven values of the M rules.

In addition, in order to evolve the CAs population it is necessary to be able to evaluate each candidate solution and assign a fitness value to it. This evaluation requires 3D (two spatial and one spectral dimension) HISs (at least one). The CAs that make up the population are evaluated using this images as well as a 2D black and white target version of the same image that provides the ground truth for the detection process. It is by comparing the images obtained through the evaluation performed by each CA to the target images that a measure of the fitness of that CA can be obtained. It is important to highlight here the fact that using different target images for a given hyperspectral training image will lead the evolutionary process towards CAs that produce different edge detection results.

There are many possible ways of measuring the difference between the results obtained by a CA and the target image. Here, one that provides balance between false positives and false negatives was chosen. This is an important aspect that will be considered later. Thus, the detection error ε has been defined, and it is calculated using Eq. (6):

$$\varepsilon = \max\left(\frac{sum\,(B_{Ideal}\&N_{CA})}{nB_{Ideal}}, \frac{sum\,(N_{Ideal}\&\mathrm{B}_{CA})}{nN_{Ideal}}\right) \tag{6}$$

where:

$sum\,(B_{Ideal}\&N_{CA})$	number of white pixels (edges) in the target image that are black in the resulting image (false negatives)
$sum\,(N_{Ideal}\&\mathrm{B}_{CA})$	number of black pixels (edges) in the target image that are white in the resulting image (false positives)
nB_{Ideal}	number of white pixels (edges) in the target image
nN_{Ideal}	number of black pixels in the target image that are black in the resulting image

The evolutionary algorithm chosen for these first tests as well as the parameters used in each evolution are described in the test section.

4.3 Experiments and Results

In this section, the results obtained by applying the ECA described above are presented. The experiments performed have been carried out in order to, firstly, validate the development method and, secondly, demonstrate how this method for obtaining CA based edge detection structures is able to adapt to several edge tracing strategies.

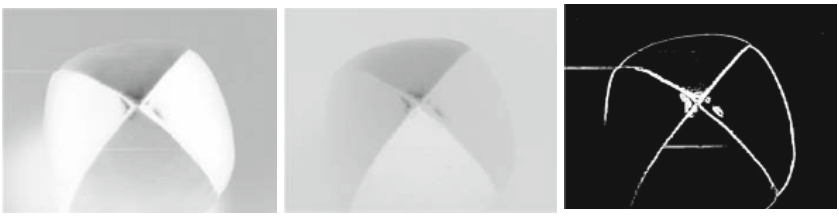
The first step of the algorithm developed consists in evolving the ECA over a number of training images to obtain the set of rules that govern its behaviour These training images consist of a HIS and its corresponding ground truth. These examples have made use of the images shown in Fig. 3, which correspond to indoor captures provided by the GIC[3]. It can be noted that these two HISs present some horizontal lines corresponding to artefacts produced by the hyperspectral camera which will also appear as artefacts in the edge detection results.

[3] Grupo de Inteligencia Computacional, Universidad del Pais Vasco (Spain).

(a) (b)

Fig. 3. 610 nm band of two indoor hyperspectral captures

To produce a ground truth image for these examples traditional edge detection algorithms have been employed. Consequently, in this case the ECA is not really being trained to produce the correct edge detection, but rather to behave like other edge detectors. The process followed to produce this ground truth consisted, first, in the application of a dimensionality reduction technique. The HIS was transformed by means of a Principal Component Analysis (PCA) procedure preserving 99% of the total variance. This means that the hyperspectral cube was reduced to the first five components, which are the ones that will be considered to calculate the parameters that will be introduced in the ECA. Figure 4 shows a representation of the training image. It displays the first principal component (Fig. 4(a)) and the second principal component (Fig. 4(b)) after the PCA transformation. Finally, a Sobel edge filter has been applied over the five principal components of the reduced hyperspectral cube. The ground truth is obtained by merging these resulting five edge images considering an infinite norm. Figure 4(c) shows this ground truth where the white pixels represent real edges and the black ones the background.



(a) (b) (c)

Fig. 4. (a) first principal components after applying a PCA transformation, (b) second principal components after applying a PCA transformation, (c) the ground truth of the training image in the first experiment (Sobel filter)

This ground truth was used in order to evolve the ECA. The Real Valued Genetic Algorithm (RVGA) algorithm (Real Valued Genetic Algorithm) was used to obtain the

optimum rule set for the ECA based edge detector. The HIS of Fig. 3(b) was chosen for the evaluation and its corresponding ground truth is that of Fig. 4(c).

As defined above, the ECA is controlled by a set of M rules, each one of them, as indicated in the previous section, defined by seven real parameters or components. In this example the maximum number of rules was limited to 20. Therefore, each individual in the population to evaluate is made up of 140 parameters.

Concerning the RVGA parameters, a population size of 450 individuals and a maximum of 100 generations of evolution were used. The objective was the minimization of the error function defined in the previous section. A summary of the parameters used in the evolutionary process is presented in Table 1.

Table 1. Parameters of the evolutionary algorithm

Type of parameters	**Real**
Number of parameters	140
Population size	450
Number of generations	100
Crossover method	Heuristic child = parent2 + 1.2 * (parent1 - parent2)
Crossover fraction	0.8
Elitism (number of individuals)	2
Mutation method	Gaussian
Selection method	Stochastic uniform
Stopping criteria	Generations

Figure 5 displays the resulting fitness evolution for the best individual and the average for the whole population during the ECA adaptation process. Throughout the evolutionary process, every individual of the population was evaluated using a fixed percentage of pixels randomly selected from the training image. Consequently, the best individual fitness may not be representative of the fitness if applied to the whole training set, especially during the first steps of evolution. For this reason, the evaluation of the fitness obtained during this evolution should take into account, not only the best individual fitness for every time step, but also the average fitness of the population.

After the evolutionary process was completed, the best individual was selected and applied to the two images. Figure 6 and Fig. 7 are some examples of the results produced by the ECA after one pass with no further processing over different images. Figure 7 corresponds to a HIS capture in an outdoor scenario. This image was captured using a hyperspectral camera designed and constructed in our research group and shows a view of a waterway in Ferrol (Spain).

In order to measure the goodness of the ECA obtained through evolution for edge detection tasks, the results obtained after applying the ECA were contrasted to the results of an edge detection process performed by means of the Sobel method for HISs (or Hyper-Sobel). The detection accuracy for the three test images is displayed in Table 2. These results are very acceptable taking into account that the training image (Fig. 3(a)) was actually obtained using a Hyper-Sobel method. In other words, an ECA was obtained that mimics the behaviour of a Hyper-Sobel filter very accurately.

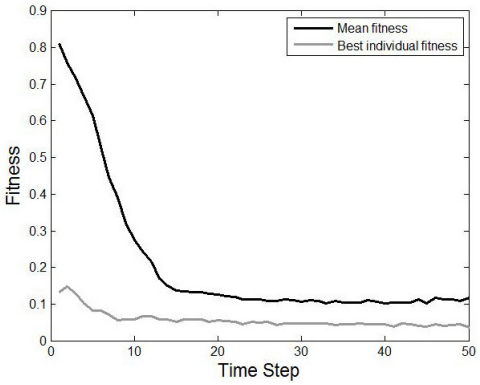
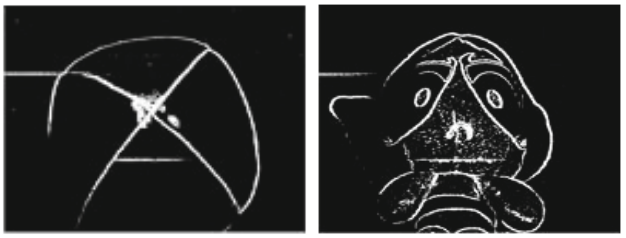


Fig. 5. Fitness evolution for the best individual and the average for the whole population during the CA adaptation process



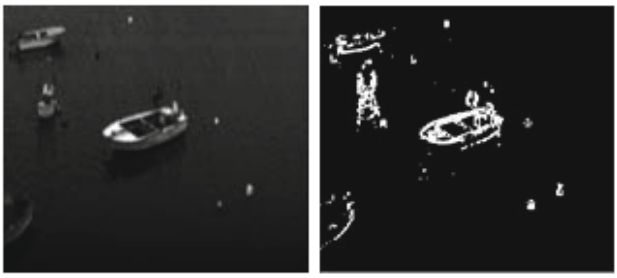
(a) (b)

Fig. 6. (a) and (b) are the resulting images for the edge detection process after applying the ECA obtained in experiment 1 to the HISs shown in Fig. 2

The results of this first experiment show the capabilities of the proposed method for achieving satisfactory edge detection by means of ECA. At this point, the ECA will be tested to know how its behaviour adapts or reproduces other edge detection strategies different from the one used in the first experiment. For this purpose, in a second experiment, the evolutionary process was repeated using the same training image but a different ground truth. In this case, a vertical Prewitt edge filter was applied to the five principal components of the reduced hyperspectral cube and the resulting five edge images were joined together by means of an infinite norm. This is what some authors have called a Hyper-Prewitt operator. Figure 8 shows the resulting edge detection after using the Hyper-Prewitt edge filter on the PCA transform of the indoor HIS set, which is the ground truth for this experiment. It can be observed that the horizontal artefacts in Fig. 6(a) and Fig. 6(b) disappear by using this edge detection method.

The same configuration parameters as those used for the first experiment (see Table 3) were chosen for the evolutionary process carried out in this second experiment.

The results of applying the algorithm to the two HISs of Fig. 3 are displayed in Fig. 9 and Table 3.



(a) (b)

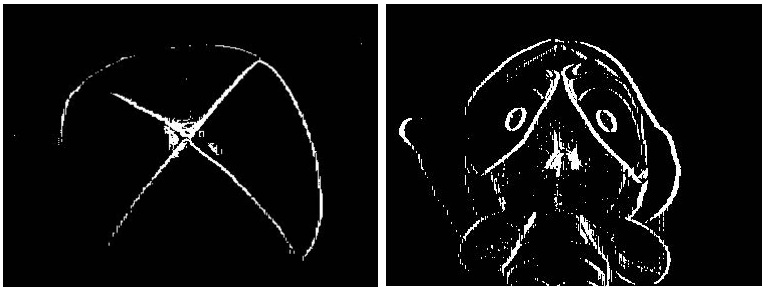
Fig. 7. (a) one band of the HIS captured, (b) edge detection after applying the ECA

Table 2. Detection accuracy in experiment 1

	TRUE NEGATIVES (%)	TRUE POSITIVES (%)
Image 1 (Fig. 3(a))	99.90%	98.02%
Image 2 (Fig. 3(b))	99.52%	97.23%
Image 3 (Fig. 7(a))	99.11 %	96.38%

(a) (b)

Fig. 8. Ground truth for the second experiment (Prewitt filter). Image(a) is the one used during the evolutionary process to train the ECA



(a) (b)

Fig. 9. Edge detection after applying the ECA obtained in experiment 2 to the HISs shown in Fig. 3

Table 3. Detection accuracy in experiment 2

	TRUE NEGATIVES (%)	**TRUE POSITIVES** (%)
Image 1 (Fig.3(a))	99.83%	97.67%
Image 2 (Fig.3(a))	99.65%	97.08%

It can be noted that this second ECA is able to satisfactorily perform the edge detection task using a different type of strategy than that of the first experiment, and the results are also very successful. In this case an ECA that mimics a Prewitt filter was obtained.

These two experiments indicate the ease with which cellular automata can be obtained to perform different edge detection functions through evolution and how these evolved ECAs can be adapted to achieve a good edge detection based on traditional techniques.

5 Evolving Cellular Automata for Segmentation

In the case of the edge detection procedure using ECA only one iteration is applied to the hyperspectral cube. After the first iteration, the hyperspectral cube is transformed into a binary image, and thus it is not necessary (nor possible) to repeat the ECA procedure. However, in the ECA based segmentation approach, the ECA is iteratively applied over the hyperspectral cube to modify the state (spectrum) of each pixel in each iteration. This modification depends, on the information of the spectra of the eight closest neighbouring pixels, on the set of rules that control the automaton and, finally, on the state of the cell over which the automaton is applied. Figure 10 displays a schematic of the operation of the ECA over the hyperspectral cube and how it has been modified.

Following the same structure as for the previous chapter, in this one, Subsection 5.1 explains the algorithm and the encoding of the ECA. Subsection 5.2 details how the rule set is evolved and Subsection 5.3 presents the experiments carried out and the results obtained.

5.1 Definition and Encoding of the ECA

Again, unlike the previous case, the information the state of the automata handles is not going to be reduced. The spectral character of the information will be preserved. In order to evaluate the difference between these states so as to be able to perform operations, the spectral angle will be used as distance measure. This angle is calculated for each of the eight surrounding pixels (normalized in the range [0,1]). Thus, for each pixel angle vector A_i is constructed as (Eq. (7)).

$$A_i = (\alpha_{i1}, \alpha_{i2}, ..., \alpha_{i8}) \tag{7}$$

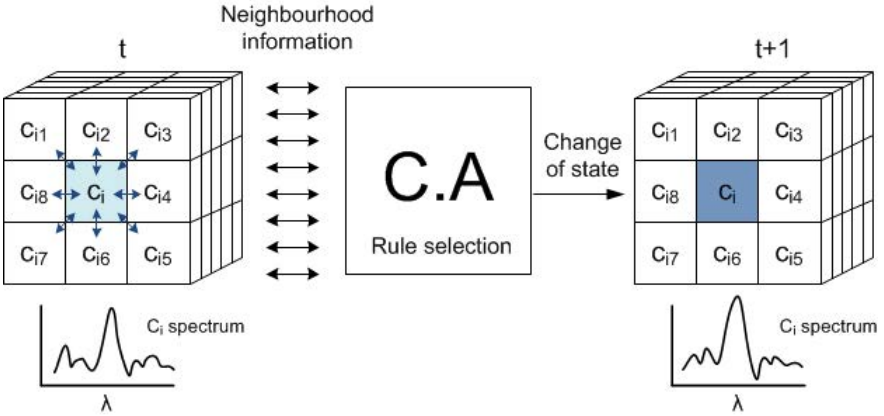


Fig. 10. Schematic of the operation of the cellular automaton over a hyperspectral cube for the segmentation task

To simplify the processing within the rules of the ECA a discretization of the spectral angles is carried out, so that they take values from -1, 0, 1. This discretization is performed according to Eq. (8).

$$\alpha_{c_{ij}} = \begin{cases} -1 \; if \quad \alpha_{ij} < \varepsilon \\ 0 \;\; if \; \varepsilon \leq \alpha_{ij} \leq \theta \\ 1 \;\; if \quad \alpha_{ij} > \theta \end{cases} \tag{8}$$

where ε and θ are two parameters of the ECA.

Thus, the cellular automata consists, in this case, of two angles, ε and θ, and a set of M rules, each one of them made up of 8+1 parameters. Thus, the automata can be encoded as a vector with (9M +2) floating point values between 0 and 1 (Eq. (9)).

$$CA = (\varepsilon, \theta, r_{11}, r_{12}, ..., r_{19}, r_{21}, r_{22}, ..., r_{29}, ..., r_{M1}, r_{M2}, ..., r_{M9}) \tag{9}$$

The first eight r_{xj} parameters of each rule are related to the information on the spectral angle of the neighbours with respect to the pixel being processed. The ninth parameter is related to the change of state of the cell and points towards one of the eight neighbours.

When the automaton is executed over a given pixel, it compares the angles of the pixel to those of all of its neighbours constructing vector A and transforms the angles α_{ij} into c_{ij} as stated above using the first two values of the automaton, ε and θ. It then finds the rule whose first eight r values are most similar to A and applies it. Basically, when a given rule is applied the spectrum of the pixel being processed is moved in the direction of the spectrum of the pixel towards which the rule points (in the case of the examples in this paper they are averaged). Thus, the change of state of each cell corresponds to a continuous variation of the spectrum of the cell. In order to make rule selection invariant to rotations and reflections four possible rotation and the vertical and horizontal flips of each rotation are taken into account.

The cellular automaton is applied to all the pixels of the HIS sequentially each iteration producing each instant a hyperspectral cube that has been modified with respect to the one corresponding to the previous instant.

5.2 Evolutionary Process

Again, after encoding the automata as a vector with (9M +2) floating point values, and as in the previous case, it is necessary to provide a set of examples over which the prospective ECAs can be evaluated during evolution. As it is very difficult to obtain correctly labelled HISs and since the distance measure used is independent of dimensionality, it is assumed that the ECA could be evolved using a lower dimensional training set for it to learn a rule set that performs the desired function and then applied to higher dimensional images and obtain the same type of results. To this end synthetic RGB images were used with the variability range sought. Each one is generated at run time for the evaluation of an individual. Basically the prospective cellular automaton is run over the image for a given number of iterations and the result is compared to a desired ground truth. It is important to note that a different image is generated for each evaluation.

Out of the different possibilities to measure the quality of the segmentation obtained after applying the automaton in order to provide a fitness value, a measure that provides a balance between the homogeneity of the segmented regions and the discrimination of the different regions within the HIS has been chosen. It is given by Eq. (10).

$$\begin{aligned} \varepsilon &= \max(A,B) \\ A &= \frac{\sum_{i=1}^{K} \max(std(\mathbf{p}_{\mathbf{i},\mathbf{1}}), std(\mathbf{p}_{\mathbf{i},\mathbf{2}}), \ldots, std(\mathbf{p}_{\mathbf{i},\mathbf{N}}))}{K \cdot 0.5} \\ B &= 1 - \frac{\sum_{i=1}^{K} \left(\sum_{l=1}^{n_{ri}} f(\alpha_{l,m_i}, \alpha_{th})/n_{ri} \right)}{K} \end{aligned} \tag{10}$$

where:

$$f(\alpha_{l,m_i}, \alpha_{th}) = \begin{cases} \alpha_{l,m_i} & if \quad \alpha_{l,m_i} > \alpha_{th} \\ 0 & f \quad \alpha_{l,m_i} \leq \alpha_{th} \end{cases} \tag{11}$$

and where:

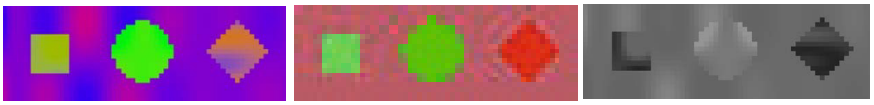
K	number of different regions in the HIS
N	number of bands in the image
$\mathbf{p}_{1,n}$	reflectance value of the n^{th} band of the pixels belonging to region i
$std(\mathbf{p}_{\mathbf{i},n})$	standard deviation of $\mathbf{p}_{1,n}$
n_{ri}	is the number of pixels of region i
$\alpha_{l,m}$	spectral angle between pixel l of region i and the average spectrum of the region i
α_{th}	threshold angle value

The next section describes a series of tests that have been carried out as well as the experimental setup.

5.3 Experiments and Results

This section provides some results of using evolution to obtain cellular automata for segmenting HISs in an unsupervised manner. The automata were obtained using the algorithm described below and were then tested over two different types of images: a set of synthetic images, and the well-known AVIRIS Salinas image for comparison purposes.

Unlike the method for detecting edges described in the previous section, here, it is necessary generate synthetic images at runtime to train the genetic algorithm. To evaluate each candidate automaton in the population a different image is created, although all of them had similar types of features corrupted by noise and artefacts. The images used for this task are similar to those shown in Fig.11.



(a) (b) (c)

Fig. 11. (a) Training RGB image, (b) Ground truth, (c) Spectral angle image with respect to an arbitrary fixed vector

The particular evolutionary algorithm chosen for obtaining the rule sets of the cellular automata in these experiments was, again, a standard Real Valued Genetic Algorithm (RVGA). Automata with a total of 20 rules were used such as those described in the previous section. As each rule has nine components and there are two additional parameters corresponding to the threshold angles, each automaton ends up with 182 real valued parameters. A summary of the parameters for this RVGA are displayed in Table 4.

Figure 12 displays the evolution of the fitness of the best individual and the average fitness of the population for one of the evolutionary processes. The fluctuations that can be observed in the fitness curves are due to the fact that the evaluation of the individuals is carried out over different images. However, the global trend followed by the curves is clear.

The first experiment consisted in testing the ECAs over synthetic images that were generated much in the same way as those used for evolution but the shapes of the areas to be segmented were completely different from anything seen during training. Figure 13 shows the result of applying the ECA obtained over one of these synthetic images. In this case, unlike the training images which were RGB images, here a series of HISs constructed using four 64 band base spectra were used. These images were spatially corrupted by noise and artefacts as represented in Fig.13(a) (as a 2D projection of the spectrum as spectral angles with respect to a fixed reference). Figure13(b) displays the segmentation obtained and Fig.13(c) the ground truth.

Table 4. Parameters of the evolutionary algorithm

Type of parameters	Real
Number of parameters	182
Population size	182
Number of generations	50
Crossover method	Scattered)
Crossover fraction	0.8
Elitism (number of individuals)	2
Mutation method	Gaussian
Selection method	Tournament (size = 4)
Stopping criteria	Generations

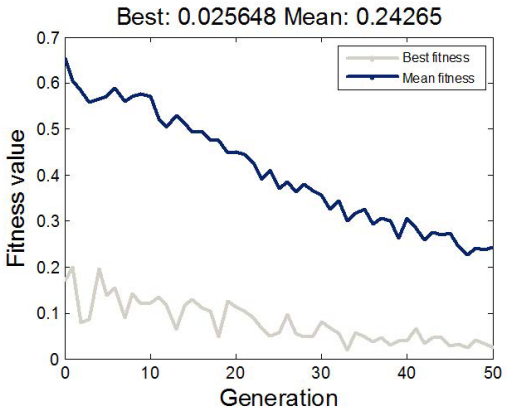


Fig. 12. Fitness function value (minimized) during one of the evolutionary processes

The results obtained in this test are quite good. This ECA, after iterating 100 times over the image, obtained an Overall Accuracy (OA) of 99.9%.

Experiments: After evolving the ECA and obtaining the rule set that governs its behaviour, it is necessary to evaluate its performance. Accordingly two tests were carried out:

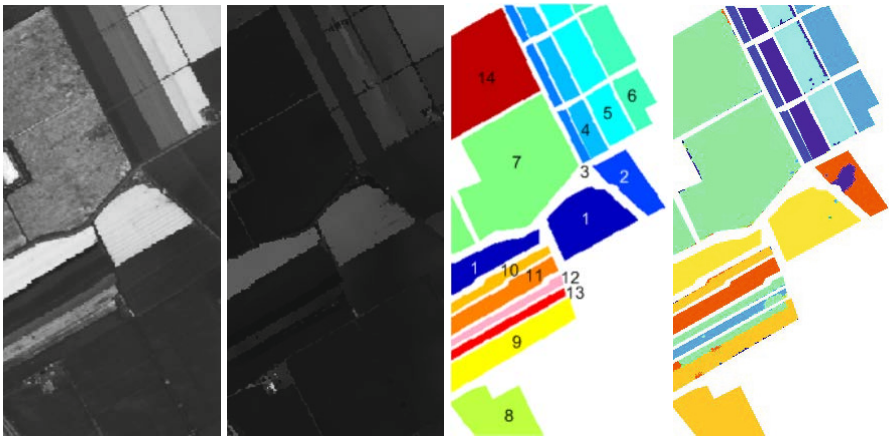
One: Test over synthetic images.

Two: Test over real images.

After testing the automata over synthetic images, it was tested over a set of real images. To show some results the AVIRIS Salinas HIS, which is a typical case of study in the hyperspectral field, was selected. The rightmost part of Fig. 14 shows the results of this segmentation.

(a) (b) (c)

Fig. 13. (a) 2D transformation of a synthetic 64 band HIS, Each pixel displays the angle between the spectrum of the pixel and a reference spectrum with all of its bands at the maximum value, (b) 2D transformation of the cube resulting from the segmentation obtained by the ECA, (c) Labelling of each area



(a) (b) (c) (d)

Fig. 14. (a) 2D angular transformation of the AVIRIS scene, (b) 2D angular transformation of the hyperspectral cube resulting from iterating the ECA over the image 100 times, (c) ground truth (the content of each area is indicated in Table 5), (d) result of the unsupervised CA based segmentation process (the colours do not correspond to labels, they just indicate homogeneous regions or segments as determined by the ECA)

These results are visually quite acceptable and show that the ECA has obtained a set of rules that allow it to clearly segment the image into the relevant areas. However, there are certain areas where the results differ from the ground truth. In areas 2 and 9 of the Fig.14(c) the segmentation presents a slight over-segmentation, for each one the ECA delimits two categories. Also, in the borders of some areas some noisy pixels are present. It is difficult to really know whether the aforementioned over-segmentation represents some relevant feature of the image as the only information available on the area is the ground truth image shown in figure 5. However it can be said that the areas are correctly segmented because they present different spectra in the original image. On the other hand, the noisy pixels in the borders of some areas have to do with imprecise

Table 5. Overall accuracy (%)

Area	Cover	OA(%)
1	Broccoli	99.7
2	Fallow	82.4
3	Fallow rough plough	100
4	Fallow smooth	98.4
5	Stubble	93.8
6	Celery	99.5
7	Grapes	98.7
8	Soil vineyard develop	100
9	Corn senesced green weeds	70.3
10	Lettuce romaine 4wk	94.8
11	Lettuce romaine 5wk	100
12	Gaussian	97.8
13	Stochastic uniform	83.4
14	Generations	99.3

delimitation of the areas in the ground truth (borders of fields that do not contain plants and are taken as paths). However, to really be able to quantify the goodness of the segmentation the OA index is used again, in this case for each one of the areas with the same label in the ground truth image. These OA values are presented in Table 5 and they lead to a global OA of 94.1%.

It is important to note that the unsupervised segmentation procedure presented here seeks to provide a single label in terms of a single spectrum to represent each area; however, the problem under study is segmentation, not classification. As a consequence, the representative spectrum may not be the one corresponding to the ground truth label. Thus, when two separate areas in an image that contain the same type of cover are segmented they may be assigned different labels (much in the same way as in the case watershed based segmentation algorithms). That is, for instance, in an area where there are very small plants each surrounded by a lot of soil, the representative spectrum obtained in an unsupervised manner may be very close to soil. Thus if one takes it at face value, without a posterior labelling or classification process, this may lead to confusion between areas that present this characteristic.

6 Conclusions

The detection of edges and segmentation of HISs are two of the most basic operations when spatially extended elements need to be obtained from the images. They often require the combined use of both the spectral and the spatial information present, leading to what is usually called spatial-spectral processing. Consequently, this task should be as efficient as possible so as to be able to perform it in real-time with adequate reliability when in the field. In this line, using CA based structures for performing spatial-spectral operations over HISs is quite a promising approach. This is due to fact that these structures are very well adapted to their implementation in very efficient high performance

processing hardware such as GPUs. Nonetheless, this approach poses the problem of the need to generate the rule set for the CA to perform the task it has been assigned. This paper describes two methods that permit obtaining the corresponding automata through evolutionary processes. These methods are competitive and, more importantly, they can adapt better to changes in the way the user wants the edge detection or the segmentation to be performed than other more traditional approaches.

The results presented were obtained using a distance metric based on the spectral angle. Even though it does consider the spectral information present in the images, it does not do it in a very detailed fashion. The reason is that it considers large classes of equivalence that may not be suitable for some applications. As a consequence, an improvement in performance could be expected if more detailed spectral information were considered in the distance metric. Actually this work is continuing and focusing on CA based approaches that take into account the detailed information provided by the spectra directly in the distance metric in order to benefit from this wealth of information.

Acknowledgement. This work was partially funded by the Spanish MICINN through project TIN2011-28753-C02-01 and the Xunta de Galicia and European Regional Development Funds through projects 09DPI012166PR and 10DPI005CT.

References

[1] Glackin, D.L., Peltzer, G.R.: Civil, Commercial, and International Remote Sensing Systems and Geoprocessing. AIAA (American Institute of Aeronautics & Astronautics) (1999)
[2] Pan, Z., Healey, G.E., Prasad, M., Tromberg, B.J.: Hyperspectral face recognition for homeland security. In: AeroSense 2003. International Society for Optics and Photonics, pp. 767–776 (2003)
[3] de Juan, A., Tauler, R., Dyson, R., Marcolli, C., Rault, M., Maeder, M.: Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis. TrAC Trends in Analytical Chemistry 23(1), 70–79 (2004)
[4] Siddiqi, A.M., Li, H., Faruque, F., Williams, W., Lai, K., Hughson, M., Bigler, S., Beach, J., Johnson, W.: Use of hyperspectral imaging to distinguish normal, precancerous, and cancerous cells. Cancer Cytopathology 114(1), 13–21 (2008)
[5] Li, Q., Wang, Y., Liu, H., Sun, Z., Liu, Z.: Tongue fissure extraction and classification using hyperspectral imaging technology. Applied Optics 49(11), 2006–2013 (2010)
[6] Okamoto, H., Lee, W.: Green citrus detection using hyperspectral imaging. Computers and Electronics in Agriculture 66(2), 201–208 (2009)
[7] Liu, L., Ngadi, M., Prasher, S., Gariépy, C.: Categorization of pork quality using gabor filter-based hyperspectral imaging technology. Journal of Food Engineering 99(3), 284–293 (2010)
[8] Chuanwu, Z.: Performance analysis of the cpld/fpga implementation of cellular automata. In: International Conference on Embedded Software and Systems Symposia, ICESS Symposia 2008, pp. 308–311. IEEE (2008)
[9] Heras, D., Argello, F., Gmez, J., Becerra, J., Duro, R.: Towards real-time hyperspectral image processing, A gp-gpu implementation of target identification, vol. 1, pp. 316–321 (2011)
[10] Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 4109(6), 679–698 (1986)

[11] Verzakov, S., Paclík, P., Duin, R.P.W.: Edge detection in hyperspectral imaging: Multivariate statistical approaches. In: Yeung, D.-Y., Kwok, J.T., Fred, A., Roli, F., de Ridder, D. (eds.) SSPR 2006 and SPR 2006. LNCS, vol. 4109, pp. 551–559. Springer, Heidelberg (2006)
[12] Tarabalka, Y., Chanussot, J., Benediktsson, J.A.: Segmentation and classification of hyperspectral images using watershed transformation. Pattern Recognition 43(7), 2367–2379 (2010)
[13] Li, J., Bioucas-Dias, J.M., Plaza, A.: Hyperspectral image segmentation using a new bayesian approach with active learning. IEEE Transactions on Geoscience and Remote Sensing 49(10), 3947–3960 (2011)
[14] Veracini, T., Matteoli, S., Diani, M., Corsini, G.: Robust hyperspectral image segmentation based on a non-gaussian model. In: 2010 2nd International Workshop on Cognitive Information Processing (CIP), pp. 192–197. IEEE (2010)
[15] Duro, R., Lopez-Pena, F., Crespo, J.: Using gaussian synapse anns for hyperspectral image segmentation and endmember extraction. Computational Intelligence for Remote Sensing, 341–362 (2008)
[16] Graña, M., Villaverde, I., Maldonado, J.O., Hernandez, C.: Two lattice computing approaches for the unsupervised segmentation of hyperspectral images. Neurocomputing 72(10), 2111–2120 (2009)
[17] Ganguly, N., Sikdar, B.K., Deutsch, A., Canright, G., Chaudhuri, P.P.: A survey on cellular automata (2003)
[18] Packard, N.H.: Adaptation toward the edge of chaos. University of Illinois at Urbana-Champaign, Center for Complex Systems Research (1988)
[19] Mitchell, M., Hraber, P.T., Crutchfield, J.P.: Revisiting the edge of chaos: Evolving cellular automata to perform computations. Complex Systems 7, 89–130 (1993)
[20] Subrata, R., Zomaya, A.Y.: Evolving cellular automata for location management in mobile computing networks. IEEE Transactions on Parallel and Distributed Systems 14(1), 13–26 (2003)
[21] Elmenreich, W., Fehérvári, I.: Evolving self-organizing cellular automata based on neural network genotypes. In: Bettstetter, C., Gershenson, C. (eds.) IWSOS 2011. LNCS, vol. 6557, pp. 16–25. Springer, Heidelberg (2011)
[22] Lee, M., Bruce, L.: Applying cellular automata to hyperspectral edge detection. In: 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2202–2205. IEEE (2010)
[23] Wang, D., Kwok, N., Jia, X., Fang, G.: A cellular automata approach for superpixel segmentation. In: 2011 4th International Congress on Image and Signal Processing (CISP), vol. 2, pp. 1108–1112. IEEE (2011)
[24] Kauffmann, C., Piché, N.: Seeded nd medical image segmentation by cellular automaton on gpu. International Journal of Computer Assisted Radiology and Surgery 5(3), 251–262 (2010)
[25] Gallego, J., Hernandez, C., Graña, M.: A morphological cellular automata based on morphological independence. Logic Journal of IGPL 20(3), 617–624 (2012)
[26] Smith, A.: Introduction to and survey of polyautomata theory. In: Automata, Languages, Development. North Holland Publishing Co. (1976)
[27] Von Neumann, J., Burks, A.W.: Theory of self-reproducing automata (1966)

Bootstrapped Dendritic Classifiers in MRI Analysis for Alzheimer's Disease Recognition

Darya Chyzhyk and Manuel Graña

Grupo de Inteligencia Computacional, UPV/EHU
daria.chizhik@gmail.com

Abstract. This paper presents an intelligent approach to classification analysis of Alzheimer disease patients. Bootstrap technique is chosen to get rid of weak point of Dendritic Classifiers (DC), which is low Specificity and improve the Accuracy at all. Bootstrapped Dendritic Classifiers (BDC) is an ensemble of weak DC trained combining their output by majority voting. Weak DCs are trained on bootstrapped samples of the train data setting varying the depth by limit number of trees and varying number of dendrites. The classification accuracies of the combined LICA-DC, Kernel LICA-DC and BDC are compared. The experimental on T1-weighted Magnetic Resonance Imaging (MRI) images indicate that the developed method can significantly improve classification results.

Keywords: Lattice Computing, Alzheimer's Disease, Dendritic Classifiers.

1 Introduction

The Random Forest is a machine learning algorithm usually used for classification and regression problems. RF algorithm is a classifier [5] that encompasses bagging [4] and random decision forests [1], being used in a variety of pattern recognition and medical image applications [2]. RF became popular due to its simplicity of training and tuning while offering a similar performance to boosting. In the BDC proposed here the weak classifiers are built using DC [3,10,11], a simple, fast, efficient biologically inspired method to build up classifiers for binary class problems. Specifically the Single Neuron Lattice Model with Dendrite Computation (SNLDC), has been proved to compute a perfect approximation to any data distribution [11], however it suffers from over-fitting problems which we have been working to overcome in previous works [6].

The role of structural changes analysis in the brain on magnetic resonance imaging (MRI) is increasing nowadays and it is also essential in the study of neurological and psychiatric diseases. The target application of our work is the classification of Alzheimer's Disease (AD) patients using features extracted from brain MRI scans. Our experimental data is a database of MRI features[1] extracted from the OASIS database of MRI scans of AD patients and controls [14,13,8].

[1] http://www.ehu.es/ccwintco/index.php/GIC-experimental-databases

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 136–142, 2013.

Specifically, we selected from the OASIS database a well balanced subset of AD patients and controls of the same sex. Then we performed a Voxel Based Morphometry (VBM) analysis to determine the location of the voxel clusters most affected by the disease. The values of the gray matter segmentation of each MRI scan in the sites of these voxel clusters were collected to compute feature vectors for classification. In this paper, the feature vectors are built gathering the mean and standard deviation of the voxel gray matter segmentation values of all and each of the detected clusters.

Section 2 introduces the BDC. Section 3 recalls the characteristics of the experimental dataset. Section 4 provides experimental results. Finally, section 5 gives our conclusions.

2 Bootstrapped Dendritic Classifiers

DC is a model inspired by biological brain cell architecture which takes into account the computation performed by dendrites in the neuron. A single layer morphological neuron endowed with dendrite computation based on lattice algebra was introduced in [11]. The response of the j-th dendrite is as follows:

$$\tau_j\left(x_i\right) = p_j \bigwedge_{k \in I_j} \bigwedge_{l \in L_{kj}} (-1)^{1-l} \left(x_{i,k} + w_{kj}^l\right), \tag{1}$$

where $l \in L_{kj} \subseteq \{0,1\}$ identifies the existence and inhibitory/excitatory character of the weight, $L_{ij} = \emptyset$ means that there is no synapse from the i-th input neuron to the j-th dendrite; $p_j \in \{-1,1\}$ encodes the inhibitory/excitatory response of the dendrite. The total response of the neuron is given by:

$$\tau\left(x_i\right) = f\left(\bigwedge_{j=1}^{J} \tau_j\left(x_i\right)\right),$$

where $f\left(x\right)$ is the Heaviside hardlimiter function. A constructive algorithm [11] obtains perfect classification of the train dataset using J dendrites.

The BDC is a collection of DCs,

$$C(x;\psi_j), j = 1, ..., N,$$

where ψ_j are independent identically distributed random vectors whose nature depends on their use in the classifier construction. Each DC classifier casts a unit vote for the most popular class of input x. Given a dataset of n samples, a bootstrapped training dataset is used to train DC $C(x;\psi_j)$. The independent identically distributed random vectors ψ_j determine the result of bootstrapping. In conventional RF they also determine the subset of data dimensions $\hat{d}$ such that $\hat{d} \ll d$ on which each tree is grown, in this paper we are not dealing with this kind of DC randomization, which will be studied elsewhere. The main parameters for the experimental evaluation of the BDC are the number of trees and the

Algorithm 1. Crossvalidation scheme for the training of the BDC

Let be $X = \{x_1, \ldots, x_n\}$ input data $x_i \in \mathbb{R}^d$, and $Y = \{y_1, \ldots, y_n\}$ the input data class lables $y_i \in \{0, 1\}$.
N is the number of DC classifiers

1. for i=1:10 (crossvalidation folds)
   (a) select disjoint train $X^e = \left\{x^e_{i_1}, \ldots, x^e_{i_{n-n/10}}\right\} \subset X$, $Y^e = \left\{y^e_{i_1}, \ldots, y^e_{i_{n-n/10}}\right\} \subset Y$ and test $X^t = \left\{x^t_{i_1}, \ldots, x^t_{i_{n/10}}\right\} \subset X$, $Y^t = \left\{y^t_{i_1}, \ldots, y^t_{i_{n/10}}\right\} \subset Y$ datasets .
   (b) For $j = 1 : N$ (construct of classifiers)
      i. Bootstrap a train dataset $X^{eb} = \left\{x^{eb}_{i_1}, \ldots, x^{eb}_{i_{n-2n/10}}\right\} \subset X^e$, $Y^{eb} = \left\{y^{eb}_{i_1}, \ldots, y^{eb}_{i_{n-2n/10}}\right\} \subset Y^e$. Out-of-bag error may be computed on the remaining training data and test $X^e - X^{eb}$, $Y^e - Y^{eb}$, disjunctions.
      ii. Apply DC to train classifier $C_j : \mathbb{R}^d \to \{0, 1\}$ on $\left(X^{eb}, Y^{eb}\right)$.
   (c) end for. Optionally compute out of bag error
   (d) Crossvalidation test, For each $x \in X^t$
      i. compute $C_1(x), \ldots, C_N(x)$
      ii. Majority voting, class $y = 0$ if $|\{j \,|C_j(x) = 0\}| > |\{j \,|C_j(x) = 1\}|$
   (e) compute accuracy, sensitivity and specificity statistics
2. end fold i

maximum depth of each DC given by the maximum number of dendrites allowed. Limiting the number of dendrites is a kind of regularization that weakens the classifier. Finally, Algorithm 1 specifies the crossvalidation scheme applied in the experiments.

3 Experimental Data

Ninety eight right-handed women (aged 65-96 yr) were selected from the Open Access Series of Imaging Studies (OASIS) database [9]. OASIS data set has a cross-sectional collection of 416 subjects covering the adult life span aged 18 to 96 including individuals with early-stage AD. We have ruled out a set of 200 subjects whose demographic, clinical or derived anatomic volumes information was incomplete. For the present study there are 49 subjects who have been diagnosed with very mild to mild AD and 49 non-demented. A summary of subject demographics and dementia status is shown in table 1.

The OASIS database has been built following a strict imaging protocol, to avoid variations due to imaging protocol which would pose big image normalization problems. Multiple (three or four) high-resolution structural T1-weighted Magnetization-Prepared Rapid Gradient Echo (MP-RAGE) images were acquired [7] on a 1.5-T Vision scanner (Siemens, Erlangen, Germany) in a single imaging session. Image parameters: TR= 9.7 msec., TE= 4.0 msec., Flip angle=

Table 1. Summary of subject demographics and dementia status. Education codes correspond to the following levels of education: 1: less than high school grad., 2: high school grad., 3: some college, 4: college grad., 5: beyond college. Categories of socioeconomic status: from 1 (biggest status) to 5 (lowest status). CDR. MMSE score ranges from 0 (worst) to 30 (best).

	Very mild to mild AD	CS
No. of subjects	49	49
Age	78.08 (66-96)	77.77 (65-94)
Education	2.63 (1-5)	2.87 (1-5)
Socioeconomic status	2.94 (1-5)	2.88 (1-5)
CDR (0.5 / 1 / 2)	31 / 17 / 1	0
MMSE	24 (15-30)	28.96 (26-30)

10, TI= 20 msec., TD= 200 msec., 128 sagittal 1.25 mm slices without gaps and pixels resolution of 256×256 (1×1mm).

The feature vector extraction processes is based on the voxel location clusters detected as a result of a VBM analysis . The VBM detected clusters are used as masks to determine the voxel positions where the features are extracted. These masks are applied to the GM density volumes result of the segmentation step in the VBM analysis. The feature extraction process computes the mean and standard deviation of the GM voxel values of each voxel location cluster, we denote these features as MSD in the result tables given below.

4 Experimental Results

We report the average accuracy, sensitivity and specificity of ten repetitions of 10-fold cross-validation of the BDC developed for AD detection computed over the OASIS data. The best results found in this computational experiment reported in previous publications for the same MSD features (24 values from each subjectMRI volume) are presented in Table 2. Notice that we are labeling as class 0 the AD patients, while in the referred papers [12,15] they were labeled as class 1. The last rows corresponds to the best results obtained with DC classifiers, while the others correspond to results with other classification algorithms. Specifically, LICA-DC and Kernel-LICA-DC refer to the application of DC after the application of some data preprocessing. In general, DC classifiers have a very low specificity because they target the modeling of the positive class of AD patients. The last row contains the best result found with the proposed BDC which are competitive with the best results found so far. These results have been found in an exhaustive exploration of the effect of the two main parameters, the number of classifiers in the ensemble and the maximum allowed number of dendrites. Figure 1 shows the response surface of the Accuracy. It seems that the number of DC classifiers in the ensemble has little effect, though the best average crossvalidation accuracy (89%) was found with large number of classifiers (64). The maximum number of classifiers appears to have some effect, when number is small - accuracy is small. Allowing more than five dendrites results in average

crossvalidation accuracies above or near 80%. Figure 2 shows the response surface of the ensemble sensitivity. The effect of the maximum number of dendrites is very strong, small number of dendrites prevents the overfitting of the classifiers, resulting in low sensitivity because DC tries to model the positive class. As the number of dendrites grows, the BDC ensemble easily reaches very high (close to 100%) sensitivities. Finally, Figure 3 shows the specificity response surface of the BDC. Specificity remains very low for almost all combinations of number of classifiers and dendrites, and it is the main cause of low classifier performance.

Table 2. Results over the MSD features computed from the OASIS data for AD detection

Classifiers	Accuracy	Sensitivity	Specificity
rbf SVM [12]	81	89	75
LVQ1 [12]	81	90	72
LVQ2 [12]	83	92	74
rbf-DAB-SVM [12]	85	92	78
rbfRVM-LVQ1[15]	87	92	73
LICA - DC [6]	72	88	56
Kernel - LICA - DC [6]	74	96	52.5
Bootstrapped DC	89	100	80

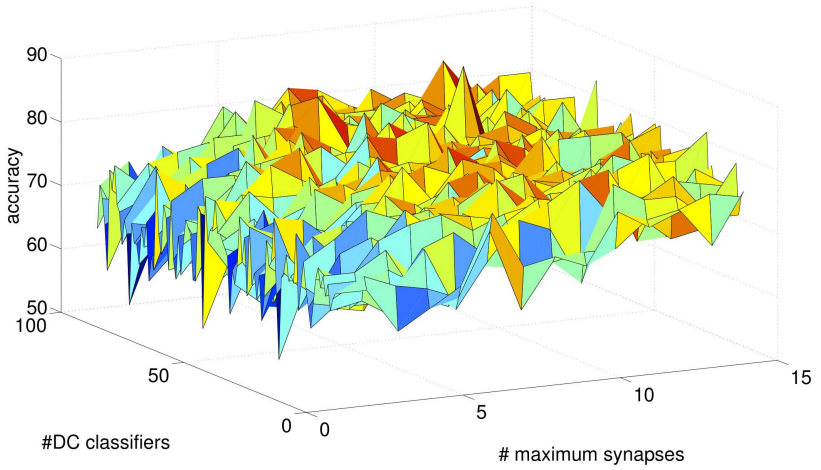


Fig. 1. Average accuracy for varying number of DC classifiers and maximum number of dendritic synapses

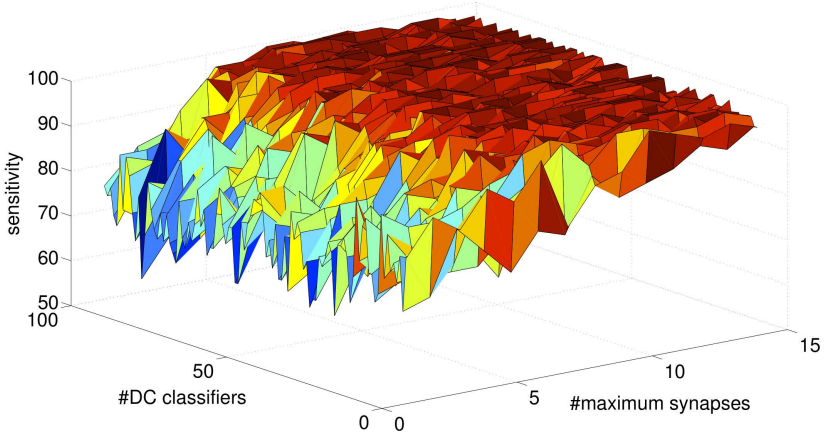


Fig. 2. Average sensitivity for varying number of DC classifiers and maximum number of dendritic synapses

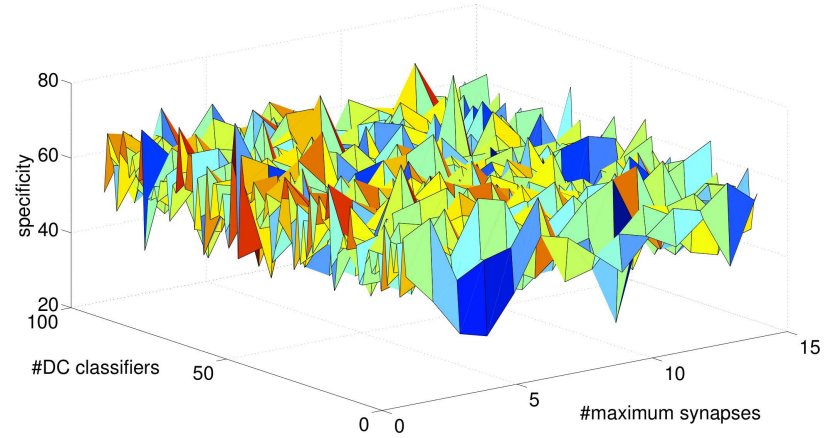


Fig. 3. Average specificity for varying number of DC classifiers and maximum number of dendritic synapses

5 Conclusions

This Chapter proposes BDC. This form of classifier is an ensemble of DC classifiers. Each DC trained on a bootstrapped training dataset. The main parameters of BDC are the number of DC classifiers and the maximum allowed number of dendrites. We have performed an extensive computational experimentation over a dataset of MRI features for AD patient versus healthy control classification. We have found results which are competitive or improve the best found in the literature for this database. The examination of the response surface shows that

the approach seems to be much more sensitive to the number of dendrites than to the number of DC classifiers in the ensemble. Further works will be addressed to more extensive experimentation with diverse conventional datasets.

References

1. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation 9(7), 1545–1588 (1997)
2. Barandiaran, I., Paloc, C., Graña, M.: Real-time optical markerless tracking for augmented reality applications. Journal of Real-Time Image Processing 5(2), 129–138 (2010)
3. Barmpoutis, A., Ritter, G.X.: Orthonormal basis lattice neural networks. In: 2006 IEEE International Conference on Fuzzy Systems, pp. 331–336 (2006)
4. Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)
5. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
6. Chyzhyk, D., Graña, M., Savio, A., Maiora, J.: Hybrid dendritic computing with kernel-lica applied to Alzheimer's disease detection in MRI. Neurocomputing 75(1), 72–77 (2012)
7. Fotenos, A.F., Snyder, A.Z., Girton, L.E., Morris, J.C., Buckner, R.L.: Normative estimates of cross-sectional and longitudinal brain volume decline in aging and AD. Neurology 64(6), 1032–1039 (2005)
8. García-Sebastián, M., Savio, A., Graña, M., Villanúa, J.: On the use of morphometry based features for Alzheimer's disease detection on MRI. In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) IWANN 2009, Part I. LNCS, vol. 5517, pp. 957–964. Springer, Heidelberg (2009)
9. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of Cognitive Neuroscience 19(9), 1498–1507 (2007)
10. Ritter, G.X., Schmalz, M.S.: Learning in lattice neural networks that employ dendritic computing. In: 2006 IEEE International Conference on Fuzzy Systems, pp. 7–13 (2006)
11. Ritter, G.X., Urcid, G.: Lattice algebra approach to single-neuron computation. IEEE Transactions on Neural Networks 14(2), 282–295 (2003)
12. Savio, A., Garcia-Sebastian, M., Chyzhyk, D., Hernandez, C., Graña, M., Sistiaga, A., Lopez de Munain, A., Villanua, J.: Neurocognitive disorder detection based on feature vectors extracted from vbm analysis of structural MRI. Computers in Biology and Medicine 41, 600–610 (2011)
13. Savio, A., García-Sebastián, M., Graña, M., Villanúa, J.: Results of an adaboost approach on Alzheimer's disease detection on MRI. In: Mira, J., Ferrández, J.M., Álvarez, J.R., de la Paz, F., Toledo, F.J. (eds.) IWINAC 2009, Part II. LNCS, vol. 5602, pp. 114–123. Springer, Heidelberg (2009)
14. Savio, A., García-Sebastián, M., Hernández, C., Graña, M., Villanúa, J.: Classification results of artificial neural networks for alzheimer's disease detection. In: Corchado, E., Yin, H. (eds.) IDEAL 2009. LNCS, vol. 5788, pp. 641–648. Springer, Heidelberg (2009)
15. Termenon, M., Graña, M.: A two stage sequential ensemble applied to the classification of Alzheimer's disease based on mri features. Neural Processing Letters 35(1), 1–12 (2012)

An Analytic Aggregation-Based Ontology Alignment Approach with Multiple Matchers

Fuqi Song, Gregory Zacharewicz, and David Chen

Univ. Bordeaux, IMS, UMR 5218,
351 Cours de la Libration 33400 Talence, France
{fuqi.song,gregory.zacharewicz,david.chen}@ims-bordeaux.fr

Abstract. A critical aspect of providing data interoperability relies on ontological alignment and the successful semantic integration. Ontology alignment is being applied to these domains as a fundamental component. Many basic matching techniques have been proposed, however in order to adapt to the diverse sources of ontology and enhance the matching ability, the crucial point facing is how to choose and combine different alignment algorithms. In this chapter, an approach with multiple alignment algorithms is described. The algorithms are proposed from lexical, semantic and structural levels of source ontology. The algorithms are chosen by a comprehensive pre-defined strategy, one or more specific algorithms will be chosen according to the features of entities to be matched. To aggregate the different matching results dynamically and automatically, Analytic Hierarchy Process (AHP) is applied innovatively based on three similarity indicators, which reflect the essential features of source ontology. The results of the benchmarking experiment suggested that the approach has strong matching ability. It obtained high precision and promising evaluation results.

Keywords: Ontology Alignment, Alignment Algorithm, Analytic Aggregation, Multiple Matchers.

1 Introduction

Nowadays information and communication systems are commonly applied everywhere, and huge amount of data is existing in these heterogeneous systems. Information integration is a way to promote the interoperability among distributed heterogeneous systems. Ontology, as a formal, explicit specification of a shared conceptualization [1], is being applied to represent domain knowledge models. In the definition, "formal" indicates that an ontology is machine-readable, which can be processed by computers. "Explicit" refers to that all the concepts and relations in an ontology are defined explicitly and directly. From this perspective, an ontology is described as a 6-uple: {C, P, H^c, H^p, A, I}, including a set of concepts C and a set of properties P. The hierarchy relationship between concepts and sub-concepts are denoted by H^c, in the same way, H^p denotes the hierarchy relations between properties and sub-properties. A is a set of axioms, while I is the set of instances of concepts and properties. The representation language of ontology are diverse [2], in the chapter, the ontology language uses OWL[1]

[1] See http://www.w3.org/TR/owl2-overview/

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 143–159, 2013.

(Ontology Web Language) , which is a standard ontology representation language created by World Wide Web Consortium (W3C).

Ontology alignment is a way to reuse existing ontology to develop new ontology, as well as a way to enable data integration between different data models. For the former case, ontology alignment is a crucial issue in the domain of semantic integration to facilitate data interoperability, which is an essential part of Enterprise Information System (EIS) interoperability [3]. Ontology alignment seeks to find semantic correspondences between a pair of ontology entities by identifying semantic relations. The definition of *correspondence* from Euzenat and Shvaiko [4] is defined as: given two ontology O and O' with associated entity languages O_L and $O_{L'}$, a set of alignment relations Θ and a confidence structure over Ξ, a correspondence is a 5-uple: $\{id, e, e', r, n\}$, where *id* is a unique identifier of the given correspondence, $e \in Q_L(o)$, $e' \in Q_{L'}(o')$, $r \in \Theta$ and $n \in \Xi$. The entities to be matched mainly include: classes, properties and instances.

This research issue has been studied for years, and the matching mechanism is developed from a single matching algorithm to multiple combined matching algorithms. Thus two facing problems are: how to choose the matching algorithms and how to combine them. As stated by Shvaiko and Euzenat [5]: matcher selection, combination and tuning is a challenge facing in the domain of ontology alignment. Ontology alignment is performed from three levels of ontology as shown in Fig. 1.

- At element level, the entity itself is treated as the object of study; the label, comment and internal information of it are investigated. The mostly used techniques are string metric [6], string similarity, domain and data type comparison [7].
- At local level, the entities and the relations linked to the studied entity are taken into account, such as, graph-based method [8] and taxonomic-based methods. Directly linked entities and the studied entity are composed into a local group, this group of entities is taken as the object of study.
- At global level, the whole ontology is taken as a context and environment. The relation and affect between the studied object is investigated. Machine Learning (ML),Artificial Neural Network (ANN) [9] are the methods applied at this level.

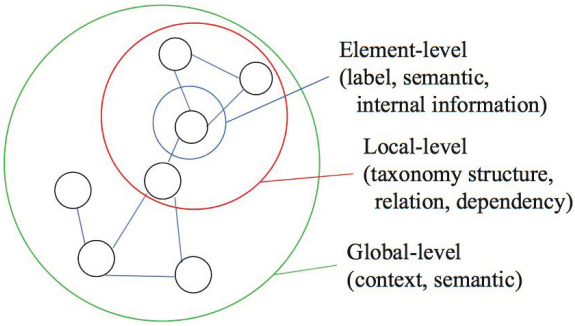


Fig. 1. Three levels of source ontology

In this chapter, three matchers are proposed from the three levels of source ontology to be matched. At element level, a lexical matcher String-Based Matcher (SBM) is given. SBM applies edit distance [10] and N-gram [11]. At local level, a structural matcher Graph-Based Matcher (GBM) by applying similarity flooding [12] is proposed. A Matcher for Semantic Matching (MSM) is designed from both element level and global level. MSM measures the semantic similarity with Lin Model [13] and WordNet [14], also it checks the similarity of homonyms in different context.

In order to combine the different matchers, a comprehensive selection strategy (see Section 3.2) is given according to different conditions. When several matchers are applied, an analytical method AHP is applied and adapted to aggregate different matching results. AHP is a method for organizing and analysing complex decisions with a structured process. It was developed by Saaty [15] based on mathematics and psychology. AHP balances various factors against the goal. First the expected goal, the criteria and the alternative are defined. Then with a strictly defined process, the alternatives are compared to each other against one criteria and a specific intensity of importance (scale) is assigned. Finally the results from each step are combined and the priority of each alternative is computed.

Normally the priority is used to decide an alternative which suits the goal best. In this article, the priority is adapted as the weight of each matchers. The intensity of importance is obtained based on three similarity indicators, which are derived from source ontology and reflect essential features from one certain aspect. Saaty and Vargas [16] analysed the drawbacks of AHP and applying conditions. Vaidya and Kumar [17] reviewed some applications using AHP in different domains. Mochol *et al.* [18] also adopted AHP for ontology matching approaches selection, however it aims to choose matching approaches or techniques from macro level.

The reminder of this chapter is organized as follows. Section 2 investigates the related work about multiple matcher-based approaches in this research area. Section 3 illustrates the structure of the approach and the details of matchers. Section 4 proposes three similarity indicators and uses them to aggregate the results with AHP. Section 5 evaluates the proposed approach and compares to the others using benchmarking test, and provides the major conclusions.

2 Related Work of Multiple Matcher-Combined Approaches

In the domain of ontology alignment, many basic matching techniques have been proposed and developed. The authors, Euzenat and Shvaiko [4], Granitzer *et al.* [19] and Alasoud *et al.* [20] gave comprehensive introduction and comparison of different basic matching techniques and their applications. For a specific matching algorithm, it works more efficiently and obtains more accurate results on some specific ontology. In the same way, the matcher needs matching conditions to achieve better results. There are no perfect matchers for specific ontology, and no ontologies could only be matched by one specific matcher. The key issue is how to combine different matching results to make them complementary. In this research area, several approaches are proposed. Some multiple matchers-based approaches are investigated and listed in Table 1, in order to see how the matchers are designed and how the weight is assigned.

Li *et al.* [21] adapted edit distance, vector distance and similarity flooding based algorithms to match. It is proposed that two similarity factors: label similarity *F_LS* and structure similarity *F_SS*, which are computed from conceptual and structural levels of source ontology. The weights of each matcher are computed based on these two factors. A strategy is applied to choose which matcher to use according to an experimental threshold value.

Pirro and Talia [22] used four matchers: lucene matcher, string matcher, WordNet matcher and structural matcher. Two affinity coefficients are proposed: lexical affinity coefficient L_a and structural affinity coefficient S_a, which also concluded from source ontology. The weight are calculated based on the two affinity coefficients with a heuristic function.

Mao *et al.* [23] proposed a concept "harmony" h to estimate the importance and reliability of similarities and used it as the weight. Tu and Yu [24] first calculated the credibility of each matcher and then used this credibility as weight of each matcher. Huang [9] applied an artificial neural network approach to learn the weight from training data. Akbari and Fathian [25] also proposed a combined approach with lexical and structural matchers. The weights are set manually according to experiments, the weight for lexical matcher is $\alpha = 0.4$ and for structural matcher is $\beta = 0.6$. Xu *et al.* [26] proposed a metric called "differentor" to integrate different similarity measurement results obtained by different matching techniques. The weights are computed based on this metric. The weights are at entity level, which means that each pair of entities matched has a different weight. The matching algorithms proposed to use are: lexical, structural, extensional and relation similarity.

From the investigation of these multiple matchers-combined approaches, there are mainly two kinds of methods to decide the weights:

- First, some factors are defined, and then the weights are computed based on these factors, such as RiMOM and UFOme;
- First, a variable is defined via a specific approach, and then this variable is used as weight, such as PRIOR+, CMC and SFS.

For each approach, it tries to choose several matching techniques, which complement each other. In this article, three matchers from the three levels of source ontology are proposed. Each of them can discover certain mappings from a different perspective. From the three levels, the matchers can discover complete alignments. The weights of each matcher are assigned automatically and dynamically with AHP and three indicators, which are learned from the source ontology. This approach allows to generate the weights, which could balance the important factors according to the specific source ontology. It will assign higher weight for important matchers after evaluating the importance of each matcher with indicator.

Adapted from the representation format of correspondence in Euzenat [27], two types of correspondences for a multiple matcher-based approach are defined: intermediate and final. Intermediate correspondence is discovered by a specific matcher, and then several intermediate correspondences are combined into a final correspondence using pre-defined strategy. Namely, final correspondences are used for aligning ontology, whereas intermediate correspondences are used for generating final correspondences. An intermediate correspondence *ic* is defined as $ic = \{mid, e, e', r, n, M\}$,

Table 1. Investigation of multiple matchers-based ontology alignment approaches

Matching Approach	Factor	Weight aggregation
RiMOM [21] -Edit distance -Vector distance -Similarity flooding	Label similarity factor: $F\_LS = \frac{\sharp ide\_conc\_l + \sharp ide\_prop\_l}{max(\lvert C_1\rvert + \lvert P_1\rvert, \lvert C_2\rvert + \lvert P_2\rvert)}$ Structural similarity factor: $F\_SS = \frac{\sharp nonl\_conc + \sharp nonl\_prop}{max(\lvert \sharp C_1\rvert + \lvert \sharp P_1\rvert, \lvert \sharp C_2\rvert + \lvert \sharp P_2\rvert)}$	$w_{name} = \frac{F\_LS}{max(F\_LS, F\_SS))}$ $w_{vec} = \frac{F\_SS}{\max(F\_SS, F\_SS)}$ $sim = \frac{w_{name}\sigma(sn) + w_{vec}\sigma(sv)}{w_{name} + w_{vec}}$
UFOme [22] -Lucene matcher -String matcher -WordNet matcher -Structural matcher	Lexical affinity coefficient : $\mathrm{L}_a(O_s, O_t) = \frac{\#common\_entities\_l}{\min(\lvert S\rvert, \lvert T\rvert)}$ Structural affinity coefficient : $\mathrm{S}_a(O_s, O_t) = \frac{\#common\_entities\_s}{\min(\lvert S\rvert, \lvert T\rvert)}$	$w_l = \frac{e^{\eta L_a} - e^{-\eta L_a}}{e^{\eta L_a} + e^{-\eta L_a}}$ $w_s = \frac{e^{\psi S_a} - e^{-\psi S_a}}{e^{\psi S_a} + e^{-\psi S_a}}$
PRIOR+ [23] -Name similarity -Edit distance -Profile similarity -Structural similarity	-NA	$harmony : h = \frac{\#s\_\max}{\min(\#e_1, \#e_2)}$ $sim = \frac{\sum_k h_k \times Sim_k(e_{1i}, e_{2j})}{n}$
CMC [24] Credibility prediction Multiple matchers	-NA	Mean Square Error: $MSE = E_F[(sim - sim_{ac})^2]$ Credibility: $c = e^{-C \times MSE}$ $similarity = \sum c \cdot sim / \sum c$
SFS [9] Similarity in concepts names(s1),properties(s2), and relationships(s3) Artificial neural network	-NA	$\sum w_i = 1$, w_i is initialized randomly, adjusted via learning $sim = \sum_{i=1}^{3} (w_i s_i)$

where e and e' are elements from ontology O and O' to be matched. Respectively, n denotes the confidence between e and e' identified by matcher M with a relation r, which could be "*equal*", "*subClassOf*", "*superClassOf*". *mid* is an unique identifier for this correspondence.

A final correspondence *fc* is similar to intermediate correspondence without the information concerning to a specific matcher M, the definition is the same with the 5-uple $fc = \{e, e', fr, fn, id\}$, where *fr* is a relation derived from relations in intermediate correspondences and *fn* denotes a combined confidence from intermediate correspondences' confidences. The semantic similarity-based approaches seek to discover the equal relation between ontologies. Hierarchical relation, such as super class, child class, and the other relations, are beyond the ability of similarity-based approaches. This paper focuses on discovering *equal* relation between ontologies, the other types of relation are not considered.

3 Ontology Matching

This section describes the matching approach in detail. First, an overview of the structure is given to show the general process of this approach. Second, the selection strategy of different matchers is described. Third, the explanation of each matcher is elaborated in detail.

3.1 Overview

Figure 2 illustrates the structure of the approach, including the main components and process. A pair of ontology is as input and alignments are obtained as output. Ontology pre-processing includes elimination and tokenization of entities. *Strategy selector* chooses a pre-defined strategy according to the features of the elements to be matched. The strategy decides which matcher(s) to use for performing matching task. Indicators are used to generate the weights of each matcher. *Aggregator* combines the intermediate correspondences into final correspondence. *Evaluator* evaluates the final correspondence with reference alignments from benchmarking data set. In this section, ontology pre-processing, matcher selection and matchers are described. Aggregation and experiments will be discussed in Section 4 and Section 5 respectively.

Before the matching task, elimination and tokenization of the labels of entities are performed as a pre-process. Most of the labels of entities are compound words rather than a single meaningful word. Elimination helps to eliminate the punctuation information which could confuse the matching task. Tokenization split the compound words into single ones. Semantic matcher will perform matching based on the tokenized words, since the compound words cannot be processed directly by computer and lexical database.

3.2 Matcher Selection

Different strategies are applied for adapting various situations. Figure 3 illustrates the overall strategy selection. A pair of elements from pre-processed ontology is as input,

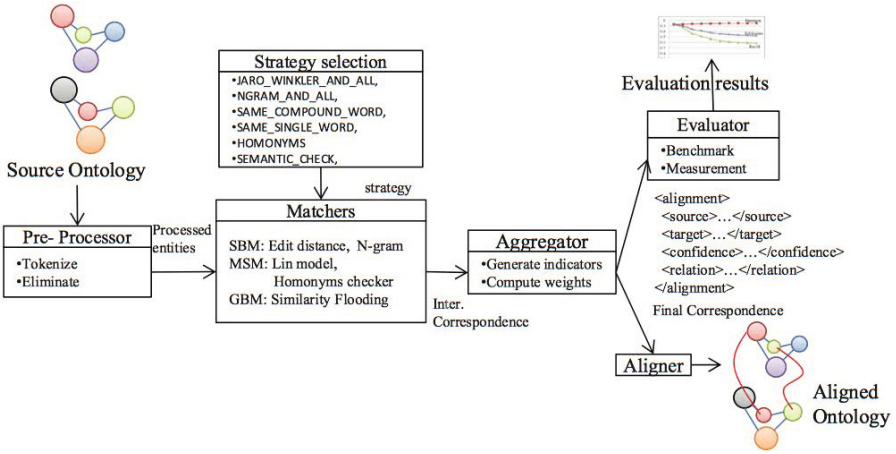


Fig. 2. General architecture of ontology alignment

and then several conditions are examined to decide which strategy to apply, namely, which matcher(s) to use. For the strategy, which uses multiple matchers, the found results will be aggregated. First, it is checked that whether the labels of elements are identical, and then it is checked whether the label is compound word. If they have identical compound label, such as "*name_of_employee*", they are considered to be equivalent as the same word. In this case, none of the other matchers will be used. If they have identical single word, then it needs to check whether they belong to the same context. For instance, "paper", it may make references to sheets used to write on or a dialogue news publication. To check this, a semantic similarity indicator (see Section 4.1) is used. If they belong to the same context, then they are regarded as the *SAME_WORD*, otherwise homonyms checker is applied to calculate similarity. In the other cases, three matchers will be applied.

3.3 Matchers

Three matchers from different levels of source ontology are proposed in this section. The matchers SBM, MSM and GBM are described as follows:

- SBM adopts Jaro-winkler [28] distance and N-gram distance function with a constraint of label length. SBM tries to find the lexical similarity matching from element level.
- MSM uses a taxonomy-based model [13] with WordNet [14] from local level, and uses a specific solution to solve the homonym issue.
- GBM applies similarity flooding [12] to learn the similarity between elements from a global level.

String-Based Matcher (SBM). String-based matchers are designed based on the label of entity, which represents the concepts, properties or annotations. These elements are

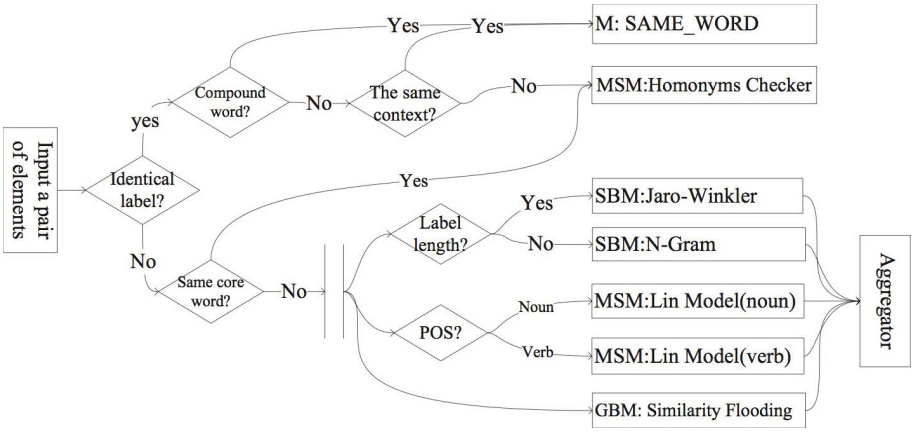


Fig. 3. General flow of matcher selection

treated only as a sequence of letters, without considering the meaning represented and the structure contained. SBM adopts two kinds of algorithms: string metric and n-gram.

Jaro-Winkler Distance. Levenshtein distance (also known as edit distance) is the mostly known distance function, in which distance is the cost of operations, including insertion, deletion and substitution, for converting s_1 to s_2 in a best sequence. Jaro-Winkler [28] distance is a broadly string metric based on Jaro distance [10]. The equations defined in Winkler [28] and Jaro [10] are as Eq. (1) and Eq. (2) respectively, where m is the number of matching character, t is half of the transportation number, and P is the length of longest common prefix of s_1 and s_2.

$$Jaro(s_1,s_2) = \frac{1}{3} \cdot \left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) \tag{1}$$

$$Jaro-Winkler(s_1,s_2) = Jaro(s_1,s_2) + \frac{\min(P,4)}{10} \cdot (1 - Jaro(s_1,s_2)) \tag{2}$$

N-Gram. Jaro-Winkler has limitations when matching between two strings which have big differences in length, such as "*member*" and "*conference_member*". When min($|s_1|$,$|s_2|$)-1 is greater than max($|s_1|$,$|s_2|$)/2, SBM applies n-gram [11] as matching technique, more specifically, tri-gram. Tri-gram (n,s) represents the set of tri-grams, where n is the length of sub-string s. The similarity TG between strings s_1 and s_2 with n-gram is defined in Eq. (3).

$$TG(s_1,s_2) = \frac{|trigram(s_1) \bigcap trigram(s_2)|}{min(s_1,s_2) - 2} \tag{3}$$

Matcher for Semantic Matching (MSM). A taxonomy-based approach proposed by Lin [13], called *Lin Model*, for semantic similarity matching is applied. Assuming that

the taxonomy is a tree, if $s_1 \in C_1$ and $s_2 \in C_2$, the commonality between s_1 and s_2 is $s_1 \in C_0 \wedge s_2 \in C_0$, where C_0 is the most specific class that subsumes both C_1 and C_2. $P(C)$ is the probability that a randomly selected object belongs to C. WordNet [14] is used as the taxonomy. The similarity is defined in Lin [13] as formula Eq. (4).

$$\mathrm{Lin}(s_1, s_2) = \frac{2 \cdot \log P(C_o)}{\log P(C_1) - \log P(C_2)} \tag{4}$$

The same word does not always represent the same meaning. A homonym is a special case in semantic matching. It represents different meanings in different contexts, such as "article" may refer to a paper or refer to a product. First, whether the two ontologies, where the homonyms occurred, belong to the same context are checked according to a semantic indicator I_{sem} defined in Section 4.1. A threshold th is set. If the indicator I_{sem} is greater than the threshold th, then the two ontology are considered as belonging to same context. In this case, the two words are considered as identical and the similarity is assigned to 1.0. Otherwise, a formula (see Eq. (5)) is applied for computing the similarity of a pair of homonyms, where *#m* is the number of different explanations (retrieved from WordNet) that the word has. In this work, the threshold is set manually as $th = 0.2$.

$$H(s) = \frac{\#\mathrm{m} - 1}{\#\ \mathrm{m}} \tag{5}$$

Graph-Based Matcher (GBM). Similarity flooding proposed by Melnik *et al.* [12] is an algorithm for matching two data schemas based on similarity propagation graph and fix point computation. The algorithm takes two graphs as input and produces the mappings between corresponding nodes of graphs. First, one ontology is converted into a similarity propagation graph. A triple node (s, p, o) represents each edge in the graph, where s and o are the source and target nodes, while p refers to the label. Then, the algorithms uses the converted graph to search correspondences. When converting an OWL ontology to a graph, the relation listed in Table 2 are used as edges to construct the graph. In Table 2, c denotes a class, p denotes a property, e denotes an entity and i denotes an instance of class.

Table 2. Relations for constructing graph node

Source node	Edge	Target node	Description
c_i	*l_super_class*	c_j	Class c_i has super class c_j
c_i	*l_sub _class*	c_j	Class c_i has sub class c_j
p_i	*l_sub_property*	p_j	Property p_i has sub property p_j
c_i	*l_object_property*	p_j	Class c_i has object property p_j
c_i	*l_data_property*	p_j	Class c_i has data property p_j
c_i	*l_domain*	e_j	Class c_i has domain e_j
c_i	*l_range*	c_j	Class c_i has range with class c_j
c_i	*l_has_individual*	i_j	Class c_i has instance i_j

A brief algorithm to show the matching process is given in Listing 1. First, two ontology O and O' are converted to two graphs G and G' with the relations defined in

Table 2. Then, an initial mapping between the two graphs is needed in order to start the similarity propagation. In GBM, the top nodes of ontology "*Thing*" are taken directly as the initial mapping nodes. The details of the computation is not elaborated in this paper, it can be followed in Melnik *et al.* [12]. In the implementation of GBM, an API[2] of similarity flooding algorithm is invoked. This Application Programming Interface (API) provides core computation of this algorithm.

```
1 G =ConvertToGraph(O), G' = ConvertToGraph(O'); initialMap =
2 StringMatch(G.thing, G'.thing); mapping = SFA(G, G',
      initialMap);
3 result = SelectThreshold(mapping); correspondence = Wrap (
      result);
```

Listing 1. Algorithm of GBM

4 Dynamic Analytic Aggregation

This section describes how to aggregate the different matching results with AHP. First three similarity indicators are proposed from different aspects of source ontology, which could reflect the essential features of source ontology. With the AHP process, the indicators are used to decide the intensity of importance with specific rules. The final obtained priority values is taken as weights of matchers.

Following the AHP process, in the first step, the goal, criteria and alternatives are defined in Figure 4. More specifically, the goal is to determine a suitable matcher by respecting to three aspects of matching ability. Besides the three criteria, the other criteria of them are assumed to be equal, such as, accuracy and performance. The alternatives are the different matchers defined in Section 3.3.

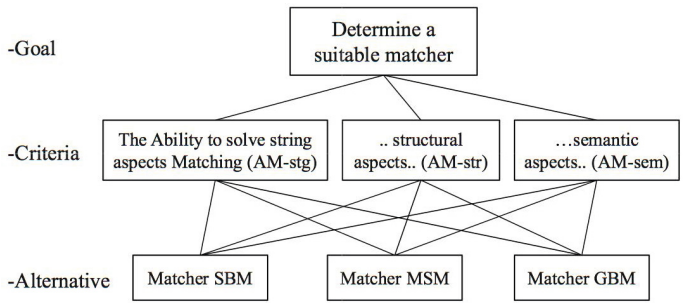


Fig. 4. Description of goal, criteria and alternatives with AHP

[2] See http://infolab.stanford.edu/~melnik/mm/sfa/

4.1 Similarity Indicators

In order to aid assign the scale of importance in AHP automatically, three indicators are proposed. The indicators are generated from a global perspective of source ontology to reflect similarity in one specific aspect. For instance, string similarity indicator I_{stg} is generated based on the identical strings in both ontology. It represents the global similarity in the term of lexicon in ontology. The other two indicators:I_{str} and I_{sem} are generated in term of structure and semantic respectively.

String Similarity Indicator. The indicator I_{stg} reflects the similarity in string between two ontologies. This indicator is adapted from lexical affinity coefficient defined in Pirro and Talia [22]. Before comparison, the stop words in label are removed. The Indicator is defined in Eq. (6), where #*ic* and #*ip* denote the number of classes and properties identified with identical labels respectively, #*tcp* denotes the total number of classes and properties of one ontology.

$$I_{stg} = \frac{\#ic + \#ip}{\min(\#tcp_1, \#tcp_2)} \tag{6}$$

Structural similarity indicator The indicator I_{str} denotes the number of nodes with similar structure by investigating its subclasses (hierarchy) and relations (dependency). First the ontology is treated as a directed graph $G = <V, E>$, V is a set of vertices (or nodes), E is a set of edges with ordered pairs of vertices (v_i, v_j) from V. A vertex in ontology is described as (*#indegree, #outdegree, #subclass*), the indicator is denoted in Eq. (7), where *#common_ds* denotes the number of vertices that have the same *#indegree*, *#outdegree*, and *#subclass*. *#niv* denotes the number of non-isolated vertices.

$$I_{str} = \frac{\#commom\_ds}{\min(\#niv_1, \#niv_2)} \tag{7}$$

Semantic similarity indicator Semantic indicator I_{sem} is computed based on the tokenized single words, not the original compound labels. For a tokenized word in entity of ontology O_1, if there is a synonym existing in the entity of ontology O_2, then *#synonym* count adds 1. It is defined in Eq. (8), where *#synonym* is the number of synonyms identified between O_1 and O_2. This indicator is also taken as an indicator for checking whether two ontologies belong to the same semantic context. It is used in MSM matcher.

$$I_{sem} = \frac{\#synonym}{\min(tcp_1, tcp_2)} \tag{8}$$

4.2 Weight Calculation

If the two ontology have high similarity in one aspect (string, structural or semantic), then the matcher, which is based on this aspect, is more important than the others in discovering correspondences. To assign the intensity of importance automatically, the similarity indicators are used as importance directly with a rule. If the matcher is in the same category with the ability to match, then an additional base ratio *br* will be added on the indicator. Otherwise, the indicator value is used directly. *br* is computed from the

indicators automatically and defined in Eq. (9), where x denotes one of the indicators, I_x denotes one of the three indicators values, and br_x denotes the basic ratio for each.

$$br_x = |\tfrac{1}{3}\sum I_x - I_x| \tag{9}$$

For example, $I_{stg} = 0.3$, $I_{str} = 0.2$, $I_{sem} = 0.7$, then $br_{stg} = 0.1$, $br_{str} = 0.4$ and $br_{sem} = 0.1$. First the criteria: the ability to solve string aspects matching (AM-stg) is evaluated. The matchers compare to each other and assign the scale in Table 3. The result is transferred to a matrix for calculating priority in Table 4.

Table 3. A sample for deriving the priorities respecting to the ability for solving string aspects matching (AM-stg)

Alter.	Scale	Alter.	Scale
SBM	0.4(br+I_{stg})	GBM	0.2 (I_{str})
SBM	0.4 (br+I_{stg})	MSM	0.7 (I_{sem})
GBM	0.2 (I_{str})	MSM	0.7 (I_{sem})

Table 4. Transfer to matrix and compute the priority of each matcher respecting to AM-stg

AM-stg	SBM	MSM	GBM	Prio.
SBM	1	4/7	4/2	0.308
MSM	7/4	1	7/2	0.538
GBM	2/4	2/7	1	0.154

With the same process, the other two criteria (AM-str and AM-sem) are evaluated and two matrices as Table 4 are generated. In AHP, the next process is to compare the importance between criteria. The criteria are set initially to be equally important. This value and the basic ratio would be adjusted according to experiment. A final process that aggregates all these data will be done and get the priority of each matcher. An illustration example is shown in Table 5. The priority of each alternative will be taken as the weight of the matcher. These steps are not detailed in this paper, the AHP process could be followed in the AHP introduction [15] and application [17].

Table 5. Overall priorities of all alternatives

	AM-stg	AM-str	AM-sem	Goal /Weight
SBM	0.1257	0.0482	0.3860	0.5599
MSM	0.0314	0.0185	0.1287	0.1786
GBM	0.1257	0.0071	0.1287	0.2615
Total:	0.2828	0.0738	0.6434	1.0000

4.3 Combination

In the final step, the priority of each matcher is obtained and assigned as weight. Therefore, the final similarity value fv is generated by Eq. (10), where v_x is the intermediate confidence obtained by each matcher, and w_x is the weight generated with the method introduced in this section.

$$fv = \sum v_x * w_x \tag{10}$$

5 Experiments and Conclusions

To test and validate the proposed approach, a software prototype was developed in Java. It uses WordNet as lexical database for checking synonyms and homonyms, postagger [29] for identifying core words. The java APIs used in implementation are JWI[3], JWS[4] and Alignment API[5].

In the section, firstly the benchmarking data set from Ontology Alignment Evaluation Initiative (OAEI) 2011[6] is described. Secondly, the measurements for evaluating the results: precision, recall and F1-measure are described. Thirdly, it compares with other approaches, who had participated in OAEI 2011 campaign. The approach described in this paper did not participate in OAEI campaign, however, the API and data used are the same in order to assure the results are compared in the same basis. Finally, some major conclusions are drawn.

5.1 Test Cases

OAEI is an initiative contributing to assess the ontology alignment systems and approaches. Since 2004, it has organized 11 campaigns using systematic testing data set. Data set **biblio** has been used since 2004 and the seed ontology concerns bibliographic references, which contains 33 named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals. The data sets are generated based on the seed ontology. The tests are systematically generated to from seed reference ontology by discarding a number of information in order to evaluate how the algorithm behaves when this information is lacking. The information includes name, comment, specialization hierarchy, instance, property and classes.

Data set are grouped into five test cases DS1 to DS5 in Table 6. Test case DS1 contains three ontology with small changes in labels and structure. Test case DS2 contains 10 ontology with same structure and different lexical labels. Test case DS3 has many variations in structure. Data set #248 to #266 (DS4) has variations in both aspects; especially the labels are randomly generated strings. Data set #301 to #304 (DS5) are four real-life ontology created by different institutions.

Table 6. Benchmarking data set "biblio"

Test case	Data set	No. of ontologies	Description
DS1	#101 - #104	3	Simple ontology
DS2	#201 - #210	10	Variations in lexical aspect
DS3	#221 - #247	18	Variations in structural aspect
DS4	#248 - #266	15	Both aspects
DS5	#301 - #304	4	Real ontology

[3] See `http://projects.csail.mit.edu/jwi/`
[4] See `http://www.sussex.ac.uk/Users/drh21/`
[5] See `http://alignapi.gforge.inria.fr/`
[6] See `http://oaei.ontologymatching.org/2011/benchmarks/`

5.2 Evaluation and Analysis

Three measurements [30] are used to evaluate: harmonic precision (*HP*), harmonic recall (*HR*) and harmonic F1-measure (*HF1*). Precision measures the ratio of correctly found correspondences over the total number of returned correspondences, and recall measures the ratio of correctly found correspondences over the total number of expected correspondences. In logical term, precision and recall are supposed to measure the correctness and completeness of method respectively. F1-measure combines and balances between precision and recall. The set of alignments identified by the approach described in this paper is denoted as A_d, and the set of reference alignments is denoted as A_r. Reference alignments are provided by the OAEI. The computer for running the test is a Dell laptop, model E5510, OS Windows 7, 2G RAM, P4500 1.87GHz. The basic measurements precision (P), recall (R) and F1-measure ($F1$) are denoted in Eqs. (11).

$$P = \frac{|A_d \cap A_r|}{|A_d|},\ R = \frac{|A_d \cap A_r|}{|A_r|},\ and\ F1 = \frac{2 \times P \times R}{P+R} \tag{11}$$

Harmonic mean measurements are denoted as Eqs. (12), where A_{di} and A_{ri} are the i^{th} set of alignment discovered by the approach described in this paper and reference alignment respectively.

$$HP = \frac{\sum_{i-1}^{m} |A_{di} \cap A_{ri}|}{|A_{di}|},\ HR = \frac{\sum_{i-1}^{m} |A_{di} \cap A_{ri}|}{|A_{ri}|},\ and\ HF1 = \frac{2 \times HP \times HR}{HP+HR} \tag{12}$$

In order to measure the results of combined method with AHP, first each single matcher is evaluated. Only one matcher and some strategies are used to run the test cases. As shown in Table 7, the first column is the number of test cases, and then the following three columns represent the results of each single matcher:SBM, MSM, and GBM. The fifth column is the simple average value of the results obtained by three matchers without applying AHP. The last column represents the results generated by combining all the matchers with AHP. The running time of all five test cases is listed in the last row of Table 7. The running time of each matcher is 2 m 33 s, 8 m 12 s, and 9 m 24 s respectively. The combined method uses 15 m 12 s.

Table 7. Comparison of results of single matcher and combined matchers

#	SBM			MSM			GBM			Simple average			with AHP		
	HP	HR	HF1	HP	HR	HF1	HP	HR	HF1	HP	HR	HF1	HP	HR	HF1
DS1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
DS2	.69	.45	.54	1.0	.38	.55	1.0	.59	.74	.89	.48	.62	1.0	.69	.82
DS3	.96	1.0	.98	1.0	1.0	1.0	.99	1.0	1.0	.98	1.0	.99	.99	1.0	1.0
DS4	.25	.02	.03	.64	.01	.02	.96	.19	.31	.62	.07	.13	.96	.29	.44
DS5	.60	.44	.51	.98	.36	.53	.72	.35	.47	.77	.39	.51	1.0	.58	.74
Avg.	.70	.58	.64	.92	.55	.69	.93	.63	.75	.85	.59	.69	.99	.71	.83
Time	2m33s			8m12s			9m24s			6m42s			15m12s		

Table 7 suggests that SBM has better matching results than MSM, and MSM has better matching results than SBM. The results on DS4 are lowest on all matchers. On

the contrast, DS1 and DS3 obtain quite high matching results, the F1 measure is approximately 1. The results of combined matchers are better than each single matcher on all test cases. Its precision remains very high and the average value is 0.99. The recalls of DS4 and DS5 are relatively low, consequently, the F1 measure on the two data sets are relatively low. The average F1-measure of all test cases is 0.83. Compared with the results of each single matcher, the value is higher of 0.19, 0.14 and 0.08 respectively.

5.3 Conclusions

It applies harmonic HF1-measure to compare with the other approaches. The data of the other approaches is from OAEI campaign 2011 [31]. There were 15 participants, and two of which, AgrMaker [32] and MapSSS [33], had no results in biblio benchmarking. The comparison is displayed in Figure 5. The last column (msOA) is the result obtained by the approach described in this paper. It ranges second (out of 13), the F1-measure is 0.83 and is lower than the first (YAM++ at 0.86) [34] by 0.03 points and equal with (CSA at 0.83) [35].

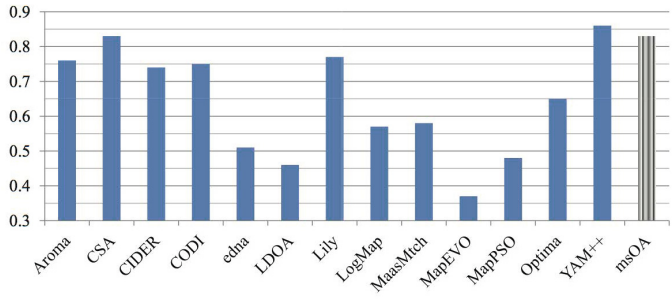


Fig. 5. Comparison of different approaches with harmonic F1-measure (HF1)

It is believed that due to the four following reasons that this approach obtained promising results from the experiments.

- The matchers are chosen by focusing on different levels of source ontology, which enables comprehensive matching and avoids ignoring important parts;
- A reasonable matcher selection strategy is applied according to the features of source ontology, which guarantees that the best suitable matchers are selected;
- AHP is applied to the aggregation of different matchers. It assigns the weight of each matcher automatically and dynamically by balancing various factors and evaluating the importance of each matcher to the specific source ontology;
- Three proposed similarity indicators reflect the essential features of source ontology. They aid to decide the matchers' selection and weight aggregation.

The paper presents an analytic method with AHP to learn the weights for multiple matcher-based ontology alignment. This method could be adopted to solve similar issues in the other areas. Concerning to the ontology alignment issues, future work will be done by adjusting matchers and the basic ratio to adapt more diverse situations.

References

1. Studer, R., Benjamins, V.R., Fensel, D.: Knowledge engineering: Principles and methods. Data and Knowledge Engineering 25(1-2), 161–197 (1998)
2. Song, F., Zacharewicz, G., Chen, D.: An ontology-driven framework towards building enterprise semantic information layer. Advanced Engineering Informatics 27(1), 38–50 (2013)
3. Song, F., Zacharewicz, G., Chen, D.: An Architecture for Interoperability of Enterprise Information Systems Based on SOA and Semantic Web Technologies. In: Proceedings of 13th International Conference on Enterprise Information Systems, Beijing, pp. 431–437. SciTePress (2011)
4. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
5. Shvaiko, P., Euzenat, J.: Ontology Matching: State of the Art and Future Challenges. IEEE Transactions on Knowledge and Data Engineering 25(1), 158–176 (2013)
6. Stoilos, G., Stamou, G., Kollias, S.: A String Metric for Ontology Alignment. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 624–637. Springer, Heidelberg (2005)
7. Ehrig, M., Staab, S.: QOM – Quick Ontology Mapping. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 683–697. Springer, Heidelberg (2004)
8. Hu, W., Jian, N., Qu, Y., Qu, Y., Wang, Y.: GMO: A Graph Matching for Ontologies. In: Proceedings of K-CAP Workshop on Integrating Ontologies, pp. 43–50 (2005)
9. Huang, J., Dang, J., Vidal, J.M., Vidal, J.M., Huhns, M.N.: Ontology Matching Using an Artificial Neural Network to Learn Weights. In: Proceedings of 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 80–85 (2007)
10. Jaro, M.A.: Probabilistic linkage of large public health data files. Statistics in Medicine 14(5-7), 491–498 (1995)
11. Brown, P.F., de Souza, P.V., Mercer, R.L., Pietra, V.D., Lai, J.C.: Class-based n-gram models of natural language. Computational Linguistics 18(4), 467–479 (1992)
12. Melnik, S., Garcia-Molina, H., Rahm, E.: Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In: Proceedings of the 18th International Conference on Data Engineering, pp. 117–128. IEEE Computer Society, Washington, DC (2002)
13. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proceedings of 5th International Conference on Machine Learning, Wisconsin, USA, pp. 296–304. Morgan Kaufmann (1998)
14. Pedersen, T., Patwardhan, S., Michelizzi, J.: WordNet Similarity: measuring the relatedness of concepts. In: Proceedings of NAACL-Human Language Technology Conference, Boston, Massachusetts, pp. 38–41. Association for Computational Linguistics (2004)
15. Saaty, T.L.: How to make a decision: The analytic hierarchy process. European Journal of Operational Research 48(1), 9–26 (1990)
16. Saaty, T.L., Vargas, L.G.: The Seven Pillars of the Analytic Hierarchy Process. In: Models, Methods, Concepts and Applications of the Analytic Hierarchy Process, 2nd edn. International Series in Operations Research and Management Science, vol. 175, pp. 23–40. Springer (2012)
17. Vaidya, O.S., Kumar, S.: Analytic hierarchy process: An overview of applications. European Journal of Operational Research 169(1), 1–29 (2006)
18. Mochol, M., Jentzsch, A., Euzenat, J.: Applying an Analytic Method for Matching Approach Selection. In: Proceedings of 1st International Workshop on Ontology Matching (2006)
19. Granitzer, M., Sabol, V., Onn, K.W., Lukose, D., Tochtermann, K.: Ontology Alignment–A Survey with Focus on Visually Supported Semi-Automatic Techniques. Future Internet 2(3), 238–258 (2010)

20. Alasoud, A., Haarslev, V., Shiri, N.: An empirical comparison of ontology matching techniques. Journal of Information Science 35(4), 379–397 (2009)
21. Li, J., Tang, J., Li, Y., Luo, Q.: RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. IEEE Transactions on Knowledge and Data Engineering 21(8), 1218–1232 (2009)
22. Pirro, G., Talia, D.: UFOme: An ontology mapping system with strategy prediction capabilities. Data and Knowledge Engineering 69(5), 444–471 (2010)
23. Mao, M., Peng, Y., Spring, M.: An adaptive ontology mapping approach with neural network based constraint satisfaction. Web Semantics: Science, Services and Agents on the World Wide Web 8(1), 14–25 (2010)
24. Tu, K., Yu, Y.: CMC: Combining multiple schema-matching strategies based on credibility prediction. In: Zhou, L.-Z., Ooi, B.-C., Meng, X. (eds.) DASFAA 2005. LNCS, vol. 3453, pp. 888–893. Springer, Heidelberg (2005)
25. Akbari, I., Fathian, M.: A novel algorithm for ontology matching. Journal of Information Science 36(3), 324–334 (2010)
26. Xu, P., Wang, Y., Liu, B.: A differentor based adaptive ontology matching approach. Journal of Information Science 38(5), 459–475 (2012)
27. Euzenat, J.: An API for Ontology Alignment. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 698–712. Springer, Heidelberg (2004)
28. Winkler, W.E.: The state of record linkage and current research problems. In: Statistical Research Division U.S. Bureau of the Census (1999)
29. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of NAACL-Human Language Technology Conference, Edmonton, Canada, pp. 173–180. Association for Computational Linguistics (2003)
30. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: Proceedings of 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, pp. 348–353. Morgan Kaufmann (2007)
31. Euzenat, J., Ferrara, A., van Hage, W.R., Hollink, L., Meilicke, C., Nikolov, A., Ritze, D., Scharffe, F., Shvaiko, P., Stuckenschmidt, H., Svab-Zamazal, O., dos Santos, C.T.: Results of the ontology alignment evaluation initiative 2011. In: The 6th International Workshop on Ontology Matching, Bonn, Germany (2011)
32. Cruz, I.F., Stroe, C., Caimi, F., Fabiani, A., Pesquita, C., Couto, F.M., Palmonari, M.: Using AgreementMaker to Align Ontologies for OAEI 2011. In: The 6th International Workshop on Ontology Matching, Bonn, Germany. CEUR Workshop Proceedings, pp. 114–121 (2011)
33. Cheatham, M.: MapSSS results for OAEI 2011. In: The 6th International Workshop on Ontology Matching, Bonn, Germany. CEUR Workshop Proceedings, pp. 184–190 (2011)
34. Ngo, D.H., Bellahsene, Z., Coletta, R.: YAM++ – Results for OAEI 2011. In: The 6th International Workshop on Ontology Matching, Bonn, Germany. CEUR Workshop Proceedings, pp. 228–235 (2011)
35. Tran, Q.-V., Ichise, R., Ho, B.-Q.: Cluster-based similarity aggregation for ontology matching. In: The 6th International Workshop on Ontology Matching, Bonn, Germany. CEUR Workshop Proceedings, pp. 142–147 (2011)

Part III

AI Technquiues

Intelligent Texture Reconstruction of Missing Data in Video Sequences Using Neural Networks

Margarita Favorskaya, Mikhail Damov, and Alexander Zotin

Siberian State Aerospace University, Krasnoyarsk, 660014, Russian Federation
{favorskaya,damov,zotin}@sibsau.ru

Abstract. The paper provides an intelligent method of texture reconstruction after removal of non-disabled objects or artifacts in video sequences. Data under subtitles, logotypes, damages of information medium or small size objects are referred to as missing data. A novel implementation of separated neural network was used to receive spatial texture estimations in missing data region. Usually several types of textures are located under removed object. A fast wave algorithm was developed for boundary interpolation between different types of texture into a missing data region. Three strategies of wave algorithm for contour optimization were suggested. A fully connected one-level neural network was applied for choice of texture inpainting method (blurring, texture tile, and texture synthesis). The proposed technique was tested for visual reconstruction of missing text regions (subtitles, logotypes) and missing objects with area less 8-12% of frame in animation and movies. In the first case, a simplified decision without stage of boundaries approximation may be applied; in the second case, the reconstruction results are significantly determined by a background complexity and motion in scene.

Keywords: Texture Reconstruction, Neural Networks, Video Sequences, Wave Algorithm.

1 Introduction

Texture reconstruction of missing data is concerned to video editor functions, and does not suppose a real-time application. In this task, the main criteria are the accuracy of reconstruction and acceptable visual effect using minimal reconstruction time and computer cost. In intelligent inpainting systems, the background complexity, motions in each scene and the frequency of scene changes will directly influence the result. At present, all existing methods and algorithms try to improve the quantity of all components during the reconstruction process. The process includes three sub-tasks: texture estimations of the surrounding region (with several types of textures), boundary interpolation into a missing data region, and a texture reconstruction of the obtained sub-regions. The novelty of this approach connects with spatio-temporal performances of missing and non-missing data from video sequences that permit the use of original processing algorithms.

A large spectrum of mathematical models for color-texture processing for single images can be found in literature. Originally heuristic algorithms provided the methodology for color-texture segmentation. Enhanced statistical and structural methods are now

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 163–176, 2013.

being being developed (including fractal theory) with a wish to extract more suitable features for texture recognition, synthesis, and reconstruction. The earliest representative papers include the works of Haralick [1] who suggested the universal method of texture analysis based on 28 statistical features and introduced the term "texton" and Funakubo [2] who proposed a region-based color-texture segmentation method for biomedical tissue segmentation. Song-Sheng and Jernigan [3] also advanced an algorithm that integrates the spatial and spectral information for the segmentation of natural scenes.

The reconstruction of multiple missing places of texture data is still being investigated. Ojeda et al. [4] suggested a segmentation algorithm that highlights the edges on an image when the original scene is divided into small regions and a 2D autoregressive model is fitted to each of these segments. Tsai et al. [5] proposed a fast high resolution image reconstruction algorithm for continuous wave in diffuse optical tomography. Iterative algorithms that reconstruct images from far-field x-ray diffraction data were developed by Quiney et al. [6]. Also the techniques of image inpainting use a wavelet-based inter- and intra-scale dependency (Cho and Bui [7], Padmavathi et al. [8]), a hierarchical level set and texture mapping (Du et al. [9]), a discrete regularization on graphs (Ghoniem et al. [10]), textured patterns via wavelet analysis (Lefebvre et al. [11]), a framework approach for non-local image inpainting (Arias et al. [12], Bugeau et al. [13]), and some others approaches. This paper details the research conducted to specify the issue of multiple texture reconstruction within video sequences based on neural network approach.

2 Related Work

At present, the great variability of spatial methods for texture features extraction is known (Hedjam and Mignotte [14], Krinidis and Pitas [15], Nammalwar et al. [16], Ilea and Whelan [17]). Pietikainen [18] represented the texture as locally sampled object by the joint distribution of the Local Binary Pattern (LBP) and contrast features. Some authors [19] introduced an algorithm that combines the color in HSV-space (Hue, Saturation, Value) and the Local Edge Pattern (LEP) using the contribution as 0.6 for color and 0.4 for texture. Garcia Ugarriza et al. [20] combined the color and texture features using region growing and a multi-resolution merging for the segmentation of natural images by an automatic Gradient SEGmentation (referred to as GSEG). Sparse representation of color image is one of ways of restoration (Fadili et al. [21], Xu and Jian [22]). Methods of spatial texture analysis demonstrate approximate results but their application is not enough for texture reconstruction in video sequences.

The temporal texture analysis is based on global motion estimations in scene. One of famous technique is a feature points' surveillance by the method of optical flow based on Sobel detector, gradient detector, SIFT-detector (Shift Invariant Feature Transformation) [23], PCA-SIFT method (Principal Component Analysis – Shift Invariant Feature Transformation). Also other models of global motion estimations are known [24] such as 2-parametric model (the simplest model which considers only vertical and horizontal shifts), 4-parametric motion model (describes vertical and horizontal shifts, rotations and scales), 6-parametric model (full affine transformation model), and 8-parametric

model (the perspective transformations, the most accurate model with a high computation cost). All methods of global motion estimations permit to replace texture fragments from adjacent frames on current reconstructed frame.

In common cases, the task of defining texton is yet to be solved. A set of filters is used for reconstruction of missing region by texture tile. Filters include spots, oriented bands in a variety of directions. They may also consider phases and scales. Algorithms based on data context, texture synthesis by Markov random fields (Vacha et al. [25]), two-directional texture functions, Gabor filters (Khan et al. [26]), texture synthesis by samples in multivariate space are concerned to alternative approaches. Al-Takrouri et al. [27] used the static image modeling and model validation to recover corrupted frames in video sequence. Each texture is characterized by such parameters as smoothness, structural performance, and isotropy. One of appropriative approach is the usage of neural networks (Favorskaya and Petukhov [28]). Multi-level neural networks of direct propagation provide three types of invariance: structural invariance, invariance to test sampling, and the usage of invariant features. The last approach is more suitable for texture analysis because methods of invariant features extraction are well developed.

A common way of background reconstruction under small sizes objects connects with exemplar-based inpainting (Anupam et al. [29], Guo and An [30], Shalini and Menaka [31], Zhang and Lin [32]), marching and block based sampling (Huan et al. [33]), space-time adaptation for patch-based image sequence (Boulanger et al. [34]). The reconstruction on super resolution level is also one of interesting way (Protter et al. [35]).

Table 1 lists a number of existing methods used to reconstruct missing data within a video sequence.

Table 1. Existing methods of missing data reconstruction

	Methods	Short description
Spatial methods:	1) Texture analysis method 2) Method of pixels compare	Information from adjacent regions in frame near a missing data region is used
Temporal methods:	1) Method of optical flow 2) Method of dynamic textures	Information about local and global motion from neighborhoods frames is used
Hybrid methods:	Any combinations of spatial and temporal methods	Information from adjacent regions in frame and neighborhoods frames is used

The outline of this paper is as follows. In Section 3, the spatial features of texture are extracted by using three separate neural networks for definition of smoothness, structural performance, and isotropy. Also it is needed to analyze temporal features in scene. Section 4 provides the interpolation algorithm of open-ended contours in a missing region by a fast algorithm of wave propagation. A possible technique of texture

reconstruction is presented in Section 5. Experimental results with animation and movies are shown and explained in Section 6, followed by a brief conclusion in Section 7.

3 Extraction of Spatio-temporal Texture Features

Extraction of spatio-temporal texture features in adjacent regions relative to a segment with a missing data is the stage which determines a success/failure of all following process. These include the extraction of spatial and temporal texture features. The calculation of spatial texture features and the definition a texture type will discussed in Section 3.1. The detection of temporal texture features will be considered in Section 3.2.

3.1 Spatial Texture Features

For spatial analysis the known Haralick [1] and then the modified version of texture descriptors were used to estimate average m, dispersion σ, homogeneity U, smoothness R, entropy e. Eqs. 1–2 were applied to calculate the modified relative smoothness R_m, and normalized entropy e_n, where L is a number of brightness levels, $L > 1$.

$$R_m = \begin{cases} -\log R\ , & \text{if } R > 0 \\ 10\ , & \text{if } R = 0 \end{cases} \tag{1}$$

$$e_n = e \big/ \log_2 L \tag{2}$$

If parameter $R = 0$ then a relative smoothness is forcibly maintained into small empirical value differing from 0, $R_m = 10$. Normalized entropy e_n indicates some equalization effect in dark and bright areas of frame.

The neural network approach was used to define the texture type (under assumption that a texture in neighborhood is a texture of a single type). The novelty of this approach consists in choice of three texture parameters: smoothness *SM*, structural performance *ST*, and isotropy *IS*. Three fully connected two-level Neural Network (NN) for each parameter were designed because these parameters do not influence on each other but a common NN would be large, complex and tangled. Eq. 3 is used to estimate the smoothness *SM* by NN with parameters:

1. f_{ac} is an activate function (sigmoid),
2. k and l are the number of neurons in zero and first hidden layers; i and j are current indexes,
3. w_{U0j}, w_{Rm0j}, w_{en0j} are weights of synapses from input parameters U, R_m and e_n to neurons j of zero hidden layer correspondingly,
4. $w_{0j,1i}$ is weight of synapse connecting neurons j of zero hidden layer to neurons i of first hidden lay,
5. w_{1iY} is a weight of synapse connecting neurons i of first hidden layer to the output *Y*.

$$SM = f_{ac}\left(\sum_{i=0}^{l-1} f_{ac}\left(\sum_{j=0}^{k-1} f_{ac}\left(Uw_{U0_j} - \log R_m w_{(R_m)0_j} + \frac{e_n}{\log_2 L} w_{e_n 0_j}\right) w_{0_j,1_i}\right) w_{1_i Y}\right) \tag{3}$$

Structural performance ST is estimated by Eq. 4, where m is an order of central moment; μ_m is a central m-order moment; $w_{\mu m 0 j}$ is a weight of synapse from input parameters μ_m to neuron j of zero hidden layer.

$$ST = f_{ac}\left(\sum_{i=0}^{l-1} f_{ac}\left(\sum_{j=0}^{k-1} f_{ac}\left(\sum_{m=3}^{6} \frac{\mu_m}{(L-1)^m} w_{\mu_m 0_j}\right) w_{0_j,1_i}\right) w_{1_i Y}\right) \tag{4}$$

Eq. 5 is applied for estimation of isotropy IS, where M is a maximum of a probability; μ_2 is a two order moment of elements difference; w_{M0j} and $w_{\mu 20j}$ are weights of synapses from input parameters M and μ_2 to neurons j of zero hidden layer correspondingly.

$$IS = f_{ac}\left(\sum_{i=0}^{l-1} f_{ac}\left(\sum_{j=0}^{k-1} f_{ac}\left(Uw_{U0_j} - M\, w_{M0_j} + e_n w_{e_n 0_j} + \mu_2 w_{\mu_2 0_j}\right) w_{0_j,1_i}\right) w_{1_i Y}\right) \tag{5}$$

Descriptors calculated by Eqs. 1 – 5 for each texture region are stored for a decision: which method will be applied for reconstruction of a missing data.

3.2 Temporal Texture Features

The temporal analysis is based on motion information receiving from previous sequential frames before the reconstruction frame. Temporal texture features characterize motion properties of moving texture regions (translation, rotation, and scaling). Any scene may be classified a scene with simple type of motion (affine motion model) and a scene with complex type of motion (objects' overlapping, periodic motions, scaling, large speed motion, etc.). Available set of sequential frames permits to use affine motion model.

The set of feature points, calculated local and global vectors of motion within a scene are the main temporal features available for texture reconstruction. A feature point is such point **p** the surroundings of which are essentially differed from surroundings of other point relatively another point **q**. Usually Harris corner detector is used for feature point detection. Harris detector [36] is a response function estimating a similarity of point surroundings on a corner for each pixel in frame. For this purpose, a gradient matrix **M** is calculated by Eq. 6, where $I(x, y)$ is an image brightness function in point with coordinates (x, y); $\partial I/\partial x$ and $\partial I/\partial y$ are partial derivatives on variables x and y.

$$\mathbf{M} = \begin{bmatrix} \left(\frac{\partial I}{\partial x}\right)^2 & \left(\frac{\partial I}{\partial x}\right)\left(\frac{\partial I}{\partial y}\right) \\ \left(\frac{\partial I}{\partial x}\right)\left(\frac{\partial I}{\partial y}\right) & \left(\frac{\partial I}{\partial y}\right)^2 \end{bmatrix} \tag{6}$$

If both two eigenvalues of matrix **M** are large then even small displacement of point with coordinates (x, y) occurs the essential brightness change; that corresponds to feature point. Eq. 7 is used to estimate the corner response function *CR* where k=0.04 is a coefficient suggested by Harris; trace(**M**) is a spur of matrix **M**.

$$CR = \det \mathbf{M} - k\,(\text{trace}\,(\mathbf{M}))^2 \tag{7}$$

Points corresponding to local maximums of corner response function *CR* are considered as feature points. The connecting task is a tracking of detected feature points in sequential frames. The task of such tracking is based on optical flow suggested by Lucas and Kanade [37]. Motion parameters and distortions of feature surroundings are reduced to definition of motion parameters and distortions of feature surroundings when the difference c is minimized. Eq. 8 calculates these differences, where W is a point surroundings; $w(x, y)$ is a weight function (its values may be equal to 1); $J(x, y)$ and $I(x, y)$ are brightness functions in current and following frames; **d** is a vector of point displacement; **A** is a matrix of affine parameters.

$$C = \int\int_W \left(J\left(|\mathbf{A}\,\mathbf{p}(x,y) + \mathbf{d}|\right) - I\,(x,y)\right)^2 w\,(x,y)\;dx\,dy \tag{8}$$

The Eq. 8 is relatively differentiated of motion parameters and is equaled to 0. Then a system of equations is linearized by expansion in CityplaceTaylor series and solved interactively by Lucas and Kanade method [37]. Such approach is based on 2-, 4-, 6-, and 8-parametric motion models that permit to build a field of local motion vectors in sequential frames. Usually a field of local motion vectors is noised. Some known methods normalizes this field but this step is not essential for reconstruction task. The following stage is a choice of global model of motion in a scene based on local vectors of motion.

Three motion approximation models were introduced according to following classification: linear, rotational, and scalable approximation models. The linear model is a singular affine transformation (transition) with motion vector $\mathbf{v} = (x_i – x_{i-1},\ y_i – y_{i-1})$ where (x_i, y_i), (x_{i-1}, y_{i-1}) are point coordinates in two sequential frames. Three sequential frames with point coordinates (x_i, y_i), (x_{i-1}, y_{i-1}), (x_{i-2}, y_{i-2}) are used in rotational model. It is necessary to determine a rotation angle α, center coordinates (x_c, y_c), and a rotating radius R. Let's solve following combined Eq. 9:

$$\begin{aligned} (x_i - x_c)^2 + (y_i - y_c)^2 &= R^2 \\ (x_{i-1} - x_c)^2 + (y_{i-1} - y_c)^2 &= R^2 \\ (x_{i-2} - x_c)^2 + (y_{i-2} - y_c)^2 &= R^2 \end{aligned} \tag{9}$$

and from a scalar vector products a rotation angle α between two sequential frames is found by Eq. 10

$$\alpha = \arccos\left(\frac{(x_i - x_c)\;(x_{i-1} - x_c) + (y_i - y_c)\;(y_{i-1} - y_c)}{\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}\sqrt{(x_{i-1} - x_c)^2 + (y_{i-1} - y_c)^2}}\right) \tag{10}$$

Eq. 11 estimates a scale coefficient k by displacement vectors lengths into sequential frames series in scalable model, where $(x_{i,j}, y_{i,j})$, $(x_{i-1,j}, y_{i-1,j})$, $(x_{i-1,j-1}, y_{i-1,j-1})$,

$(x_{i,j-1}, y_{i,j-1})$ are points coordinates of a displacement vector; i and j are sequential couple of frames.

$$k = \frac{\sqrt{\left(x_{i-1,j-1} - x_{i,j-1}\right)^2 + \left(y_{i-1,j-1} - y_{i,j-1}\right)^2}}{\sqrt{\left(x_{i-1,j} - x_{i,j}\right)^2 + \left(y_{i-1,j} - y_{i,j}\right)^2}} \tag{11}$$

Counters accumulate the number of motion vectors according to each motion model by a voting method. Type of motion is chosen by acounter which stores a maximum value. A linear approximation model implies that angle of motion vectors near a region with missing data tends to zero. An angle of motion vectors near a region with missing data will be more zero for a rotation approximation model. Intersections of motion vectors in the center of coordinate system indicate about a scale approximation model.

4 Boundary Interpolation into the Missing Data Region

One application of the wave algorithm is the formation of vector descriptions in a gradient image. Let's suppose that a set of pre-processed gradient frames is available. In initial point, a wave is generated and propagated according to the determined rules (the passed points are labeled by a number of steps) up to a target point. The number of steps will be a normalized distance between initial and target points. There are two constraints from pixel structure of image:

1. Space resolution.
2. Direction resolution (90°for 4-coupled wave and 45° for 8-coupled wave).

For 4-coupled case, wave propagation is going as a rhomb; for 8-coupled case, wave propagation is going as a square. The generation of spherical wave occurs as combination of 4-couples and 8-coupled waves and is going as an octagon. On each step of wave algorithm which tracks a central (contour) line, the displacement of mass center is analyzed with or without considering its previous locations. The proposed wave algorithm includes two stages:

1. Contour building (skeleton) on current wave step by spherical wave propagation.
2. Contour optimization by updated vector directions which were received on previous wave steps.

On each wave step, a complex vector description containing a set of small fragments is received that may decrease reconstruction accuracy. Vector lengths are smoothed by analysis a set of ribs between nodes. A rib is removed if its deviation from interpolation line exceeds the determined threshold; so the ribs with high deviation will not consider in the result vector description. A value of determined threshold is chosen according to a line width. Three strategies of contour optimization were realized:

1. The simplest case when the direction of last vector become a predictable direction.
2. A predictable direction is determined as approximation of two vector directions from two last steps of wave algorithm.

3. A predictable direction is determined as approximation of three vector directions from three last steps of wave algorithm.

The scheme of spherical wave propagation is presented on Figure 1, and the examples of contour optimization according to designed three strategies are indicated on Figure 2.

The proposed approach of boundaries interpolation into a missing data region does not possess a high accuracy but enough for texture reconstruction. The main advantage is a fast interpolation with a low computational cost

5 Texture Inpainting into the Missing Data Region

Texture inpainting within video sequences has also an aspect of spatio-temporal process. For a choice of texture reconstruction method into a missing data region, the decision rule as a function of the first order [24] is represented by Eq. 12 , where *ES* is size of missing data region; *SS* is a scene stability and absence of complex cases; $w_0, \ldots, w_5$ are a weighting coefficients.

$$DF\,(SM, ST, IS, ES, SS) = w_0 + w_1 SM + w_2 ST + w_3 IS + w_4 ES + w_5 SS \tag{12}$$

For an approximation of decision rule, function of which is a polynomial of the first order, let's use a fully connected one-level NN. The learning stage permits to receive weighting coefficients of the decision function is calculated by Eq. 13, where $w_{MS,j}$, $w_{IS,j}$, $w_{IS,j}$, $w_{ES,j}$, $w_{SS,j}$ are weights of synapses from input parameters *SM*, *ST*, *IS*, *ES*, and *SS* to neurons *j* of hidden layer correspondingly, *DS* is an output of NN, and $w_{j,DS}$ is a weight of synapse connecting neurons *j* of hidden layer to the output *DS*.

$$DS = f_{ac}\left(\sum_{j=0}^{l-1}\left(f_{ac}\left(SMw_{SM,j} + STw_{ST,j} + ISw_{IS,j} + ESw_{ES,j} + SSw_{SS,j}\right) w_{j,DS}\right)\right) \tag{13}$$

Learning stage permits to receive weighting coefficients in Eq. 12. Characteristics of NN for a choice of reconstruction method are situated in Table 2.

Experimental investigation shows that methods from Table 3 can be successfully applied for texture reconstruction.

Blurring methods reconstruct missing pixels by mixing of brightness and color components: the anisotropic blurring is realized by Gaussian pyramid, the isotropic blurring suggests an assignment weighed values for all known texture pixels. Texture tiles are applied if a texton is successfully segmented. Texture synthesis is used only in the case when a type of texture can not be determined.

Three motion approximation models in scene permit to find the correspondence feature points in adjacent frames for reconstruction of missing data region. In the case of the linear motion model in scene point coordinates (x_n, y_n) in reconstruction frame *n* are calculated by Eq.14:

$$\begin{bmatrix} x_n \\ y_n \end{bmatrix} = (n-1) \times \begin{bmatrix} x_i - x_{i-1} \\ y_i - y_{i-1} \end{bmatrix} \tag{14}$$

Table 2. Characteristics of NN for a choice of reconstruction method

Parameter	Value
Number of NN inputs	5
Number of NN outputs	1
Number of hidden layers of NN	1
Number of neurons in hidden layer	10
Activation function	Sigmoid function
Number of samples	1,500
Samples for learning	1,200
Samples for testing	300
Number of learning cycles	1,000,000
Recognition accuracy (learning sampling)	97.8 %
Recognition accuracy (testing sampling)	99.2 %

Table 3. Classification of texture synthesis methods

Type of methods	Methods	Output range of NN
Blurring	Anisotropic blurring	0 - 49
	Isotropic blurring	50 - 99
Texture tile	Anisotropic tile	100 - 149
	Isotropic tile	150 - 199
Texture synthesis	Statistical methods	200 - 249
	Superposition of textures	250 - 299

Eq. 15 estimates a transition of known point from the adjacent frame being reconstructed in the case of a rotational model with a view $\mathbf{X}_n$, where

1. $\mathbf{X_n}$ is a matrix of homogeneous coordinates of reconstructed point,
2. **T** is a matrix of transition of coordinate system into a rotating center,
3. **R'** is a rotating matrix of point,
4. **S** is a matrix of transition of coordinate system into initial center,
5. **X** is a matrix of homogeneous coordinates of point in adjacent frame,
6. n is a displacement in frames.

$$\mathbf{X}_n = \mathbf{T} \cdot \left(\prod_{i=1}^{n-1} \mathbf{R}'_i \right) \cdot \mathbf{S} \cdot \mathbf{X} \tag{15}$$

Let's suppose that a scale coefficient k is a constant for several sequential frames. Then a transition of known point (x_{i-1}, y_{i-1}) from some frame in the point (x_n, y_n) of reconstructed frame (in the case of scalable model) is described by Eq. 16, where **K** is a scale matrix.

$$\mathbf{X_n} = \mathbf{T} \cdot \left(\prod_{i=1}^{n-1} \mathbf{K}_i \right) \cdot \mathbf{S} \cdot \mathbf{X} \tag{16}$$

For approximation of any type motion, the analysis of long frames series is recommended.

6 Experimental Results

The accuracy of experimental results estimated as a relation of right reconstructed background pixels to their common amount into missing data regions. Video sequences (animation and movies) had a different resolution. In Tables 4 and 5, results of a spatial reconstruction (blurring) and a temporal reconstruction (a linear motion model) are presented. As one can see, the results for static scenes are better then for dynamic scenes, and blurring is a worse method as compared with a linear model.

Table 4. Estimation results of reconstruction in static scenes

Video sequences	Subtitles	Static logotype	Artifact	Object
		Blurring		
Animation	0.88	0.84	0.90	-
Movies	0.83	0.85	0.93	-
		Linear motion model		
Animation	0.92	0.93	0.97	0.84
Movies	0.95	0.97	0.97	0.89

Table 5. Estimation results of reconstruction in dynamic scenes

Video sequences	Subtitles	Static logotype	Artifact	Object
		Blurring		
Animation	0.70	0.72	0.85	-
Movies	0.72	0.75	0.88	-
		Linear motion model		
Animation	0.89	0.85	0.90	0.69
Movies	0.82	0.83	0.90	0.78

On Figure 3 a frame with complex background which was taken from “CSI TV show” is sequentially processed. One can see results of only spatial reconstruction (Figure 3 (d)); they are significantly worse then result of spatio-temporal reconstruction (Figure 3 (e)).

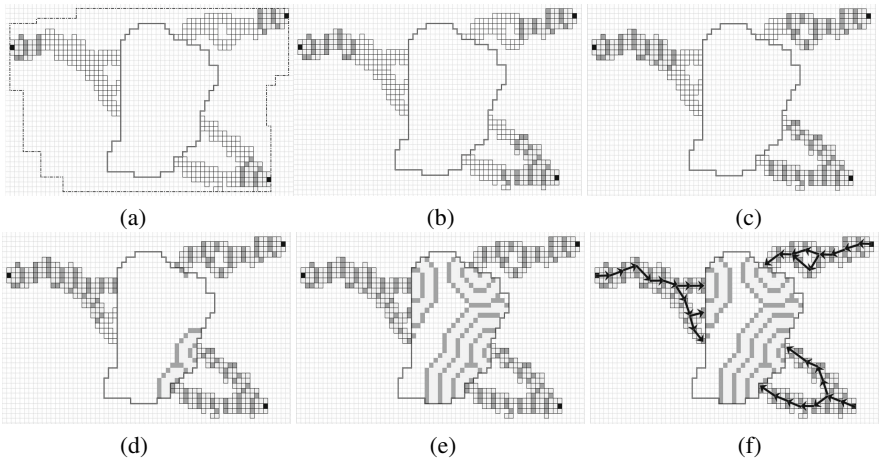


Fig. 1. The scheme of spherical wave propagation: (a) an initial fragment (black points are the centers of wave propagation); (b); (c); (d); (e) the sequential steps of wave propagation under the initial conditions; and (f) a vectors representation in a current step

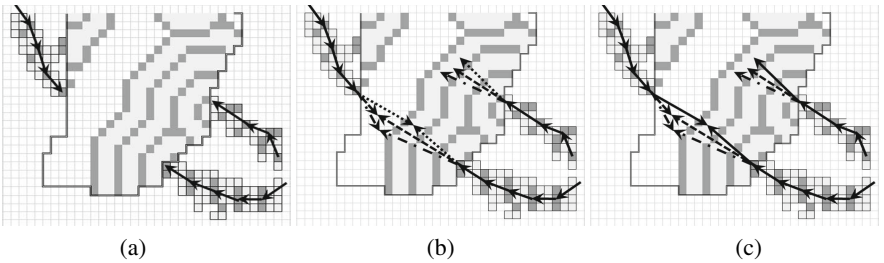


Fig. 2. The examples of contour optimization: (a) an initial fragment; (b) the vectors building (first strategy – dot-and-dash line, second strategy – large-scale dash line, third strategy – small-scale dash line); and (c) the chosen variant of contour optimization

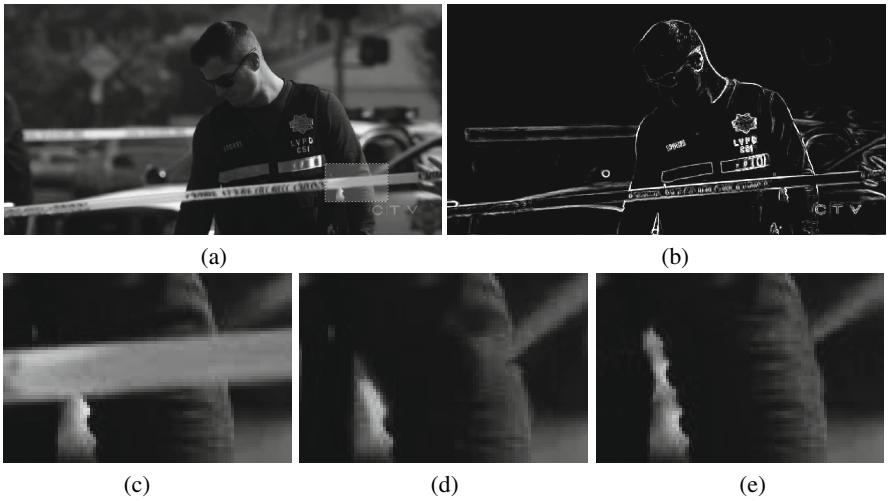


Fig. 3. The examples of small region reconstruction: (a) the initial frame in grey-scale performance; (b) Laplacian performance; (c) the increased fragment; (d) result of spatial reconstruction; and (e) result of spatio-temporal reconstruction

7 Conclusion

In this paper, an intelligent method of texture reconstruction of missing data in video sequences is proposed. The neural network approach is used to define the texture type (spatial features). Three fully connected two-level neural networks were designed to estimate texture smoothness, structural performance, and isotropy. Temporal features are based on feature points extraction and motion detection in each scene by using linear, rotational, and scalable motion models.

A fast algorithm of wave propagation accomplishes a boundaries' interpolation in a missing data region. This process is necessary for artifacts and objects removal. Three strategies of contour optimization based on the analysis of previous motion vectors were realized. Spatial texture inpainting is realized by using a fully connected one-level neural network for choice one of following methods: blurring, texture tile, and texture synthesis. Temporal features simplify the texture inpainting process. Also experimental estimations of objects reconstruction in animation and movies were received. Experimental results confirm the adequacy of proposed intelligent reconstruction technique.

References

1. Haralick, R.M.: Statistical and structural approaches to texture. Proceedings of the IEEE 67(5), 786–804 (1979)
2. Funakubo, N.: Region segmentation of biomedical tissue image using color texture features. In: Proceedings of the 7th Int. Conf. on Pattern Recognition, vol. 1, pp. 30–32 (1984)

3. Song-Sheng, L., Jernigan, M.E.: Texture Analysis and Discrimination in Additive Noise. Comput. Vision, Graph., Image Process. 49(1), 52–67 (1990)
4. Ojeda, S., Vallejos, R., Bustos, O.: A new image segmentation algorithm with applications to image inpainting. Computational Statistics and Data Analysis 54, 2082–2093 (2010)
5. Tsai, J.-J., Chen, N.-J., Fang, W.-C., Chen, J.-S.: A fast image reconstruction algorithm for continuous wave diffuse optical tomography. In: IEEE/NIH Life Science Systems and Applications Workshop (LiSSA), pp. 92–95 (2011)
6. Quiney, H.M., Nugent, K.A., Peele, A.G.: Iterative image reconstruction algorithms using wave-front intensity and phase variation. Optics Letters 30(13), 1638–1640 (2005)
7. Cho, D., Bui, T.D.: Image inpainting using wavelet-based inter- and intra-scale dependency. In: International Conference on Pattern Recognition (ICPR), Tampa, FL, pp. 1–4 (2008)
8. Padmavathi, S., Priyalakshmi, B., Soman, K.P.: Hierarchical Digital Image Inpainting Using Wavelets. Signal & Image Processing: An International Journal (SIPIJ) 3(4), 85–93 (2012)
9. Du, X., Cho, D., Bui, T.D.: Image segmentation and inpainting using hierarchical level set and texture mapping. Signal Processing 91(4), 852–863 (2011)
10. Ghoniem, M., Chahir, Y., Elmoataz, A.: Nonlocal video denoising, simplification and inpainting using discrete regularization on graphs. Signal Processing 90, 2445–2455 (2010)
11. Lefebvre, A., Corpetti, T., Moy, L.H.: Estimation of the orientation of textured patterns via wavelet analysis. Pattern Recognition Letters 32, 190–196 (2011)
12. Arias, P., Caselles, V., Sapiro, G.: A variational framework for non-local image inpainting. In: Cremers, D., Boykov, Y., Blake, A., Schmidt, F.R. (eds.) EMMCVPR 2009. LNCS, vol. 5681, pp. 345–358. Springer, Heidelberg (2009)
13. Bugeau, A., Bertalmio, M., Caselles, V., Sapiro, G.: A comprehensive framework for image inpainting. IEEE Transactions on Image Processing 19, 2634–2645 (2010)
14. Hedjam, R., Mignotte, M.: A hierarchical graph-based Markovian clustering approach for the unsupervised segmentation of textured color images. In: Proceedings of the International Conference on Image Processing (ICIP 2009), pp. 1365–1368 (2009)
15. Krinidis, M.: Pitas, placeI.: Color texture segmentation based on the modal energy of deformable surfaces. IEEE Transactions on Image Processing 18(7), 1613–1622 (2009)
16. Nammalwar, P., Ghita, O., Whelan, P.F.: A generic framework for color texture segmentation. Sensor Review 30(1), 69–79 (2010)
17. Ilea, D.E., Whelan, P.F.: Image segmentation based on the integration of color–texture descriptors – A review. Pattern Recognition 44, 2479–2501 (2011)
18. Pietikainen, M., Maenpaa, T., Viertola, J.: Color texture classification with color histograms and local binary patterns. In: Proceedings of the Second Int. Workshop on Texture Analysis and Synthesis, Copenhagen, Denmark, pp. 109–112 (2006)
19. Chen, K.M., Chen, S.Y.: Color texture segmentation using feature distributions. Pattern Recognition Letters 23(7), 755–771 (2002)
20. Garcia Ugarriza, L., Saber, E., Vantaram, S.R., Amuso, V., Shaw, M., Bhaskar, R.: Automatic image segmentation by dynamic region growth and multi- resolution merging. IEEE Transactions on Image Processing 18(10), 2275–2288 (2009)
21. Fadili, M., Starck, J.L., Murtagh, F.: Inpainting and zooming using sparse representations. Computer Journal 52, 64–79 (2009)
22. Xu, Z., Jian, S.: Image inpainting by patch propagation using patch sparsity. IEEE Transactions on Image Processing 19, 1153–1165 (2010)
23. Lowe, D.: Distinctive Image Features from Scale-invariant Keypoints. Int. J. of Computer Vision 60, 91–110 (2004)
24. Favorskaya, M., Zotin, A., Damov, M.: Intelligent Inpainting System for Texture Reconstruction in Videos with Text Removal. In: Proceedings of Int. Congress on Ultra Modern Telecommunications and Control Systems, ICUMT 2010, pp. 867–874 (2010)

25. Vacha, P., Haindl, M., Suk, T.: Colour and rotation invariant textural features based on Markov random fields. Pattern Recognition Letters 32, 771–779 (2011)
26. Khan, J.F., Adhami, R.R., Bhuiyan, S.M.A.: A customized Gabor filter for unsupervised colour image segmentation. Image and Vision Computing 27(4), 489–501 (2009)
27. Al-Takrouri, S., Savkin, A.V.: A model validation approach to texture recognition and inpainting. Pattern Recognition 43, 2054–2067 (2010)
28. Favorskaya, M.N., Petukhov, N.Y.: Recognition of natural objects on air photographs using neural networks. Optoelectronics, Instrumentation and Data Processing 47(3), 233–238 (2011)
29. Anupam Goyal, P., Diwakar, S.: Fast and Enhanced Algorithm for Exemplar Based Image Inpainting. In: 4th Pacific-Rim Symposium on Image and Video Technology (PSIVT), pp. 325–330 (2010)
30. Guo, H., An, J.: Image Restoration with Morphological Erosion and Exemplar-Based Texture Synthesis. In: 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM), pp. 1–4 (2010)
31. Shalini, S.S., Menaka, D.: Exemplar Based Image and Video Inpainting. International Journal of Communications and Engineering 04(4), 112–119 (2012)
32. Zhang, Q., Lin, J.: Exemplar-Based Image Inpainting Using Color Distribution Analysis. Journal of Information Science and Engineering 28, 641–654 (2012)
33. Huan, X., Murali, B., Ali, A.L.: Image restoration based on the fast marching method and block based sampling. Computer Vision and Image Understanding 114, 847–856 (2010)
34. Boulanger, J., Kervrann, C., Bouthemy, P.: Space–time adaptation for patch-based image sequence restoration. IEEE Transactions on PAMI 29, 1096–1102 (2007)
35. Protter, M., Elad, M., Takeda, H., Milanfar, P.: Generalizing the non- local-means to super-resolution reconstruction. IEEE Transactions on Image Processing 18(1) (2009)
36. Harris, C., Stephens, M.: A combined corner and edge detection. In: Proceedings of The Fourth Alvey Vision Conference, Manchester, UK, pp. 147–151 (1988)
37. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: International Joint Conference on Artificial Intelligence, pp. 674–679 (1981)

Classification Based on Prototypes Generated with Fuzzy C-means Clustering and Differential Evolution

Joanna Jędrzejowicz[1] and Piotr Jędrzejowicz[2]

[1] Institute of Informatics, Gdańsk University,Wita Stwosza 57, 80-952 Gdańsk, Poland
jj@inf.ug.edu.pl
[2] Department of Information Systems, Gdynia Maritime University, Morska 83, 81-225 Gdynia, Poland
pj@am.gdynia.pl

Abstract. In this paper we propose a simple and effective combined classifier based on the data reduction carried-out through applying fuzzy C-means clustering and differential evolution techniques. The idea is to produce clusters from the training set instances applying fuzzy C-means algorithm. In further step cluster centroids are used as seeds in the differential evolution algorithm to construct prototypes, each representing a single cluster. Simple distance-based weak classifiers are then used to produce the AdaBoost combined classifier. The approach has been validated experimentally. Computational experiment results confirm good quality of the proposed classifier.

Keywords: machine learning, combined classifier, fuzzy C-means clustering, evolutionary algorithms, differential evolution.

1 Introduction

In this paper we propose a simple and effective combined classifier based on the data reduction carried-out through applying fuzzy C-means clustering and differential evolution techniques. The problem of selecting relevant information remains an important focus of research in the field of machine learning. In the machine learning context selection of information is called data reduction. Data reduction is expected to result in decreasing a quantity of data required without loss of the important information.

In other words the purpose of data reduction is to find out or construct a subset of prototypes representing the original training set in such a manner that the performance of a classifier induced from the set of prototypes is better than, or at least not worse than a classifier induced from the original dataset [11]. This means that data reduction techniques aim at selecting informative instances and, finally, producing a minimal set of instances or prototypes to represent a training set and presenting the reduced dataset to a machine learning algorithm [14].

Data reduction algorithms can be broadly classified into two categories: prototype selection and prototype extraction. Prototype selection is concerned with choosing a subset of reference vectors from the original set, including a selection of relevant attributes,

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 177–188, 2013.

whereas prototype extraction deals with construction of entirely new dataset, smaller than the original dataset. Data reduction can be also achieved by feature construction based on some transformation of original features, such that the resulting dataset is smaller than the original one [3]. Reviews of the data reduction techniques can be found in [10] and [4]. Unfortunately, data reduction belongs to the class of the NP-hard problems [7]. Hence, techniques used to reduce data are usually based on various approximation algorithms. Another technique often useful in improving classification results accuracy is combining several classifiers into the, so called, combined classifier. Benett [2] argues that computational gain can be realized by partitioning the data and applying an instance of the classifier to each subset. In other situations, combining classifiers can be seen as a way of extending the hypothesis space or relaxing the bias of the original base classifier. Among methods based on classifier combination one can mention stacking [16] and boosting [13], as well as adaptive boosting known as the AdaBoost [5]. In this paper we propose integrating data reduction with the adaptive boosting. The idea is to produce clusters from the training set instances applying fuzzy C-means algorithm. In further step cluster centroids are used as seeds in the differential evolution algorithm to construct prototypes, each representing a single cluster. Simple distance-based weak classifiers are then used to produce the AdaBoost combined classifier.

This paper is organized as follows. Section 1 contains introduction. In Section 2 the proposed combined classifier is described and explained. Section 3 presents results of the computational experiment. Finally, Section 4 includes conclusions and ideas for future research.

2 Classification Using Data Reduction Techniques

2.1 Data Classification Problem

In this paper the general data classification problem is considered. In what follows, C is the set of categorical classes which are denoted $1,\ldots,|C|$. We assume that the learning algorithm is provided with the learning instances $LD = \{< d,c > \mid d \in D, c \in C\} \subset D \times C$, where D is the space of attribute vectors $d = (w_1^d,\ldots,w_n^d)$ with w_i^d being numeric values, n - the number of attributes. The learning algorithm is used to find the best possible approximation $\bar{f}$ of the unknown function f such that $f(d) = c$. Then $\bar{f}$ can be used to find the class $\bar{c} = \bar{f}(\bar{d})$ for any $\bar{d}$ such that $(\bar{d},\bar{c}) \notin LD$, that is the algorithm will allow to classify instances not seen in the process of learning. The set of learning instances LD consists of two subsets $LD = TD \cup TS$ that is TD-training set and TS - testing set.

The first step in learning is the preprocessing of data from the training set in order to find reference vectors $RV^c \subset TD$ for each class $c \in C$, or in other words, best possible representations for each class. Since the size of the set of reference vectors $\bigcup_{c \in C} RV^c$ will be considerably smaller than that of TD this will allow for a speed-up of the next learning step.

The following metrics were used (and compared) in the experiments. For x, y being attribute vectors of dimension n we have:

- Euclidean metrics: $d_e(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$

- Manhattan metrics: $d_m(x,y) = \sum_{k=1}^{n} \mid x_k - y_k \mid$
- discrete metrics: $d_d(x,y) = \sum_{k=1}^{n} g_k$, where

$$g_k = \begin{cases} 0 & x_k = y_k \\ 1 & \text{otherwise} \end{cases}$$

When describing the algorithms, we assume the metrics, being one of the above and denoted $\| \ \|$. The preprocessing stage makes use of two techniques:

- fuzzy C-means clustering (see Dunn [6]), and
- differential evolution (see Storn and Price [12]).

Fuzzy C-means clustering is used to replace data vectors by a fixed number of cluster centroids for each class. To continue prototype selection the next step uses Differential Evolution (DE) to increase diversity. Note that both steps are of prototype extraction type and an entirely new dataset is constructed.

2.2 Fuzzy C-means

Fuzzy C-means is a clustering method which allows one row of data to belong to two or more clusters. The method is based on minimization of the objective function

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{noCl} u_{ij}^m \| x_i - c_j \|$$

where m is a fixed number greater than 1 (in the experiments the value was fixed and equal 2), N is the number of data rows, $noCl$ is the number of clusters, c_j is the center of the j-th cluster and u_{ij} is the degree of membership of the i-th data row x_i in cluster j. Fuzzy C-means clustering is an iterative process with the update of membership factors u_{ij} and cluster centers c_j defined by Eqs. 1 and 2:

$$u_{ij} = \frac{1}{\sum_{k=1}^{noCl} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{1}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{2}$$

The initial values in matrix $u = (u_{ij})$ are random numbers between 0 and 1. The algorithm of fuzzy C-means clustering is shown as Algorithm 1. Note that the computational complexity of Algorithm 1 is $O(noIt \cdot N \cdot |CL|)$, where $|CL| = |C| \cdot noCl$ is the number of clusters, $N = |TD|$ is the number of datarows in the training set and $noIt$ is the number of iterations.

2.3 Differential Evolution

Differential evolution is a parallel direct search method which utilizes a fixed number of n-dimensional parameter vectors $\{w_i : i = 1, \ldots m\}$ as a population in each generation.

Algorithm 1. Fuzzy C-means clustering

Require: training data $TD = \bigcup_{c \in C} TD^c$, *noCl* - number of clusters, accuracy ε
Ensure: for each class $c \in C$ the collection of *noCl* cluster centers
1: **for all** $c \in C$ **do**
2: initialize the matrix $U^{(0)} = [u_{ij}]$, $i = 1, \ldots |TD_c|$, $j = 1, \ldots, noCl$
3: **repeat**
4: **for** $j = 1$ to *noCl* **do**
5: calculate the cluster center c_j according to (2)
6: **end for**
7: update the matrix U using (1)
8: **until** $max_{ij} |u_{ij}^{(k+1)} - u_{ij}^{(k)}| < \varepsilon$
9: **end for**
10: **return** *noCl* cluster centers for each class c

The population size, m, does not change in the process. Usually, the initial population is chosen randomly and should cover the entire parameter space. In each iteration step, DE generates new parameter vectors by adding the weighted difference between two population vectors to a third vector. This operation is called mutation. The mutated vector parameters are then mixed with the parameters of another predetermined vector, the target vector, to define the so-called trial vector. This operation is a cross-over. If the trial vector yields a lower fitness function value than the target vector, the trial vector replaces the target vector in the following generation. Each population vector has to serve once as a target vector so that m competitions take place in each generation.

To describe the application of differential evolution in our case we start with the definition of fitness function which is vital in the process. Let CL stand for the population of cluster centers (i.e reference vectors) and TD - for training set, as before. The set CL forms the initial population for DE. Thus the fixed population size is equal to the number of clusters $|CL|$. Then fitness of TD with respect to CL is the relative number of correctly classified vectors from TD, where the class assigned to a vector x is from the cluster $cl \in CL$ nearest to x.

Formally, let $CL = \bigcup_{c \in C} CL^c$, where

$$CL^c = \{< u, c >: \ u \text{ is n-attribute vector}\}$$

and CL^c contains a fixed number of elements. For $< x, y > \in TD$ let

$$c = arg \min_{c \in C} \{\| x - u \| : < u, c > \in CL^c\}$$

$$fitness_{CL}(TD) = \frac{\sum_{<x,y> \in TD} sg(c = y)}{|TD|}$$

where

$$sg(\varphi) = \begin{cases} 1 & \text{if } \varphi \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

The abbreviation $fitness_{CL[c \to d]}(TD)$ will stand for the fitness of training set TD with respect to cluster centers CL where one specified cluster c is replaced by d.

The basic strategy of DE is to change the population of $noCl \cdot |C|$ cluster centers in a fixed number of generations, performing in each generation two transformations for each cluster:

- mutation, and
- crossover.

For each population member x (cluster center) a mutant vector $Mut(x)$ is generated by adding the weighted difference (defined by a parameter F) between two population members x_1, x_2 to a third population member x_3. Since the randomly chosen vectors are expected to be all different and different from the mutated one x, the size of the population should be at least 4.

The crossover transformation uses vector x and its mutant $Mut(x)$ to generate the final result $DE(x,F)$ of differential evolution. The values of attributes of vector $DE(x,F)$ are randomly chosen from either x or $Mut(x)$, though it is ensured that at least one attribute value is taken from $Mut(x)$. Differential evolution is more demanding than C-means clustering in terms of computational complexity. It is $O(noIt \cdot |CL|)$ since in each step it requires the recalculation of fitness function, and the distance $||x-u||$ is counted for each datarow x and cluster centroid u. Fuzzy C-means clustering and differential evolution are both used in both classification algorithms evaluated in the paper.

2.4 AdaBoost with Weak Classifiers Generated via Prototype Selection

Notice that Algorithm 1 and Algorithm 2 produce a classifier, as an outcome (defined by a set of cluster centers). This type of a classifier is used in Algorithm 3, then applied in AdaBoost as a weak classifier and shown in Algorithm 4.

Algorithm 2. Differential evolution

Require: training data TD, the population of cluster centers $CL = \bigcup_{c \in C} CL^c$, $genN$ - number of generations, F- weight,
Ensure: new population of cluster centers CL^{new}

for all $c \in C$ **do**
  $ss = fitness_{CL}(TD)$
  $it = 0$
  repeat
    $it = it + 1$
    for all $cls \in CL^c$ **do**
      $trial = DE(F, cls)$, as defined above,
      $sn = fitness_{CL[cls \to trial]}(TD)$
      if $sn > ss$ **then**
        $ss = sn$
        $cls = trial$
      end if
    end for
  until $it = genN$
end for
$CL^{new} = CL$

Algorithm 3. FDE- fuzzy C-means clustering with differential evolution

Require: training dataset TD, testing dataset TS, number of clusters $noCl$, accuracy ε, F-weight in DE
Ensure: qc-quality of classification
1: use Algorithm 1 to generate the population CL of $noCl$ cluster centers for each class $c \in C$
2: use Algorithm 2 to transform CL to CL^{new} via differential evolution
3: in the testing stage use testing dataset TS to estimate the quality of classification $qc = fitness_{CL^{new}}(TS)$

Algorithm 4. AdaFDE - AdaBoost constructed from Algorithm 3

Require: training data $TD = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N\}$ of size N, test dataset TS, integer T - number of iterations, integer $M \leq N$ - size of the selected dataset
Ensure: qc quality of the Adaboost classifier.
1: initialize the distribution $D_1(i) = \frac{1}{N}, i = 1, \ldots, N$
2: **for** $t = 1$ to T **do**
3: for the current distribution D_t select a training dataset $S_t \subset TD$ of size M,
4: call Algorithm 3 for the dataset S_t, receive the classifier C_t
5: using the majority voting for C_t calculate the error $\varepsilon_t = \sum_{C_t(\mathbf{x}_i) \neq y_i} D_t(i)$
6: **if** $\varepsilon_t > 0.5$ **then**
7: abort
8: **else**
9: $\beta_t = \varepsilon_t / (1 - \varepsilon_t)$
10: **end if**
{ update the distribution}
11: **for** $i = 1$ to N **do**
12: **if** $C_t(\mathbf{x}_i) = y_i$ **then**
13: $D_t(i) \leftarrow D_t(i) \times \beta_t$
14: **end if**
15: normalize the distribution $D_{t+1}(i) = D_t(i)/Z_t$, $Z_t = \sum_i D_t(i)$
16: **end for**
17: **end for**
{test the ensemble classifier $C_1, C_2, \ldots, C_T$ in the test dataset TS}
18: $qc \leftarrow 0$
19: **for all** $(\mathbf{x}, \mathbf{y}) \in$ **TS do**
20: $V_i = \sum_{C_t(\mathbf{x})=i} \log(1/\beta_t)$, $i = 1, \ldots, |C|$
21: $c \leftarrow arg \max_{1 \leq j \leq |C|} V_j$
22: **if** $c = y$ **then**
23: $qc \leftarrow qc + 1$
24: **end if**
25: **end for**
26: $qc \leftarrow qc/|TS|$
27: **return** qc

The idea of Algorithm 4 is that in each iteration t the distribution weights of those instances that were correctly classified are reduced by a factor β_t and the weights of the misclassified instances stay unchanged. After the normalization the weights of instances misclassified are raised and they add up to $1/2$, and the weights of the correctly classified instances are lowered and they also add up to $1/2$. What is more, since it is required that the weak classifier has an error less than $1/2$, it is guaranteed to correctly classify at least one previously misclassified instance. In the ensemble decision those classifiers which produced small error and β_t is close to zero, have a large voting role since $1/\beta_t$ and logarithm of $1/\beta_t$ are large.

The computational complexity of AdaBoost is $O(T \cdot |weakCl|)$, where T is the number of iterations in AdaBoost and $|weakCl|$ is the computational complexity of the weak classifier used in the algorithm.

3 Computational Experiment Results

To evaluate the proposed approach computational experiment has been carried out. The experiment involved the following datasets from the UCI Machine Learning Repository [1]:

- Wisconsin Breast Cancer (WBC),
- Diabetes,
- Sonar,
- Australian Credit (ACredit),
- German Credit (GCredit),
- Cleveland Heart (Heart),
- Hepatitis,
- Ionosphere.

Basic characteristics of these sets are shown in Table 1. In the reported experiment the following classification tools have been compared:

- Fuzzy C-means with differential evolution (FDE),
- AdaBoost constructed from the FDE-induced weak classifiers (AdaFDE).

both proposed in this paper, against the following:

- Rotation Forest with gene expression programming induced expression trees (RF-GEP) proposed in [9],
- Cellular GEP with AdaBoost (GEPC-ADA) proposed in [8],
- 15 well-known classifiers from WEKA Environment for Knowledge Analysis v. 3.7.0 [15], including classic Rotation Forest, Naive Bayes, Bayes Net, Logistic Regression, Radial Basis Function Network, AdaBoost, Support Vectors Machine, Ensemble Selection, Bagging, Classification via Clustering, Random Committee, Decision Table, FT Tree, Random Forest and C4.5.

In the reported experiment FDE and AdaFDE classifiers have been run with the following settings: number of clusters equals to $[N/100]$ where N denotes the number of

Table 1. Datasets used in the computational experiment

Name	Data Type	Attribute Type	No. Instances	No. Attributes
WBC	multivariate	integer	699	11
Diabetes	multivariate, time-series	categorical, integer	768	9
Sonar	multivariate	real	208	61
ACredit	multivariate	categorical, integer, real	690	15
GCredit	multivariate	categorical, integer	1000	21
Heart	multivariate	categorical, real	303	14
Hepatitis	multivariate	categorical, integer, real	155	20
Ionosphere	multivariate	integer, real	351	35

instances in the respective dataset. In the differential evolution value of the parameter F was set to 0.2. Computations involving RF-GEP and GEPC-ADA have been run with settings reported in [9] and [8], respectively. In case of all classifiers from WEKA environment, the default parameter settings have been used. Tables 2 and 3 show computation results of the five highest ranked classifiers out of the 15 considered, compared with the results of FDE and AdaFDE classifiers. All results have been averaged over 10 repetitions of the 10-cross-validation scheme. Performance measures include classifier accuracy shown in Table 2 and the area under the ROC curve calculated as the Wilcoxon-Mann-Whitney statistic shown in Table 3.

To evaluate the performance of the proposed classifiers, the Friedman nonparametric test using ranks of the data has been applied. For each of the two - classification accuracy and ROC area the following hypotheses were tested:

- Null Hypothesis H0: All of the 19 population distribution functions are identical.
- Alternative Hypothesis H1: At least one of the populations tends to yield larger observations than at least one of the other populations.

Analysis of the experiment results shows that in each case, that is for the population of the classification accuracy observations and the population of the ROC area observations, the null hypothesis should be rejected at the significance level of 0.05. The average Friedmans ranks for the classification accuracy and ROC area are shown in

Table 2. Comparison of the classifier accuracy (%)

Classifier	WBC	Diab.	Sonar	ACr.	GCr.	Heart	Hep.	Ion.
RF-GEP	97,65	76,99	80,79	88,04	76,27	82,35	86,46	**93,73**
GEPC-Ada	95,86	77.21	81.24	86.52	**77.37**	**83.84**	**87.13**	91.35
Rotation Forest	96.99	76.69	84.13	87.25	74.80	80.74	82.58	93.73
SVM	96.99	**77.34**	75.96	84.92	75.10	84.07	85.16	88.60
Bayes Net	97.14	74.35	80.28	86.23	75.50	81.11	83.22	89.46
FDE	97.04	73.66	84.13	86.41	72.25	80.72	72.24	86.17
AdaFDE	**98.10**	76.00	**85.57**	**88.12**	74.87	83.18	79.88	89.98

Table 3. Comparison of the area under the ROC curve

Classifier	WBC	Diab.	Sonar	ACr.	GCr	Heart	Hep.	Ion.
Bayes Net	0.992	0.806	0.877	0.921	0.780	**0.902**	**0.882**	0.948
Rotation Forest	0.986	**0.810**	**0.925**	0.921	0.757	0.883	0.826	0.967
Random Committee	0.987	0.785	0.912	0.900	0.761	0.866	0.848	**0.976**
Random Forest	0.988	0.779	0.912	0.914	0.748	0.854	0.835	0.956
AdaBoost M1	0.989	0.801	0.841	**0.929**	0.723	0.878	0.851	0.944
FDE	0.996	0.750	0.890	0.893	0.727	0.841	0.791	0.882
AdaFDE	**0.997**	0.780	0.892	0.904	**0.813**	0.879	0.827	0.942

Fig.1 and Fig. 2, respectively. With respect to accuracy it can be also observed that considering the proposed AdaFDE classifier together with the five best performing out of the remaining 17 classifiers all of the seven population functions have identical distribution functions. Similar analysis for the ROC area observation allows to state that at least one out of the six populations considered tends to yield larger observations than at least one of the other populations.

Particular advantage of the proposed AdaFDE combined classifier is its very good performance in terms of precision, understood as the fraction of instances correctly labeled

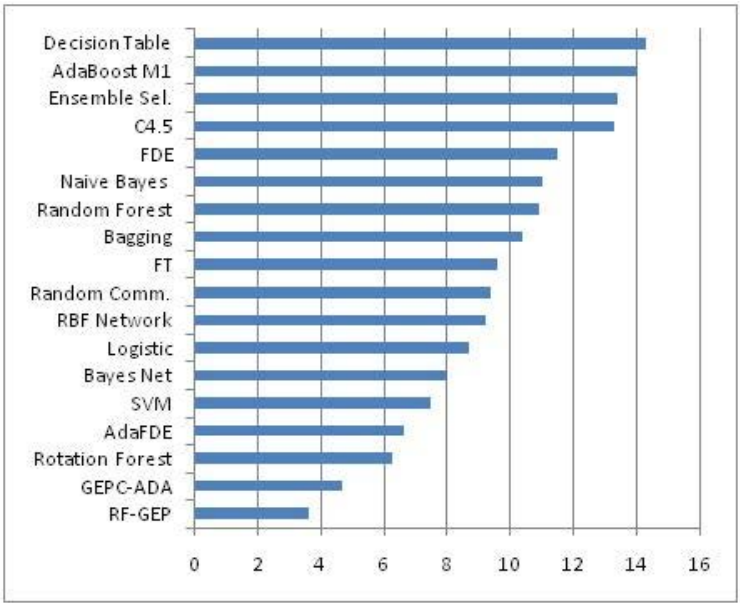


Fig. 1. The average Friedmans ranks for the classification accuracy of different classifiers (the lower value the better)

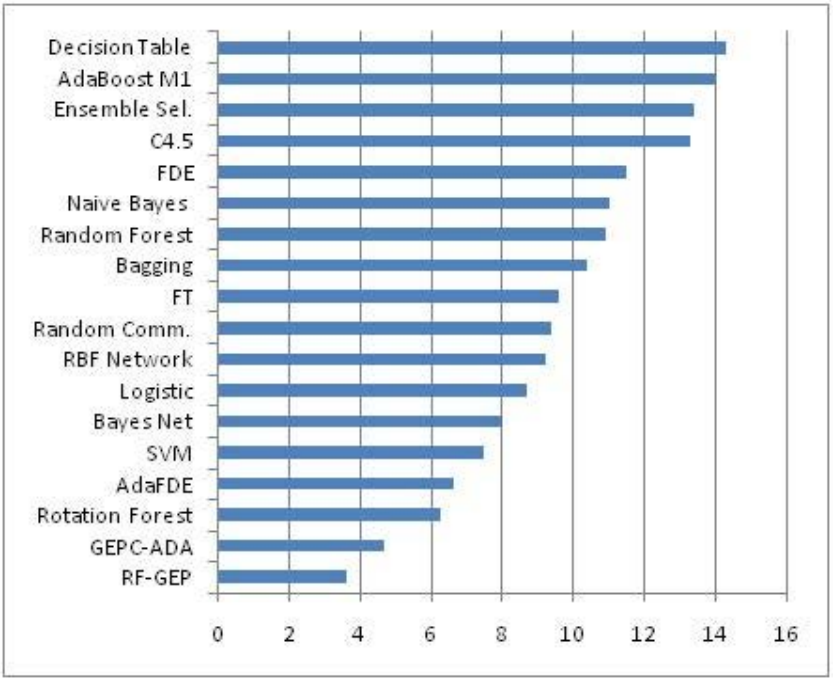


Fig. 2. The average Friedmans ranks for the ROC area (the lower the better)

Table 4. Precision of the AdaFDE versus precision of five classifiers with the highest accuracy

Classifier	WBC	Diab.	Sonar	ACr.	GCr.	Heart	Hep.	Ion.	F-rank
AdaFDE	0.975	0.855	0.899	0.914	0.810	0.794	0.872	0.963	19
GEPC-ADA	0.947	0.860	0.793	0.880	0.892	0.804	0.930	0.943	23
RF-GEP	0.919	0.908	0.782	0.893	0.880	0.681	0.920	0.968	25
Bayes Net	0.972	0.741	0.805	0.864	0.746	0.835	0.845	0.894	32
RotationForest	0.960	0.761	0.843	0.873	0.736	0.822	0.817	0.937	33.5
SVM	0.960	0.769	0.759	0.861	0.738	0.843	0.847	0.891	35.5

as belonging to the positive class. Table 4 shows average value of precision obtained in the course of the reported experiment. Ada FDE precision is compared in this table with precisions of the five classifiers with highest accuracy. Last column of the Table 4 shows value of Friedmans rank for each of the classifiers.

4 Conclusions

The paper proposes a simple and effective combined classifier based on the data reduction carried-out through applying fuzzy C-means clustering and differential evolution techniques. The idea has been to produce clusters from the training set instances applying fuzzy C-means algorithm. In further step cluster centroids have been used as seeds

in the differential evolution algorithm to construct prototypes, each representing a single cluster. Simple distance-based weak classifiers have been then constructed from prototypes with a view to produce the AdaBoost combined classifier. The proposed combined classifier denoted as AdaFDE has been validated through an extensive computational experiment described in Section 3. Analysis of its results allows to formulate the following conclusions:

- AdaFDE performs very well in terms of the classification accuracy. It has achieved best correct classification ratio in case of 2 out of 8 considered benchmark datasets. The average Friedman rank places AdaFDE on the 4th place among 18 well known and best performing classifiers known from the literature.
- Considering AdaFDE performance in terms of the ROC area it can be observed that high accuracy of the proposed classifiers does not always necessary assure best performance from the point of view of the area under the ROC curve. Nevertheless, AdaFDE can be classified among five best performing algorithms out of 18 considered with respect to the size of the area under the ROC curve.
- Particular advantage of the proposed AdaFDE combined classifier is its very good performance in terms of precision, understood as the fraction of instances correctly labeled as belonging to the positive class. It has achieved best precision ratio in case of three out of eight considered benchmark datasets. The average Friedman rank places AdaFDE on the first place among 18 well known and best performing classifiers known from the literature considering their precision. Consequently, it can be suggested that AdaFDE should be used whenever a field of application requires highest possible precision.

Although the proposed approach is conceptually simple, computational complexity inherent to combined classifiers in general and Adaboost approach in particular, makes AdaFDE suitable in case of applications where time needed for deciding on classification result is not critical which usually is the case in data streams analysis. Future research should focus on increasing efficiency of the approach through diversification of classifiers induced from the reference vectors obtained through data reduction followed by their modification process. Another direction of research should focus on decreasing computation time needed for classification through parallelization of the computation.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, University of California, School of Information and Computer Science (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
2. Bennett, P.N.: Building Reliable Metaclassifiers for Text Learning, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh (2006)
3. Bezdek, J.C., Kuncheva, L.I.: Nearest Prototype Classifier Design: An Experimental Study. International Journal of Intelligence Systems 16(2), 1445–1473 (2000)
4. Czarnowski, I.: Distributed Learning with Data Reduction. In: Nguyen, N.T. (ed.) TCCI IV 2011. LNCS, vol. 6660, pp. 3–121. Springer, Heidelberg (2011)
5. Freund, Y., Schapire, R.: A decision theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)

6. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57 (1973)
7. Hamo, Y., Markovitch, S.: The COMPSET Algorithm for Subset Selection. In: Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence, Edinburgh, pp. 728–733 (2005)
8. Jędrzejowicz, J., Jędrzejowicz, P.: Cellular GEP-Induced Classifiers. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part I. LNCS, vol. 6421, pp. 343–352. Springer, Heidelberg (2010)
9. Jędrzejowicz, J., Jędrzejowicz, P.: Rotation Forest with GEP-Induced Expression Trees. In: O'Shea, J., Nguyen, N.T., Crockett, K., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2011. LNCS, vol. 6682, pp. 495–503. Springer, Heidelberg (2011)
10. Liu, H., Motoda, H.: Instance Selection and Construction for Data Mining. Kluwer Academic Publisher (2001)
11. Nanni, L., Lumini, A.: Particle Swarm Optimization for Prototype Reduction. Neurocomputing 72(4-6), 1092–1097 (2009)
12. Storn, R., Price, K.: Differential Evolution - A Simple and Effocient Heuristic for Global Optimization over Continuous Spaces. Journal of Global Optimization 11, 341–359 (1997)
13. Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S.: Boosting the margin: A new explanation for the effectiveness of voting methods. The Annals of Statistics 26(5), 1651–1686 (1998)
14. Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-based Learning Algorithm. Machine Learning 33(3), 257–286 (2000)
15. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
16. Wolpert, D.H.: Stacked generalization. Neural Networks 5, 241–259 (1992)
17. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
18. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
19. Ferreira, C.: Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. Complex Systems 13(2), 87–129 (2001)
20. Kuncheva, L.I., Rodríguez, J.J.: An Experimental Study on Rotation Forest Ensembles. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 459–468. Springer, Heidelberg (2007)
21. Rodrigeuz, J.J., Kuncheva, I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(10), 1619–1630 (2006)

Using Multi-Agent Systems to Enhance the Level of Autonomy in Unmanned Vehicles

Jeffrey W. Tweedale[1,2]

[1] Defence Science and Technology Organisation, Edingurgh SA 5111, Australia
Jeffrey.Tweedale@dsto.defence.gov.au
[2] University of South Australia, Mawson Lakes SA 5095, Australia
Jeff.Tweedale@unisa.edu.au

Abstract. This paper describes preliminary work performed to gain an understanding of how to implement a Multi-Agent System (MAS) that can be used to improve the Level of Automation (LOA) within interfaces used by operators when controlling Unmanned Air Vehicles (UAVs). The unmanned systems market is estimated to exceed $12 billion annually by 2017. Hence there is pressure to build an autonomous fleet of vehicles that can be supervised by one human using *mission level intent*. This concept can be supported by a framework composed of interoperable MAS components acting as a team. A user interface will enable humans to select the mode of operation based on their desired level of trust. Unfortunately the term autonomy is often used synonymously with automation and this habit often confuses the topic. It is true that technology can be used to provide system automation, but increased autonomy is currently constrained by limitations in the ability to program the desired decision-making capabilities within the machine. Hybridised Computational Intelligence (CI) techniques using cognitive architectures may enable science to embed autonomy into existing mission systems. The ability to provide cognitive processing (to provide on-board intelligence) is yet to be realised. Additional capabilities are required to achieve cognition, contextual orientation, rationality, reasoning and considered decision making within the context of the situation. Hence, a sound and scalable cognitive architecture may be considered as the next step in achieving autonomy. As such, further research is required in order to evolve and adapt these concepts. Therefore, this research concentrates on the current concerns affecting the domain.

Keywords: Agents, Automation, Autonomy, Control, Machine Intelligence, Multi-Agent System, Unmanned Air Vehicle.

1 Introduction

The international market for autonomous systems will soon exceed $12 billion annually [2]. Most suppliers currently provide turn-key solutions that are tailored to achieve specific goals. They also provide disparate launch, recovery and control systems. Each system is generally supported with proprietary skills and specialist training. Industry is relying on these services because flying an Unmanned Air Vehicle (UAV) is more challenging than a child physically controlling a Radio Controlled (RC) aircraft and

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 189–208, 2013.

there is a growing demand to comply with legislative requirements. A number of modern platforms have autonomous flight systems, but many of these still require a human operator to provide the pilots skills and knowledge. An example would be the Ground Control Station (GCS) used to operate many of the larger UAVs currently being supplied. This notion applies to controlling a variety of Unmanned Vehicles (UMVs), although it may also be used to simply provide a virtual-presence on-board a platform[1]. This could be achieved with one or more humans 'in-the-loop' to successfully deploy and operate these vehicles. Smaller platforms typically offer less functionality, hence external functions must be augmented. This adds system complexity and increases the operator workload. It also introduces one or more communication node(s) in the control loop. Any additional functionality or control threads may also interrupt or inhibit existing control functions. Currently designers compensate by injecting more human support externally. This provides insurance that the technology will operate in a safe and timely manner. Unfortunately it also increases the coordination and cooperation demand, which further increases the operators work load and possibly distract them from controlling the platform, augment the machines intelligence and/or conducting payload support.

A current aim for researchers is to control more complex UAVs with a lower operator to vehicle ratio. At present a UAV control station requires a minimum of two operators: one operator to control the platform and another to control the payload of sensors. Many of these payload sub-systems retain a divergent range of functionality. This requires additional human cognitive processing, increased operator knowledge and system complexity. This added complexity increases the operator workload at a time when there is an expectation that technology will enable humans to build tools that will enable them to do more with far less effort. The military is now requesting a paradigm shift that imposes a reversal of roles, to facilitate a single operator being able to control multiple platforms.

As engineers increasingly seek to reduce the need for a human in the aircraft cockpit, they are being forced to provide operators access to traditional piloting functionality with interfaces that are easy to use. The United States Air Force (USAF) have pioneered an number of interfaces and have focused on methodology to improve the Human Machine Interface (HMI) for operators. A recent example includes experiments that seek to measure the effectiveness of several levels of automation within a customised GCS [1]. These experiments attempt to isolate cognitive processes and how human sensory information is gathered within the target environment. The results highlight that existing systems introduce temporal delays in control functions. Due to physical limitations of

[1] This virtual concept relies on mechanisms that provide a tele-presence. The process involves a form of communication that works well in controlled environments, with a fixed message or data-dump format, however the intent is often blurred or fails when the situation becomes dynamic or multi-dimensional. Examples range from a simple telephone call (turn taking), a presentation using video conferencing facilities (typically an information dump) or trying to maintain good Situational Awareness (SA) within an aerial vehicle (maintenance of environmental information as temporal knowledge to a pilot using all sensors when on-board). In a dynamic environment rich information sources are derived from numerous places and using all five human sensors. When constrained through televisory networks, this is typically limited to visual information.

communication systems, constantly changing geographical displacement and the constrained ability to share situation awareness (originally provided via an on-board presence). These factors also restrict the war-fighters ability to remotely request imagery and reduces the extent of immediate oversight a human pilot would normally provide. To address issues related to telemetry and physical access, the UAV community increasingly devolves this functionality using machine interfaces that enable the human operator to interact with the platforms on-board systems. To improve efficiency, these interfaces must also be integrated into traditional enterprise systems and compensate for the constraints associated with telemetry and cognitive delays.

Researchers need to identify enablers of the goal of providing autonomous operation by providing human like intelligence on the platform. They also need to develop tools to automate the support functions. These include: system logistics, mission planning, execution and even routine tasks like health monitoring or resource management. In order to successfully locate the human outside-of-the-loop[2], the support community still needs to improve the level of Machine Intelligence (MI) to a level that enables cognitive processing on-board. Although this technology may not be available on-board in the foreseeable future (possible up to 25 years), enabling research must be able to create the ability to recognize objects within the environment, reason, understand situational context and evaluate knowledge. Prior to this achievement, a human *must* remain in the decision loop. Demand is mounting for science to orchestrate a paradigm shift that moves society away from 'controlling' processes to a more autonomously 'operated' perspective [2]. This begins with automation and builds to achieve autonomy. It is technically feasible to facilitate higher levels of automation, however to succeed, researchers need to improve existing operator control techniques and HMIs. It is possible to augment some of the existing HMI functions with technology to automate them so they require less effort and progressively achieve improved autonomy. When they do, this will enable a successful transition where humans 'assist' machines. From this position there will be further scope to evolve to a point where humans simply 'manage' one or more capabilities.

This paper describes concepts surrounding methods of improving the Level of Automation (LOA) used by operators when controlling UAVs. Given a clear understanding of the automated system, it may then be possible to embed the knowledge required to invoke more autonomy on-board. Section 2 provides a brief background about concepts relating to both platform control and the LOA required in control systems. These topics are expanded in Sections 3 and 4, while the development of autonomy and cognitive architectures are discussed in Sections 5 and 6. The methodology for testing these concepts is also described in Section 7, which is followed by the conclusion and summary of future effort.

2 Background

UAVs are increasingly being used to conduct the dull, dirty and dangerous or repetitive tasks previously conducted by humans. Industry is focusing on leveraging autonomous

[2] Where existing autonomous systems attempt to isolate the human from the *decision-loop* in order to achieve self-governance with respect to any other connected systems.

technologies to achieve similar gains to those achieved through production lines in manufacturing. By increasing the endurance, persistence and precision of unmanned systems, revolutionary changes are expected to rapidly infiltrate existing domains, especially when operators adapt or find innovative ways of employing these capabilities. They are already becoming a viable option for use in civil applications. One recent example is *Fedex's* attempt to champion an autonomous freight system[3]. Another is the use of *Scan Eagle*[4] drones to "map the leading edge of wild fires in Alaska" using its infrared imaging capability[5]. More autonomous systems are technically and economically viable, but current safety legislation mandates the need for a human pilot behind the stick in shared airspace. This has nothing to do with the operators trust, merely concerns about avoidance, providing de-confliction and cognitive responses during emergencies. It will be difficult to convince the public and governing bodies to deliver the changes required to succeed without significant scientific and engineering evidence.

By contrast, modern manned platforms now use complex Information Technology (IT) infrastructure to achieve interoperability. These normally consist of multiple communication systems that facilitate enterprise-level, registry based and discovery compliant data mediated services (such as the Joint Strike Fighter (JSF) [3]). These systems rely heavily on corporate infrastructure to facilitate this service oriented functionality. There is an economising desire for industry to automate existing systems and achieve seamless integration in order to achieve 'Autonomic Logistic' functionality. These enhancements will be limited by the available technology and by the level of intelligence being integrated into component systems. Within these constraints, increased automation within existing systems is feasible, although many legacy systems still rely on external stimuli to succeed. For instance UAVs still rely on augmented systems to assist with launch, recovery, planning and many mission related activities. To reduce the human effort required in sustaining these activities, there is pressure to provide more integrated control systems, with intuitive HMI, that are accompanied by equally autonomous mission planning and execution tools. Examples include effort by Taylor on 'Human Automation Integration for Supervisory Control of UAVs' [4] and Calhoun et al. on the 'Effects of Levels of Automation in Unmanned Aerial Vehicle Routing Task' [5].

The Defence Science and Technology Organisation (DSTO) is focused on achieving more using less people, both in the cockpit and on mission systems. This research commonly facilitates improved HMI using Advanced Information Processings (AIPs). A limited subset of decision making rules have been incorporated into automated components using Artificial Intelligence (AI) and Computational Intelligence (CI) techniques. The resulting improvements in automation have benefited both the pilot and crew. A recent example is the increased level of automation provided on-board the Airborne Early Warning and Control (AEW&C). At present there is a growing need for more

[3] See 'Fred Smith: FedEx wants UAVs' at `http://diydrones.com/profiles/blogs/fred-smith-fedex-wants-uavs`

[4] See the overview at `http://www.boeing.com/defense-space/military/scaneagle/`

[5] 'Drones over Alaska' ccan be downloaded from `http://www.anchoragepress.com/news/drones-over-alaska/article.html`

mature levels of MI that enable increased autonomy within and between platforms[6]. This process is stimulating a paradigm shift that is moving away from automation and interoperability to focus on technologies that provides real autonomy in future systems.

At present the aviation community still views the UAV platform as an UMV [6]. As the term suggests, traditional manned platforms are controlled by humans who have the *roles* and *responsibilities* conducted by the pilot in the cockpit. This includes managing the physical, cognitive and social activities encountered during a mission. The roles of a pilot can be described as: mission commander, navigator, communications officer, local air-traffic manager, Intelligence, Surveillance, Reconnaissance (ISR) expertise, mission knowledge manager and health status monitoring or resource management. Pilots can process large amounts of environmental information to maintain both *short-term* and *long-term* memory maps of the situation, and frame this within the context of the mission and their own beliefs. This dynamic pool of cognitive knowledge provides the pilot with a level of on-board *situated intelligence* that is independent of sensor data. That *knowledge* is generally only shared upon request or as required by the pilot(s). These cognitive processes enable them to maintain a working level of immediate SA and ultimately contributes to intelligent decision making on-board. Unfortunately SA is contextual [7,8] and cognitive information contributes to how internal beliefs evolve[7]. At present humans interpret changes in their environment and share this knowledge through traditional Command and Control (C^2) or Command, Control, Communications, Computers and Intelligence (C^4ISR) mechanisms. While this enables them to share a working level of immediate SA it also contributes to intelligent decision making. Many of the more intuitive functions gained through human sensory systems are sacrificed when the pilot is isolated from the cockpit[8]. Simultaneously, temporal delays and geographical displacement issues are introduced, while the pilot's ability to employ instinctive reactions is inhibited or completely disconnected. Over time the AI community may provide sufficient machine intelligence to facilitate an appropriate LOA that can achieve the community's goals relating to autonomy, however that research is still evolving.

The above research focuses on automating the 'piloting' roles to remotely control the platform. A number of achievements provide automation include the auto-pilot or take-off and landing systems, but these require off-board support mechanisms (such as satellites and the Instrument Landing System (ILS) infrastructure). Autonomy may only be realised after the community acknowledges that the pilot is more than just arms and legs. Computers do not successfully capture cognitive knowledge or many of the more intuitive functions gained through human sensory systems (many of which are sacrificed when the pilot is isolated from the cockpit). At present any ad-hoc problem

[6] The level of MI should be measured as a Machine Quotient (MQ) that is normalised against the human Intelligence Quotient (IQ) score for comparative purposes.

[7] Note that once a belief is formed the human psyche resists change without significant or dramatic stimuli. Hence more usable interfaces, demanding less effort to employ command and control functionality are required.

[8] Mica Endsley described SA as "*the perception of the elements in the environment within a volume of time and space*". The observer must also "comprehend the right meaning of these elements and project their status into the near future" [9].

solving can be achieved by the pilot who can react intuitively, concentrate, or exploit stored knowledge. A recent study of racing car drivers confirmed that muscle memory can also provide an instinctive response that requires little cognitive processing. Simultaneously, *temporal delays* and *geographical displacement* issues are introduced were the operators ability to employ instinctive reactions is inhibited or completely disconnected. Again we know that SA is contextual and human sensory information provided from within the platform contributes to how internal beliefs evolve. Once a belief is formed the human psyche resists change without significant or dramatic stimuli. There are many issues, but the AI community may soon provide an acceptable level of MI capable of facilitating virtual cognition that will achieve the communities goals relating to autonomy on-board platforms.

Most military forces have recently increased the tempo in moving from manned to unmanned systems because the traditional approach of detaching the pilot has resulted in the need to augment the roles provided by a human on-board. The cost of achieving a reasonable level of situated intelligence has been significant. Examples include the 170 personnel for a *Reaper* Combat Air Patrol (CAP) and 300 for a *Global Hawk* [2]. The additional personnel are required to augment human functionality provided throughout the whole mission. Initial attempts relied on automating existing functions, like navigation. This approach enabled the platform to complete a pre-planned flight path and little else. The platform must also be able to autonomously take-off or land, deal with environmental influences and variation in the mission plan. Realistically, existing systems simply decentralises the process of maintaining a database of way-points and control. Operators can change plans dynamically, but they can only respond to the stimuli provided. For instance, given the situation where Global Positioning System (GPS) is lost or communications become unreliable a platform becomes vulnerable and the mission may be compromised.

The field of unmanned systems already treats the concept of managing an interconnected networks of intelligent peripheral as capabilities. This collection of discrete components[9] can be invoked using a distributed or service-based network. Based on the human need to functionally organise work flows using hierarchies, computers need to contain sympathetically structured systems. These should contain an interconnected series of control links that are governed using a set of operating instructions, based on human-like rules and procedures. Translating this concept into a working system poses a number of challenges for researchers, especially with respect to the coordination and cooperation of a dynamically evolving organization. Prior to determining how to represent the structure of a personified machine-based organisation, the community must determined the utility, costs and rewards. Designers need to consider which trade-off's can be supported after establishing the value, type and effectiveness of these future systems. Research is required to coordinate an appropriate responses from a team of experts that currently operate across multiple domains. Initial work could concentrate on automating the roles conducted by one or more humans. Prior to analyzing or enhancing selected capabilities within existing systems, a review of existing control systems and human interface automation is required. In time, MI will be capable of perceiving

[9] Software modules can be bundled as components that embody specific function and/or capabilities.

the environment, performing cognitive recognition, and rationalising context, but until then, humans need to remain-in-the-loop. This evolution is also true for autonomy and cognitive architectures.

3 Control

The concept of control is an active topic of research[10], however it is generally accepted that a hierarchy of commands does exist. In order to provide a single operator with the ability to control multiple platforms, researchers need to examine the interactions between the whole system, its supporting structure and how the organization manages its people. It would be essential to automate existing work flows and be able to capture or manage any bonds or links as they evolve. Figure 1 depicts the perceived human-machine continuum. This ranges from the most primitive or physical machine level motions through to human commands, *thought or intent*: in this case the *low level machine commands* (at the base of the triangle) to *human-like behaviour* (at the top). The concept indicates that the ideal level of control would resemble the ability to provide mission level *intent* commands, that one day may even be replaced using pure *thought*[11]. This concepts makes a distinction between the concept of human *Command* and machine *Operation*, therefore this contribution only focuses on the concepts relating to the direct *physical control* of a machine as opposed to remote or *supervised control* of the platform. All initial experiments will concentrate on autonomic logistics through the automation of positional control mechanisms.

As for the control continuum, Parasuraman et al. believe that trust and the LOA can also be categorized into ranges. They argue that the human workload increases as the degree of automation increases [12][12]. Figure 2 depicts a theoretical representation of this concept. The curve may not be linear and is only provided to delineate the perceived relationship. The human aspect depicts the relationship between the level of control and the cognitive effort required, while the machine aspect displays the level of automation required for the same functionality. By removing the pilot, you isolate the on-board intelligence, which can exacerbate the level of control and constrains the functionality available to operate the platform. Traditionally the human in the cockpit significantly adds value to its survival and the missions success. This results in the need to retain the Human in the Loop (HIL) using remote control systems. By removing the pilot, designers have effectively removed any on-board cognitive processing and replaced it with a geographically isolated and time-shifted tele-presence. This severely compromises the local source of cognitive interaction and any immediate reaction or instinctive responses.

[10] An example includes Autonomy Levels for Unmanned Systems (ALFUS) [10].

[11] The industry is already providing cognitively functional Electroencephalogram (EEG) helmets for games with limited functionality. Veritas Scientific is attempting to exploit P300 Event Related Potentials (ERPs) to provide improved human-machine interfaces. See `http://www.veritasscientific.com/`

[12] This factor was confirmed by the Air Force Research Laboratory (AFRL) recently through human-in-the-loop experiment's involving operators controlling UAVs [5].

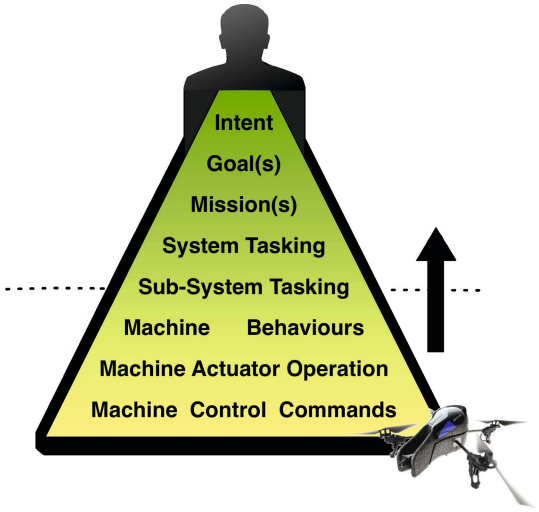


Fig. 1. Continuum of Command and Control (Most Primitive Bottom) [11]

Given the maturity of Attitude and Heading Reference System (AHRS) available, researchers and hobbyists[13] can safely ignore issues related to platform attitude controls[14]. Hence the automation required to achieve this functionality is assumed and only positional information is discussed. It is assumed that navigation control systems also process commands that adjust speed, altitude and position to maintain the platforms orientation or flight objectives, although a single control system must catalogue the kinematics of each platform, in order to provide the appropriate pre-sets for height, velocity and any variations in turn rate. These factors can be used by a virtual Proportional, Integral and Derivative (PID) controller to integrate a basic seeking function and provide positional navigation controls based on referential feedback (GPS derived). The basic algorithm for this function uses an iterative loop that includes:

- obtain the current GPS location;
- compare it with the intended way-point or leg coordinate;
- calculate any differential offsets (distance between two points);
- determine an appropriate heading;
- calculate the inertial corrections required; and
- issue commands to correct the bearing.

Existing control mechanisms must also be examined to assess the technical feasibility and cost effectiveness in providing more autonomy. Experts must also review HMI interfaces to facilitate more efficient processes to reduce the existing workload. Interface designs must also factor in the skills and experience of operators, to enable each the ability to customize the desired LOA to enable them to adjust while learning or until they develop an acceptable level of trust.

[13] See `http://www.robotgear.com.au/Product.aspx/Details/519`
[14] Where attitude is the platforms ability to autonomously maintain flight stability.

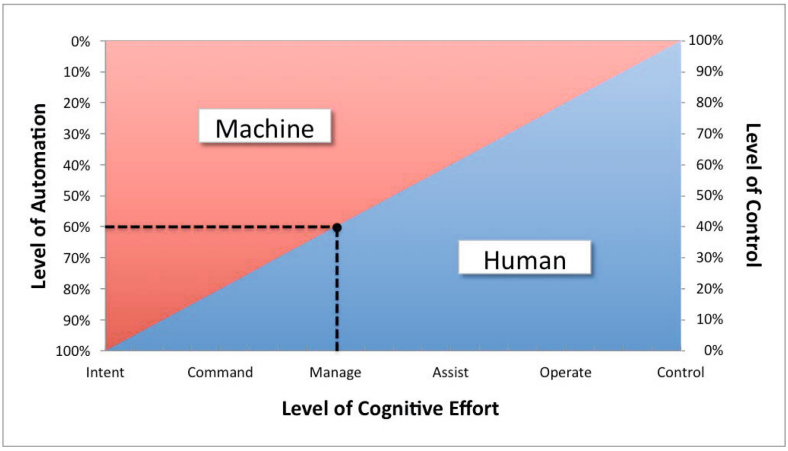


Fig. 2. Level of Effort Vs Human Control

4 Automation

The term automation describes techniques associated with operating machines with reduced or no human intervention. Engineers have increasingly invented machines that fuse a series of sensors and mechanical actuators to seamlessly implement the processes or functionality previously performed manually by humans. This enables the human to operate machines and achieve tasks more efficiently with less effort and vigilance. Using computers and AIPs, the rules of logic associated with basic operator decisions are being integrated into machines to create automated production lines. Early attempts used purely mechanical systems that were operated electronically, however more sophisticated machines are transforming controllers into operators. Modern engineers view automation as the *science of integration* and today's systems rely heavily on electronic sub-systems to mechanise multiple complex human processes. UAV design engineers are currently focused on improving the LOA to enable operates to achieve more with less effort. Research indicates that the operational effectiveness of several levels of automation has already been tested by observing GCS operators during a series of predefined scenarios [5]. During each scenario, the operator was required to control a fixed number of platforms (one or four) while a number of additional routing tasks were progressively exercised with a low, medium and high LOA. The data collected revealed the operator commonly re-worked the computer recommendations. These delay(s) often affect their ability to effectively maintain vigilance on the whole scenario and these additional demands increase the risk of failure.

Designers can also provide interfaces that provide *clumsy automation*. These interfaces force operators to re-evaluate automated recommendations to derive information they have been trained to observe[15]. This issue could ultimately result from human control issues relating to 'management-by-consent' and 'management by-exception' [4].

[15] Trust may not be the only issue. Miller believes there is a natural resistance to automation, which is often manifested by humans backing their own knowledge or judgement [13].

This may stem from the existing perception that humans control machines, but manage people. The machine is currently accepted as a tool that can be used to help humans become more efficient and technology has enabled industry to progressively automate the dirty, dull and dangerous activities. Examples include sorting, assembling and even highly skilled functions such as welding. This evolution has progressed from the need to physically control machines, through to completely automating complex processes with a view of achieving more cognitively derived concepts.

Interested parties must cooperate in order to provide a series of trusted tools that can be scripted or concatenated dynamically as the operator gains confidence in each subsystem. Each component needs a simple interface, with a common underlying description language to facilitate interoperability and automation. Existing tools need to be revised to include links for the autonomous functionality in order to compensate for the bottlenecks being introduced through human inaction or delays in processing unknown facts, rules and/or intent. Until these tools are delivered, scientific experts must continue to investigate how to seamlessly transition knowledge interpretatively across both domains so that people can interact with machines more efficiently. Engineers must also provide more innovative solutions, by embedding more cognitive capacity into machines. Confidence in this knowledge can only be derived through validation, experimentation and simulation.

Using agent-based systems [14], designers can provide scripted capabilities with a personalized model approach to control. For instance the operator should be able to progressively develop his/her trust while experiencing an increasing number of simple tasks by simply choosing to escalate the level of automation as desired (or vice versa). For instance using 'state' engines, the agent can recognize a customized level of autonomy for component groups of functionality. This concept was described in Figure 2. The approach may be as simple as an operator adopting a role and then selecting the level of desired assistance. The operator may also choose to change this setting based on the tempo of the scenario or perceived threat of a particular entity being tracked. For instance the operator conducting ISR activities might choose a variety of LOA setting as either 'control', 'operate' or 'assist'.

These categories currently represent the levels of control based on cognitively equivalent concepts applied to managing people. Given that many control systems focus on individual processes, abstracting these mechanisms for the whole system will be challenging. It would represent a evolution in control systems and enable humans to impose their will as intent or corrective feedback. Unfortunately every individual has an undefined *locus of control*. Their level of tolerance is dependent on a number of cognitive and emotional factors. Those factors translate directly into the type and amount of feedback provided. Weiner believes this voluntary response changes with respect to an individuals emotions, learning and motivation [15]. Training and familiarisation are used to equip individuals with the processes required to successfully conduct predefined tasks[16]. Practice increases the operators proficiency and provides muscle memory to aid more instinctive responses to unexpected situations.

[16] Training also introduce human bias towards automated system, because it challenges the prescribed order of existing work flows.

Endsley discusses a number of factors that benefit the LOA [16]. This level of control is often described in terms of cognitive function associated with effort. A typical example is manually controlling tools. As shown, this concept extends from the physical through to the human belief or intent. The scale used often varies, depending on the community and the environment in which these tools are being used. Sheridan originally introduced a scale to describe automation with ten discrete processes relating to marine vehicles [17]. The ALFUS framework has been recently associated with UAV research [10]. Calhoun also discusses an number of factors relating to the effectiveness of increased LOA [1]. The problems associated with automation require non-trivial solutions. Even this combination of issues clearly indicate that more effort is required in order to provide usable and instinctive forms of automation for operator control systems. It is also clear that automation can be used within autonomous systems, but this alone will not provide the self-governance required to achieve the sentience or independence currently desired.

5 Autonomy

In order to appreciate the meaning of any terminology you should examine its origins. The term stems from the Greek word *autonomia* and describes the view of something being autonomous or 'having its own laws' [18]. Therefore 'Autonomy' has historically been associated with a country and the sovereignty of its people. At present it is being extended to describe automation within a specific context. There may be a need to debate the meaning, however there is no doubt, that *choice* and *decision-making* will feature in the primary arguments. Ultimately humans have a desire to do more with less effort, although they instinctively want to maintaining control. Initially *tools* were employed to increase production and efficiency. Now they are being fielded to reduce the risk of harm to human operators. Today UAVs are being recognised as effective force multipliers and are being adaptive to conduct more innovative roles. Engineers often recognise an unmanned system as a *Complex – System of Systems (CSoS)* delivering a prescribed service. Unfortunately, when you adapt any system, unintended responses or emergent behaviour can manifest. Similarly, the goal of achieving *autonomy* through *independent thought* or *actions* will inevitably require methodical nurturing in order to deliver a *self-governing* entity with (human-like) responses. Industry is currently focused on delivery of the platform, although Defense employs each as a *capability*. In this context an automated system could be defined as one that is controlled and operated by others or by outside forces, while an autonomous system should be independent, in mind or judgement. This suggests that autonomous systems are self-directed and contain mechanisms to achieve governance without undue compulsion or restraint. Unfortunately the community frequently uses both words synonymously and this creates a substantial level of confusion. Similarly, the *utility* of autonomous systems also clouds the definition[17].

In the aviation domain, engineers have been designing automated systems that help operators to achieve better outcomes. For instance, there are smaller flight crews on

[17] This argument is the same as that which distinguishes a tool and a machine, where a machine can be used as a tool.

modern aircraft with integrated systems for navigation, automated take-off and landing. Pilots can be removed from the platform and successfully operate these remotely. These systems operate on a stimuli-response paradigm with event-based triggers. They continually transition between an analogue (human) and digital (machine) representations, using both sequential and parallel processes. Both these concepts need to be considered and integrated to existing automated systems to support the new breed of UAV operators. This will revolutionise the way systems are being employed. Unfortunately there is a cost associated with existing capabilities because the current solution is to add people to augment any gaps. Given that real autonomous systems should use mission level intent commands, designers need to entertain a paradigm shift. There is obviously still a need for increased automation to deliver better interfaces; however, the move to real platform autonomy may only be delivered by providing on-board intelligence. This capability would provide machines with sufficient human-like abilities that enable them to: perceive, recognise, contextualise, rationalise, reason and make sound decisions.

At present there is a need to frame autonomy as a supervised capability that provides benefits as a result of humans coordinating and collaborating with machines. It is clear that existing systems are not trusted, integrated or efficient [19]. The platform is a physical entity, where autonomy is predominantly virtual and achieved through software. This virtual functionality needs to be considered independently during the design and decision process. Figure 3 displays a new framework proposed by the American Defense Science Board (DSB). This model encompasses three views that span the hierarchy of control are also discussed in Section 3. The reach begins with the campaign level, moves down to the echelon and then the pilot or operator. The model catalogues a number of primary functions that could be achieved within each level. This provides a number of challenges that are associated with: Interoperability, Autonomy, Airspace Integration, Communications, Training, Propulsion (Energy) and cooperative Manned–Unmanned Teaming (MUT) [20]. Researchers must begin by defining the *levels' of intent* and aligning the *corpus of research* to concentrate on achieving incremental gains.

Autonomy is a key element in the puzzle required to solve system automation [21]. It cannot be considered in isolation and should be considered to be more than a set of delegated requests aimed at completing a task. As stated, "the utility of an autonomous capability is a function of the *ecology* of the specific mission needs, the *operating environment*, the *user* and the *vehicle*" **within** the **context** of the planned activity [2]. The technology may be available, however unless factored into the larger system, design flaws will continue to swell manpower figures in order to address the added complexity. Hence there is a need to focus on developing perceptual processing, planning, learning, human-machine-robotic interfaces and multi-agent coordination.

Each concept provides non-trivial technical challenges that requires a diversity of Subject Matter Experts (SMEs), however an integrated approach requires fresh holistic planning with a coordinated suite of guided innovative research thrusts. The key challenge in moving forward and increasing the effectiveness of automation within autonomous systems, is to focus on improving on-board capabilities. They need to acquire, process and transform data into knowledge for decision making. On-board cognitive abilities will promote the platforms ability to autonomously complete decomposable goals, without the need for a human companion or supervisor. Research thrusts

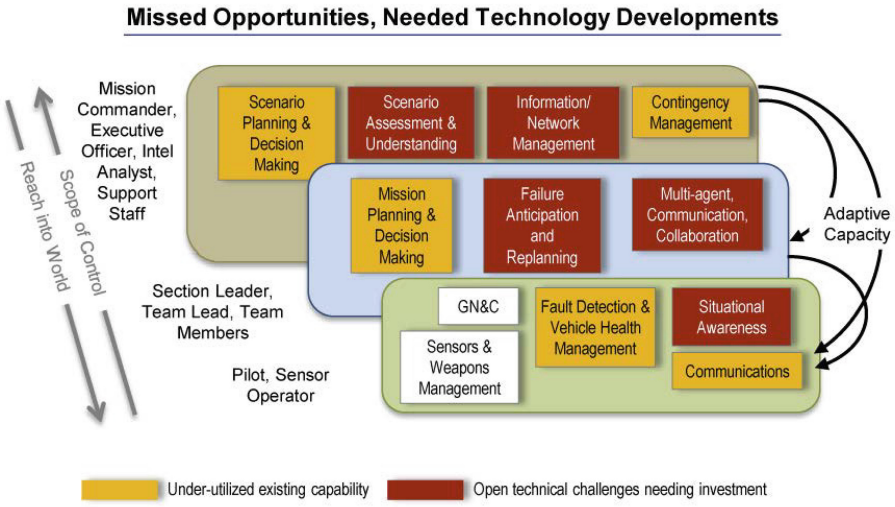


Fig. 3. Remaining Challenges and Prospective Research Program

could include improving trust, integration and autonomy while embedding a capacity to self heal, self organise and generate adaptive or cooperative responses within a dynamic environment.

6 Cognitive Architectures

Research into machines and there level of perceived intelligence has evolved as a scientific domain of study. The originally concept focused on a means of embodying the human conscience within a computer. Over time this concept has led to the development of systems capable of achieving logical decision making on-board. This field originally relied on statistics, heuristics and eventually employed hybridised AI techniques. Each of these domains began optimistically and eventually bogged down, because researchers were not able to understand or implement certain human-like functions (such as recognition and cognition). A number of researchers have embedded a series of constrained solutions to provide intelligent functionality within complex machines. Examples include the original computerised Chess Champion on *Deep Blue* [22] and more recently Neural Spiking Exchange experiments on *Blue Gene/P* [23]. Current estimates indicated that science still needs a microprocessor with at least 25 GHz of processing power to emulate the whole human brain. By realising this form of virtual cognitive processing, it will provide a significant step towards achieving the goals that inspired AI and that of achieving autonomy. Unfortunately this estimate is based on the Harvard style or transactional approach within the Von Neumann architecture. This architecture fails to represent a structure that supports the natural human thought process. This relies on a series of parallel symbolic associations that require in-line processes for pattern recognition and to be effective, both are emulated using parallel processing [24]. There

is also the need to create a mature family of techniques to achieve reliable knowledge-based engineering. Examples of attempts to evolve AI techniques to facilitate intelligent applications include: creating human-like applications [25] and using agent-oriented paradigms [14]. Preceding paradigms include: Blackboard systems, Rule-based systems [26], Case-based reasoning systems [27], Model-based reasoning systems [28], Bayesian networks [29,30], Artificial Neural Network (ANN) [31], Evolutionary computing [32,33], Fuzzy logic systems [34], Knowledge-based systems [35,36], Hybrid systems [25,37] and even Beliefs, Desires, Intentions (BDI) based agent-oriented systems [38]. No single technique has dominated in the General Problem Solving (GPS) domain, hence the need for a multi-disciplined approach still exists.

Cognitive knowledge can be stored and tested against the dynamic nature of environmental variables. This information can be moderated against rules in order to derive the best course of action available to the system. At present machines can solve problems or achieve human-like functionality; however they are not intelligent, they are merely making smart decisions. Transforming the cognitive processes and storage architecture in machines is a non-trivial problem and has been the focus of research for many years. Existing cognitive architectures have been polluted to solve specific problems using dated hardware. It is time to take stock of these changes and reflect on how this functionality could be achieved using new systems and modern technology. A paradigm shift is required to transform our understanding or existing automation techniques (that are *controlled* by humans) in order to achieve the concept of a truly autonomous system that can be *managed* by people using traditional coordinated and cooperation skills.

Dutch et al. presented a simplified taxonomy that can be used as a cognitive architecture in machines [39]. It was based on two emergent streams of research that promote virtual cognition. These include *phylogenetic* and *ontogenetic* systems. They rely on a hybridised symbolic approach as shown in Figure 4. This research should stimulate others to achieve meaningful mental models that support a virtual human mind. In order to promote this form of embedded *human-like* intelligence, there is a need to focus on both low-level perceptual elements of knowledge (microscopic view) and the higher-level social or cultural aspects (macroscopic view) of the environment. For instance, De Vee previously developed a mathematical model to describe brain-like processing [40], while Sendhoff extended this notion in 2000 [41]. More recently Sporns formalized an intuitive approach of exchanging environmental information with machines using complex networks of intelligent agent systems [42]. Agents enable designs to create an architecture that abstracts system complexity to enable the focus to remain on problem solving [43]. It is believed that cognitive processing will eventually be achieved using a combination of MI techniques hosted in a agent environment. The modular nature of some designs has already been incorporated into agency frameworks, however it may take another 25 years to champion the ultimate solution (a analysis of existing designs is the focus of a follow-on paper). To be successful in a complex and dynamic environment, the resulting structure must facilitate interoperability and be scalable. It should also facilitate the concept of memory (short-term and long-term) which would be represented as state, while the rules and functionality can be processed using dynamically linked agent capabilities. Even with these designs, it is still difficult to inject knowledge into machines. Knowledge must be gleaned from the environment using an iterative

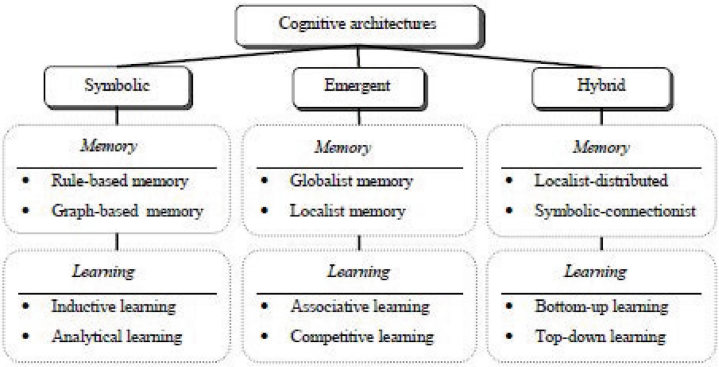


Fig. 4. Simplified Taxonomy for Cognitive Architecture [39]

approach within any given context. Once collected it must be verified and validated within the system being supported. Attempts are being made to capture this information from humans in real-time while conducting controlled activities. In 2009, Deco and Rolls linked a series of microscopic neural network models configured to study macroscopic effects within a test subject [44]. They observed the spiking activity within each network and used Weber's psycho-physical law (stimuli and response) to minimise noise [45]. Unfortunately there are many complex structural and functional networks within the brain. Hence, refining this process will take time, demand extremely resource intensive testing and makes little sense unless researchers create an appropriate machine oriented model capable of embedding the results. Experiments of this nature are also occurring at the University of South Australia.

7 Test Methodology

As suggested, a Multi-Agent System (MAS) framework has been used to create a UAV control system. Details of this model were published in two previous papers [46,11]. Both describe a MAS that is composed of an interconnected series of dynamically linked components that can be used to control one or more independent capabilities simultaneously. Each agent capability responds in accordance with the desired *mode* of autonomy selected by its user. These components include, but not limited to: control, navigation, planning, re-planning and logistic autonomic functionality. Although each component is modelled independently, a MAS can instantiate as many capabilities as required. In this case a MAS was developed to control a single UAV, however the test scenario shown in Figure 5 is designed to exploit a variety of applications in order to focus on the control of one or more platform types and activities.

The scenario depicted in Figure 5 has been formulated to initially test one controller and a single platform during a controlled series of experiments. The scenario displays a simple reconnaissance mission that requires a UAV to visually record activities at a number of given coordinates. The start and finish points are shown at the bottom left,

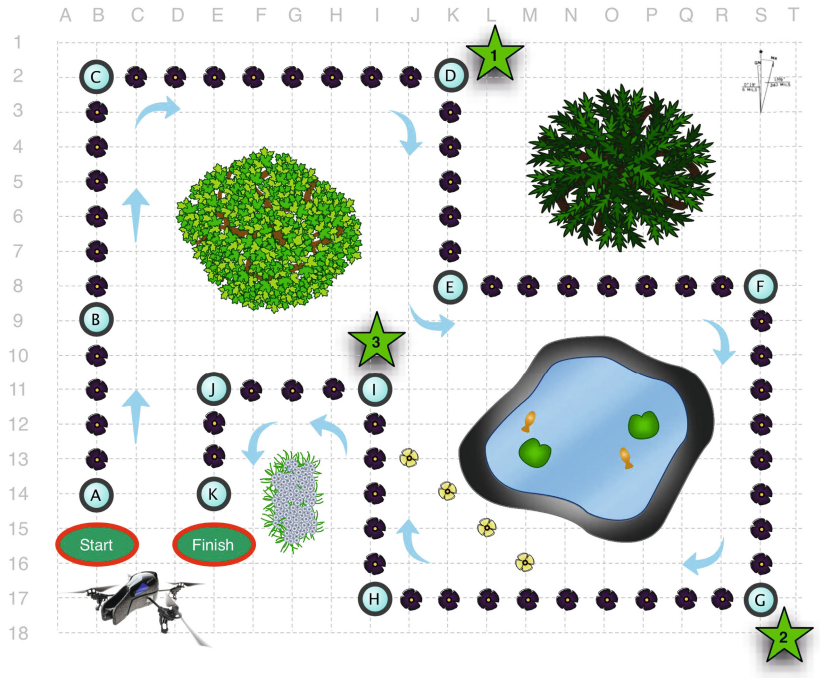


Fig. 5. Scenario Intent is to Successfully Complete the Mission

with three Way-points (WPs), identified as '1', '2' and '3'. The grid reference can be calculated from the underlying scale (with north facing up the page). For examples, the start point is located at 'B14' and WP '1' at 'K2'. The journey to each WP has been preplanned using a number of Check Points (CPs) as shown. For instance the route to WP '1' starts at CP 'A' and finishes at CP 'D' (This is were the activity for WP '1' occurs). A number of staging points have been identified to aid navigation at pre-set geographical distances. The system will be used to implement the *seeking* functionality within the navigation component of the MAS in order to traverse each *leg* of its journey. Movement of the platform will physically rely on inertial data between legs, however new control commands will be provided either by the operator or the autonomous system being developed. As the level of automation progressively increases, new processes may be recommended and eventually scripted. At present the LOA will be determined by the operator by selecting the mode of individual components via the HMI, however in the future, on-board cognitive processes could be used to reduce or even eliminate this requirement. An example of initial research would test the user's ability to progressively transition the LOA between 'recommend', 'assist' and even 'automate' modes. Each mode is supported by an underlying dynamically linked capability, while the processes and scenario flow are linked using a scripted process and task supervisor (agent). For instance in one experiment, a controlled scenario using series of operator capabilities could issue control commands in order to complete a task. Alternate experiments

could be used to obtain measures from human operators or their scripted equivalents. Given this base-line data, a series of agent techniques could be evaluated in subsequent testing. Automated capability could then be developed, customized and inserted as required. Future research involves plans to implement changes using *reflection* or *introspection* to automate the linking functionality of the framework. For instance each capability added could automatically be registered and dynamically appear as an option to the operator during run-time. This effort will be published as the project evolves.

8 Conclusions

The field of autonomy is complex and is often confused with the ability to achieve automated functionality. As discussed during this brief review, both domains are distinct and it is clear that more analysis is required to determine a clear course of action. It is proposed that a refined decision-making architecture be developed. It should take advantage of modern technologies and new AI techniques to enhance the level of MI over the next 18 months. In summary, it is important to acknowledge that *Automation* focuses on *automating* specific processes for humans, where *autonomy* is about the machine achieving *self-governance* through *independent* decision-making. This eventually means there is a need to embed some form of MI within the system. The proposed MAS focuses on the concept of managing an interconnected network of intelligent components that supports experiments aimed at progressively achieving improved LOA. Given a series of optimised components, these experiments would focus on cognitive architectures and eventually on-board cognitive processing. The scenario decentralizes the complexity of testing the control systems and enables a variety of discrete functionalities to be invoked using a distributed or service-based context. It will enable the team to study how the hierarchies evolves and the interconnected series of links that govern the system. A set of operating instructions based on rules or procedures can be developed to study the coordination and cooperation of this dynamically evolving organization. Before determining the value, type and effectiveness of this level of personified organization, research is required to coordinate a collaborative response using a team of experts, operating across multiple domains. Hence, future research is required to provide a coherent and efficient means to seamlessly integrate the context and role dependent functionality, traditionally conducted by one or more humans. In the short term, people are still required to remain in-the-loop until MI is able to perceive the environment, perform cognitive recognition, or rationalize within a given context. Until then, research into autonomy must evolve. There is more to be done and more effort is required to examine a broader range of contributions. Without this knowledge it is difficult to analyse a definitive direction for this research. For instance, a review of cognitive architectures is overdue and it is believed that researchers should consolidate the lessons learned to assist engineers in providing more human-like decision-making on-board UAVs.

References

1. Calhoun, G., Ruff, H., Draper, M., Wright, N., Mullins, B.: Levels of automation in multi-UAV control allocation and router tasks. In: AIAA Infotech@Aerospace Conference, Seattle, Washington, April 6-9, pp. 1–13. American Association for Artificial Intelligence (2009)

2. Murphy, R., Shields, J.: Defense science board - task force report – the role of autonomy in DOD systems. Technical Report CD 1172, Office of the Under Secretary of Defense for Acquisition, Technology and Logisitics, Washington, DC (July 2012)
3. Weigel, J., Jahner, T.: JSF interoperability initial capabilities and beyond - Presentation, pp. 1–25. Lockheed Martin Aeronautics Company, Fort Worth (2009)
4. Taylor, R.M.: Human automation integration for supervisory control of UAVs. In: Virtual Media for Military Applications. Number Proceedings of the RTO-MP-HFM-136, pp. 1–10. RTO, Neuilly-sur-Seine, France, Defence Science Technology Laboratory, Farnborough, UK, Her Majesty's Stationary Office (2006)
5. Calhoun, G., Ruff, H., Draper, M., Wright, N., Mullins, B.: Effects of levels of automation in unmanned aerial vehicle routing task. In: Procedings of the Human Factors and Ergonomics Society Annual Meeting. HFES, vol. 53, pp. 197–201. Sage Publications, Thousand Oaks (2009)
6. Ruff, H.A., Calhoun, G.L., Draper, M.H., Fontejon, J.V., Guilfoos, B.J.: Exploring automation issues in supervisory control of multiple UAVs. In: 2nd Human Performance, Situation Awareness, and Automation Technology Conference, Daytona Beach, FL, pp. 218–222 (2004)
7. Endsly, M.: Automation and situation awareness. In: Parasuraman, R., Mouloua, M. (eds.) Automation and Human Performance: Theory and Applications, pp. 163–181. Lawrence Erlbaum Associates, Mahwah (1966)
8. Klien, G.A., Calderwood, R., McGregor, D.: Critical decision method of eliciting knowledge. IEEE Transactions on Systems, Man and Cybernetics 19, 462–472 (1989)
9. Endsley, M.: Design and evaluation for situation awareness enhancement. In: Human Factors Society 32nd Annual Meeting, Santa Monica, CA, pp. 97–101 (1988)
10. Huang, H.M., Pavek, K., Albus, J., Messian, E.: Autonomy levels for unmanned systems (ALFUS) framework: An update. In: SPIE Defense and Security, Orlando, FL, USA, vol. 5804, pp. 439–448. The International Society for Optical Engineering (2005)
11. Tweedale, J.W.: Using mutli-agent systems to improve the level of autonomy for operators controlling unmanned vehicles. In: Graňa, M., Toro, C., Posada, J., Howlett, R.J., Jain, L.C. (eds.) Advances in Knowledge-Based and Intelligent Information and Engineering Systems. Frontiers in Artificial Intelligence and Applications, vol. 243, pp. 1666–1675. IOS Press, Amsterdam (2012)
12. Parasuraman, R., Sheridan, T., Wickens, C.: A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 30(3), 286–297 (2000)
13. Miller, C.A., Parasuraman, R.: Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. Human Factors 49(1), 57–75 (2007)
14. Tweedale, J., Jain, L.C.: Embedded Automation in Human-Agent Environment. Adaptation, Learning, and Optimization, vol. 10. Springer, Heidelberg (2011)
15. Weiner, B.: An Attributional Theory of Motivation and Emotion. Springer, Berlin (1986)
16. Endsley, M.R., Kaber, D.B.: Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42(3), 462–492 (1999)
17. Sheridan, T.B., Verplank, W.L.: Human and computer control of undersea teleoperators (man-machine systems laboratory report. Technical Report N00014-77-C-0256. MIT, Cambridge, MA, ONR, Arlington, Virginia (1978)
18. Simpson, J. (ed.): Oxford English Dictionary. Oxford University Press, New York (2012)
19. Murphy, R., Woods, D.: Beyond asimov: The three laws of responsible robotics. IEEE Intelligent Systems 24(4), 14–20 (2009)

20. McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., Weld, D., Wilkins, D.: PDDL - The Planning Domain Definition Language. Technical report, CVC TR-98-003/DCS TR-1165, Yale Center for Computational Vision and Control, New Haven, CT (1998)
21. Rogers, E.M.: Diffusion of Innovations, 3rd edn. MacMillan, New York (1983)
22. Hsu, F.: IBM's Deep Blue chess grandmaster chips. IEEE Micro, 70–81 (March 1999)
23. Hines, M., Kumar, S., Schürmann, F.: Comparison of neuronal spike exchange methods on a blue gene/p supercomputer. Frontiers in Computational Neuroscience 5(49) (2011)
24. Anderson, H.C.: Why artificial intelligence isn't (yet). AI Expert Magazine, 1–9 (July 1987)
25. Jain, L., Jain, R. (eds.): Hybrid Intelligent Engineering Systems. World Scientific Publishing Company, Singapore (1997)
26. Urlings, P.J.M., Spijkervet, A.L.: Expert systems for decision support in military aircraft. In: Murthy, T.K.S., Münch, R.E. (eds.) Computational Mechanics, pp. 153–173. Springer, Heidelberg (1987)
27. Maher, M.L., Garza, A.G.: Case-based reasoning in design. In: Brown, D.C., Birmingham, W.P. (eds.) AI in Design, vol. 12(2), pp. 34–41 (April 1997)
28. Simmons, R.: Generate, test and debug: A paradigm for combining associational and causal reasoning. In: Krivine, D., Simmons, R. (eds.) Second Generation Expert Systems, pp. 79–92. Springer, Berlin (1993)
29. Pear, J.: Probabilistic reasoning in intelligence systems. Morgan Kaufmann, San Mateo (1988)
30. Hall, D.L.: Mathematical techniques in multi-sensor data fusion. Artech House, Norwood (1992)
31. Haykin, S.: Neural networks: A comprehensive foundation. Macmillan (1994)
32. Rooij, A.V., Jain, L.C., Johnson, R.P.: Neural Network Training Using Genetic Algorithms. World Scientific Publishing Company, Singapore (1996)
33. Vonk, E., Jain, L.C., Johnson, R.P.: Automatic Generation of Neural Networks Architecture Using Evolutionary Computing. World Scientific Publishing Company, Singapore (1997)
34. Sato, M., Sato, Y., Jain, L.C.: Fuzzy Clustering Models and Applications. Springer, Germany (1997)
35. Buchanan, B.: New research on expert systems. In: Hayes, J., Michied, D. (eds.) Machine Intelligence, vol. 10, pp. 269–299. Wiley, London (1982)
36. Ryan, P., Zalcman, L.: The DIS vs HLA Debate: What's in it for Australia? In: SimTect 2003. Simulation Industry Associatioin of Australia, pp. 1–6 (2003)
37. Urlings, P.: Teaming Human and Machine: A conceptual framework for automation from an aeronautical perspective. PhD thesis, University of South Australia, School of Electrical and Information Engineering (2004)
38. d'Inverno, M., Luck, M., Georgeff, M., Kinny, D., Wooldridge, M.: The dMARS architecture: A specification of the distributed multi-agent reasoning system, vol. 9, pp. 5–53. Kluwer Academic Publishers, Netherlands (2000)
39. Duch, W., Oentaryo, R.J., Pasquier, M.: Cognitive architectures: Where do we go from here? In: Wang, P., Goertzel, B., Franklin, S. (eds.) AGI, vol. 171, pp. 122–136. IOS Press, Amsterdam (2008)
40. De Vree, J.K.: A note on information, order, stability and adaptability. Biosystems 38(2/3), 221–227 (1996)
41. Sendhoff, B., Pötter, C., von Seelen, W.: The role of information in simulated evolution. In: Proceedings of the International Conference on Complex Systems on Unifying Themes in Complex Systems, pp. 453–472. Perseus Books, Cambridge (2000)
42. Sporns, O.: From complex networks to intelligent systems. In: Sendhoff, B., Körner, E., Sporns, O., Ritter, H., Doya, K. (eds.) Creating Brain-Like Intelligence. LNCS (LNAI), vol. 5436, pp. 15–30. Springer, Heidelberg (2009)

43. Tweedale, J., Ichalkaranje, N., Sioutis, C., Jarvis, B., Consoli, A., Phillips-Wren, G.E.: Innovations in multi-agent systems. J. Network and Computer Applications 30(3), 1089–1115 (2007)
44. Deco, G., Rolls, E.T., Romo, R.: Stochastic dynamics as a principle of brain function. Progress in Neurobiology 88(1), 1–16 (2009)
45. Kang, J., Wu, J., Smerieri, A., Feng, J.: Webers law implies neural discharge more regular than a poisson process. European Journal of Neuroscience 31, 1006–1018 (2010)
46. Tweedale, J.W.: Fuzzy control loop in an autonomous landing system for unmanned air vehicles. In: WCCI 2012: 2012 IEEE World Congress on Computational Intelligence (2012)

Experts' Agreement Support for Distributed Engineering Knowledge Modelling

Ricardo Mejía-Gutiérrez, Alejandro Cálad-Álvarez, and Daniel Zuluaga-Holguín

Design Engineering Research Group (GRID). Universidad EAFIT
{rmejiag,acaladal,dzulua19}@eafit.edu.co

Abstract. Knowledge sharing among partners in collaborative and concurrent engineering processes is becoming more critical for decision-making in Product Design Engineering (PDE). The approach presented in this article proposes a Fuzzy Logic (FL) based semi-automatic support for reaching agreements in the definition of variable's domain under a Constraint Satisfaction Problem (CSP) modelling process. A negotiation model is proposed, including suggestions about preferred domains for a variable that it is used for more than one expert. Competence of the users involved in the negotiation is measured with a set of developed indexes. A Fuzzy Inference System (FIS) is used to calculate user's indexes and suggestions about the domain are made based on the proposed domains and competence index of the experts. The proposal is included in a Multi-Agent System (MAS) for supporting distributed knowledge modelling.

Keywords: Collaborative Engineering, Constraint Satisfaction Problem CSP, Variable's Domain, Fuzzy Logic, FIS, MAS.

1 Introduction

At present it is common for people to become involved in complex design projects that are geographically located in different places around the world. This aspect is making the Product Design Engineering (PDE) process a collaborative interaction among partners from a distributed engineering team [7], which work in a parallel and concurrent way, facilitating to take into account different points of view since the beginning to the end of the product life cycle. Under this environment, distributed teams have to work together to find the best solution for a specific PDE problem. This collaboration includes, among others, the action of making agreements between different experts with different backgrounds on many design topics. In order to guaranty the organization of the experts' work it is necessary to develop tools and techniques to assist in provide and coordinate the entire infrastructure. This infrastructure must support them in decision making during the collaborative, concurrent work process. and to improve the design in all of its stages since requirement definition to detailed design and production [2].

To support Collaborative Product Design (CPD) some approaches have been developed [5, 8]. This paper focuses on Constraint Satisfaction Problem (CSP). From engineering design theory, an "Engineering Problem" can be mathematically modelled using a CSP tuple denoted $P(\mathcal{X},\mathcal{D},\mathcal{C})$. This is composed of three different sets:

$\mathcal{X}$ A finite set of n variables $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$.

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 209–223, 2013.

$\mathcal{D}$ A finite set composed by n ranges of possible values that elements from the $\mathcal{X}$ set can take (called "domains" in CSP theory) and denoted $\mathcal{D} = \{D_1, D_2, \ldots, D_n\}$, where $x_1 \in D_1, x_2 \in D_2, \ldots, x_n \in D_n$.

$\mathcal{C}$ A set of p constraints $\mathcal{C} = \{C_1, C_2, \ldots, C_p\}$, that determines relationships among variables and the values that they can take simultaneously through a mathematical function [12].

This Knowledge Modelling (KM) approach and the resulting solutions for PDE are shown in Figure 1.

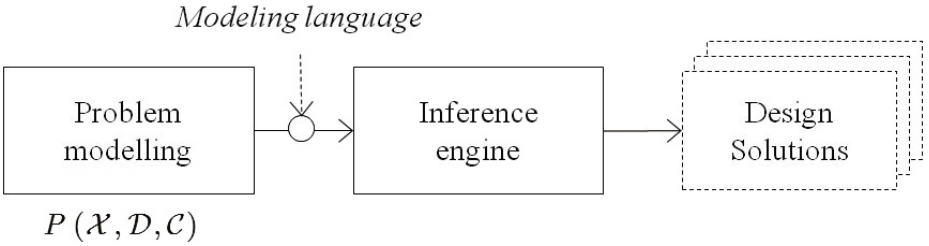


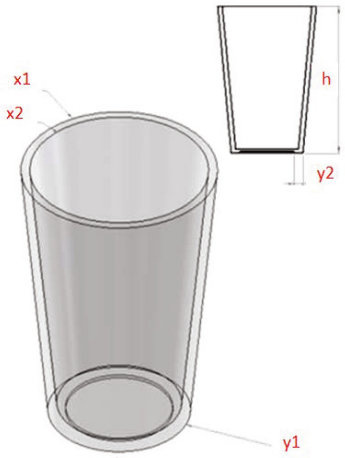
Fig. 1. Knowledge Modelling and solution approach

During this chapter the problem about the choosing of the domain for a variable defined by more than one user in a distributed engineering environment is treated. The approach uses fuzzy logic, specifically a Fuzzy Inference System (FIS) combined whit a Multi-Agent System (MAS) to define the experience of each user and weigh its desire about the domain that the variable must have. In section 2 the previous work is presented, contextualizing the process of KM in PDE, how to model it using CSP and the communication with a MAS. Section 3 shows the proposed representation for the users' preferences. Sections 4 and 5 presents the fuzzy processing to determine the weigh of each expert in the final decision. In Sections 6 and 7 tests before implementation and the actual implementation are presented. Finally conclusions and further research are presented in Section 8.

2 Background

Figure 2 shows a typical example to graphically contextualize the process of creating a KM in PDE. This models enable designers to evaluate different product configurations through the assignation of different values to design variables. Those values belong to the variable's domain and that is why shared variables among different Subject Matter Expert (SME) in collaborative design modelling processes are critical during a distributed domain definition.

Having individuals that are expected to have a greater than normal expertise in a particular discipline (SME) [11] developing a CSP model collaboratively, can cause some problems (e.g. data redundancy). It may occur that two or more users define different variables to represent the same design aspect. This specific problem should be treated on




```
#### VARIABLES #########

\vi : h  8...22 ;   #### height
\vi : x1 8...11 ; ### Upper external diameter
\vi : x2 7...10; ### Upper internal diameter
\vi : y1 8...11 ; ### Lower external diameter
\vi : y2 1...2 ; ### Low border thickness
\vr : p 0.1 [1,1.4] ; ### Scale factor


#### CONSTRAINS ######

\ci : P1 , x1 = p*y1 ;
\ci : P2 , h = 2*y1 ;
\ci : P3 , y2 = 1 ;
\ci : P4 , x2 = x1-1 ;
```

Fig. 2. Example of a Product Knowledge Model

time because it generates inconsistent results after running the model in the inference engine. An approach to solve this problem was presented by Meja-Gutierrez et. al [10]. A variable is then enriched with more information, being constituted by a set of attributes that define its characteristics (`ID`, `Name`, `Symbol`, `Type`, `Data Type`, `Expert`, `Unit Prefix`, `Lifecycle stage`, `Discipline`, `Step`, `Domain`, `Information Source` and `Measure parameter`). In this proposal, experts in an MAS have the possibility to become users of an already declared variable instead of creating a new one. Due to this, every variable in the system has an "owner" (creator) and could have one or more "users". In order to give experts the option to adopt a variable, some aspects of the already defined variable can be modified (e.g. values for `Lifecycle stage`, `Unit Prefix` and `Discipline`) in order to adapt it to the new user. Changing these aspects does not affect the core of the variable and give additional information.

Once a variable x_n is defined into the system, the process continues with the definition of its respective domain $D_n = D(x_n)$. Some CSP approaches, such as Fuzzy CSP (FCSP), enables to assign preferences to constraints and particularly the software CON'FLEX allow preferences assignment $\mu(x_i)$ to variable's Domains, where $\forall x_i \in D_i, \mu(x_i) \in [0,1]$, and they can be entered by α-*cuts*. In the PDE context, the domain assignment in early stages is one of the most critical problems. It is also the most uncertain phase and obtaining precise information upon which basing decisions is usually impossible [1]. The fuzzy set concept was introduced by Zadeth [14] as a classs of object with a continuum grade of membership between zero and one. Since that definition, some research in fuzzy logic and fuzzy sets principles have been successfully implemented to support the process of decision making in PDE and to deal with automated negotiation and its inherent imprecision. [1, 4]. However, the complexity increases with distributed design teams. In a non-distributed scenario an SME can define the variable's domain without conflicts. Under a collaborative environment experts can still create variables, but it may happen that one SME who is going to used

an existing variable, differs from some properties of the current variable. This happen due to he may have certain knowledge that may differ from the one set by the variable creator. The proposed MAS allows experts to suggest a new domain for a variable in the system, but also trying to achieve negotiated agreements. The aim of this article is to present a supporting method that helps variable owners and users for domain trade-off achievement by calculating a resulting domain that fits as much as possible both experts' preferences.

3 Preference Indexes as Fuzzy Inputs

Usually domain definition is a negotiation process between experts involved (variable owner and a new user). It requires multiple interactions to obtain a domain that fits everyone's needs, including boundaries and preferences of those values. The aim is to replace this negotiation by a semi-automatic process, where preferences (represented by Fuzzy Sets) play a key role. In order to achieve this, experts are asked about domain values and preferences with standard terms like: The variable is `equal /different/greater than/less than` a certain value . Therefore, the problem is treated by taking the linguistic variables and assigning grades of membership into fuzzy sets that represent the experts' preferences. The key is to "combine" two different domains by maximizing owner satisfaction (represented by Fuzzy preferences). This helps to reduce the design's solution space, starting from the model definition, in order to avoid high computing time during the model execution in the CSP solver. How much change an expert can accept in his proposed domain, is a topic that is directly related with the pre-negotiation phase and the knowledge problem [3]. It is assumed for this work that this acceptance depends directly on how much knowledge or information (competence) the expert has about the variable. To measure this aspect a Competence Index is proposed. This index will help to give preferences while deciding a collaborative variable's domain. The index is obtained using a FIS based on the Mamdani model [9]. A FIS is a method that interprets the input values and depending on a set of rules calculate the output. The implemented FIS is showed in Fig. 3 and its main structure is composed of three inputs, 27 rules and the competence index as output.

3.1 Crisp Inputs

The crisp inputs for the FIS are three normalized sub-indexes, defined according to variables' properties (described in Section 2), more particularly according to those that different experts can modify from a shared variable: `Lifecycle stage`, `Units` and `Discipline`.

Unit index ($\mathcal{I}_U$)

When an expert is entering/using a variable, the index represents the competence that the expert has according to variables using the same units. For example, when an expert works with morphological variables (longitudinal units, meters), this index is going to be higher than the same index for somebody who is an expert in economic variables (e.g. costs). Its calculation is based on a two components ratio: Possible units variables ($\mathcal{P}_U$), refer to the number of variables in the system

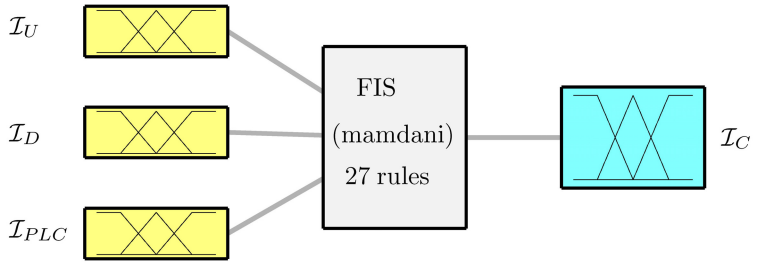


Fig. 3. Implemented FIS

that uses the same units than the Evaluating Variable (EV) and Real units variables ($\mathcal{R}_U$), represent the number of variables in the system with the same units and from which the expert is creator or user.

Discipline index ($\mathcal{I}_D$)
In an equivalent manner as previous index, $\mathcal{I}_D$ is useful to know how much and expert is closed with variables who have the same technical classification (Discipline). Its components are: Possible discipline variables ($\mathcal{P}_D$), is the number of variables in the system from the same discipline than the EV and Real discipline variables ($\mathcal{R}_D$), is the number of variables in the system that are from the same discipline than the EV and of which the expert is creator or user.

PLC index ($\mathcal{I}_{PLC}$)
This represents how much an expert knows about the life-cycle stage to which the EV belongs. Its components are: Possible PLC variables ($\mathcal{P}_{PLC}$), is the number of variables in the system that belongs from the same PLC stage as theEV. Real PLC variables ($\mathcal{R}_{PLC}$), is the number of variables in the system who belong to the same PLC stage than the EV and of which the expert is creator or user.

4 Linguistic Levels for Qualifying Preferences

Three linguistic variables were defined for the different levels that the FIS's inputs (indexes) can take. Each variable has three different levels (fuzzy sets) represented by the names of "Low", "Medium" and "High", representing a membership function $\mu(\mathcal{I}_x)$ as showed in Fig. 4.

Every index $\mathcal{I}_x$ is normalized, where $\mathcal{I}_x \in [0,1]$ and can be described by the fuzzy sets "Low" = $\tilde{L} = \mu_L(\mathcal{I}_x)$, "Medium" = $\tilde{M} = \mu_M(\mathcal{I}_x)$ and "High" = $\tilde{H} = \mu_H(\mathcal{I}_x)$ or any possible combination of these. The fuzzy sets used to represent each linguistic variable correspond to the desired behavior onto the system. For that reason the $\tilde{M}$ set uses a triangular function $f^{\triangle}(\mathcal{I}_x;\{0.2,0.5,0.8\})$ who only has the higher membership degree (1) when the index is exactly 0.5. The functions for $\tilde{L}$ and $\tilde{H}$ are $f^{Z}(\mathcal{I}_x;\{0.2,0.5\})$ and $f^{S}(\mathcal{I}_x;\{0.5,0.8\})$ respectively. The Gauss functions were discarded because the way in

which the function makes the transition between the valley and the leaning parts doesn't affect in a great way the functionality of the model and by using trapezoidal functions the calculations are significantly lower.

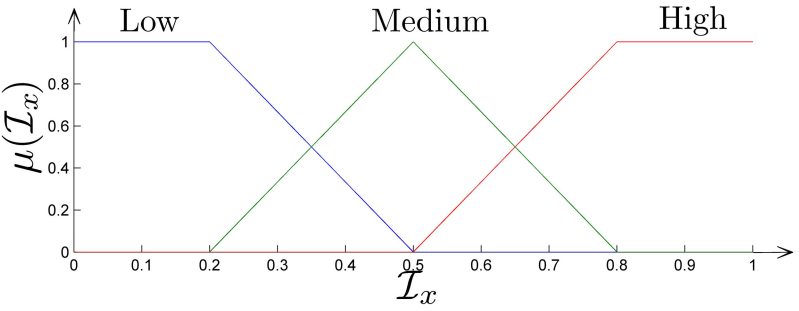


Fig. 4. Fuzzy Sets for FISs Inputs

These three indexes are calculated according in "Eqs. 1–3".

$$\mathcal{I}_U = \frac{\mathcal{R}_U}{\mathcal{P}_U} = \frac{\textit{Real units variables}}{\textit{Possible units variables}} \tag{1}$$

$$\mathcal{I}_D = \frac{\mathcal{R}_D}{\mathcal{P}_D} = \frac{\textit{Real discipline variables}}{\textit{Possible discipline variables}} \tag{2}$$

$$\mathcal{I}_{PLC} = \frac{\mathcal{R}_{PLC}}{\mathcal{P}_{PLC}} = \frac{\textit{Real PLC variables}}{\textit{Possible PLC variables}} \tag{3}$$

5 Fuzzy Preferences Processing

In this subsection an explanation about how the experts' preferences are processed using the FIS is presented. In the subsection 5.1, the fuzzy rules necessary to do the processing and the combination of the possibilities are explained. The defuzzification process and the technique used to obtain the crisp output are shown in the subsection 5.2.

5.1 Fuzzy Rules Base

For the implementation, it was necessary to use 27 fuzzy rules to cover all three inputs (indexes) with all their three possibilities (fuzzy sets) in all their combinations, which means $3^3 = 27$. These fuzzy rules were created based on the IF...THEN structure and evaluated using conjunction ($\wedge$) as logic connector. This operator was selected in order to use the lower index as a reference, aiming to avoid an unnecessary increment in the system response and using only as a relevant rule for the defuzzification phase, the rules in which all values are different from zero.

5.2 Defuzzification and Crisp Output

For the defuzzification stage the Centroid technique was used. The output of the systems is an index between 0 and 1. This index ($\mathcal{I}_C$) represents the competence of the evaluated expert around the EV. The generic response surface $\mathcal{I}_C = f(\mathcal{I}_D, \mathcal{I}_U)$ is showed in Fig. 5, but the plot is equivalent for $\mathcal{I}_C = f(\mathcal{I}_D, \mathcal{I}_{PLC}) = f(\mathcal{I}_U, \mathcal{I}_{PLC})$.

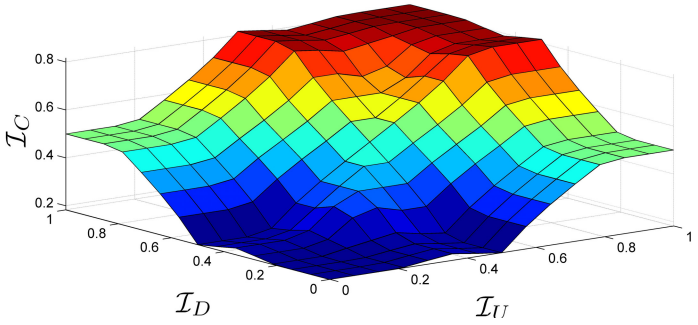


Fig. 5. FIS Response surface

The Competence Index ($\mathcal{I}_C$) is calculated for each expert, every time that he/she creates a variable or becomes user of an existing one. Depending on $\mathcal{I}_C$ levels, the system has the possibility to make decisions according to the relative expertise of experts in the EV's negotiation process. Under these conditions, when having two domains for the same variable (this means two experts have to agree on a domain for the EV) the semi-automatic negotiation model proposes a common domain for the variable. This domain is calculated by applying the t-norm (conjunction) to the two fuzzy domains. In addition to the t-norm, some boundaries can be fixed (according to $\mathcal{I}_C$). These constraints are applied to the final domain before it is presented to the experts. Finally, experts are able to decide if they agree or not with the domain proposal.

6 Functional Test

An experiment with four engineering experts was developed to test the semi-automatic negotiation model. The experiment was created using the Fuzzy Logic Toolbox of Matlab. The composition of the experiment is explained next:

1. Four engineering experts, that are involved in the negotiation.
2. Two control Agents, that are manually configured to represent both, the most and the less experimented users in the system.
3. *Control 1* is always the user who has used the system most and therefore, the one that has to have the highest $\mathcal{I}_C$.
4. *Control 2* agent in the other hand represents the one with the less interaction with the system, that is one with the lowest $\mathcal{I}_C$.

Table 1. Scenario configuration

Scenario	SystemStage		Expert 1	Expert 2	Expert 3	Expert 4	Control 1	Control 2
Variable 1 x_1	Discipline	48	36	39	1	13	48	0
	Units	20	16	11	19	20	20	0
	PLC stage	37	13	8	17	24	37	0
	$\mathcal{I}_C$		0.639	0.519	0.5	0.639	0.817	0.183
Variable 2 x_2	Discipline	32	26	18	11		30	5
	Units	45	35	15	13		43	6
	PLC stage	58	38	6	15		56	7
	$\mathcal{I}_C$		0.79	0.291	0.269		0.817	0.183
Variable 3 x_3	Discipline	12	12	11			11	2
	Units	5	5	4			4	1
	PLC stage	8	3	7			7	1
	$\mathcal{I}_C$		0.669	0.817			0.817	0.183
Variable 4 x_4	Discipline	56	20				50	5
	Units	80	43				75	8
	PLC stage	35	5				32	3
	$\mathcal{I}_C$		0.273				0.817	0.183

As seen in Table 1, four scenarios were proposed, one variable per scenario. Each scenario was defined with different conditions: Number of experts involved in the negotiation, possible/real number of variables for `Discipline`, `Units` and `PLC stage` of the chosen variable.

The "System stage" column shows the total number of variables in the system for each of the attributes of the selected variable (`Discipline`, `Units` and `PLC stage`). For the x_1 example, there were 48 variables with the same discipline, 20 variables using the same unit and 37 from the same PLC stage of the selected variable. In this scenario the participation of four experts was simulated and their corresponding values for the real Discipline, Units and PLC variables associated with them, were randomly assigned. With this data $\mathcal{I}_D$, $\mathcal{I}_U$ and $\mathcal{I}_{PLC}$ were calculated for each expert. This information was used as input for the FIS presented in Fig. 3 to calculate the Competence Index ($\mathcal{I}_C$) for each expert. To assign a domain to x_i four different people were associated with one, two, three or four variables. Then they were asked to define D_i without knowing the domain previously defined by others to the same variable (x_i) but only to prevent new users to get influenced by domains defined by others and being able to have special cases (only for experimentation purposes). To illustrate the experiment in a better way, a detailed explanation is presented about variable 2 (x_2) scenario. This variable belongs to the "Sensors" discipline, the measurement units are "Tesla" and the PLC stage is "Control". As presented in Table 1 the system contains 32 variables of the sensor discipline, 45 variables that are measured with Tesla units and 58 belonging to the Control PLC stage. The number of each expert represents the order in which they arrived to the negotiation process.

Having in mind that $\forall x_i \in D_i, \mu(x_i) \in [0,1]$, the process starts once Expert 1 defines (2,3) as the domain for Variable x_2 (Fig. 6). As there is not another associated domain at this moment for the variable, the proposed domain is accepted.

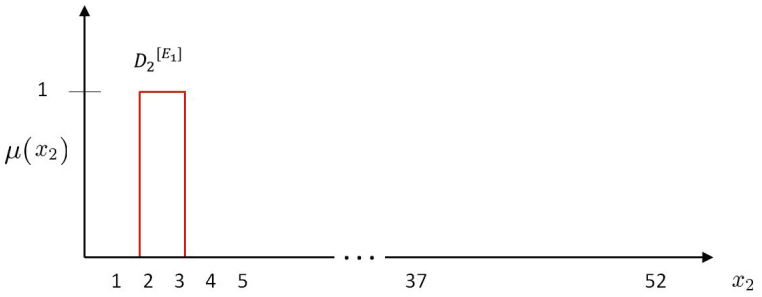


Fig. 6. Expert 1 proposed domain

Then, Expert 2 proposes a different domain for variable x_2 which is (37,52) (Fig. 7). The system validates that there is already a domain assigned to that variable.

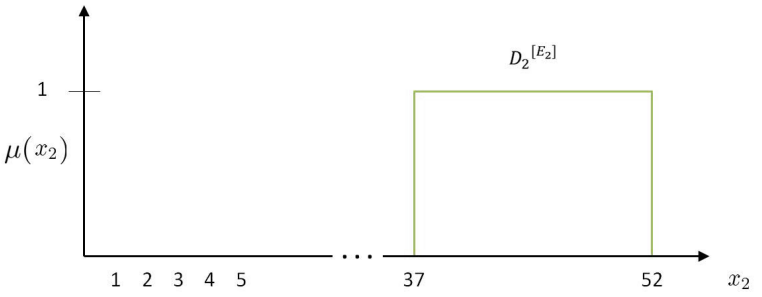


Fig. 7. Expert 2 domain

When this happens a conjunction between the domains presented by Expert 1 and Expert 2 is made to find if there is common range of values for both experts. In this example, the resultant set is empty and then, the system must proceed to calculate $\mathcal{I}_C$ for both experts. The results of the FIS were 0.79 and 0.291 for Expert1 and Expert 2 respectively. In this case the system must show an alert in order to inform experts that the new domain for variable 2 will keep being [2, 3] because Expert 1's $\mathcal{I}_C$ over variable 2 is bigger than the Expert 2's $\mathcal{I}_C$. If the experts want to change the domain they can do so but they need arrange a personal negotiation to fix it.

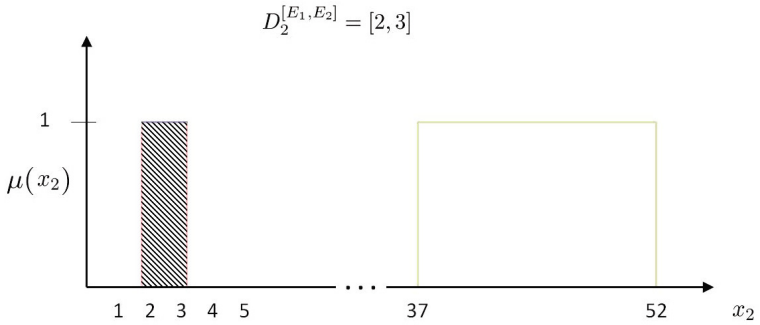


Fig. 8. Final domain between experts 1 and 2

When Expert 3 comes to the negotiation with a new proposed domain [1, 5] by executing the conjunction, the solution is the same domain [2, 3] and as no other Expert defines a different domain for variable 2 this is the final domain, which is $D_2^{[E_1,E_2,E_3]} = [2,3]$ and is shown in Fig. 9.

The previous explanation describes the model functionality. The other three scenarios exposed different aspects to have into consideration for the implementation process that will be described in Section 7.

7 Implementation

The implementation of the semi-automatic negotiation is part of a MAS proposed by Mejía-Gutiérrez et al. [10] to support collaborative design. The negotiation process is developed to support the domain definition during the collaborative construction of a

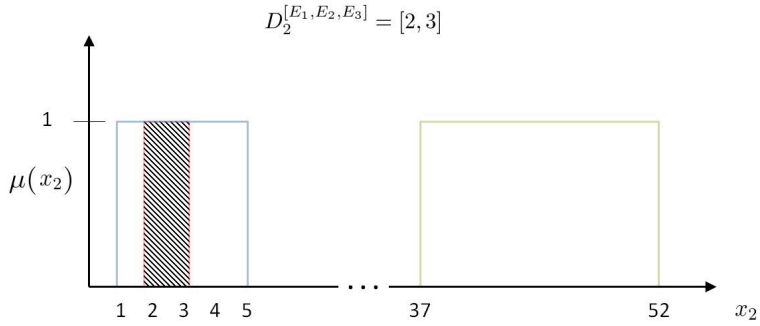


Fig. 9. Final domain for variable 2

CSP model. In this section it is presented how the semi-automatic model works during the different scenarios evidenced in the functional test and how it was implemented.

7.1 Capturing Experts' Preference

(LA) Domains are treated as an object that not only defines a valid range of values but also, that stores the user's preferences for them. These preferences are captured by a set of questions. First they are asked about the lower and upper limits of the range. By doing this, experts are forced to think about the design constraints without defining them as such. Once the range is defined, experts select one number of the defined interval to use as a reference for the definition of his preferences. Domain's preferences reflect the users' expertise. If a user knows about the engineering problem and has had interaction with similar projects, he will have a better idea of the value that might be used in the final model. This number allows the expert to have a reference value and to select one of the options shown at Fig. 10.

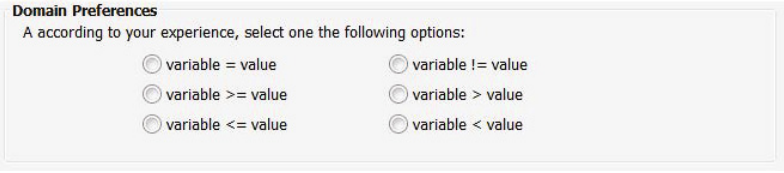


Fig. 10. Domain's preferences options

Each of these options is represented graphically as shown in Fig. 11 and converted to a fuzzy set that can be entered into the system. Each option can be plotted by four values within the domain and every value represents an inflection point of the preference. The two boundaries of the interval (value "a" and "d") are the lower and the upper value of the domain. Values outside this interval are considered with null preference.

7.2 Negotiation Support through the MAS

For the implementation of the described functionality it was necessary to add different behaviors to the already defined agents in the MAS by Mejía-Gutiérrez et al. [10], as well as define new agents. The original system includes two agents and it was necessary to develop a new agent for the semi-automatic negotiation model. It is labeled a *Logical Agent (LA)* and its main role is to propose the resultant domain. It has two main tasks: Do the conjunction between the fuzzy sets that represents the domains in conflict and when a null set is found, based on the method explained in Section 5.2 the calculation of the $\mathcal{I}_C$ is done. For the MAS the contribution is the definition of the LA and the aggregation of new behaviors to the existing agents in order to support the semi-automatic negotiation by the usage of fuzzy operations.

The LA was modeled using a combination of the two most known methodologies for MAS: MAS-COMMONKADS and GAIA [6, 13]. In the Table 2, the Roles Model of GAIA Methodology is presented for the LA.

Table 2. Schema for role of DomainAdvisor based on GAIA Methodology

Rol Schema	DomainAdvisor
Description	Supports the experts during the negotiation of an unique domain for a variable
Protocol and Activities	CheckIntersection, performFuzzyProcess, calculateSubIndex, CalculateCompetenceIndex, compareComptenceIndex, CreateDomain, suggestDomain, inform, assignDomain, createSuggestion

Permissions	`Reads`
	Proposed_Domains *//Proposed domains*
	Variables *//Real and Possible Variables*
	Domain *//Current Domain*
	`Changes`
	SubIndex *//For each expert*
	Competence_Index *//For each expert*
	Domain *//Domain of the variable*

Responsibilities	`Liveness`
	DomainAdvisor =
	$(checkIntersection.suggestDomain.inform.assignDomain)^{\omega}$ |
	PerformFuzzyProcess
	PerformFuzzyProcess =
	calculateSubIndex. calculateCompetenceIndex
	CompareCompetenceIndex. createDomain. createSuggestion.
	CreateSuggestion =
	createDomain. suggestDomain. inform. assignDomain
	`Safety`
	Every time an expert proposes a new domain execute
	DOMAINADVISOR
	The DataBaseAgent is the only one who can write in the database

During the functional test, it was identified that after having defined a domain for a variable (range and preference) there are two cases in which the negotiation process may occur. The first case is when a user of a variable proposes a new domain and the conjunction of the current and the proposed domain is not null. In this situation the negotiation model will present the calculated conjunction as the resultant domain to the experts involved in the negotiation. Here the $\mathcal{I}_C$ is not necessary. The second situation is presented when the conjunction of the domains is null. For this case, it is necessary to use the $\mathcal{I}_C$ in order to give preference to the domain proposed by the expert with the highest $\mathcal{I}_C$. The system will present the domain of the expert with highest as the new domain. In every case the system just gives a suggestion about which must be the final domain for a variable but the experts always can interact between them to define which the appropriate domain for the variable in conflict is.

The negotiation process is conducted between several agents. They include: the `TutorAgent (TA)`, `DataBaseAgent (DBA)` and `Logical Agent (LA)`, using the concepts defined in the Call-for-Proposal (CFP) Communicative act. This is defined as part of the Foundation for Intelligent Physical Agents (FIPA) Communicative Act Specifications. The Fig. 12 represents the Coordination Model for the negotiation.

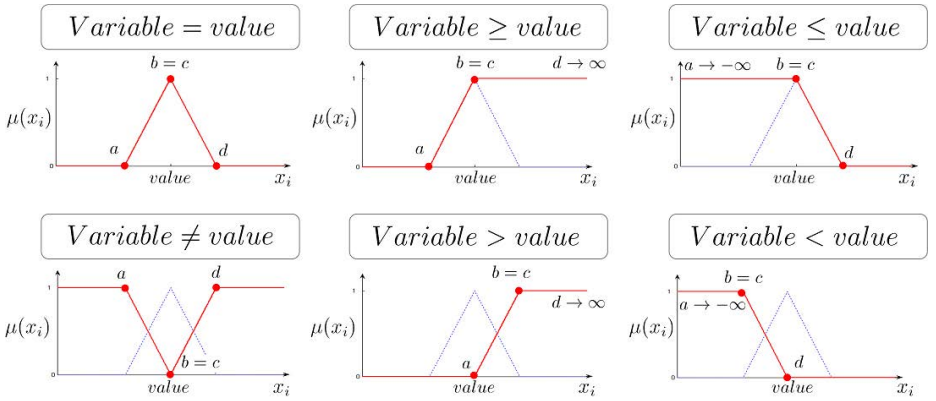


Fig. 11. Domain Preferences Representation

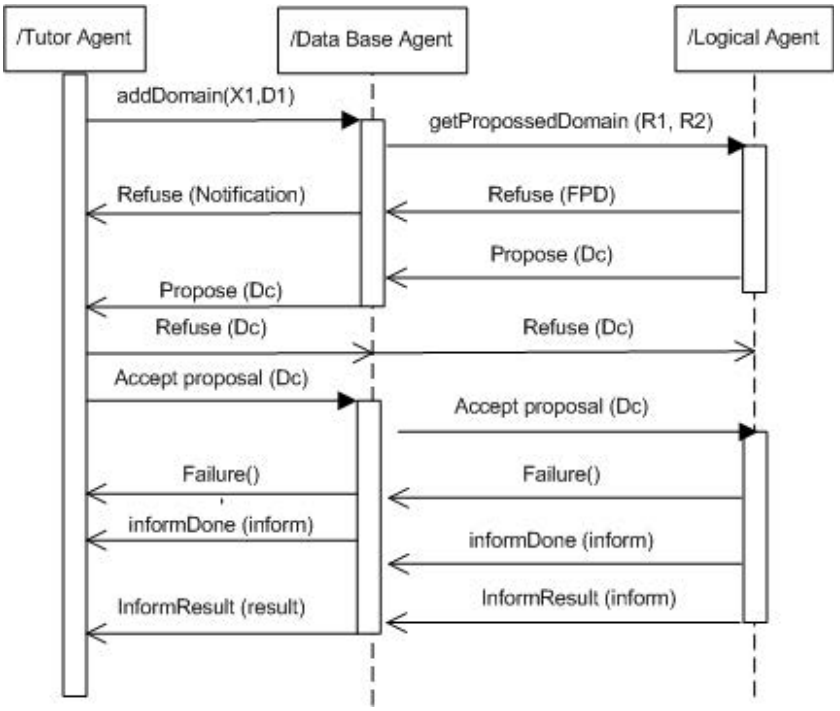


Fig. 12. Coordination Model of MAS-COMMONKADS

According to this process the communication for suggesting new domains is performed as follows:

1. A TA issues a CFP to the DBA to assign the Domain D_1 to the Variable x_1. Once the DBA receives the message, it queries the DBA to get the *Real Number of Variables*

for the users involved and the *Possible Number of Variables* in the systems. Along with this information and the domains a new CFP to the LAis issued. The CFP request the calculation of the Conjunction Domain D_c.
2. The LA performs the conjunction between the fuzzy sets that represents both domains. If $D_c \neq \texttt{null}$ a Propose message is sent to the DBA. In case of $D_c = \texttt{null}$, LA calculates $\mathcal{I}_C$ for both experts and refuses the proposal by sending the Domain of the expert with the highest $\mathcal{I}_C$.
3. When DBA receives a Refuse as answer, it writes the domain calculated by the LA in the DB and sends a Refuse message to the TA to inform users about the process. Otherwise when a propose message is received, it sends the same message to the experts' TA involved in the negotiation.
4. After the message is back to the TAs, a "Reject" or "Accept" proposal message is sent back to the DBA. If at least one of the messages is Rejected, the DBA keeps the domain that the variable had before the negotiation started. Otherwise when all messages are acceptances, the DBA store D_c in the Data Base.
5. DBA transmits the response of the users and sent the respective message to the LA in order to finish the negotiation.

8 Conclusions and Further Research

A semi-automatic negotiation method was proposed to support the domains definition within a MAS approach to support the CSP modelling in distributed PDE. To achieve this, an index was developed to measure the competence that experts have about each variable and it is calculated with a Fuzzy approach. The functional test showed that having this quantitative measure, it is helpful for decisions making and enables inferences during negotiations in a distributed knowledge modelling environment. With the proposed semi-automatic negotiation model it is possible to find empty domains before the execution of the CSP model in the solver, reducing processing time and improving quality of the CSP results. It also helps to identify conflicts in early stages (during the model construction), avoiding the execution of a non-consistent model with domain problems and also finding where these possible problems are. As further research, the system is intended to be enriched by giving relevance (weight) to CSP constraints, helping to support conflicts resolution during constraints definition. Finally, the number of indexes used for the calculation may increase depending on the context and specially if it is necessary to give relevance to another variable's properties.

References

[1] Antonssonz, E., Ottox, K.: Imprecision in engineering design (1995)
[2] Baxter, M., et al.: Product design: A practical guide to systematic methods of new product development. Chapman & Hall, London (1995)
[3] Faratin, P., Sierra, C., Jennings, N.: Using similarity criteria to make issue trade-offs in automated negotiations. Artificial Intelligence 142(2), 205–237 (2002)
[4] Grabisch, M.: The application of fuzzy integrals in multicriteria decision making. European Journal of Operational Research 89(3), 445–456 (1996)

[5] Hao, Q., Shen, W., Zhang, Z., Park, S., Lee, J.: Agent-based collaborative product design engineering: An industrial case study. Computers in Industry 57(1), 26–38 (2006)
[6] Iglesias Fernández, C.: Definición de una metodología para el desarrollo de sistemas multi-agente. PhD thesis, Universidad Politécnica de Madrid (1998)
[7] Larsson, A., Törlind, P., Mabogunje, A., Milne, A.: Distributed design teams: embedded one-on-one conversations in one-to-many. In: Proceedings of the Design Research Society International Conference, pp. 5–7. Citeseer (2002)
[8] Li, W., Qiu, Z.: State-of-the-art technologies and methodologies for collaborative product development systems. International Journal of Production Research 44(13), 2525–2559 (2006)
[9] Mamdani, E., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7(1), 1–13 (1975)
[10] Mejía-Gutiérrez, R., Cálad-Álvarez, A., Ruiz-Arenas, S.: A multi-agent approach for engineering design knowledge modelling. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS, vol. 6882, pp. 601–610. Springer, Heidelberg (2011)
[11] Pace, D.K., Sheehan, J.: Subject matter expert (SME)/peer use in M&S V&V. In: Proc. of the Foundations (2002)
[12] Tsang, E.: Foundations of constraint satisfaction, vol. 289. Academic Press, London (1993)
[13] Wooldridge, M., Jennings, N., Kinny, D.: The gaia methodology for agent-oriented analysis and design. Autonomous Agents and Multi-Agent Systems 3(3), 285–312 (2000)
[14] Zadeh, L.: Fuzzy sets. Information and Control 8(3), 338–353 (1965)

Teams of Agents for Solving the Resource-Constrained Project Scheduling Problem

Piotr Jędrzejowicz and Ewa Ratajczak-Ropel

Department of Information Systems
Gdynia Maritime University
Morska 83, 81-225 Gdynia, Poland
{pj,ewra}@am.gdynia.pl

Abstract. In this chapter the idea of Asynchronous Team (A-Team) is extended and used to solve one of the hard combinatorial optimization problems. The Team of A-Teams (TA-Teams) architecture for solving the Resource-Constrained Project Scheduling Problem (RCPSP) is proposed and experimentally validated. The RCPSP belongs to the Non-deterministic Polynomial hard (NP-hard) problem class. To solve this problem a team of parallel cooperating A-Teams is proposed. Such terms consist of asynchronous agents, implemented using JADE based A-Team (JABAT) multiagent system. Java Agent Development Framework (JADE) is a popular framework dedicated to implement agent based systems. In the proposed TA-Teams from one to four A-Teams and four kinds of optimization agent have been implemented and used. Each optimization agent represents an optimization algorithm. The proposed optimization algorithms are based on several known metaheuristics, such as a simple local search, tabu search and path relinking as well as the crossover and special heuristics dedicated to solving the RCPSP. Computational experiment involved evaluation of the proposed approach in respect of the different parameters settings controlling the working and migration strategies used in the suggested TA-Teams approach.

Keywords: Project Scheduling, Resource-Constrained Project Scheduling, RCPSP, Optimization, Agent, Multiagent System, A-Team.

1 Introduction

This chapter proposes an agent based approach to solving instances of the resource constrained project scheduling problem known as Resource-Constrained Project Scheduling Problem (RCPSP). The chapter is an extended version of the paper [11]. RCPSP problem have attracted a lot of attention and many exact and heuristic algorithms have been proposed for solving it [1, 8, 9]. The current approaches to solve this problem produce either approximate solutions or can only be applied for solving instances of the limited size. Hence, searching for more effective algorithms and solutions to the RCPSP problem is still a lively field of research. One of the promising directions of such research is to take advantage of the parallel and distributed computation solutions, which are the common feature of the contemporary multiagent systems [16].

Modern multiagent system architectures are an important and intensively expanding area of research and development. There is a number of multiple-agent approaches

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 224–235, 2013.

proposed to solve different types of optimization problems. One of them is the concept of an Asynchronous Team (A-Team), originally introduced by Talukdar et al. [15]. The idea of A-Team was used to develop the environment based on Java Agent Development Framework (JADE) for solving a variety of computationally hard optimization problems called JADE based A-Team (JABAT) [2, 12]. JABAT supports the construction of the dedicated A-Team architectures that is based on the population. The mobile agents used in JABAT promote the decentralization of computation across multiple hardware platforms. Parallel processing results in more effective use of the available resources and ultimately, a reduction of the computation time.

An extended version of JABAT has been proposed by Barbucha et al. [3]. The idea of integrating the team of asynchronous agents with an island-based genetic algorithm was introduced by Cohoon et al. [7]. The resulting Team of A-Teams (TA-Teams) architecture provided two levels of agent cooperation. Lower level cooperation takes place within a single A-Team. Solutions stored in the common memory of such an A-Team are being forwarded to optimization agents belonging to the team and send back after an attempted improvement. The improved solutions may replace some other solutions in the common memory. Cooperation takes place through sharing solutions from the common memory. Cooperation at the upper level involves communication, that is information exchange, between cooperating A-Teams belonging to the TA-Teams.

In this chapter the TA-Teams architecture is used for solving instances of the RCPSP problem. It is expected that introducing two-levels of cooperation between A-Teams and running them in parallel will result in obtaining high quality solutions in an efficient manner. The approach is experimentally validated.

Optimization agents used to produce solutions to the RCPSP instances represent metaheuristic algorithms such as the tabu search or path relinking algorithm. A behavior of the single A-Team is defined by the, so called, working strategy and cooperation between A-Teams by the migration topology and migration strategy. The approach extends the earlier research results described in [3, 9, 13].

The chapter is constructed as follows: Section 2 of the chapter contains the RCPSP problem formulation. Section 3 gives some information on extended JABAT environment. Section 4 provides details of the proposed dedicated TA-Teams architecture designed for solving the RCPSP instances. Section 5 describes settings of the computational experiment carried-out with a view to validate the proposed approach and contains a discussion of the computational experiment results. Finally, Section 6 contains conclusions and suggestions for future research.

2 Problem Formulation

A single-mode resource-constrained project scheduling problem consists of a set of n activities, where each activity has to be processed without interruption to complete the project. The dummy activities 1 and n represent the beginning and the end of the project. The duration of an activity j, $j = 1,\ldots,n$ is denoted by d_j where $d_1 = d_n = 0$. There are r renewable resource types. The availability of each resource type k in each time period is r_k units, $k = 1,\ldots,r$. Each activity j requires r_{jk} units of resource k during each period of its duration, where $r_{1k} = r_{nk} = 0$, $k = 1,...,r$. All parameters are non-negative integers. There are precedence relations of the finish-start type with a

zero parameter value ($FS = 0$) defined between the activities. In other words activity i precedes activity j if j cannot start until i has been completed. The structure of a project can be represented by an activity-on-node network $G = (SV, SA)$, where SV is the set of activities and SA is the set of precedence relationships. SS_j (SP_j) is the set of successors (predecessors) of activity j, $j = 1,\ldots,n$. It is further assumed that $1 \in SP_j$, $j = 2,\ldots,n$, and $n \in SS_j$, $j = 1,\ldots,n-1$. The objective is to find a schedule S of activities starting times $[s_1,\ldots,s_n]$, where $s_1 = 0$ and resource constraints are satisfied, such that the schedule duration $T(S) = s_n$ is minimized.

The above formulated problem as a generalization of the classical job shop scheduling problem belongs to the class of Non-deterministic Polynomial hard (NP-hard) optimization problems [5]. The considered problem class is noted as PS$|prec|C_{max}$ [6]. The objective is to find a minimal schedule in respect of the makespan that meets the constraints imposed by the precedence relations and the limited resource availabilities.

3 The Extended JABAT Environment

JABAT is an environment that facilitates the design and implementation of the A-Team architecture for solving various combinatorial optimization problems. The problem-solving paradigm on which the proposed system is based can be best defined as the population-based approach.

JABAT produces solutions to combinatorial optimization problems using a set of optimization agents, each representing an improvement algorithm. Each improvement (optimization) algorithm when supplied with a potential solution to the problem at hand, tries to improve this solution. The initial population of solutions (individuals) is generated or constructed. Individuals forming the initial population are, at the following computation stages, improved by independently acting optimization agents. The main functionality of the proposed environment includes organizing and conducting the process of search for the best solution.

Up to now JABAT has been used to solve instances of the following problems: the Resource-Constrained Project Scheduling Problem (RCPSP), the Traveling Salesman Problem (TSP), the Clustering Problem (CP), the Vehicle Routing Problem (VRP) and the Euclidean Planar Travelling Salesman Problem (EPTSP) [2, 3].

In the extended version of JABAT the TA-Teams can be implemented and used in a similar way as a single A-Team. A behavior of the single A-Team is defined by the, so called, working strategy and cooperation between A-Teams by the migration topology and migration strategy. The strategies are defined by the user. Each A-Team in the TA-Teams uses one population of individuals and a fixed number of optimization agents. All optimization agents within one A-Team work together to improve individuals from its population in accordance with the working strategy. Individuals can migrate between A-Teams in accordance with the migration strategy. The earlier experiments using the TA-Teams architecture were described by Barbucha et al. [3].

To implement the proposed architecture, the most important are the following classes of agents:

SolutionManager represents and manages one A-Team. For example the set of optimization agents and the population of solutions stored in the common memory.

MigrationManager manages the communication between A-Teams represented by solution mangers.

OptiAgent optimization agent that represents a single improving algorithm (for example: local search, simulated annealing, tabu search, genetic algorithm.).

Other important classes in JABAT and extended JABAT include: Task representing an instance or a set of instances of the problem and Solution representing the solution. To initialize the agents and maintain the system the TaskManager and PlatformManager classes are used. Objects of the above classes also act as agents.

In the extended version of JABAT the MigrationManger supervises the process of communication between solution managers with their common memories where populations of solutions are stored. The migration is asynchronous. With a given frequency the MigrationManager sends messages to SolutionManagers pointing out to which SolutionManager messages with the best solution or solutions should be send to. Then each thus informed SolutionManager resends the best current solution or solutions to the respective common memory. A single SolutionManager controls the process of solving a single problem instance (task) in accordance with the working strategy. Working strategy is a set of rules applicable to managing and maintaining a population of current solutions in the common memory.

JABAT has been designed and implemented using JADE (Java Agent Development Framework), which is a software framework proposed by TILAB [4] supporting the implementation of the multiagent systems. More detailed information about the JABAT environment and its implementations can be found in [2, 3, 12].

4 Dedicated TA-Teams Architecture

JABAT was successfully used by the authors for solving the RCPSP, MRCPSP and RCPSP/max problems [3, 9, 10]. In the proposed approach, several modifications and improvements with respect to agents, classes describing the problem and ontologies have been implemented within a dedicated TA-Teams architecture and used to solve instances of the RCPSP problem.

Classes describing the problem are responsible for reading and preprocessing the data and generating random instances of the problem. The discussed set includes the following classes:

RCPSPTask inheriting from the Task class and representing the instance of the problem,

RCPSPSolution inheriting from the Solution class and representing the solution of the problem instance,

Activity representing the activity of the problem,

Mode representing the activity mode,

Resource representing the renewable resource.

The second set includes classes describing the optimization agents. Each of them includes the implementation of an optimization algorithms used to solve the RCPSP problem. All of them are inheriting from OptiAgent class. These implement specialist

algorithms: LSA, TSA, CA and PRA described below. The prefix Opti is assigned to each agent with its embedded algorithm. In the proposed dedicated TA-Teams this set includes the following classes:

OptiLSA implementing the Local Search Algorithm (LSA),
OptiTSA implementing the Tabu Search Algorithm (TSA),
OptiCA implementing the Crossover Algorithm (CA), and
OptiPRA implementing the Path Relinking Algorithm (PRA).

The algorithms, earlier described by the authors [9, 11], have been modified and improved.

The LSA is a local search algorithm which finds local optimum by moving each activity to all possible places in the schedule. For each combination of activities the value of possible solution is calculated. The best schedule is remembered and finally returned.

The TSA is an implementation of the tabu search metaheuristic. It finds local optimum by exchanging each two activities in the schedule (the neighborhood of the schedule). The best move from the neighborhood of the solution, which is not tabu, is chosen and performed. The best schedule is remembered and finally returned.

The PRA is an implementation of the path-relinking algorithm. For a pair of solutions a path between them is constructed. The path consists of schedules obtained by carrying out a single move from the preceding schedule. The move is understood as moving one of the activities to a new position in the schedule. For each schedule in the path the value of the respective solution is checked. The best schedule is remembered and finally returned.

The CA is an algorithm based on the idea of the one point crossover operator. For a pair of solutions one point crossover is applied. The parameter $step$ determines the frequency the operation is performed. The best schedule is remembered and finally returned.

All optimization agents (OptiAgents) work in parallel improving solutions from their A-Team common memory managed by the SolutionManager. An individual is represented as schedule of activities S. The final solution is obtained from the schedule by forward or backward Serial Generation Scheme (SGS) procedure [14]. The working strategy of SolutionManager has been defined as follows:

- All individuals in the initial population of solutions are generated randomly and stored in the common memory.
- Individuals for improvement are selected from the common memory randomly and blocked, which means that once selected individual (or individuals) cannot be selected again until all other individuals have been tried.
- Returning individual replaces the first found worse individual. If a worse individual cannot be found within a certain number of reviews (where review is understood as a search for the worse individual after an improved solution is returned) then the worst individual in the common memory is replaced by a randomly generated one, representing a feasible solution.
- The computation time of a single A-Team is defined by the no improvement time gap set by the user. If in this time gap no improvement of the current best solution has been achieved, the A-Team stops computations.

All A-Teams (managed by the SolutionMangers) exchange the best solutions according to the migration strategy carried-out by the MigrationManager. In this chapter three migration strategies based on the topologies proposed by Jędrzejowicz et al. [13] to solve the EPTSP problem are considered. The three topologies which were occurred the best ones in solving EPTSP are used to solve RCPSP problem. There are *One Way Ring*, *Two Way Ring* and *Randomized* topology.

In the migration strategy based on the *One Way Ring* topology the MigrationManager cyclically sends message to each SolutionManger and asks it to send its best solution to its next SolutionManager. SoltionManagers are ordered according to their creation sequence.

In the case of migration strategy based on the *Two Way Ring* (or *Ring*) topology the MigrationManager cyclically sends message to each SolutionManger and asks it to send its best solution to its two adjacent SolutionManagers.

The third migration strategy is based on the *Randomized* topology in which the one (source) SolutionManager asks for a new solution. SolutionManager asks for a new solution when the current best solution in its common memory has not been changed by a fixed part of no improvement time gap. The source SolutionManager sends appropriate message to the MigrationManager. It choose randomly one other (target) SolutionManager and ask it to send its best solution to the source one.

For all mentioned above migration strategies the following settings are used:

- In each case one individual is sent from the source to the target SolutionManager.
- The best solution taken from the source SolutionManager replaces the worst solution in the common memory of the target one.
- All A-Teams stop computation, regardless of recent improvements in their best solutions, when one of them stops due to its working strategy.

5 Computational Experiment

This experiment highlights computational evidence supporting the theories used to satisfy the resource constraints typically encountered while scheduling projects. The experiment settings are discussed in Section 5.1 and the results in Section 5.2.

5.1 Settings

To evaluate the effectiveness of the proposed approach and compare the results, depending on the number of SolutionManagers used, the computational experiment has been carried out using benchmark instances of RCPSP from Project Scheduling Problems LIBrary (PSPLIB)[1] - test sets: sm30 (single mode, 30 activities), sm60, sm90, sm120. Each of the first three sets includes 480 problem instances while set sm120 includes 600. The experiment involved computation with the fixed number of optimization agents, fixed population size, and the limited time indicated by the no improvement time gap.

[1] See PSPLIB at `http://129.187.106.231/psplib`

In the experiment four sets of parameters have been used as presented in Table 1. In each set the total number of individuals in all populations is 80 and the total number of optimization agents working for all SolutionManagers is 32. The no improvement iteration gap has been set to two minutes, and the fixed part of no improvement time gap after which the SolutionManager asks for new solution is half of that time. The number of reviews in the working strategy is five.

Table 1. Parameters setting

#SolutionManagers	#OptiAgents for one SolutionManager	Population size for one SolutionManager
1	8x4	80
2	4x4	40
4	2x4	20
8	1x4	10

In the case of 4 and 8 A-Teams (SoltionManagers) three migration strategies have been considered: *One Way Ring*, *Two Way Ring* and *Randomized* described in section 4. This inlcudes four kinds of optimization agents representing the LSA, TSA, CA and PRA algorithms.

The experiment has been carried out using nodes of the cluster Holk of the Tricity Academic Computer Network built of 256 Intel Itanium 2 Dual Core 1.4 GHz with 12 MB L3 cache processors and with Mellanox InfiniBand interconnections with 10Gb/s bandwidth. During the computation one node per eight optimization agents was used.

5.2 Results

During the experiment the following characteristics of the computational results have been calculated and recorded: Mean Relative Error (MRE) calculated as the deviation from the optimal solution for sm30 set or from the Critical Path Lower Bound (CPLB) for sm60, sm90 and sm120 sets, Mean Computation Time (MCT) required to find the best solution and Mean Total Computation Time (MTCT). Each instance has been solved five times and the results have been averaged over these solutions. In the case of one and two SolutionManagers there is no migration strategy used. For two SolutionManagers the best solution is sent between them.

The computational experiment results are presented in Tables 2-5. In each case the 100% of feasible solutions has been obtained. The *Randomized* migration strategy has been occurred the best one for solving RCPSP problem for settings proposed in this approach (see Fig. 1). However the other two migration strategies has not been significantly worse. In all cases significantly better results has been obtained using more than two A-Teams (SolutionManagers).

The presented results are compared with the results reported in the literature. In Table 6 the results obtained by the heuristics algorithms compared in [1, 8] are presented. The comparison includes the results with known computation times and processor clocks

Table 2. Results for benchmark test set sm30 (RE from optimal solution)

Migration strategy	#SolutionManagers	MRE	MCT [s]	MTCT [s]
-	1	0.028 %	6.43	72.62
-	2	0.021 %	6.32	70.29
One Way Ring	4	0.015 %	11.24	68.76
Two Way Ring	4	0.019 %	11.40	69.06
Randomized	4	0.014 %	10.39	70.73
One Way Ring	8	0.015 %	17.15	72.86
Two Way Ring	8	0.013 %	18.02	72.23
Randomized	8	0.009 %	15.16	70.53

Table 3. Results for benchmark test set sm60 (RE from CPLB)

Migration strategy	#SolutionManagers	MRE	MCT [s]	MTCT [s]
-	1	11.44 %	32.70	75.56
-	2	11.34 %	28.36	75.44
One Way Ring	4	11.11 %	31.78	69.56
Two Way Ring	4	11.04 %	35.87	71.03
Randomized	4	11.01 %	34.45	71.42
One Way Ring	8	11.02 %	35.83	68.79
Two Way Ring	8	11.01 %	34.66	72.92
Randomized	8	10.98 %	36.72	67.23

Table 4. Results for benchmark test set sm90 (RE from CPLB)

Migration strategy	#SolutionManagers	MRE	MCT [s]	MTCT [s]
-	1	11.38 %	36.05	74.51
-	2	11.31 %	35.13	71.08
One Way Ring	4	10.89 %	43.83	80.45
Two Way Ring	4	10.99 %	42.49	82.21
Randomized	4	10.75 %	40.33	83.06
One Way Ring	8	10.92 %	45.10	80.23
Two Way Ring	8	10.84 %	44.24	72.56
Randomized	8	10.70 %	42.87	62.12

Table 5. Results for benchmark test set sm120 (RE from CPLB)

Migration strategy	#SolutionManagers	MRE	MCT [s]	MTCT [s]
-	1	34.43 %	78.38	137.10
-	2	34.52 %	78.52	126.27
One Way Ring	4	33.15 %	89.31	135.12
Two Way Ring	4	32.88 %	89.69	123.78
Randomized	4	32.83 %	87.41	145.31
One Way Ring	8	33.05 %	88.98	146.23
Two Way Ring	8	32.79 %	90.16	133.53
Randomized	8	32.73 %	89.23	134.12

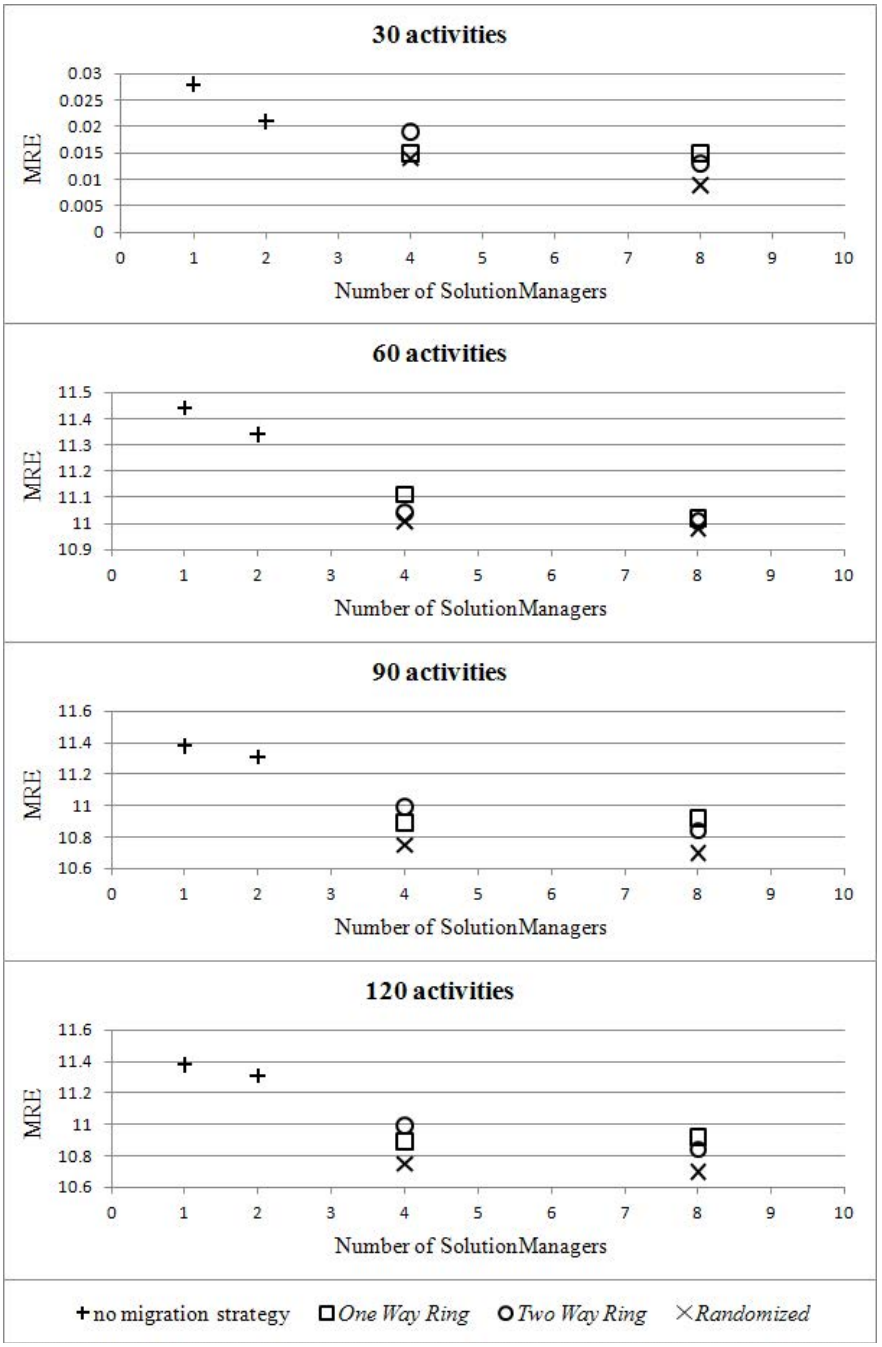


Fig. 1. The graphical representation of the results from Tables 2–5: Mean Relative Errors (MRE) for 1, 2, 4 and 8 SolutionManagers and different migration strategies

Table 6. Literature reported results [1, 8]

Set	Algorithm	Authors	MRE	MCT [s]	Computer
sm30	Decompos. & local opt	Palpant et al.	0.00	10.26	2.3 GHz
	VNS–activity list	Fleszar, Hindi	0.01	5.9	1.0 GHz
	Local search–critical	Valls et al.	0.06	1.61	400 MHz
sm60	PSO	Tchomte et al.	9.01	–	–
	Decompos. & local opt	Palpant et al.	10.81	38.8	2.3 GHz
	Population–based	Valls et al.	10.89	3.7	400 MHz
	Local search–critical	Valls et al.	11.45	2.8	400 MHz
sm90	Filter and fan	Ranjbar	10.11	–	–
	Decomposition based GA	Debels, Vanhoucke	10.35	–	–
	GA–hybrid, FBI	Valls at al.	10.46	–	–
sm120	Filter and fan	Ranjbar	31.42	–	–
	Population-based	Valls et al.	31.58	59.4	400 MHz
	Decompos. & local opt.	Palpant et al.	32.41	207.9	2.3 GHz
	Local search–critical	Valls et al.	34.53	17.0	400 MHz

mainly. However in the case of the agent based approaches it is difficult to compare computation times as well as the number of schedules, which is another widely used measure of the algorithm efficiency. In the proposed agent-based approach computation times as well as number of schedules differ between nodes and optimization agent algorithms working in parallel. The results obtained be a single agent may or may not influence the results obtained by the other agents. Additionally the computation time includes the time used by agents to prepare, send and receive messages.

The experiment results show that the proposed implementation is effective and combining A-Teams within the proposed TA-Teams architecture is beneficial. The results obtained using more solution managers are in most cases better than the results obtained using only one SolutionManager with a similar parameter settings.

6 Conclusions

The computational experiment results show that the proposed dedicated TA-Teams architecture is an effective and competitive tool for solving instances of the RCPSP problem. Presented results are comparable with solutions known from the literature and in some cases outperform them. It can be also noted that they have been obtained in a comparable time. However, in this case time comparison may be misleading since the proposed TA-Teams have been run using different numbers and kinds of processors. In case of the agent-based environments the significant part of the time is used for agent communication which has an influence on both - computation time and quality of the results.

The presented experiment should be extended to examine the TA-Teams behavior for longer no improvement time gaps and different sets of parameters, especially more solution managers, different population sizes and different numbers of optimization

agents. Future research will concentrate on implementing more sophisticated optimization agents, as well as on searching for the best configuration of the heterogenous agents used during computations.

Since JABAT has a possibility to run more than one copy of each agent it is interesting which agents should or should not be replicated to improve the results. Moreover, testing and adding to JABAT more different optimization agents and improving the existing ones will be considered.

Acknowledgments. The research has been supported by the Ministry of Science and Higher Education grant no. N N519 576438 for years 2010–2013. Calculations have been performed in the Academic Computer Centre TASK in Gdansk.

References

[1] Agarwal, A., Colak, S., Erenguc, S.: A Neurogenetic Approach for the Resource–Constrained Project Scheduling Problem. Computers & Operations Research 38, 44–50 (2011)

[2] Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: e-JABAT - An Implementation of the Web-Based A-Team. In: Nguyen, N.T., Jain, L.C. (eds.) Intelligent Agents in the Evolution of Web and Applications. SCI, vol. 167, pp. 57–86. Springer, Heidelberg (2009)

[3] Barbucha, D., Czarnowski, I., Jędrzejowicz, P., Ratajczak-Ropel, E., Wierzbowska, I.: Parallel Cooperating A-Teams. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part II. LNCS, vol. 6923, pp. 322–331. Springer, Heidelberg (2011)

[4] Bellifemine, F., Caire, G., Poggi, A., Rimassa, G.: JADE. A White Paper, Exp. 3(3), 6–20 (2003)

[5] Błażewicz, J., Lenstra, J., Rinnooy, A.: Scheduling subject to resource constraints: Classification and complexity. Discrete Applied Mathematics 5, 11–24 (1983)

[6] Brucker, P., Drexl, A., Möhring, R., Neumann, K., Pesch, E.: Resource-Constrained Project Scheduling: Notation, Classification, Models, and Methods. European Journal of Operational Research 112, 3–41 (1999)

[7] Cohoon, J.P., Hegde, S.U., Martin, W.N., Richards, D.: Punctuated Equilibria: A Parallel Genetic Algorithm. In: Proceedings of the Second International Conference on Genetic Algorithms, pp. 148–154. Lawrence Erlbaum Associates, Hillsdale (1987)

[8] Hartmann, S., Kölisch, R.: Experimental Investigation of Heuristics for Resource-Constrained Project Scheduling: An Update. European Journal of Operational Research 174, 23–37 (2006)

[9] Jędrzejowicz, P., Ratajczak-Ropel, E.: New Generation A-Team for Solving the Resource Constrained Project Scheduling. In: Proc. the Eleventh International Workshop on Project Management and Scheduling, Istanbul, pp. 156–159 (2008)

[10] Jędrzejowicz, P., Ratajczak-Ropel, E.: Solving the RCPSP/max Problem by the Team of Agents. In: Håkansson, A., Nguyen, N.T., Hartung, R.L., Howlett, R.J., Jain, L.C. (eds.) KES-AMSTA 2009. LNCS, vol. 5559, pp. 734–743. Springer, Heidelberg (2009)

[11] Jędrzejowicz, P., Ratajczak-Ropel, E.: Team of A-Teams for Solving the Resource-Constrained Project Scheduling Problem. In: Grana, M., et al. (eds.) Advances in Knowledge-Based and Intelligent Information and Engineering Systems, pp. 1201–1210. IOS Press (2012)

[12] Jędrzejowicz, P., Wierzbowska, I.: JADE-Based A-Team Environment. In: Alexandrov, V.N., van Albada, G.D., Sloot, P.M.A., Dongarra, J. (eds.) ICCS 2006. LNCS, vol. 3993, pp. 719–726. Springer, Heidelberg (2006)
[13] Jędrzejowicz, P., Wierzbowska, I.: Impact of Migration Topologies on Performance of Teams of A-Teams. In: Grana, M., et al. (eds.) Advances in Knowledge-Based and Intelligent Information and Engineering Systems, pp. 1161–1170. IOS Press (2012)
[14] Kölisch, R.: Serial and parallel Resource-Constrained Project Scheduling Methods Revisited: Theory and Computation. European Journal of Operational Research 43, 23–40 (1996)
[15] Talukdar, S., Baerentzen, L., Gove, A., de Souza, P.: Asynchronous Teams: Co-operation Schemes for Autonomous, Computer-Based Agents. Technical Report EDRC 18-59-96. Carnegie Mellon University, Pittsburgh (1996)
[16] Wooldridge, M.: An Introduction to MultiAgent Systems, 2nd edn. John Wiley and Sons (2009)

Part IV

Applications

Integrating Ultra Mobile Devices in Tactical Defence Environments through Middleware

Kate Foster

Defence Science and Technology Organisation, PO Box 1500, Edinburgh SA 5111, Australia
Kate.Foster@DSTO.defence.gov.au

Abstract. This paper is concerned with integrating mobile technology (such as smartphones and tablets) into tactical defence environments. The motivation for and key technologies involved in this research are discussed in detail. The vehicle for conducting this research is the Defence Science and Technology Organisation's Net Warrior initiative. The initial work conducted on this research program under Net Warrior has investigated how distributed object and publish/subscribe middleware can be incorporated into a smartphone to achieve information interoperability with tactical platforms and systems.

1 Introduction

Mobile computing devices, such as smartphones and tablets, are inexpensive and sophisticated electronic devices. Recent advances have included graphical displays, keyboards and other native sensor and networking capabilities. Already mobile technology is contributing to some of the tactical edge coordination activities of high value Australian Defence Force (ADF) assets (fighter, weapon and surveillance platforms) to a limited degree. In the future, these devices might also be used by warfighters to access or provide critical operational information, such as intelligence, live video, terrain descriptions, maps or asset descriptions. The following is an example of the recent experience of the United States armed forces: "In Afghanistan, Special Operations Forces are lightly armed, but very well connected to networks. Our fighting forces are themselves sensors and they are connected to weapons systems and platforms that are capable of delivering enormous fire power" [1, slide 14].

Organisations implementing a net centric approach aim to achieve effective and efficient outcomes by capitalising on information sharing for better situational awareness, improved decision making and enhanced collaboration. The main driver for net centricity has been the recent progress achieved in information and communication technologies [2]. These technologies can be considered as essential *enablers* for net centric systems and organisations will be required to adapt their structure and processes in order to exploit them. Network Centric Warfare (Network Centric Warfare (NCW)) applies the idea of net centricity to military operations and it is *networking* that underlies the information advantage that NCW may provide.

Over recent years, the United States Department of Defense has begun to incorporate smartphones and tablets into its NCW environment [3]. For example, the Defense Advanced Research Projects Agency (DARPA) Transformative Apps program [4] aims

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 239–253, 2013.

to provide warfighters with the mobile applications they need when they are needed. This will involve developing a diverse range of national security applications, creating a military application marketplace, and implementing an innovative development and acquisition process. The US Army has also established the Connecting Soldiers to Digital Applications (CDSA) and Applications (Apps) for the Army programs. The purpose of CDSA [5] is to determine the value of providing soldiers with applications on mobile devices. The applications developed will be for both administrative tasks and tactical operations. The Apps for the Army program [6] aims to reduce the time it takes to deploy applications so that they are available within ninety days after being requested.

The Australian NCW Concept [7] focuses on an effects-based approach with the aim of increasing operational tempo and improving agility by using information to maximise operational effect and facilitate collaboration. The human (or organisational and sociological) dimension is concerned with training, education, doctrine, organisation and leadership and requires trust to enable effective collaboration. The network (or technological) dimension connects engagement, sensor and command systems. A third component, networking, describes how the ADFs human and network dimensions will collaborate to build a system of systems. Therefore, the Australian focus is on the adaptation of military structure, tactics and concept of operations to net centric environments so that greater improvement can be achieved. Five premises (Figure 1) have been developed to explain how the human dimension, the network dimension and networking will produce a warfighting advantage. The following elements have been proposed in order to achieve self-synchronisation (premise 5) and deliver the desired operational effects:

1. A sensor grid, which consists of sensors and intelligence sources.
2. A C2 grid and an engagement grid will use information from the sensor grid to achieve more effective command, control and targeting.
3. An information grid, which is a network that better connects elements of Defence and protects its information.

Each of these grids consists of a human dimension and a network dimension along with a networking component. Figure 2 illustrates how these grids will interact.

The Australian Defence Organisation (ADO) is currently implementing the Australian NCW Concept. In order for this to be successful, the ADO requires advice regarding the underlying technologies that enable NCW and, specifically, how mobile computing devices can be integrated into its NCW environment.

The Defence Science and Technology Organisation (DSTO) Net Warrior initiative was established in late 2005. It aims to mitigate the risk involved in implementing NCW and exploit the opportunities NCW presents. This paper discusses the initial work conducted under Net Warrior to investigate how mobile technology can be integrated into tactical defence Service Oriented Architecture (SOA) environments. The motivation for conducting this work is discussed in Section 2. The Net Warrior initiative, its key middleware technologies and two relevant demonstrations are discussed in Section 3. Section 4 presents the initial work conducted on integrating a smartphone into Net Warrior. Section 5 discusses related work and Section 5 summarises this paper.

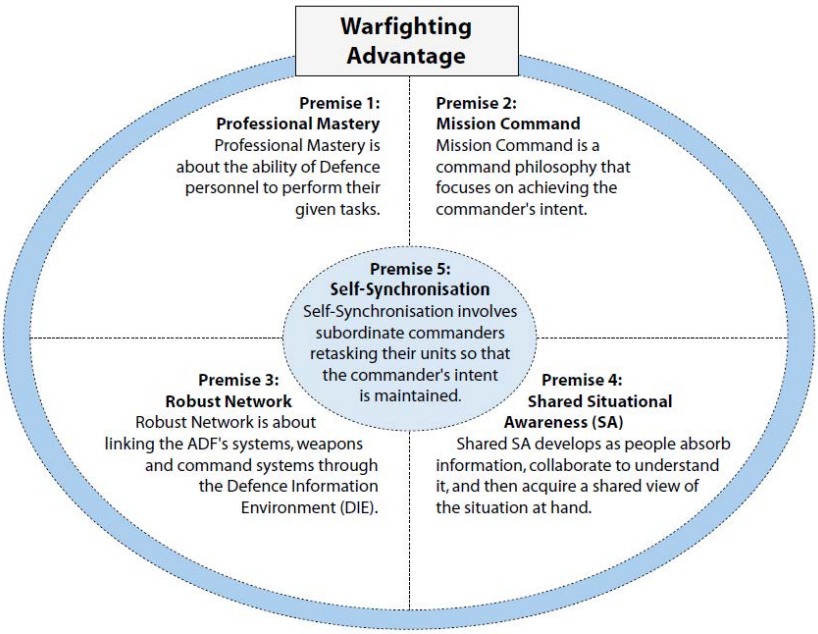


Fig. 1. The five premises of the Australian NCW Concept [8, p. 10]

2 Motivation

In order to implement the Australian NCW concept, Defence must have a dependable, secure and integrated information environment that supports both military operations and the Defence enterprise environment. The Chief Information Officer Group (CIOG) within the Australian Department of Defence is responsible for ensuring the successful implementation of the Defence information environment. CIOG requires science and technology support to achieve its goals. In the 2009 Defence White Paper the Australian Government stated that Defence will maintain its strategic capability advantage partly through "self-reliant defence research, development and innovation, and collaborative programs with scientifically and technologically capable partners" [10, paragraph 8.57]. One of these partners is DSTO, which is the Australian Government's primary research and development agency that provides support for Australia's defence and security requirements.

The Defence White Paper also states that DSTO will "increase its investigation and application of key enabling technologies which will provide significant returns for development of the future force" [10, paragraph 17.18]. Two of these technologies are integrated Intelligence, Surveillance and Reconnaissance (ISR) and networked systems, which are highly relevant to the core business of CIOG. Consequently, CIOG released its first Science and Technology (S&T) plan in 2011 [11] to provide guidance to DSTO regarding where it can provide assistance to CIOG.

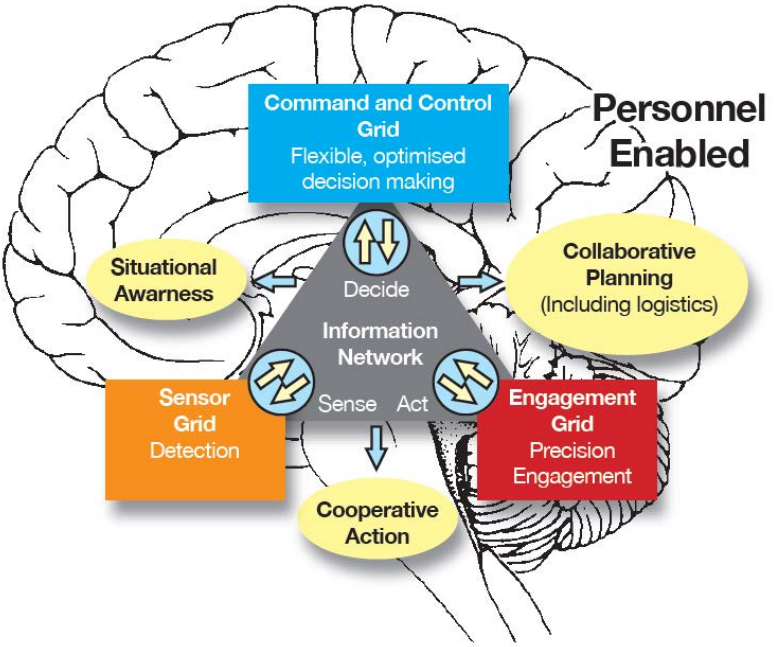


Fig. 2. Interaction between the key elements of Australian NCW [9, p. 6]

Program Information and Communications Technology (ICT)-1 in the CIOG S&T plan concerns future information technologies. The objectives of this program include understanding potentially disruptive information technologies and developing information technology options to achieve a capability edge for the future ADF. One of the requirements under this program is to research emerging wireless communication technologies. The problem is described as: "Defence needs to use advanced distributed and mobile technologies for global operations. Defence's strategic intentions are to use commercial network technologies to enhance its military capability. Seamless mobility – any where, any time – has a high potential value to Defence if it can rapidly adapt wireless and mobile technologies in order to achieve fitness for purpose" [11, p. 15]. One of the work components to be covered under this requirement is experimentation with commercially derived wireless and mobile information technologies for Defence use.

Program ICT-3 in the CIOG S&T plan concerns SOA. One of the objectives of this program is to develop an understanding of how to apply SOA techniques to areas of interest to Defence, including integrated ISR, systems integration and NCW. The deployed SOA requirement under this program involves providing advice on how SOA techniques should be used in deployed, operational and tactical environments. CIOG is currently delivering the SOA backbone for Defence's Single Information Environment [12] and requires specialist advice in a number of areas, including SOA interoperability and near real-time SOA architecting and implementation. The CIOG S&T plan envisages that work on this requirement will build on the Net Warrior *D10* demonstration conducted in 2010 (see Section 3.4), as well as other work conducted by DSTO.

This section has shown that there is strong motivation for DSTO to investigate the applicability and vulnerabilities of using mobile devices in the tactical domain. Further, with the increasing incorporation of middleware-based SOA systems into the Defence environment [13], mobile devices seem well suited to providing limited situational awareness and other interactive coordination and communications capabilities to the tactical edge (chat, tactical white-boarding and other augmented applications). An ideal vehicle for demonstrating this research is the Net Warrior initiative.

3 Net Warrior

3.1 Purpose

In alignment with the ADO's approach to implementing NCW through 'learning by doing', the DSTO Net Warrior initiative addresses new and evolving net centric capabilities and mission system technologies to enhance ADF joint warfighting capabilities [14]. Net Warrior activities are conducted through a program of events categorised as infrastructure events (NW-I#), demonstration events (NW-D#) and experimentation events (NW-E#). Technology demonstrations and experimentation are conducted with real systems, testbeds and simulators across DSTO and, potentially, wider Defence. These activities are applied to the operational, systems and technical elements of NCW and enable Net Warrior to provide advice to the ADO regarding the extent to which it needs to consider and implement particular NCW concepts and technologies.

The aim of Net Warrior demonstration events is to exhibit the integration of legacy and future platforms and technologies to support NCW implementation. The challenges faced by Net Warrior are the horizontal integration of middleware technologies that have been designed and implemented for operation in specific domain environments and the integration of heterogeneous systems developed by multiple suppliers.

4 Key Middleware Technologies

Seamlessly sharing information within and between systems in a timely manner is essential in order to successfully conduct NCW. Defence capability procurement has typically concentrated on platforms which has resulted in stove-piped systems that satisfied a capability gap. In net centric environments capabilities need to be acquired with the ability to interoperate with other systems. NCW is underpinned by a range of standards and technologies that support interoperability. Such standards and technologies of particular importance to Net Warrior include the real time Common Object Request Broker Architecture (Real Time - CORBA (RT-CORBA)) specification, component-based systems, Web Services, SOA, and the Data Distribution Service (Data Distribution Service (DDS)) for Real-Time Systems Specification. These standards and technologies are central to the approach taken by a range of other organisations, including the Open Architecture program in the US Navy [15]. This program was developed to address weapon system affordability, interoperability and performance for the current fleet and the Navy after next.

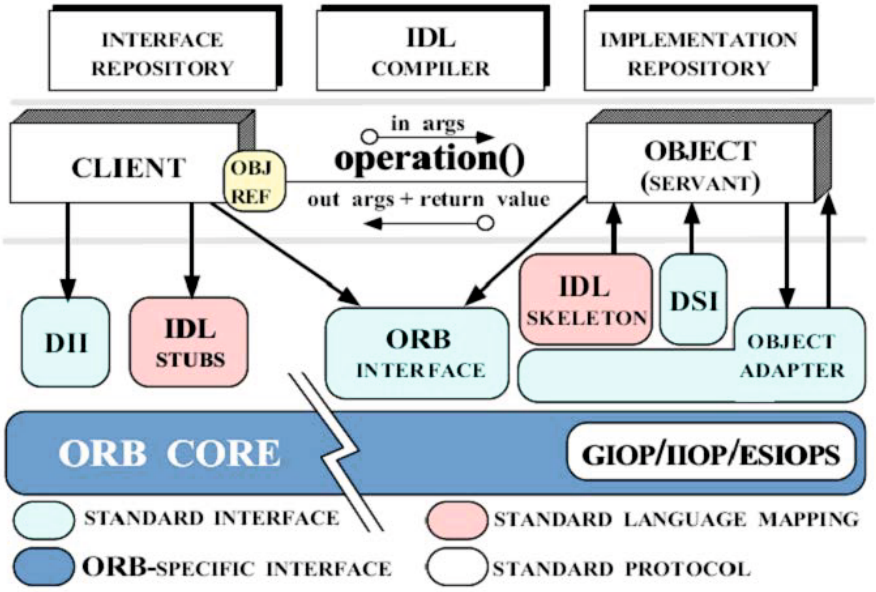


Fig. 3. RT-CORBA features [18]

The RT-CORBA specification [16,17] developed by the Object Management Group (OMG) supplies a set of abstractions and services to address interoperability in distributed heterogeneous computing environments. This reduces the reliance on specific programming languages, operating systems, communication protocols and hardware. In a distributed object system developed using RT-CORBA (See Figure 3), *requestors'* of services (clients) are separated from providers of services (servants) by a standardised published interface. An Object Request Broker (ORB) provides the mechanism for clients and servants to communicate. Within a CORBA-based system, objects can be located anywhere on the network, implemented in different programming languages, run on different computing platforms and communicate through various networking technologies.

Software components are self-contained and deployable elements that form applications when assembled with other components. Bachmann et al. [19] propose that components should exhibit the following properties: opaquely implement functionality; be able to be reused to form new composites; and conform to a component model. Component-based systems, when supported by middleware and software frameworks, are able to satisfy the design needs of applications to produce stable mission and net centric systems.

In creating a single information environment for Defence ICT, CIOG will be implementing Defence business processes as services (e.g. intelligence analysis, discovery and collaboration) [12]. One of the enabling technologies is Web Services [20], which is deliberately adapted to the Web unlike other distributed computing technologies. In the defence domain Web Services is an appropriate technology for enterprise, ISR and

headquarters environments. However, it cannot address the near real-time requirements of tactical systems (e.g. call for fires). Tactical systems require targeted standards and technologies.

RT-CORBA, component-based systems and Web Services enable sophisticated distributed systems to be developed. They also contribute to the mechanisms and structure for services to be developed for a SOA-based capability [13]. SOA is an architectural approach that aligns an organisation's information technology with its business processes. It is generally seen as well suited to managing integration and interoperability in large complex systems. In NCW environments, SOA enables flexible and adaptable operational effectiveness by enabling better integration of disparate systems and capabilities. The Defence Information and Communications Technology Strategy has committed to implementing a Defence wide SOA infrastructure. This implementation will "drive efficiency without compromising effectiveness across Defence by delivering reusable, granular, modular and interoperable services" [21, p. 50].

Figure 4 illustrates the SOA approach. Each middleware layer provides a new level of functionality with higher layers becoming more application specific and providing an interface upwards. Each application is developed to integrate down to the layer it supports and the layered architecture enables it to interoperate with other applications. The dotted lines are logical connections between middleware that is hosted on different hardware platforms or operating systems. The difference in the size of the services indicates the amount of effort required to develop services that integrate at different layers (i.e. services that integrate with lower layers require much more developer effort than those that leverage the functionality provided by higher layers).

In net centric environments, tactical information management systems are required to deliver the "right information to the right user at the right time" [23, p. 192] and satisfy quality of service (QoS) requirements in heterogeneous environments. To address this, the OMG released the DDS for Real-Time Systems Specification [24] in 2005, which incorporates a Quality of Service (QoS) function based on a data-centric middleware platform. DDS ensures interoperability through the use of middleware with interfaces defined by standards. DDS enables applications to communicate by asynchronously "publishing" information they have and "subscribing" to information they need in a timely manner and within the constraints imposed by definable QoS mechanisms. The use of middleware technologies in DDS provides location independence and consequently anonymity for publishers and subscribers. According to Schmidt et al. [25, p. 24], "DDS is an important distributed software technology for mission-critical Department of Defence (DoD) net-centric systems". The DDS programming model separates the active aspects of subscribing and publishing from control and data elements (Figure 5). This separation enhances scalability as there is no need for an active and centralised distribution entity. Therefore, a large population of subscribers and publishers can be supported.

The distributed object client/server model provided by RT-CORBA is complementary to the DDS publish/subscribe model. RT-CORBA is based on a client /server architecture and is best suited to applications in which one software component (the servant) is supplying a service to one or more other interacting components (clients). DDS is based on a publish/subscribe architecture and is best suited to applications in which one

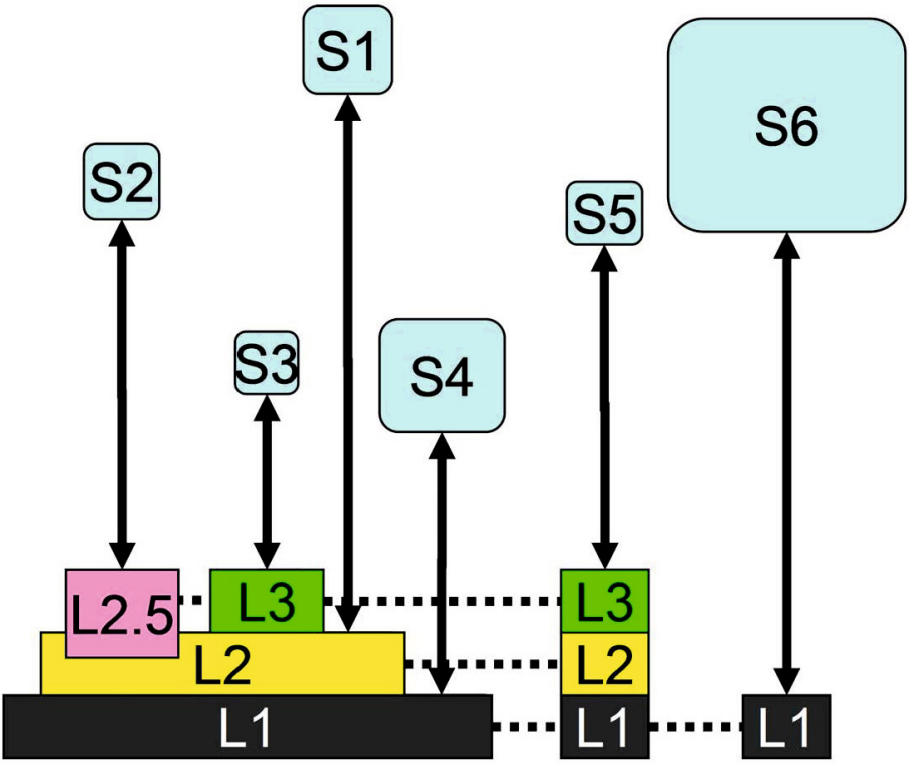


Fig. 4. Connecting systems through a SOA protect [22] 22

or more data sources (publishers) need to communicate information to one or more data users (subscribers). While RT-CORBA is used to distribute processing across a computing network, DDS shares data across the network. Many applications have requirements to distribute both processing and data, and therefore they have adopted RT-CORBA and DDS together. RT-CORBA and DDS both support heterogeneous and scalable systems. They are well suited for integrating applications that have to run on diverse hardware (ranging from servers to embedded systems) and need to incorporate real-time capabilities.

Figure 6 depicts a layered reference architecture used in Net Warrior to support nodes in the development of their middleware-based SOA environments. The platform environment represents the computing hardware and operating system. The platform abstraction environment encapsulates and augments the capabilities of the platform environment to provide a homogeneous interface to the layers above. The communications layer sits above the platform abstraction environment and provides standard transport and protocol support. Above the communications layer is the distributed object middleware environment that provides distribution standards and services to support networked applications. The publish/subscribe middleware environment is layered on top of the distributed object middleware environment and provides asynchronous dissemination

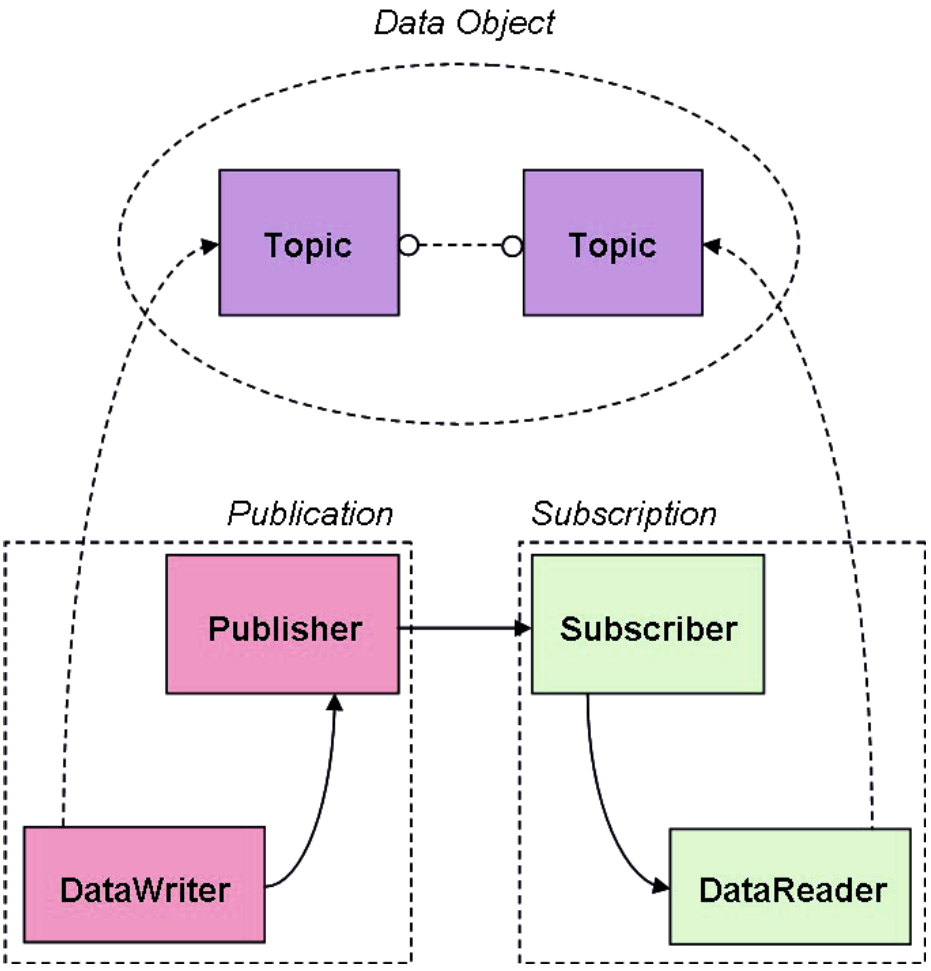


Fig. 5. Conceptual view of DDS

of data between applications. Specific domain environments at the top level provide a framework layer to host applications. Applications are developed and deployed on top of the domain environment by extending the framework and may also interact directly with the middleware environments or, in the case of legacy applications, lower layers.

The intention in Net Warrior is, where possible, for each node to establish a domain specific SOA based on the reference architecture in Figure 6 from which to provide services to other nodes in the network. Examples of nodes that have implemented this approach are the Airborne Early Warning and Control (AEW&C) and ISR nodes.

The AEW&C node has been developed by the Air Operations Division at DSTO and includes the Wedgetail Integration and Research Environment (WIRE) and the AEW&C Mission System Testbed (MST) [14]. The WIRE includes the real mission system of the AEW&C platform, which provides the ADF with a surveillance and control capability.

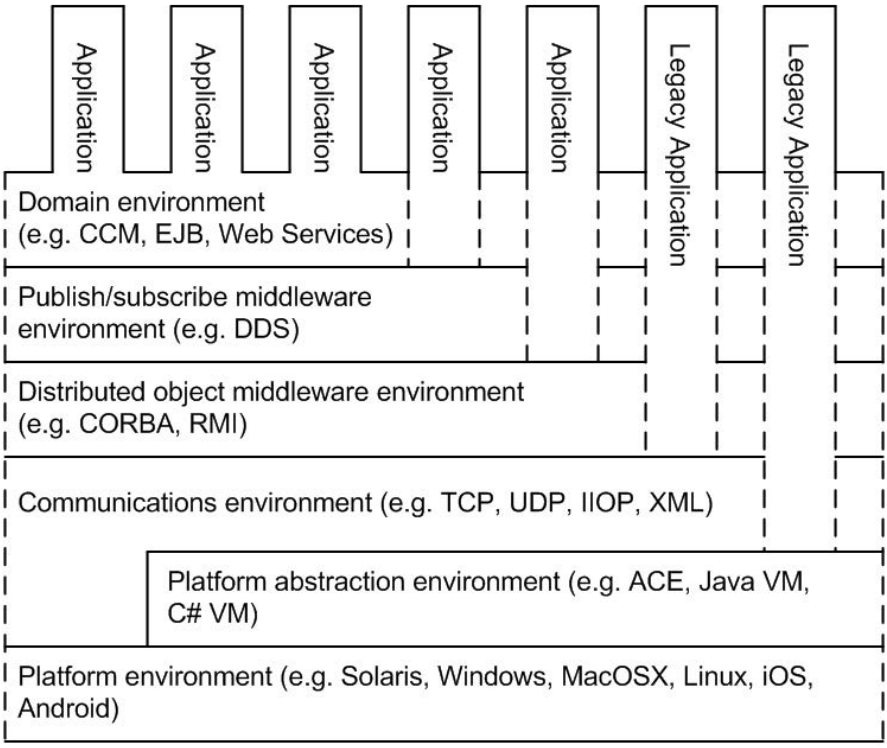


Fig. 6. Net Warrior reference architecture

The AEW&C MST is a generic SOA mission system and has been developed to support evaluation of Wedgetail mission computing while providing the freedom to explore the integration of NCW enabling technology into Wedgetail and other platforms. In Net Warrior, the AEW&C MST primarily acts as a middleware gateway for adaptation and bridging to the other nodes. Both the WIRE and AEW&C MST are built on a component-based distributed computing environment and employ a SOA approach.

The ISR node has been developed by the ISR Division at DSTO. The main research focus in the ISR node is the Australian Defence Force ISR Integration Backbone (ADIIB). The ADIIB is a SOA system based on Web Services technology and integrating it with other Net Warrior nodes provides a continuous interoperability measure for developing the ADIIB.

Two Net Warrior demonstrations have been conducted in recent years and are discussed in Sections 5 and 6.

5 NW-D3

The NW-D3 event, Demonstration of Information Interoperability using Middleware Technologies in Dynamic Environments, was held in 2008 [26]. Two nodes were

involved: the AEW&C node and the Future Operations Centre Analysis Laboratory (FOCAL) located within Command Control Communications and Intelligence Division at DSTO. This event involved demonstrating SOA, component-based systems, middleware, distributed computing, DDS, adapters and visualisation techniques. A DDS framework was developed in order to incorporate DDS into the AEW&C MST. One of the aims of NW-D3 was to demonstrate the integration of these technologies and methodologies with legacy systems. NW-D3 involved asynchronously publishing tracks from the AEW&CMST (using the DDS framework) and sinking these tracks to a number of visualisation devices across the network. Visualisation was provided through the Solipsys Tactical Display Framework, Google Earth and the Virtual Battlespace software in FOCAL, which enables 3D stereo visualisation of geo-spatial and track data.

6 NW-D10

The NW-D10 event [13,27] was held in 2010 and extended the NW-D3 event. Partners in NW-D10 were the AEW&C, ISR, FOCAL and Airborne Systems Connectivity Environment Laboratory (ASCEL) nodes. NW-D10 involved publishing airborne tracks and intelligence information derived from multiple sources to visualisation devices across the network. The aims of NW-D10 included demonstrating SOA in deployed systems, including those that operate at the tactical edge (AEW&C), and integrating SOA based systems that operate in difference domains (enterprise, real-time synchronous control and real-time publish/subscribe). The primary challenge was the horizontal integration of SOA technologies that were designed and implemented for operation in specific domains. Also, these technologies had to be integrated with both legacy and emerging mission systems.

NW-D10 fostered a small community of cross-disciplinary applied research that is of operational significance. The outcomes from NW-D10 are relevant to the capability integration and tactical SOA questions being considered by CIOG and other parts of Defence. The event illustrated that RT-CORBA, DDS and SOA technologies can be used to integrate components at a node, system-of-nodes and systems-of-systems levels. Furthermore, these technologies can potentially support system requirements from mobile computing devices to large scale operational platforms, providing an interoperable range of solutions at all levels within the system and network architecture. One of the current activities in the Net Warrior initiative is to integrate mobile technology to support CIOG's deployed SOA and emerging wireless communication technology requirements. Section 4 discusses the initial work conducted in this area.

7 Smartphone Integration in the Net Warrior Initiative

In late 2010 research under the Net Warrior initiative was extended to investigate how mobile computing devices could be integrated into tactical SOA environments (e.g. with restricted bandwidth and unreliable links). The goals of this research program are to investigate:

1. How mobile devices could be integrated into the Australian NCW environment.

2. Information interoperability between mobile devices and high value Defence assets (initially represented by an airborne mission system testbed) to enable increased situational awareness.
3. The utility of mobile devices for air-surface integration.

A smartphone (specifically a 16 GB iPhone 4) is being used as a representative mobile device but this could easily be a smartphone or tablet from a different manufacturer and with a different operating system (e.g. Android or Windows Mobile). The initial work conducted has investigated how distributed object and publish/subscribe middleware can be incorporated into a smartphone to achieve information interoperability. The middleware technologies chosen are the OMG RT-CORBA and DDS specifications discussed in Section 3.2. RT-CORBA and DDS are suited to low bandwidth tactical environments because utilisation of the underlying communication bearers can be tightly controlled. The initial implementations chosen to incorporate RT-CORBA and DDS on the smartphone are The Ace Orb (ACE)/Adaptive Communication Environment (TAO) and OpenDDS respectively because they are open source and there is significant experience with these technologies under the Net Warrior initiative.

ACE is the ADAPTIVE Communication Environment [28], an open-source object oriented software framework developed by the Centre for Distributed Object Computing at Washington University. ACE is infrastructure middleware that encapsulates and augments the capabilities of a wide range of operating systems and hardware architectures. It is particularly useful for developing portable, high performance and multithreaded applications, examples of which are airborne mission system software and smartphone applications. TAO [29], The ACE ORB, is distribution middleware built on top of ACE. TAO is a high performance RT-CORBA ORB that enables communication between distributed applications. OpenDDS [30] is an open-source C++ implementation of the DDS specification and builds upon ACE/TAO.

Current work on this research program involves investigating whether ACE and TAO are suitable for developing iOS applications. iOS is the iPhone and iPad operating system. Registered iOS developers can develop iOS applications using the iOS software development kit (Software Development Kit (SDK)) and run these applications on the iPhone simulator and hardware. Initially, several simple iOS applications were developed to become familiar with the iOS SDK. These applications were run on the iPhone simulator, however it was discovered that the simulator has limited utility as it does not emulate the iPhone hardware (i.e. the iPhone simulator is not an emulator). This means that testing on both the iPhone simulator and hardware is essential when developing applications. ACE and TAO have both been built successfully for the iPhone simulator and hardware and testing has shown that ACE and TAO operate as expected. A simple distributed application has been developed using ACE and TAO and its operation is currently being tested in the iPhone simulator and on the hardware.

In the near future, work will commence on building and testing OpenDDS for the iPhone simulator and hardware. A simple DDS application will be developed and its operation will be tested in the iPhone simulator and on the hardware. Once this is achieved, the focus will shift towards integrating the iPhone with an airborne mission system testbed through DDS to provide a simple air picture. Further research activities

could focus on end user interactive application development (i.e. chat) and incorporating camera and voice capabilities into future ISR applications.

8 Related Work

There has been a vast range of middleware developed for sensor networks. It is often argued that the kind of middleware applied to different types of sensor networks depends on the requirements of each network (e.g. computational power, battery life, how connected the network is). However, it is worrying that most middleware solutions for sensor networks are not aligned with any standards and there does not appear to be convergence on either middleware standards or implementations [31]. A small number of middleware implementations developed for sensor networks are based on well established standards, like RT-CORBA in [32] and RT-CORBA and DDS in [33]. However, rather than use existing middleware implementations (ACE/TAO and OpenDDS) these were either solely or partly handcrafted.

Also, much of the work on middleware for sensor networks has focussed on sensors that are homogeneous and have greatly constrained computational power and battery life [31]. These assumptions are not valid when considering sensor network applications in NCW environments [3,34]. For example, defence sensor networks can include large-scale surveillance systems (such as an AEW&C capability or over-the-horizon radar), unattended ground sensors and smartphones used for intelligence gathering and targeting.

Over recent years the computation, communication and sensing capabilities in smartphones and tablets has become increasingly powerful. While it is still important to be aware of resource consumption when employing middleware on smartphones, they are now relatively powerful mobile computing platforms [35]. Therefore, it is now appropriate to investigate the application of existing high performance middleware technologies (ACE/TAO and OpenDDS) to smartphones. These technologies are based on open standards and are either in use or interoperable with middleware technologies in operational ADF platforms and systems. This is important for information interoperability in NCW environments because it reduces the risk of integrating disparate sensors and systems. While some approaches implement custom middleware for sensor networks in NCW environments [36], this should not be necessary when widely accepted high performance middleware standards like RT-CORBA and DDS exist. Therefore, investigating information interoperability between mobile devices and high value ADF platforms and systems using existing high performance middleware technologies is an important research problem.

9 Conclusion

This paper has argued that there is strong motivation for investigating interoperability between mobile technology and defence tactical platforms and systems in support of NCW implementation. Further, it has been shown that there is a need to investigate how existing high performance middleware technologies based on open standards can be incorporated into mobile computing devices. In support of this, research under the

Net Warrior initiative has been extended to incorporate the investigation of smartphone integration into the Australian NCW environment.

Acknowledgement. The author thanks Derek Dominish of DSTO for the many discussions regarding the technical content of this paper.

References

1. Stein, F.: Network-centric warfare principles, new initiatives and operational examples. In: Proceedings of NCW Conference (2010)
2. Cebrowski, A.K., Garstka, J.J.: Network-centric warfare: its origin and future. U.S. Naval Institute Proceedings Magazine (January 1998)
3. Hosmer, C., Jeffcoat, C., Davis, M., McGibbon, T.: Use of mobile technology for information collection and dissemination. Technical Report 518055, The Data and Analysis Center for Software, DACS (2011)
4. DARPA: Broad agency announcement: Transformative apps. Technical Report DARPA-BAA-10-41, DARPA (2010)
5. Gould, J., Hoffman, M.: Army sees many smart phones playing important role. Army Times, Gannett Government Media, Springfield, VA (2010)
6. Lopez, C.T.: 'Apps for Army' to shape future software acquisition. Army Article, Media Relations Division, Chief of Public Affairs Washington, DC (2010)
7. Directorate of Future Warfighting: Enabling future warfighting: Network centric warfare. Technical Report ADDP-D.3.1, Department of Defence, Canberra, Australia (2004)
8. Director of General Capability and Plans: Explaining NCW. Technical report, Department of Defence, Canberra, Australia (2006)
9. Director of General Capability and Plans: NCW Roadmap. Technical report, Department of Defence, Canberra, Australia (2007)
10. Department of Defence: Defending Australia in the Asia Pacific Century: Force 2030, Defence White Paper. Technical report, Department of Defence, Canberra, Australia (2009)
11. Chief Information Officer Group: Science and Technology Plan FY 2011-2012. Technical report, Department of Defence, Canberra, Australia (2011)
12. Department of Defence: Single Information Environment (SIE): Architectural Intent. Technical report, Department of Defence, Canberra, Australia (2010)
13. Dominish, D., Sioutis, C., Baillie, C., Temple, P., Foster, K.: Net Warrior D10 techology report: AOD's AEWC and data link nodes. Technical Report DSTO-TR-2567, DSTO, Edinburgh, Australia (2011)
14. Foster, K., Iannos, A., Lawrie, G., Temple, P., Tobin, B.: Exploring a net centric architecture using the Net Warrior AEW&C node. Technical Report DSTO-TR-2093, DSTO, Edinburgh, Australia (2007)
15. Strei, T.J.: Open architecture in naval combat system computing of the 21st century: network-centric applications. In: Proc. of the 8th ICCRTS (2003)
16. OMG: Common Object Request Broker Architecture Specification. Object Management Group, version 3.0 (2002)
17. Schmidt, D., Kuhns, F.: An overview of the real-time corba specification. Computer 33(6), 56–63 (2000)
18. Schmidt, D.: Overview of CORBA, Washington University, St. Louis (2006), `http://www.cse.wustl.edu/~schmidt/corba-overview.html`

19. Bachmann, F., Bass, L., Buhman, C., Comella-Dorda, S., Long, F., Robert, J., Seacord, R., Wallnau, K.: Volume II: technical concepts of component-based software engineering. Technical Report CMU/SEI-2000-TR-008, ESC-TR-2000-007, Carnegie Mellon Software Engineering Institute, Pittsburgh (2000)
20. Newcomer, E.: Understanding Web Services. Addison-Wesley (2002)
21. Department of Defence: Defence Information and Communications Technology Strategy. Technical report, Department of Defence, Canberra, Australia (2009)
22. Sioutis, C., Dominish, D.: Developing intelligent agents with distributed computing middleware. In: Proc. KES-IDT (2011)
23. Alberts, D.C., Hayes, R.E.: Power to the Edge. DoD C4ISR Cooperative Research Program, Washington DC, United States (2003)
24. OMG: Data Distribution Service for Real-Time Systems Specification. Object Management Group, version 1.1 (2005)
25. Schmidt, D., Corsaro, A., van't Hag, H.: Addressing the challenges of tactical information management in net-centric systems with DDS. CrossTalk, 24–29 (March 2008)
26. Sioutis, C., Foster, K., Dominish, D., Temple, P.: Achieving information interoperability using data distribution middleware. In: Proc. MilCIS, Canberra, Australia (2008)
27. Kilmartin, D.: Net Warriors avert airborne threat. DSTO Connections 155, 7 (2011)
28. Schmidt, D.: The ADAPTIVE Communication Environment (ACE). Washington University, St. Louis (2011), `http://www.cse.wustl.edu/~schmidt/ACE.html`
29. Schmidt, D.: Real-time CORBA with TAO (The ACE ORB). Washington University, St. Louis (2011), `http://www.cse.wustl.edu/~schmidt/TAO.html`
30. Object Computing: OpenDDS. St. Louis, MO (2011), `http://www.opendds.org`
31. Henricksen, K., Robinson, R.: A survey of middleware for sensor networks: state-of-the-art and future directions. In: Proc. MidSens 2006, pp. 60–65 (2006)
32. Gill, C., Subrarnonian, V., Parsons, J., Huang, H., Torri, S., Niehaus, D., Stuart, D.: ORB middleware evolution for networked embedded systems. In: Proceedings of the Eighth International Workshop on Object-Oriented Real-Time Dependable Systems (WORDS), pp. 169–176 (2003)
33. Boonma, P., Suzuki, J.: Toward interoperable publish/subscribe communication between wireless sensor networks and access networks. In: 6th IEEE Consumer Communications and Networking Conference, CCNC 2009, pp. 1–6 (2009)
34. Chong, C.Y., Kumar, S.: Sensor networks: evolution, opportunities, and challenges. Proceedings of the IEEE 91(8), 1247–1256 (2003)
35. Riva, O., Kangasharju, J.: Challenges and lessons in developing middleware on smart phones. Computer 41(10), 23–31 (2008)
36. Moon, Y.W., Jung, H.S., Jeong, C.S.: Context-awareness in battlefield using ubiquitous computing: Network centric warfare. In: International Conference on Computer and Information Technology, pp. 2873–2877. IEEE Computer Society (2010)

Computational Approach for Measuring the Tear Film Break-Up Time in an Unsupervised Manner

Lucía Ramos[1], Noelia Barreira[1], Antonio Mosquera[2], Manuel Currás[1], Hugo Pena-Verdeal[3], María Jesús Giráldez[3], and Manuel G. Penedo[1]

[1] VARPA Group, Department of Computer Science, Univ. of A Coruña, Spain
{l.ramos,nbarreira,manuel.curras,mgpenedo}@udc.es
[2] Artificial Vision Group, Department of Electronics and Computer Science, Univ. of Santiago de Compostela, Spain
antonio.mosquera@usc.es
[3] Optometry Group, Dept. Applied Physics, Univ. Santiago de Compostela, Spain
hugo.pena@rai.usc.es, mjesus.giraldez@usc.es

Abstract. Dry eye syndrome is a common disorder of the tear film, affecting a significant percentage of the population. The Break-Up Time (BUT) is a clinical test used for the diagnosis of this disease. In this research, it is proposed an automatic methodology to evaluate the BUT test. This methodology locates the different measurement areas from a video of the tear film, extracts the Region Of Interest (ROI) and performs the BUT test in each measurement area. Furthermore, it is independent of some specific features of each video such as the eye size, the intensity variation, or the starting point of the measurement frame sequence. This methodology has been tested on a dataset composed of 18 videos that have been annotated by four different experts. The average difference between the automatic measurement and the experts' measures is on the acceptable range considering the high inter-observer variance.

Keywords: Tear Film, Dry Eye Syndrome, Break-Up Time (BUT) Test, Video Analysis, Image Processing.

1 Introduction

The tear film is a trilaminar structure comprising an outer lipid layer, a middle aqueous layer and an inner mucous layer [1]. It provides a smooth optical surface by compensating for the micro irregularities of the corneal epithelium and plays an essential role in the maintenance of ocular integrity by removing foreign bodies from the front surface of the eye. It contains bacteriostatic substances that inhibit the growth of microorganisms, and reduces the surface friction associated with eyelid blinking and eye movement. Each tear film layer has a specific role in the formation and stability of the tear film. The quality and thickness of each layer, as well as their adequate interaction, are important in order to have a stable tear film. Abnormalities in any of the layers can cause tear dysfunction problems, as the Dry Eye Syndrome (DES).

The DES is a common disorder of the tear film, which affects a significant percentage of the population, especially among contact lenses users [2, 3]. Furthermore, it worsens

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 254–267, 2013.

with age. The prevalence of this syndrome has been increasing in recent years, affecting up to 10-15% of normal population, and 18-30% of contact lenses users. Several factors, such as adverse environmental conditions, use of certain medications, or visual tasks that reduce blink rate, have contributed to that increment [4, 5]. This disease results in symptoms of discomfort and visual disturbance affecting the common vision related daily activities. Pain and irritate symptoms disrupts negatively with the welfare of the patient. Also, DES affects ocular and general health. Perception and visual function may be reduced, which impact on visual performance. The diagnosis of this condition is extremely difficult due to its multifactorial etiology and the variability of the clinical tests [6].

The study of the tear film stability is essential for the dry eye characterization [7]. There are several clinical tests to assess the quality and stability of the tear film on the ocular surface [8–10]. The Break-Up Time (BUT) measure is one of the tests most commonly used in clinical practice. To perform this test, sodium fluorescein is instilled into the eye, and the tear film is observed with the help of cobalt-blue filter attached to a slit-lamp biomicroscope, while the patient avoids blinking. The BUT is measured as the time elapsed until a dark area appears on the tear film. It corresponds to a thinning of the tear film, which identifies the break-up. This test presents a considerable variability, so the automation would reduce its subjective character.

The automation of the BUT measure is a little explored field. Yedidda et al. [11, 12] proposed a multi-step algorithm for this purpose. In this approach, they locate the iris in the first frame, then, they align the consecutive frames, and, finally, they scan the aligned video to find the dry areas. A preliminary approach was conducted in [13], which consists of locating the different measurement areas, extracting the Region Of Interest (ROI), and performing the BUT test in each measurement area.

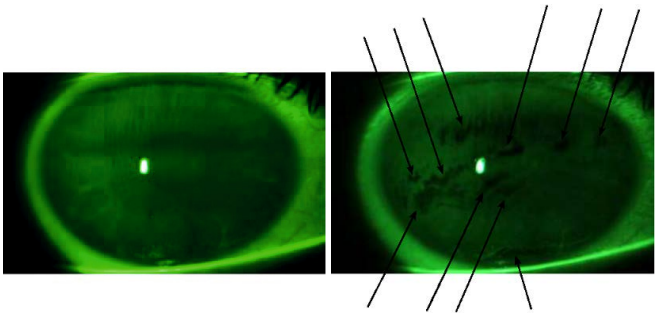
In this work, this approach is extended with several improvements oriented to obtain a fully automatic methodology, independent of the size and shape of each patient's eye. Furthermore, this procedure should provide real time results to be used in the clinical practice. This methodology has been tested on a dataset with tear film videos provided by the Escuela Universitaria de Óptica y Optometría of the Universidad de Santiago de Compostela. These videos have been recorded with a Topcon DV-3 camera attached to the Topcon SL-D4 slit lamp.

This paper is organized as follows. Section 2 describes the methodology as well as the improvements proposed to automatize the computation. Section 3 details the case study considered. Section 4 summarizes the results obtained with this methodology. Finally, Section 5 presents the conclusions and future work.

2 Techniques

This methodology has been tested on tear film videos where the tear is stained by green due to fluorescein, as shown in Fig. 1 (a). Each video has a duration of several minutes and contains different BUT tests. For each of these tests it is analyzed the appearance of dark spots, as shown in Fig. 1 (b), which correspond to the break-up and indicate tear film disruption.

The methodology for measuring the BUT has several steps [13, 14], as shown in Fig. 2. First, the different measurement areas contained in a tear film video are automatically



(a) (b)

Fig. 1. (a) Tear film stained by fluorescein. (b) Formation of dark spots points related to the tear film break-up.

located. Then, the ROI is extracted within each frame. Finally, the BUT test is conducted in each measurement area. The different stages are explained in more detail in the following sections.

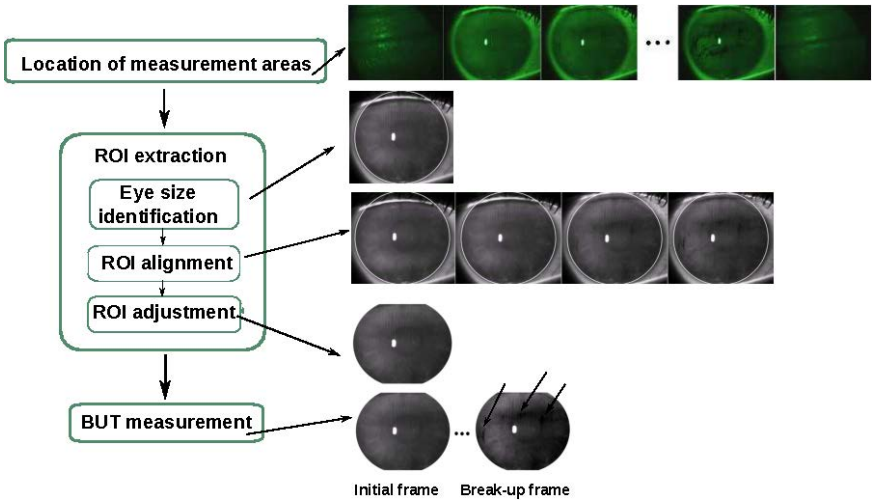


Fig. 2. Steps of the methodology for the automatic BUT measurement

2.1 Location of Measurement Areas

Each tear film video contains several measurement areas separated by blinks. The blinks delimit the beginning and the end of those measurement areas. These blinks are related to frames with lower intensities than those corresponding to the open eye. Therefore, the detection of blinks is based on calculating the finite differences of mean values of gray between consecutive frames, and then applying a threshold. Sometimes the blink

occurs gradually and the differences between successive frames may not be enough to detect it. In order to detect all blinks, the symmetric finite differences every two frames are computed and added to the differences between consecutive frames, as calculation in Eqs. (1) and (2), where $\overline{G}_i$ is the mean value of gray for the frame i.

$$d_i = dif(i,1) + dif(i,2) \tag{1}$$

$$dif(i,d) = \overline{G}_{i+d} - \overline{G}_i \tag{2}$$

This sum emphasizes that the intensity changes are produced during a blink. On these differences, a negative peak represents the beginning of a blink, since there is a transition from a lighter frame (open eye) to a darker frame (closed eye). Similarly, at the end of the blink there is a transition from a darker frame to a lighter frame, producing a positive peak, as shown in Fig. 3. In order to identify these peaks, it is get

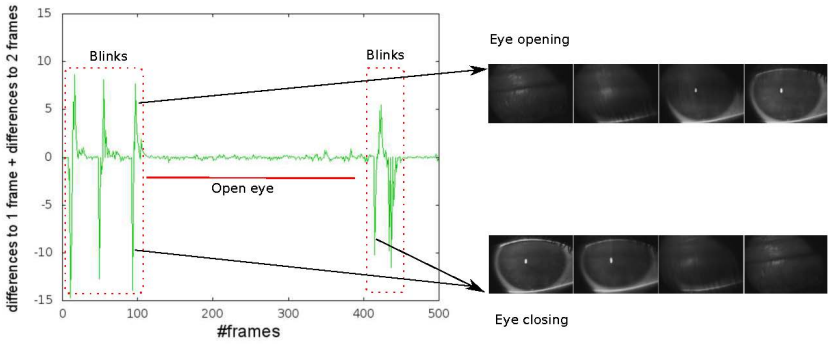


Fig. 3. Symmetric finite differences of mean values of gray. Peaks correspond to blinks and the flat areas are related to frames where the eye is open.

an adaptive threshold t_b according to the differences obtained for each video. To this end, it is computed the mean of the sum of differences discarding those values minor than 1, as calculation in Eq. (3), since they are related with slight variations.

$$t_b = mean\{d_i, 0 \leq i < frames, d_i > 1\} \tag{3}$$

The range $(-t_b, t_b)$ obtained from this threshold is used to identify transitions from open to closed eye and vice versa. Thus, values outside this range are related to blinks, while values inside this range correspond to the areas where the eye is open.

Therefore, the measurement areas are identified as those intervals starting with a positive difference and ending with a negative difference. Also, they should exceed a threshold to ensure that the measurement area has a minimum of frames. Sometimes there could be two consecutive differences with the same sign. This occurs when the lamp is off or when there are semi-blinks in the measurement area. In this case, the lowest absolute values are removed until all pairs of consecutive blinks have opposite signs. Fig. 4 shows the areas automatically detected by this approach in a tear film video sequence.

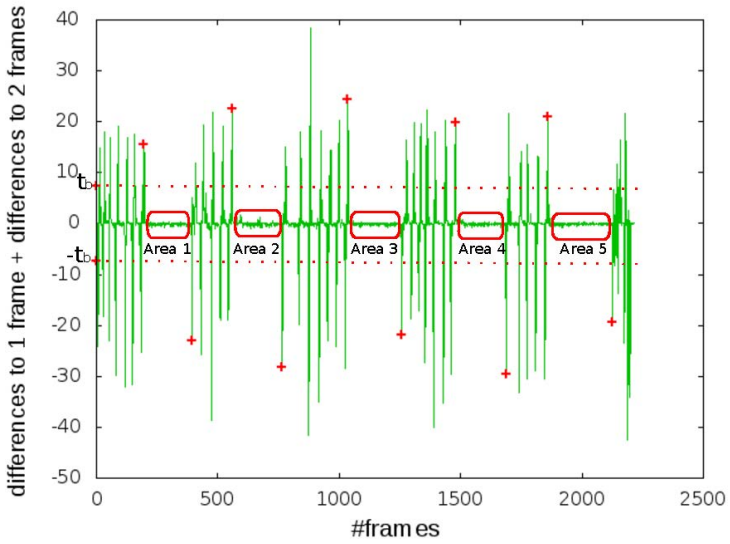


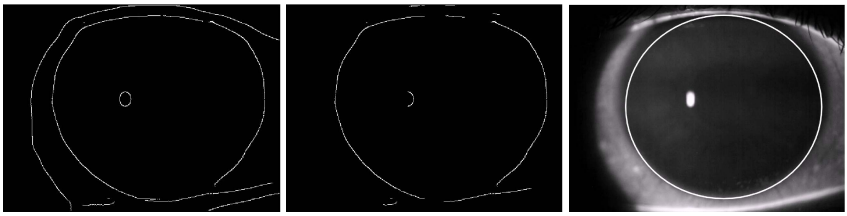
Fig. 4. Automatic detection of the different measurement areas from the sequences between peaks with opposite signs

2.2 ROI Extraction

Once the measurement areas are located, the next stage is extracting the ROI within each frame to discard regions without relevant information. This step begins with the eye size identification. To this end, the frames placed at 25%, 50% and 75% of each measurement area are selected to have different samples of the eye throughout the sequence. It is applied the Canny edge detector [15] to the green component of the Red, Green, Blue (RGB) image of these frames and then it is computed the correlation in the frequency domain between each edge image and a set of masks.

The natural structure of the eye causes the original frames contain a circular region surrounding the iris. In order to avoid mismatches with this region, it is considered the orientation of the gradient of the Canny edges. Thus, instead of using the original edge image, as Fig. 5 (a) shows, the outgoing edges in relation with the center are used, as Fig. 5 (b) shows. Since the iris has almost a circular shape, circles and ellipses are used as masks. It is selected a range of radius covering the typical eye sizes to create a subset of circular masks. In addition, each of these radii is slightly extended in the horizontal axis to create another subset of elliptical masks in order to cover the maximum possible ROI. Then, the maximum correlation in the frequency domain among the three edge images and the different masks is analyzed to determine the best fit. The largest value of these correlations corresponds to the optimal radius and, therefore, the size of the eye as Fig. 5 (c) shows. This size is fixed throughout the sequence.

The eye presents slight motions throughout the video, so it is necessary to register the ROI in each frame. It is performed an alignment by correlation in the frequency domain of the edge images of each frame in relation with the immediately preceding frame. After this, the mask selected in the previous stage is applied to the aligned frames. This



(a) (b) (c)

Fig. 5. ROI extracted by correlation. (a) Edge image obtained by Canny edge detector (b) Edge image discarding incoming pixels (c) Eye size obtained from the best correlation value.

way, the methodology is independent of slight motions of the ROI, since since the edge alignment between consecutive frames produces a good match.

In some cases, the eye is not fully open and the ROI contains outer parts like eyelids or eyelashes. These elements can disrupt the results so it is performed a ROI adjustment. Due to each eye has a different shape and openness degree, it is proposed an adaptive adjustment taking into account the features of each eye to discard regions without relevant information. Therefore, an upper and a lower limit are calculated to crop the ROI at the top and at the bottom, respectively. The number of edges is computed in each row i, as calculation in Eq. (4). Then it is selected the upper limit as the furthest row in the upper half of the image that accumulates more edge points than a variable threshold, as calculation in Eq. (5). In the same way, the lower limit is the closest row in the lower half of the image that accumulates more edge points than a variable threshold, as calculation in Eq. (6).

$$Acc(i) = \sum_{j=i}^{cols} edges(i,j) \tag{4}$$

$$l_{upper} = max\{i, Acc(i) > t_{upper}\} \tag{5}$$

$$l_{lower} = min\{i, Acc(i) > t_{lower}\} \tag{6}$$

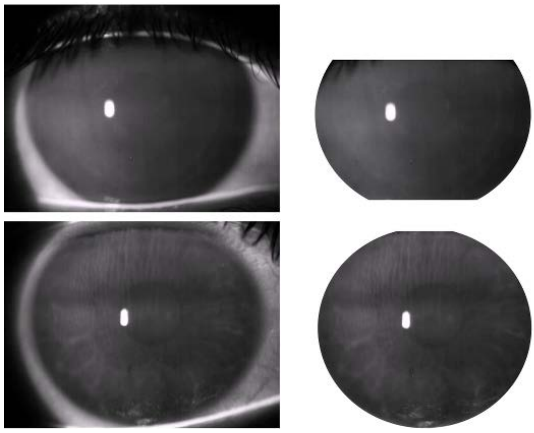
The thresholds are computed as a percentage of the maximum number of edge points found in each half of the image. Given the nature of the eye, the upper eyelid and eyelashes usually invade more ROI than the lower, so the parameters α and β are used as percentages for the upper and lower threshold, as calculation in Eqs (7) and (8).

$$t_{upper} = \alpha max\{Acc(i), 0 \leq i < rows/2\} \tag{7}$$

$$t_{lower} = \beta max\{Acc(i), rows/2 \leq i < rows\} \tag{8}$$

Furthermore, the radius is slightly reduced to get rid of noise at the boundaries of the iris.

Figure 6 shows the adjustment of the ROI for two measurement areas, one with the eye slightly closed and the other with the eye fully open. The first column shows the



(a) (b)

Fig. 6. Adjustment of the ROI. (a) Original frames. (b) The adaptive adjustment discards irrelevant elements.

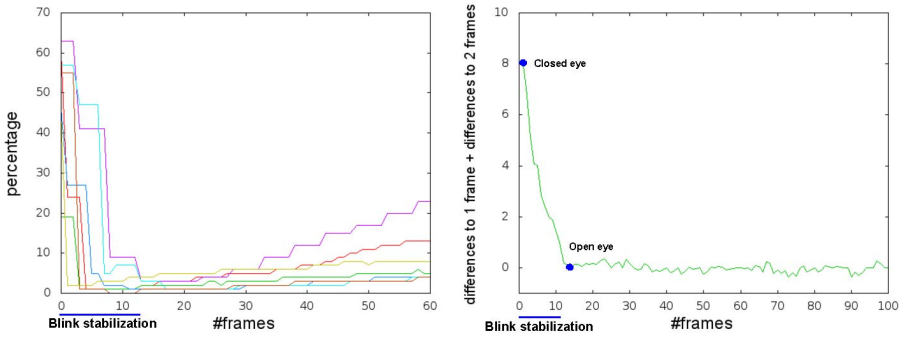
original images, whereas the second is the result of the adaptive adjustment. In the first case, the eye is a bit closed and the the adaptive adjustment discards these outer parts. However, in the second case the eye is fully open, so the adaptive adjustment covers more ROI.

2.3 BUT Measurement

The last step of the methodology is the computation of the BUT measure, that is, the time elapsed since the last blink, until a dark area appears on the surface of the eye. The BUT is obtained from the evolution curves of the percentage of black through of each measurement area, from the eye fully open until the final blink.

At the beginning of each measurement area there are several frames after the initial blink in which the patient is still opening the eye. In order to discard these frames of the intensity analysis, a reference frame is located at the beginning of the the measurement area, after the first blink. The blink duration can vary several frames, as can be seen in Fig.7 (a), so an adaptive reference frame is selected for each measurement area. To this end, it is used the sum of differences computed for locating the measurement areas. Only are considered the values for a subset of frames at the beginning of the measurement area since they are related to the end of the blink and the stabilization of the tear, as Fig.7 (b) shows. The evolution of the curve begins with a peak related to the eye fully closed. The value decreases while the eye is opening. It reaches a minimum, value that corresponds to the eye fully open. The frame related to the minimum of this curve is selected as the reference frame. That is the point where the analysis of the evolution of fluorescein in the tear film begins.

Each tear film video presents variations in the illumination and the amount of fluorescein instilled, so not all measurement areas have the same intensity levels. Furthermore, the dark pixels at the break-up vary in a range of values close to zero according to lighting conditions, but not exactly zero. For these reasons, a threshold is set to determine



(a) (b)

Fig. 7. Stabilization after a blink. (a) The number of frames involved in the stabilization varies in the measurement areas. (b) Evolution of the sum of differences during a blink.

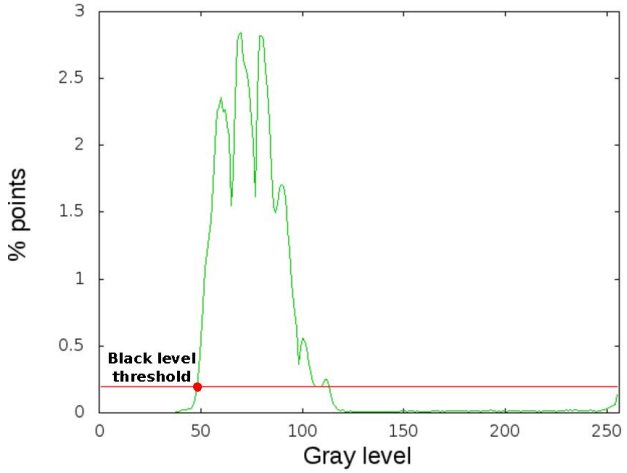
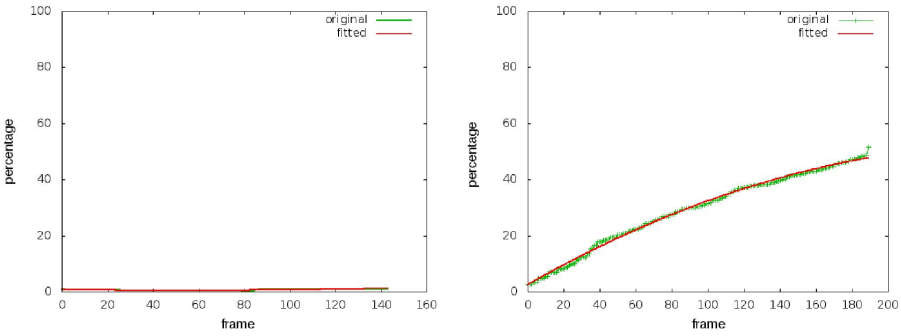


Fig. 8. Histogram of the reference frame. The black threshold is obtained as the highest gray level of a percentage of the darkest pixels.

the range of black in each measurement area, that is, the maximum intensity for a pixel to be considered as black. To this end, it is created the histogram of the reference frame of each measurement area and the gray levels of a percentage of the darkest pixels are analyzed, as shown in Fig. 8. The black level threshold corresponds to the largest value of these gray levels. For example, pixels with values below this threshold area are considered as black.

Therefore, the evolution curve is built using the black level threshold to get the percentage of black for each consecutive frame. Small variations produce curves with irregular slopes, so a curve fitting is performed to discard these fluctuations. A second order polynomial function has been selected for this adjustment since it provides a good fit,

as shown in Fig. 9. In some cases, this curve is virtually zero because the tear does not break-up in the interval as shown in Fig. 9 (a). On the contrary, if there is a measure, the percentage of black increases with the time since the fluorescein is not regenerated, as shown in Fig. 9 (b). In order to determine the BUT measure, another threshold is obtained from a percentage of the total height of the evolution curve. The BUT measure is computed as the time elapsed from the beginning of the measurement area until the curve exceeds this threshold.



(a) (b)

Fig. 9. Fitted evolution curve by second order polynomial function. (a) There is no measure. (b) The black area increases with the time.

3 Case Study

The methodology has been tested on a dataset of tear film videos. The dataset consists of 18 videos from healthy patients with ages ranging from 19 to 33, varying from very dry eye to an eye with no visible dryness. These videos have been manually annotated by four different experts.

The videos are divided in two subsets, one for training and the other for testing. The first contains 12 videos with 44 positive measures and 20 areas where there is no measure. The second set contains the remaining 6 videos, with 23 areas with measure and 30 areas without measure.

This methodology has been developed in C++ using the Open Computer Vision library (OpenCV) for performing some video and image processing operations. The development and testing operations have been conducted in a Linux operating system running on an Intel Pentium Quad processor at 2.33GHz and 4 GB of RAM. In order to validate the methodology, first it is tested the detection of the different measurement areas. Then, the accuracy of the BUT detection is validated in relation with the values provided by the four different experts. Finally, a time analysis is conducted to check the methodology can be performed in real time.

4 Analysis of Results

The videos of the training set have been used to adjust the different parameters of the methodology. Therefore, the values of α and β for adjusting the ROI were experimentally set at 0.65 and 0.8, respectively. The percentage of black pixels considered as black was set at 0.2%, and the percentage of the maximum in the evolution curve to identify the break-up was set at 60%.

The location of the measurement areas has been validated by comparing the areas detected by the system with the areas delimited by the experts. Table 1 shows the confusion matrix for the location of the different measurement areas on the entire dataset. As can be seen, the automatic location works well detecting areas with BUT measurement such as discarding areas where there is no measure. Furthermore, this approach defines with accuracy the beginning and the end of each measurement area.

Table 1. Confusion matrix between measurement areas automatically detected and those delimited by the experts

		System	
		Measure	No measure
Expert	Measure	94%	6%
	No measure	8%	92%

In order to validate the accuracy of the BUT detection, differences between the value annotated by each expert and the average of all them have been analyzed. Figure 10 shows the dispersion between the expert average and each individual expert measure. The error is located mainly in an interval of ± 2.5 seconds due to the subjectivity of this measure.

Figure 11 shows the dispersion between the expert average and the BUT measurement obtained with the automatic methodology. This value is in agreement with the manual detection, since it is within the same range as among the experts themselves. The 96% of the differences among the experts are in an interval of ± 2.5 seconds. In the automatic methodology, the percentage of values within this range was 94% and the mean deviation was 1.26 seconds.

Besides the tear film break-up time, the evolution curves provide additional information about the amount of break-up region on each measurement area. Figure 12 shows examples of three different measurement areas, showing the frames placed at 50%, 75% and 100% as well as the evolution curve associated to the entire sequences. In the first two instances, the break-up area is small so the percentage of black is low. However, in the third case, there is a massive break-up so the black percentage is high and the evolution curve increases fast.

Each tear film video has a duration of several minutes and contains an average of four different BUT tests. The execution time of each entire video is about 3.5 minutes, in which the detection of the measurement areas takes about 16 seconds and the BUT computation for each measurement area lasts about 50 seconds. Therefore, this method provides real time results that can be applied in the clinical practice.

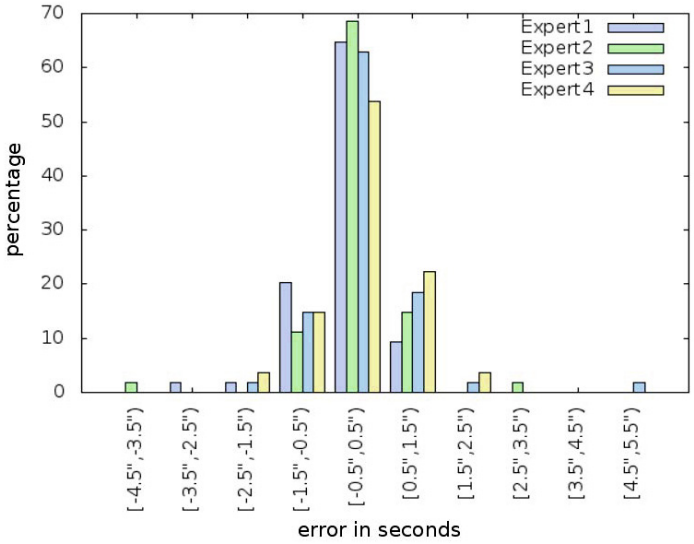


Fig. 10. Differences between the value annotated by each expert and the average expert BUT measurement. Most of the measures have an error in an interval of ± 2.5 seconds.

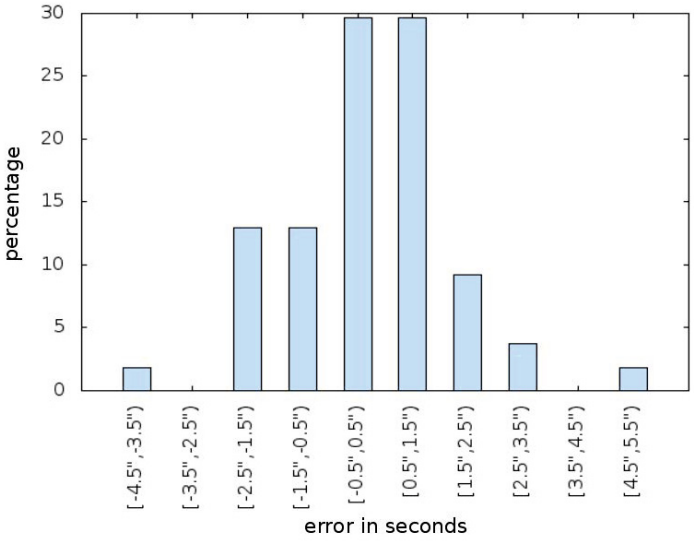


Fig. 11. Difference in seconds between the expert average and the automatic BUT measurements.

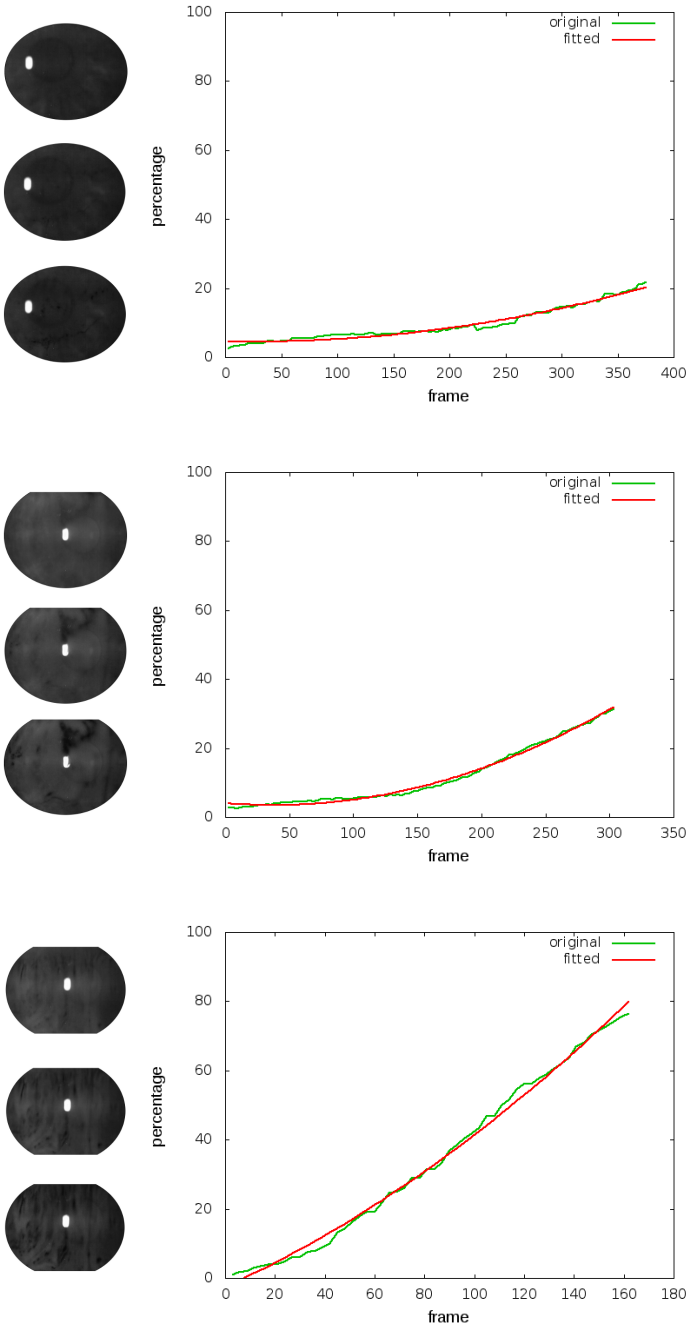


Fig. 12. Black evolution curves of three different measurement areas. The height of the curve is related to the amount of break-up area on the ROI. Higher curves correspond to large tear film break-up zones and vice versa.

5 Conclusions and Future Work

In this work, it is presented a methodology for the automatic computation of the break-up time test in tear film videos. This approach is adaptive according to some specific features of each video such as lighting conditions as well as the shape and eye size. The procedure first identifies the different measurement areas located between consecutive blinks. Then, the ROI is extracted within each frame taking into account the eye features. Finally, the BUT measure is computed by analyzing the evolution of the intensities through the sequence. This methodology has been tested on a dataset consisting of 18 tear film videos achieving results on an acceptable range considering the high inter-observer variance.

Future work in this field includes a local analysis to study the dryness degree at each area of the eye. In order to analyze other parameters for the diagnosis of the dry eye syndrome, such us position, shape or evolution of the break-up of the tear film.

Acknowledgement. This work has been partially funded by the Ministerio de Ciencia e Innovacin and the Instituto de Salud Carlos III through the research projects PI10/00578 and TIN2011-25476.

References

[1] Holly, F., Lemp, M.: Tear film physiology and dry eyes. Surv. Ophthalmol 22, 69–87 (1977)
[2] Lemp, M.: Advances in understanding and managing dry eye disease. Am. J. Ophthalmol. 46, 350–356 (2008)
[3] Lowther, G.: Dryness, tears and contact lens wear. Reed Elsevier, USA (1977)
[4] Fenga, C., Aragona, P., Cacciola, A., Spinela, R., Di Nola, C., Ferreri, F., Rania, L.: Melbonian gland dysfunction and ocular discomform in video display terminal workers. Eye 22, 91–95 (2008)
[5] García-Resúa, C., Lira, M., Yebra-Pimentel, E.: Evaluación superficial en jóvenes universitarios. Rev. Esp. Contact 12, 37–41 (2005)
[6] Khanal, S.: Dry eye diagnosis. Invest. Opthalmol. Vis. Sci. 49, 1907–1914 (2008)
[7] Guillon, J.P.: Non-invasive tearscope plus routine for contact lens fitting. Contact Lens and Anterior Eye 2, S31–S40 (1998)
[8] Jin Hak Lee, M.D., Chang Won Kee, M.D.: The significance of tear film break-up time in the diagnosis of dry eye syndrome. Kor. J. Opthalmol. 21, 69–71 (1988)
[9] Cho, P., et al.: Reliability of the tear break-up time thechnique of assessing tear stability and the locations of thear break-up in hong kong chinese. Optom. Vis. Sci. 69, 879–885 (1992)
[10] Cho, P., Brown, B.: Review of the tear break-up time and a closer look at the tear break-up time of hong kong chinese. Optom. Vis. Sci. 70, 30–38 (1993)
[11] Yedidya, T., Hartley, R., Guillon, J.P.: Automatic detection of pre-ocular tear film break-up sequence in dry eyes. In: DICTA, pp. 442–448 (2008)
[12] Yedidya, T., Carr, P., Hartley, R., Guillon, J.-P.: Enforcing monotonic temporal evolution in dry eye images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part II. LNCS, vol. 5762, pp. 976–984. Springer, Heidelberg (2009)

[13] Cebreiro, E., Ramos, L., Mosquera, A., Barreira, N., Penedo, M.G.: Automation of the tear film break-up time test. In: ISABEL (2011)
[14] Ramos, L., Barreira, N., Mosquera, A., Currás, M., Pena-Verdea, H., Giráldez, M.J., Penedo, M.G.: Adaptive parameter computation for the automatic measure of the tear break-up time. In: 16th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES 2012), vol. 243, pp. 1370–1379 (2012)
[15] Canny, J.F.: A computational approach to edge detection. IEEE PAMI, 679–698 (1986)

Analysis of the Job Categories of the New Japanese Information Technology Skills Standards

Rasha El-Agamy and Kazuhiko Tsuda

[1] Department of Risk Engineering, Faculty of Systems and Information Engineering, University of Tsukuba, Japan
r24644001@risk.tsukuba.ac.jp
[2] Graduate School of Business Sciences, University of Tsukuba, Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
tsuda@gssm.otuka.tsukuba.ac.jp

Abstract. Organizations operate in an increasingly competitive environment, which drives a need for continuous employee skill development. The rapid pace of technological change requires everyone to engage in life-long learning. The importance of the information technology services industry is growing year by year, and it would not be exaggerating to say that it is playing a major role in placecountry-regionJapan's industry. So, the Japanese government has published the documents that define the required knowledge about IT. These documents are called the Information Technology Skill Standards (ITSS) . The ITSS documents define 11 job categories and 35 special fields. In order to learn efficiently, it is indispensable to discern what is important for targeted for learning. This paper analyzes the Japanese skill standards using text mining methods. These methods were used to extract the keywords and to compute the similarity between the different job categories of skill standards. This type of analysis has not made intensively, such as clustering the skill standards' job categories and the required skills to change engineer's career. For these backgrounds, the authors made an intensive research with the eleven job categories of the Japanese information technology skill standards published by the Japanese ministry of economy, trade and industry. From the results of the research, the authors have succeeded in proposing a method that enables the engineers to identify the required keywords to move from job category to another. Also, high weighted keywords were used to sort the required learning courses for any job category. The authors think that this method will make it easy to the engineers to know the priority of the required learning courses in every job category.

Keywords: Information Technology, Job Categories, Clustering, Cosine Similarity, Learning, Skill Standards.

1 Introduction

IT is now an integral part of economic activities and underpins the social infrastructures. Competitive pressures are an issue for employees and employers alike. To be successful, IT workers must make themselves as valuable as possible to be hired by companies. So, due to the increase in use of IT in the society and increase of importance of software

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 268–280, 2013.

and services in the IT industry, the necessity for developing human resources with high level IT skills has been rising [1,2,3][1-3]. In order to support the competitiveness of the IT industry, which derives the information society, the Japanese government promoted a set of IT Skill Standards Information Technology Skill Standards (ITSS) that clarify and systemize the required skills-set of IT service personnel[1].

This paper analyzed the documents of ITSS using text mining techniques [4,5][5-6]. In order to improve accuracy, it is necessary to define the weight of the extracted keywords. The higher weighted keywords indicate the significance and priority of learning the associated skills. Keywords with a high weight are important for identifying the required skill for different job categories. So, it is necessary to learn the skills related to the high weight keywords. Using these high weight keywords, the required learning courses for ITSS job categories were sorted. This sorting allows IT engineers to know what are the most important courses have to learn first, so they can concentrate on studying the most needed courses for their career instead of wasting time in studying not important courses for them. However, the clarification of the required skills for each job category is important for IT human resources, the clarification of the skills required to move between different job categories is also important. The authors have suggested a method that derives the required keywords to move between different job categories.

This paper is organized as follows. Section 2 explains ITSS. Section 3 discusses the process outline. Section 4 presents the experiments and results. Finally, Section 5 concludes the paper.

2 ITSS Description

Under the current circumstances where Information Technology (IT) is widely recognized in society as an infrastructure essential for economic activities and people's lives, country-regionJapan faces an urgent issue, development of advanced IT human resources who will play a leading role in enhancing the international economic competitiveness of country-regionplaceJapan and supporting the healthy development of social systems. In response to the awareness of this issue, "Skill Standards for IT Professionals" (ITSS) was published by the Ministry of Economy, Trade and Industry (METI) in December 2002 [4]. Since then, it has come into widespread use as indices of skills of human resources among companies in the IT service industry. In order to promote spread and utilization of ITSS, the IT Skill Standards Centre was established in the Information Technology Promotion Agency (IPA) by METI in July 2003 [7]. With cooperation of Professional Committee, IPA continues to reinforce enrichment of ITSS, training and development, and assessment of IT professionals by utilizing ITSS. METI establishes IT skill standards as a measure to clarify and systematize actual ability necessary for providing IT services, and promotes them as a framework for human resource development in both private companies and schools.

Specifically, ITSS is a set of systematic indices that clarify and systemize the skills needed for people working in the IT services industry. IT skill standards define the professional job-related knowledge, skills, and abilities required to succeed in the digital-age

[1] See `http://www.ipa.go.jp/english/humandev/forth_download.html`

workplace. They can be used as a foundation tool for developing educational curriculum, profiling jobs, recruiting and evaluating employees, and designing academic and professional certification. They can be used alone or in conjunction with other input, such as that from a subject matter expert, industry advisory committee, professional organization, existing academic or vendor-specific curriculum, or accrediting organization. ITSS is utilized as a tool for developing professional human resources to implement corporate strategies. Organized into a career framework, ITSS classifies the information services industry into eleven job categories and 35 specialty fields. In each field, there are seven levels based on individual experience and results as shown in Figure 1.

3 Research of ITSS and Distance Computation

This research focused on the clustering of ITSS job categories. Then making a skill-up road map that determined the study or education (skills/abilities) required changing the career path of any human resource or which skills required moving between different job categories. In this research each job category is represented by a text document. As

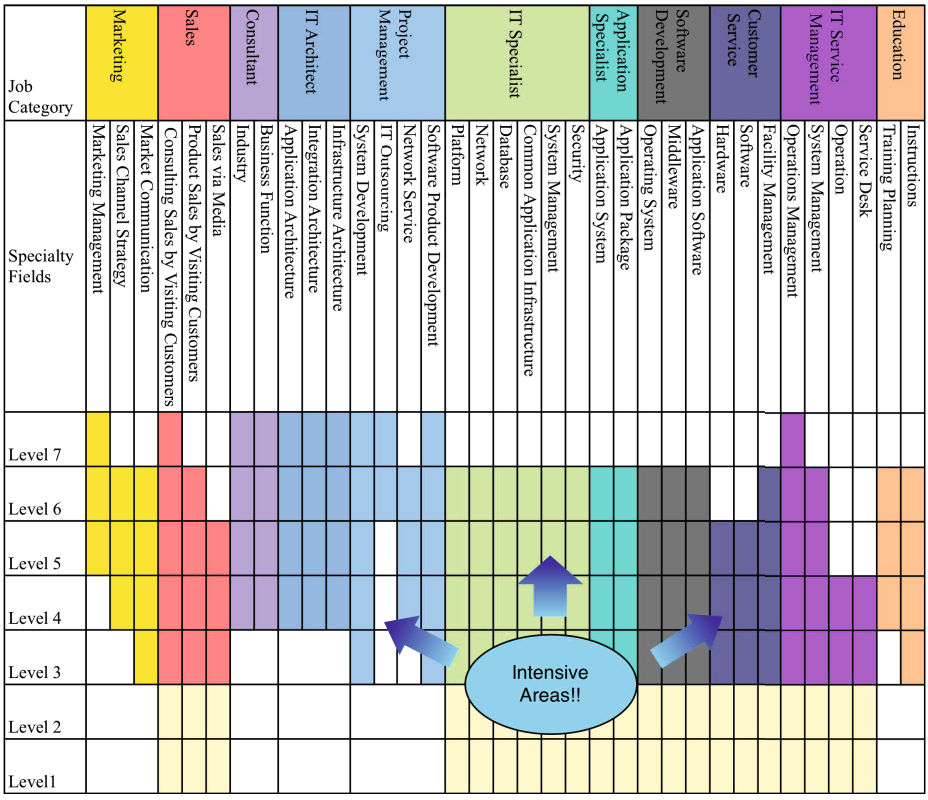


Fig. 1. ITSS career framework

Table 1. Job categories names abbreviations

Categories Names	Abbreviation
Marketing	Mrk
Sales	Sal
Consultant	Cnslt
IT Architect	IT-Arc
Project Management	ProMng
IT Specialist	IT-Spl
Application Specialist	ApSpl
Software Development	SwDpt
Customer Service	CstSvc
IT Service Management	IT-SM
Education	Edu

mentioned, have 11 job categories, generating 11 text documents. These documents are published by IPA [7]. Here after, this paper describes the 11 job categories as shown in Table 1.

3.1 Processing Outline

The outline of the process is shown in Fig 2. The input of this process is eleven text documents. To be able to analyze these text documents, applying the following steps:

Extraction Process. Keywords play a crucial role in extracting the correct information as per user requirements. Everyday thousands of books, papers are published which makes it very difficult to go through all the text material; instead there is a need of a good information extraction or summarization methods which provide the actual contents of a given document. As such effective keywords are a necessity. Since keyword is the smallest unit which expresses meaning of entire document, many applications can take the advantage of it such as automatic indexing, information retrieval, clustering, and classification [8]. There are a lot of approaches for keyword extraction. In this paper the authors used AnalogX keyword extraction tool [9]. AnalogX Keyword Extractor extracts all of the keywords of a webpage, then sorts and indexes them based off of their usage and position; once indexed, you can adjust search-engine specific weighting factors and keyword criteria to get the best possible view of how a search engine sees your site. AnalogX Keyword Extractor can load up both local files as well as files off other websites, can work through a proxy, and can have separate configurations for as many search engines as you choose to enter.

Pre-processing Step. This step applies the following four rules:

rule a If the keyword is a stop keyword then delete it. Stop words are very common words as, prepositions and non-content bearing words. The

list of stop words differs from a research to another and they are typically used to filter out non scientific English words that carry low domain-specific information content. In this research we used, the default English stop words list [10]. Table 2 shows part of the removed stop words.

rule b Truncate suffixes and trailing numerals so that words having the same root are collapsed to the same word for frequency counting.

rule c If the frequency of the keyword=0 in only two categories or less then delete this keyword. This rule filtered the set of keywords from the words that appeared so frequent in most documents.

rule d If the frequency of the keyword=0 in nine categories or more then delete this keyword. This rule removes the keywords that are rarely appearing in the documents.

Table 2. Part of removed stop words

Keyword	Mrk Fre	Sal Fre	Cnsl Fre	IT-Arc Fre	Pro Mng Fre	IT-Spl Fre	Ap Spl Fre	Sw Dpt Fre	Cst Svc Fre	IT-SM Fre	Edu Fre
About	0	4	0	0	2	38	6	3	5	0	0
All	0	16	43	0	0	0	0	0	3	0	2
Among	0	0	0	0	0	2	0	0	0	0	0
And	0	294	303	523	652	594	233	322	258	553	144
Annual	0	0	80	0	65	0	0	0	0	0	0
Are	0	2	0	0	2	2	2	2	2	2	0
Area	42	39	32	58	76	130	50	79	99	72	23
Around	0	0	0	12	0	0	0	0	9	0	0

In Table 3 some of the keywords deleted after applying *rule c* and *rule d* in the preprocessing step. The "*No of Zero*" column expresses numbers of the documents do not contain a specific keyword. As an example, the word "data" appears in 10 documents. This means that this word is so frequent. Therefore it was deleted after applying *rule c*. Also, the word "accident" appears in only 1 category. So, the word "accident" is deleted after applying *rule d* because it appears in only one document (IT-Spl document).

3.2 Mathematical Representation

There are several ways to model a text document. For example, it can be represented as a bag of words. These words are extracted from the texts of the documents to be used for content identification. This chapter used the vector space model [11] to represent the text documents in n-dimensional space. Each individual document D_w is represented as a vector of terms as shown in equation 1,

$$D_w = < w_1, w_2, ..., w_m > \quad (1)$$

Table 3. Part of the keywords deleted after applying *rule c* and *d*

keywords	Mrk Fre	Sal Fre	Cnsl Fre	IT-Arc Fre	Pro Mng Fre	IT-Spl Fre	Ap Spl Fre	Sw Dpt Fre	Cst Svc Fre	IT-SM Fre	Edu Fre	No of zero
Abroad	0	0	0	0	0	0	0	0	0	0	2	10
Accessibility	0	0	0	0	0	0	0	3	0	0	0	10
Accident	0	0	0	0	0	2	0	0	0	0	0	10
accounting	0	0	6	0	0	0	1	0	0	0	0	9
Data	1	0	8	3	1	10	4	3	7	7	1	1

And each term w_i in a document D_w is assigned a weight w_i which represents its importance. . The weight of a word expresses the importance of each word in every document. Several terms weighting schemes have been proposed to compute the importance of a term in a document [12-16]. These weighting schemes assign a value to keywords based on how useful they are likely to be in determining the relevance of a document. In this research the weight of the keywords was computed using equation 2.

$$w_i = \frac{tf_i}{\sum_{i=1}^{m} tf_i} \tag{2}$$

where tf_i is the frequency of the i^{th} term in document D, m is the number of keywords in document D.

3.3 Similarity Measure

A similarity measure is a function which computes the degree of similarity between a pair of text objects. There are a large number of similarity measures proposed in the literature such as cosine similarity, Euclidean distance and the Jaccard correlation coefficient [17-20]. This research uses the cosine similarity measure to compute the closeness between every pair of documents. The cosine similarity between D_i and D_j is defined by the angle between their feature vectors which are in our case w_i as shown in equation 3.

$$sim(D_i, D_j) = \frac{D_i * D_j}{\|D_i\| * \|D_j\|} \tag{3}$$

Where " * " denotes the dot product of the two vectors D_i and D_j and D_i denotes the length or norm of a vector D_i.

3.4 Transferring between Job Categories

Companies and organizations operate in an increasingly competitive environment, which drives a need for continuous employee skill development. To be successful, workers must make themselves as valuable as possible to be hired by companies. So, the clarification of the required skills and abilities for every job category is important for IT human resources. In addition, the clarification of the required skills to move between

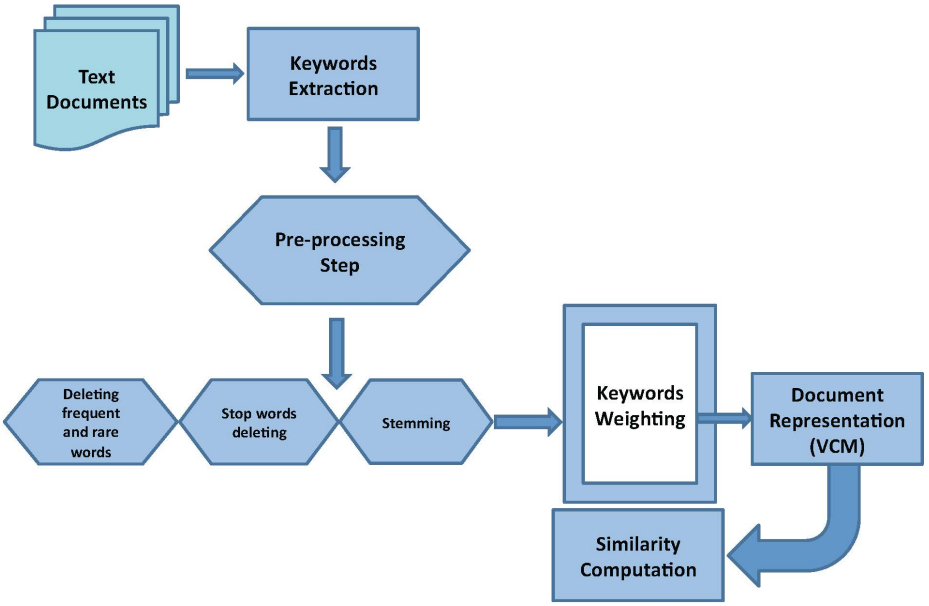


Fig. 2. Process Outline

different job categories is very important. This chapter proposed a method for deriving the required keywords to transfer from a job category to another. The main issue in this method is the keywords of each job category. Any two job categories have two types of the keywords: Common keywords and Specific keywords. Common keywords are the words that appeared in the both job categories, while, the specific keywords are the words concerned to only one of the two job categories. Every type of the keywords has its weight formula as shown in equation 4 and 5. The weight of special keywords is computed by the formula 4:

$$S.W_k = \log(tf_{i,k}, average(tf_{i,1}, tf_{i,2}, ..., tf_{i,n}) \tag{4}$$

Where, $S.W_k$is the weight of a special keyword k in the document i, $k = 1, 2, ..., n$is the keywords in a document, $i = 1, 2, ..., 11$is one of the 11 text documents.

The weight of the common keywords is computed by equation 5:

$$\begin{aligned} C.W_{i,k} &= wt_{i,k} * P(wt_{i,k}) \\ C.W_{j,k} &= wt_{j,k} * p(wt_{j,k}) \end{aligned} \tag{5}$$

Where, $wt_{i,k}$, $wt_{j,k}$ is the weight of the word k in the documents i, j respectively, $p(wt_{i,k})$, $P(wt_{j,k})$ are the probability of $wt_{i,k}$ and $wt_{j,k}$ respectively as shown in equations 6, 7, 8.

$$P(wt_{j,k}) = \frac{wt_{i,k}}{wt_{i,k} + wt_{j,k}} \tag{6}$$

$$P(wt_{j,k}) = \frac{wt_{j,k}}{wt_{i,k} + wt_{j,k}} \tag{7}$$

$$P(wt_{i,k}) + P(wt_{j,k}) = 1 \tag{8}$$

Now, for any two documents i and j, they have two sets of keywords: special keywords set and common keywords set. The required keywords to move from document $i$to document j are:

1. The special keywords for document j and,
2. The common keywords, for document j, those have$R \geq \alpha$. Where$R = \frac{C.W_{j,k}}{C.W_{i,k}}$, $\alpha$is a threshold.

For more explanation, figure 3 shows the process of transition from document D_1 and document D_2.

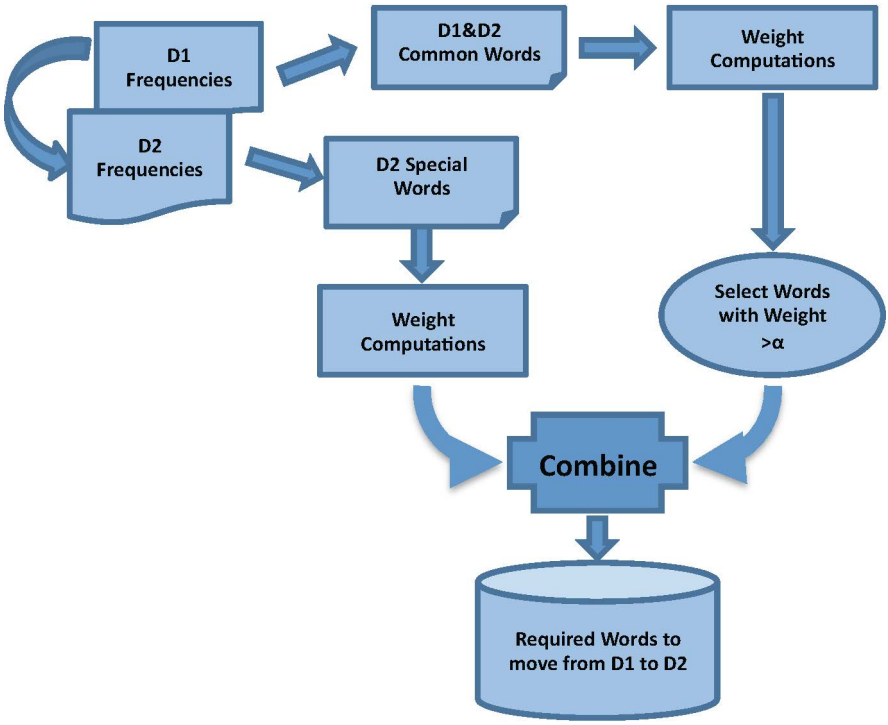


Fig. 3. D1 to D2 Moving Process

4 Experiment and Results

For this experiment the proposed method was applied on eleven text documents. Each document represents one job category of ITSS. These documents were published by API. The first step is the keywords extraction process. To extract key words, AnalogX keyword extraction tool has been used to extract the keywords from every document.

The number of extracted keywords is 449 keywords. This set of extracted keywords contains many repeated words and a lot of stop words and so on. So, it is necessary to perform the pre-processing step. This step applies four rules. Firstly, is the deleting the stop words manually using the default English stop words list. Secondly, is the reducing of all the keywords to their roots. Thirdly and fourthly, is deleting of the frequent and rarely appeared words. After applying this step, resulting in 83 keywords. The third step is the weight computation. Equation 2 to compute the weight of keywords. Now, the documents are ready to be mathematically represented using vector space model. At this point every document is represented by 83 dimensions vector. Finally, the cosine similarity was computed. Data from Table 4 show the result of cosine similarity.

Table 4. Cosine Similarity Results

keywords	***Mrk***	***Sal***	***Cnsl***	***IT-Arc***	***ProMng***	***IT-Spl***	***ApSpl***	***SwDpt***	***CstSvc***	***IT-SM***	***Edu***
Mrk	1	0.714	0.519	0.018	0.068	0.086	0.036	0.059	0.249	0.191	0.073
Sal	0.714	1	0.768	0.016	0.03	0.029	0.009	0.014	0.039	0.027	0.014
Cnsl	0.519	0.768	1	0.098	0.142	0.106	0.139	0.032	0.05	0.074	0.081
IT-Arc	0.018	0.016	0.098	1	0.225	0.265	0.449	0.306	0.091	0.092	0.022
ProMng	0.068	0.03	0.142	0.225	1	0.459	0.493	0.355	0.358	0.227	0.024
IT-Spl	0.086	0.029	0.106	0.265	0.459	1	0.803	0.418	0.336	0.543	0.146
ApSpl	0.036	0.009	0.139	0.449	0.493	0.803	1	0.562	0.391	0.585	0.092
SwDpt	0.059	0.014	0.032	0.306	0.355	0.418	0.562	1	0.546	0.272	0.074
CstSvc	0.249	0.039	0.05	0.091	0.358	0.336	0.391	0.546	1	0.564	0.011
IT-SM	0.191	0.027	0.074	0.092	0.227	0.543	0.585	0.272	0.564	1	0.104
Edu	0.073	0.014	0.081	0.022	0.024	0.146	0.092	0.074	0.011	0.104	1

4.1 Example for the Transition between Job Categories

To explain the results for deriving the required keywords to move between different job categories, the two job categories ProMng and IT-Spl are used as an example. Equation 2 was applied to compute the weight of the keywords and then used that weight to represent ProMng and IT-Spl mathematically. ProMng and IT-Spl are two vectors with 83 dimensions. By using Equation 3 the cosine similarity between the two documents is 0.459. The required keywords to move from ProMng <–> IT-Spl are shown in Table 5. Equations 3 and 4 were used for computing the weight of special and common words weight. Table 5 shows special words, common words and the required keywords to move between ProMng and IT-Spl.

For more declaration Figure 4 shows the required keywords to move between some special fields as Mrk, Sal, Cnsl, ProMng, AppSpl, IT-Spl. Figure 4 is composing of 3 blocks: B1, B2,B3 according to the similarity values. The blocks are B1=(Mrk, Sal, Cnsl) , B2=(ProMng, AppSpl,IT-Spl) and B3=(Edu). The distance between the job categories in the same block is smaller than the distance between job categories in different blocks. This because the cosine similarity values between the job categories inside any block is bigger than the values of the cosine similarity between the job categories in

Table 5. Transformation between ProMng and IT-Spl

ProMng S.W	ProMng wt *1000	IT-Spl S.W	IT-Spl wt *1000	ProMng & IT-Spl C.W	ProMng wt *1000	IT-Spl wt *1000	ProMng–>IT-Spl	IT-Spl->ProMng
Adaptation	0.71	Measure	1.44	assessment	0.12	3.02	Measure	Adaptation
Corporate	1.42	Transaction	1.44	consultant	13	14.9	Transaction	corporate
Equipment	1.42	Migration	2.17	definition	22.6	1.46	Migration	Equipment
Intellectual	1.42	advice	4.33	Design	2.2	18.2	assessment	intellectual
Symposia	2.13	operating	4.33	dictionary	0.35	0.36	advice	symposia
Solution	4.96	Internet	5.05	engineering	0.69	18.1	operating	solution
Structure	7.09	regulation	5.05	improvement	3.14	1.1	Internet	Structure
Execution	12.8	science	5.05	Inspection	0.70	0.73	regulation	execution
Administrate	14.2	Structure	5.05	installation	53.3	4.05	science	Administrate
Performance	14.2	architecture	7.94	ITEE	0.35	0.36	Structure	Performance
Verification	28.3	accurately	8.66	maintenance	3.82	14.8	architecture	service
Estimating	42.5	preparation	8.66	manage	6.32	6.55	accurately	definition
Qualitie	46.0	Computer	11.54	method	1.56	7.42	preparation	verification
Activitie	56.7	collaboration	12.99	network	26.0	32.8	Computer	Estimating
Control	96.4	Test	14.43	peak	62.2	62.9	support	qualitie
Contract	119	guidance	20.20	policy	1.41	1.46	collaboration	installation
		scale	23.81	Practice	8.26	9.28	Test	Activitie
		Technologie	25.97	product	4.75	18.7	engineering	control
		Infrastructure	26.70	security	0.04	44.8	Design	Contract
		Platform	33.19	service	15.7	0.11	guidance	
		Specialist	57.72	Software	40.1	12.4	scale	
		responsibility	101.7	successfully	48.1	1.53	Technologie	
				support	0.14	11.8	Infrastructure	
				System	4.21	94.1	Platform	
							security	
							Specialist	
							System	
							responsibility	

different blocks Also, according to the similarity results the job category ‘Education‘ is far from all the other job categories.

4.2 High Weight Keywords and the Required Learning

The key point of the author’s analysis here is the keywords. The keywords were used to represent the text documents to can apply text mining methods. Now, the keywords will be used to sort the required courses for each ITSS job category. But, here only high weight keywords will be used. The higher weighted keywords indicate the significant and priority of learning the associated skills. So, in this research the high weight keywords were used to estimate the most important learning courses for each job category by a very simple method. The required courses for job categories [2], published by IPA, were used in this method.

Table 6 contains the higher weighted keywords in each job category. These keywords have the highest weight in each document. The authors noticed that these keywords are important in realty for their job categories. For example, the words” *products, sales, and strategy*” are very important for *Sal* and *Mrk* jobs.

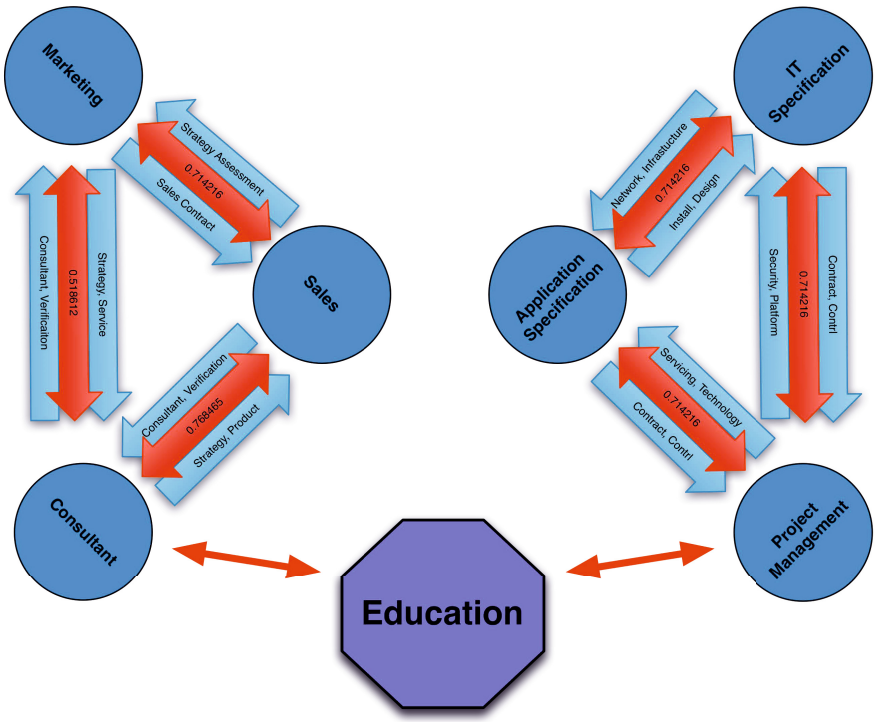


Fig. 4. Transferring between different Job Categories

4.3 Proposed Method for Sorting Learning Courses Using High Weight Keywords

- ***Input***: High weight words (K), ITSS courses names(C), Common skills to job category(S).
- ***Output***: Sorted Courses.
- ***Process***:

1. For each job category, search if K exists in C, S.
2. If exist, count how many times appears.
3. If not, ignore.
4. Now, every K has a value (how many times appears in S, C).
5. Sum the keywords values for each C.
6. Finally, Sort the courses according their values in step 5.

At step 5 every course had a value. This value (point) indicates how importance the course is. So, the course with Cityplacehigh point should be learned first because this course is important for this job. The obtained result in this research is very important for human resources to determine education recommended for their careers.

Table 6. The keyword list of the high weight for every job category

Job Category	High Weight Keywords						
Mrk	CityplaceSale	Strategy	Product	Segment	Methodology	Service	company
Sal	Assign	Company	Product	Sales	Segment	Strategy	Solution
Cnsl	Consultant	User	Maintaince	Quality	Sales	Definition	Strategy
IT-Arc	Architect	Definition	Design	Maintaince	Peak	User	Solution
ProMng	Peak	contract	Control	Definition	Manage	Quality	Maintaince
IT-Spl	Maintaince	Platform	Network	Responsible	Security	Specialist	System
ApSpl	Design	Maintaince	Peak	Technology	Security	Specialist	System
SwDpt	Software	Design	System	Peak	Product	Operating	Middleware
CstSvc	Hardware	Install	Maintaince	Segment	Service	Software	System
IT-SM	Maintaince	Network	Security	Service	Strategy	Support	User
Edu	Practice	Methodology	Assign	Company	Goal	Implement	Infrastructure

Table 7 shows some of the required courses for Application Specialist after sorting them. Here there are 7 high weight keywords. Every keyword has had a value corresponding to each course. Also, each course name had a value by summing the values of all keywords belonged to this job category. According to this value, the courses were sorted. The course with Cityplacehigh point seems to be an important course. Therefore, it is good to learn from courses have Cityplacehigh point. For ApSpl job category, the courses, Security system component Technology, Application system design, Distributed computing system component Technology, System Design Fundamentals, should be learned first.

Table 7. Required courses for appSpl after sorting

Course Name/Highest weight keywords	Design	Maintaince	Peak	Technology	Security	Specialist	system	Total
Security system component Technology	0	0	0	29	9	0	4	42
Application system design	25	0	0	6	0	0	7	38
Distributed computing system component Technology	1	0	0	37	5	0	3	35
System Design Fundamentals	22	0	0	7	0	0	5	34
Application system operations /Maintaince	0	1	0	11	3	0	11	25
System Operations/Maintaince	0	1	0	11	4	0	7	23
System Management infrastructure component technology	0	0	0	17	0	0	8	23
Database Component Technology	0	0	0	18	1	0	4	23
Network component Technology	0	0	0	19	0	0	3	22
Platform component technology	0	0	0	14	0	0	6	20
Application package fundamentals	5	0	0	11	0	0	4	20

5 Conclusion

Accelerated investment and innovation in information technology requires a high level of foundation and technical skills in the workforce. Educational and training institutions must restructure themselves to better prepare this new workforce. One effective tool for this restructuring is application of information technology skill standards (ITSS). The Japanese ITSS was promoted by IPA to develop the human resources that will lead and support country-regionplaceJapan in the future. This study analyzed the Japanese skill standards' job categories. In this chapter the keywords were extracted, the documents were mathematically represented, the similarity function between the eleven

job categories was computed, the high weighted keywords were identified, and finally, a method for clarifying the required keywords to move from one category to another was proposed. Moreover, the high weighted keywords were used to sort the required learning courses for any job category according to the priority of learning. Future work could focus on analyzing the specialty fields of the Japanese information technology skill standards.

References

1. Evans, N.: Information technology jobs and skill standards. In: Hawkins, B.L., Rudy, J.A., Wallace, W.H. (eds.) Technology Everywhere: A Campus Agenda for Educating and Managing Workers in the Digital Age, pp. 25–38. Jossey-Bass, A Wiley Company (2002)
2. Broderick, R.F., Boudreau, J.W.: Human resource management, information technology, and the competitive edge. Academy of Management Executive 6(2), 7–17 (1992)
3. Gardner, S.D., Lepak, D.P., Bartol, K.M.: Virtual hr: The impact of information technology on the human resource professional. Journal of Vocational Behavior 63, 159–179 (2003)
4. Manning, C.D., Raghavan, P., Schutze, H.: An Introduction to Information Retrieval. Cambridge University Press (2009)
5. Salton, G., McGill, M.: Introduction to modern information retrieval. McGraw-Hill, New York (1983)
6. Atlam, E.S.: A new approach for text similarity using articles. Int. J. Information Technology and Decision Making 7(6), 23–34 (2008)
7. Deisy, C., Gowri, M., Baskar, S., Kalaiarasi, S., Ramraj, N.: A novel term weighting scheme midf for text categorization. Journal of Engineering Science and Technology 5(1), 94–107 (2010)
8. Salton, G.M., Yang, C.: On the specification of term values in automatic indexing. Journal of Documentation 29(4), 351–372 (1973)
9. Salton, G.M., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing&Management 24, 513–523 (1988)
10. Salton, G.M., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing&Management 24, 513–523 (1988)
11. Salton, G.M., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM 18, 613–620, ISSN: 0001-0782 EISSN: 1557-7317
12. Atlam, E.S., Fuketa, M., Morita, K., Ichi Aoe, J.: Documents dissimilarity measurement using field association terms. Information Processing and Management 39(6), 809–824 (2003)
13. Gupta, V., Lehal, G.S.: Features selection and weight learning for punjabi text summarization. International Journal of Engineering Trends and Technology (2011)
14. Huang, A.: Similarity measures for text document clustering. In: Computer Science Research Student Conference, New Zealand, Christchurch (2008)
15. Salton, G.M.: Automatic text processing. Addison-Wesley Longman Publishing Co., Inc., Boston (1988) ISBN:0-2:1-1227-8
16. Veni, R.: Effects of similarity metrics on document clustering. Master's thesis, (School of Computer Science Howard R. Hughes College of Engineering) (2009)

An Efficient Method of Characterization of the Bad Debt Customers in the Mail Order Industry

Masakazu Takahashi[1], Hiroaki Azuma[2], Masanori Ikeda[3], and Kazuhiko Tsuda[3]

[1] Graduate School of Innovation and Technology Management, Yamaguchi University
2-16-1, Tokiwadai, Ube, Yamaguchi 755-8611, Japan
masakazu@yamaguchi-u.ac.jp
[2] Hazs Corporation, 6-2-6 Kojimachi, Chiyoda, Tokyo 102-0083, Japan
hirokiazuma2002@hazs.biz
[3] Graduate School of Business Sciences, University of Tsukuba, Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
masanori-ikeda.fu@u, tsuda@gssm.otuka.tsukuba.ac.jp

Abstract. This paper present investigating for analyzing customer characteristics from the bad debt list of a mail order corporation that aims to understand the characteristics of the bad debt customers. This type of investigations have not made intensively, such as the private default risks so far and the conventional method for predicting such kind of risks mostly depend on the employees of working experiences in shipping. In order to be expansion of the mail order industry, such bad debt customers are also unavoidable as long as most of the Japanese mail order industry adopts the reversionary payment systems. It is something of side-effect factor so far. For these backgrounds, the authors made an intensive research with the bad debt customer list and the sales data gathered from a mail order company. The analytical results suggest that the one of the characteristics of the bad debt customers in the mail order industry, it turns out that the bad debt customers order repeatedly and big-ticket. This result will make it use for the decision support knowledge for screening customer at the order-received phase in the mail order industry.

Keywords: Mail Order, Customer Analysis, Bad Debts, Random Forest, Machine Learning, Service Science and Management Engineering.

1 Introduction

The Mail order industry is one of the most promising methods of expanding sales, even during deflation condition, with a long-term slump in Japanese retail business. This kind of business has characteristics of delivering the items to the customers of hand whereas they have to make transactions with the absence of face-to-face contact. A survey from the Japan Direct Marketing Association says the percentage of credit losses of a mail order company is estimated about 0.5% of net sales. Therefore, it is important for the mail order company to predict risk exposures in customers' credit control domain. It is because too large risk exposure leads to high default risk and too small risk exposure misses business opportunities. In this paper, The authors carry out the data analyses

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 281–290, 2013.

for the purpose of the customer behavior through the bad debt list from a mail order company.

The rest of the paper is organized as follows: Section 2 discusses the backgrounds of the research and related works; Section 3 briefly summarize the gathered data on the target mail order company; Section 4 describes analytics of the data and presents analytical results; and Section 5 gives some concluding remarks and future works.

2 Backgrounds and Related Works

While the domestic retail business is in the long term slump, the mail order industry is continuing the sales expansion. One of the conclusive factors for the sales expansion among the mail order industry is the diversification of payment methods. Not only the reversionary system that is settles accounts after the order item arrives with which improving customers' conveniences also appeal the transactional safety from the view point of the customers which expect the large increase both of the sales and the profits for the mail order company.

On the other hand, there are no credit criteria at the time of start trading. Therefore, there are many transaction of scam or fraud by vicious customers making use of the reversionary systems. It is important to realize the safe dealings in the mail order industry, that provides redistribution to the customers and reduce sunken security costs. According to the mail order industry sales survey, it was estimated amount to 4,310 billion yen in total of the 2009 fiscal year which increased by 170 billion yen 4.1% raise compared with previous year and the highest on record since the research started [1] . Table 1shows the component of payment methods in the mail order industry [1]. The percentage of the customers who used the convenience store or the postal transfer as the reversionary payment systems reached 39.7% in 2009. As far as these reversionary payment systems continue, the fraud transactions will increase in number.

Table 1. The Primary Payment Methods in 2009

Payment Methods	%
Cash On Delivery	28.3%
Convenience Store	23.8%
Credit Card	23.2%
Postal Transfer	15.9%
Bank Transfer	8.2%
Others	0.5%
No answer	0.1%

As for the related works on the mail order industry, it separates into the activity of before order and of after order received. The related works of the before order activities were focused on the elaboration of the order received in consideration of the time lag from customers such as the demand predictions [2–5]. On the other hand, at the

phase for the after order received, most of the researches have been made for customer analysis with the purchase history. There are lots of customer analyses with the data-mining methods [6, 7]. For example, the order history has been used for such as trend analyses of customers and sales promotion strategies as the retailers' decision support tool. As long as the reversionary system is adopted in mail order industry, it is important to collect the payment from the customer as soon as possible, whereas the related works on the collecting the payment was less studied than that of customer analysis, especially collecting the bad debts. Concerning to the research on the bad debts, most of them have been done about the bankruptcy prediction of the corporation [8–10]. It is mainly used for used for the preliminary screening to hedge risks with the machine learning methods [11–14].Moreover, machine learning is used for finding new business partners, fraud claim and credit scoring [15–17]. One of the characteristics of the mail order system is delivered to the customer hand even both the name and the address are written correctly. Therefore, most of the customer lists from the mail order corporation are not fulfilled all the customer attributes. Besides, the random forest is one of the promising methods for such an insufficient data prediction such as an authorship identification, a software production workload estimation, promising customers' classifications, and credit risk evaluations and so on [18–24]. From the backgrounds and the related works, The authors find out that the customer relation management among the mail order industry is mainly used for the sales expansion, so far. As for the payment collection phase, since each payment is petty, so that they have to take into consideration for trade-off between the collection cost and the invoice. Therefor, they are taking a financial policy for counting up the allowances for the future bad debts as the advance risk hedge for the bad debts. As long as they are taking the reversionary system as one of the diversification of paying methods to make their market expand, there need to understand of the characteristics of the bad debts customers more than used to be, beforehand.

3 Data Summary

The authors make an intensive research to figure out the characteristics of the bad debt customers with the following data. Table 2 shows the summary for the debt list of a mail order corporation. This data was financially processed into the bad debts for Fiscal year of 2010 in March 2011.

Grand total of the debts amount to $350,571.94 and composed of 9,357 bad debt customers. Average sales price per customer amount to $235.57, Average sales price per item reached $109.49, and Average sales price per transaction scored $148.68 respectively.

Table 3 and Table 4 indicate the number of transactions by category and the amount of debts by category, respectively. Both figures indicate that Lady's items shared a majority.

Table 5 indicates the proportion of the debt customer attributes by sex and age. From the results, around the four thirds of the customers are female with around 40's. This result indicates the image of the primary customer. Meanwhile, there should be suspected the age replies that are both less than 10 years old and over 80 years old. It is hard to

Table 2. Debts Summary

Fiscal Year	2010
Number of Customers	9,357
Number of Invoices	20,132
Average Sales Price per Customer	$235.57
Average Sales Price per Item	$109.49
Average Sales Price per Transaction	$148.68
Grand total of the Debt	$350,571.94

Table 3. Number of the Transaction by Category

Category	Transaction	%
Lady's	10,629	52.80%
Men's	1,198	5.95%
Others	8,305	41.25%
Total	20,132	100.00%

imagine that those who the age below 10 or over 80 could purchase items through the mail order without supporting the other person.

From the basic research from the bad debts list, The authors figure out the characteristics of the attributes and the item categories. In the system of the mail order industry, the ordered items will deliver to the customer hands at least both the name and the address is written exactly. Therefore, the mail order company needs to predict the potential customer who might fall into the dad debts list.

4 Data Analysis

Since this data is not cover all the transaction, at first, the authors focus on the percentage of transaction and customer per 10,000 populations by prefecture in order to see the local characteristics. As for the population, The authors make use of Census as of October 1st, 2010 [25].

Table 4. Amount of Bad Debts by Category

Category	Debts	%
Lady's	$1,078,040.20	48.91%
Men's	$94,517.18	4.29%
Others	$1,031,640.08	46.80%
Total	$2,204,197.47	100.00%

Table 5. Customer Attributes

	Number	%	Age Answered	Age N.A.	Min. Age	Max. Age	Ave. Age
Male	2,443	26.11%	2,084	359	5	90	48.13
Female	6,894	73.68%	5,960	934	3	97	42.62
NA	20	0.21%	12	8	22	88	48.25
Total	9,357	100.00%	8,056	1,301	3	97	44.05

Figure 1 indicates percentage of debt transaction per 10,000 populations by prefecture and Figure 2 also indicates percentage of debt customer per 10,000 populations by prefecture, respectively. From these figures, Hokkaido, Osaka, and Fukuoka scored high ratio of bad debt both of the transaction and customer compared with another prefecture. As for this bad debt list, it is insufficient information of the customer, so that we will make intensive research with the random forest to characterize the bad debt customer.

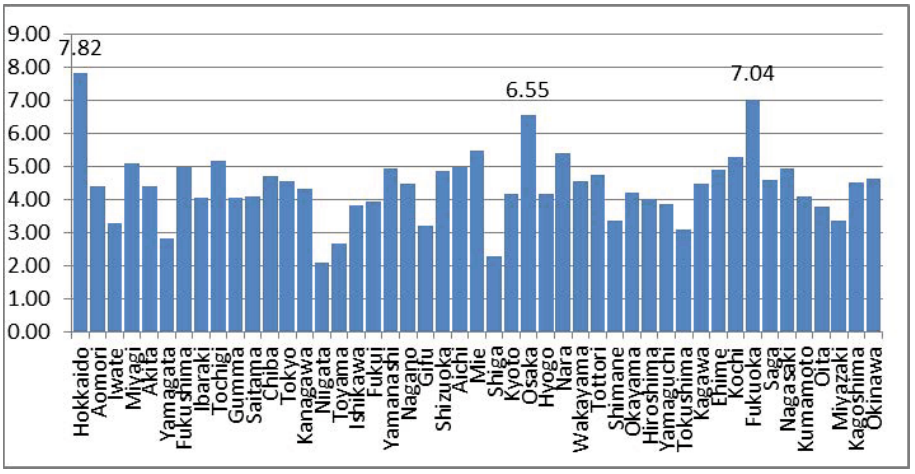


Fig. 1. Percentage of Debt Transaction per 10,000 Population by Prefecture

Random forest has some merits such as insufficient data prediction. In the system of the mail order industry, the ordered items will be delivered to the customer hands at least both the name and the address is written exactly. The authors summarize the bad debt list into 20,132 records by order date and customer ID. The authors make characterization five type customer conditions with eleven attributes in Table 6 and Table 7.

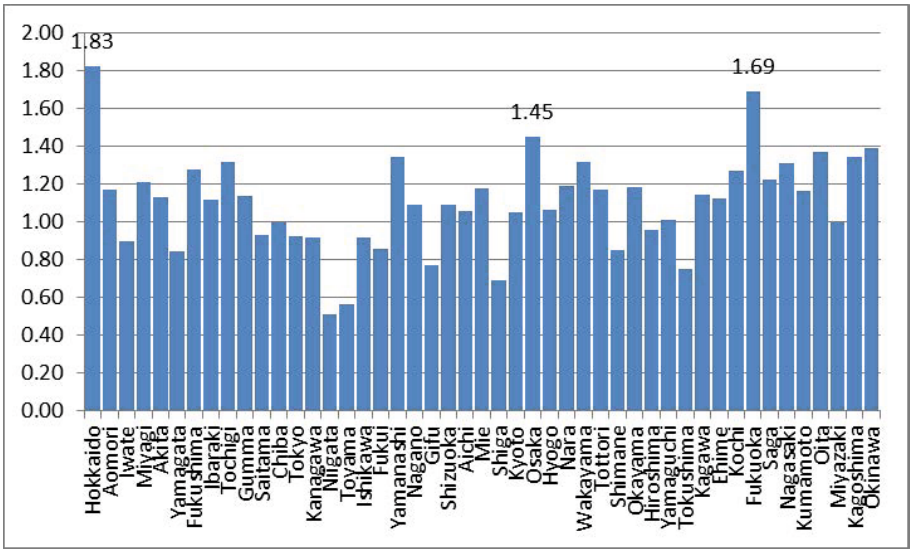


Fig. 2. Percentage of Debt Customer per 10,000 population by Prefecture

Table 6. Types of Customer Conditions

Costomer Conditions
Claim difficulty
Collecting difficulty
Death with no inheritance
Missing
Personal Bankruptcy

Table 7. Types of CustomerAttributes

Costomer Attributes
Customer ID
Prefecture
Sex
Age
Debt Processing Period
Payment Method
Number of Payment
Number of Remittance
Phone Call Condition
Number of Purchased Items
Amount of Sales

Among the summarized data, The authors set sixty percent of data as training data (12,079 records) and rest of forty percent of the data as test data (8,053 records) with random number assignment in Table 8.

Table 8. Types of Data Configuration

Data	Records	%
Test data	12,079	60%
Training data	8,053	40%
Total	20,132	100%

As a result, 500 trees from Three parameters classified customer conditions from six parameters with class errors in Table 9. And the error rate for the test with OOB (out-of-bag) data results 4.97%. This score means that the cluster calculated with the classifier scored 4.97% error compared with the cluster contained in the original data. OBB data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

Table 9. Result of Random Forest Trial

Type of random forest	Classification
Number of trees	500
Number of variables tried at each split	3
OOB (Out Of-Bag) estimate of error rate	4.97%

Table 10 indicates the class error by customer conditions. From the table, except the condition for collecting difficulty, Random forest predicts the customer conditions. Since conditions for the customer are decided by in charge of bad debt collection through the negotiations with the customers, so that it is difficult to predict for the condition.

Table 10. Class Error by Customer Condition

Customer Conditions	Class error
Claim difficulty	0.617187500
Collecting difficulty	0.001083907
Death with no inheritance	0.788461538
Missing	0.527859238
Personal Bankruptcy	0.486206897

Table 11 indicates the mean decrease gini by customer attributes. This index means contribution of each parameter. The parameter of prefecture contributes characterize the customer next to the parameter of customer ID and Debt processing period.

Table 11. Mean Decrease Gini by Customer Attributes

Parameters	Mean Decrease Gini
Customer ID	651.91084
Prefecture	333.90053
Sex	59.35230
Age	203.55312
Debt Processing Period	651.05821
Payment Method	23.60707
Number of Payment	116.68290
Number of Remittance	213.72330
Phone Call Condition	136.34009
Number of Purchased Items	110.25428
Amount of Sales	307.83575

From the result of the analyses, it figure out the bad debt customers' characteristic that the regionality is an important factor, in addition to the long term evasion and big-ticket purchase. This result means that the destination of the ordered item is an important factor for the fraud transaction. Furthermore, the amount of money for order and ID are also important factor. These parameters mean that the bad debt customers order repeatedly and big-ticket.

5 Concluding Remarks

This paper present investigating for analyzing customer characteristics from the bad debt list of a mail order corporation which aims to understand the characteristics of the bad debt customers. The authors describe the research background, research method, and analytical results. A conventional screening method of the potential bad debt customers, it depends on the heuristic knowledge based on the staffs' working experiences. In order to be expansion of the mail order industry, such bad debt customers are also unavoidable. It is something of side-effect factor so far. The analytical results suggest that the one of the characteristics of the bad debt customers in the mail order industry, it turns out that prefecture is an important factor. Furthermore, the amount of money for order and ID are also important factor. These parameter mean that the bad debt customers order repeatedly and big-ticket. This result will make it use for the decision support knowledge for screening customer at the order received phase in the mail order industry. The authors' future work include as follows;

1. Customer analysis with not only the black list but the white list of order to predict customer condition that fall into bad debt situation.

2. Comparison search regarding to the machine learning method to fit prediction of the customer characterization.
3. Finding new characteristics of the customer with another machine learning method.

These works will require algorithm investigations and pay attention to hear the views of the rank-and-file employees for prediction and further survey studies.

Acknowledgement. The authors wish to express our gratitude of the cooperation from the mail order corporation to our analysis.

References

[1] The Japan Direct Marketing Association, 17th Annual National Mail Order Survey Report (2010) (Japanese)
[2] Simester, D.I., Sun, P., Tsitsiklis, J.N.: Dynamic Catalog Mailing Policies. Management Science 52(5), 683–696 (2006)
[3] Kimijima, M.: A Study on Measuring Input-Output Process on Order-Getting Costs for Direct Marketing. Yokohama National University Departmental Bulletin Paper 16(1), 21–39 (2010) (Japanese)
[4] Conlin, M., O'Donoghue, T., Vogelsang, T.J.: Projection Bias in Catalog Orders. American Economic Review 9(4), 1217–1249 (2007)
[5] Matsuda, Y., Ebihara, J.: Forecasting Model in the Mail-Order Industry. UNISYS Technology Review 71, 52–68 (2001) (Japanese)
[6] Motoda, H., Washio, T.: Perspective of Data Mining, System/Control/Information, the Institute of Systems. Control and Information Engineers 46(4), 169–176 (2002) (Japanese)
[7] Ishigaki, T., Motomura, Y., Chan, H.: Consumer Behavior Modeling Based on Large Scale Data and Cognitive Structures. IEICE Technical Report, NC2008-157, 108(480), 319–324 (2009) (Japanese)
[8] Yamashita, S., Tsuruga, T., Kawaguchi, N.: Consideration and Comparison about a Valuation Method of a Credit Risk Model, Discussion Paper Series (11), Financial Research Center, Financial Service Agency (2003) (Japanese)
[9] Yano, J., Ikai, M., Nakagawa, K., Takahashi, S., Namatame, T.: A Study on Credit Card Users' Default Prediction. Journal of Operations Research of Japan 51(2), 104–110 (2006) (Japanese)
[10] Sunayama, W., Yada, K.: Modeling and Analysis of Persuading Process using Conversation Logs. In: Proc. of the 20th Annual Conference of the Japanese Society for Artificial Intelligence, 3C3-3 (2006) (Japanese)
[11] Tanabe, K., Kurita, T., Nishida, K.: Prediction of Corporate Credit Ratings by Support Vector Machine. Journal of Japan Society for Management Information 20(1), 23–38 (2011) (Japanese)
[12] Min, J.H., Lee, Y.C.: Bankruptcy Prediction Using Support Vector Machine with Optimal Choice of Kernel Function Parameters. Expert Systems with Applications 28(4), 603–614 (2005)
[13] Abe, N., Melville, P., Pendus, C., Reddy, C.K., Jensen, D.L., Thomas, V.P., Bennett, J.J., Anderson, G.F., Cooley, B.R., Kowalczyk, M., Domick, M., Gardinier, T.: Optimizing Debt Collections Using Constrained Reinforcement Learning. In: Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 75–84. ACM (2010)

[14] Hashimoto, M., Yoshida, K.: A Study on the Evaluation Technique of the Claim Assessment Model in a Retail Financial Sector. In: Proc. of The Annual Conference of The Japan Society for Management Information, pp. 63–66 (2004) (Japanese)
[15] Mori, J., Kajikawa, Y., Kashima, H., Sakata, I.: Machine Learning Approach for Finding Business Partners and Building Reciprocal Relationships. Expert Systems with Applications 39(12), 10402–10407 (2012)
[16] Viaene, S., Derrig, R.A., Baesens, B., Dedene, G.: A Comparison of State-of-the-art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection. The Journal of Risk and Insurance 69(3), 373–421 (2002)
[17] Huang, C.L., Chen, M.C., Wang, C.J.: Credit Scoring with a Data Mining Approach based on Support Vector Machines. Expert Systems with Applications 33, 847–856 (2007)
[18] Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
[19] Jin, M., Murakami, M.: Authorship Identification using Random Forests. Proc. of the Institute of Statistical Mathematics 55(2), 255–268 (2008) (Japanese)
[20] Jin, M.: Estimation of When the Works were Written: With the Works of Ryunosuke Akutagawa as Examples. Behaviormetrics 36(2), 89–103 (2009) (Japanese)
[21] Kobayashi, Y., Tanaka, S., Tomiura, Y.: Classification and Assessment of English Scien-tific Papers Using Random Forests. IPSJ SIG Technical Reports, 2011-CH-90(6), 1–8 (2011) (Japanese)
[22] Konishi, F., Uchida, S., Toda, K., Monden, A.: An Approach of Software Development Effort Estimation by Random Forests. In: Proc. of the 2011 IEICE General Conference, Information/System(1), p. 17 (2011) (Japanese)
[23] Ohashi, K., Toyoda, H., Kubo, S.: A Study on Classification and Prediction of the Potential Customers. Journal of Operations Research of Japan 56(2), 71–76 (2011) (Japanese)
[24] Umezawa, Y., Mori, H.: Credit Risk Evaluation of Power Market Players with Random Forest. Trans. of the IEEJ, Power and Energy Society 128(1), 165–172 (2008) (Japanese)
[25] The Statistics Bureau, Population Census (2010), `http://www.stat.go.jp/data/nenkan/zuhyou/y0203000.xls` (accessed March 15, 2013)

Wearable Smart System for Physical Activity Support

Paweł Świątek, Piotr Klukowski, Krzysztof Brzostowski, and Jarosław Drapała

Institute of Computer Science, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{pawel.swiatek,piotr.klukowski,krzysztof.brzostowski,jaroslaw.drapala}@pwr.wroc.pl

Abstract. Wearable smart systems that measure and process a human vital sings have become important tool in eHealth, rehabilitation and recreation. This research was about to develop a system that aims to support recreational physical activity by delivering communication and computational services. The system architecture facilitates implementation of advanced data processing methods in the form of computational services and easily include additional services according to the user needs. An eHealth application was developed to demonstrate the system capabilities. The application makes use of expert's knowledge together with measurement data in order to generate the optimal training plan. The system performs such tasks as: modelling of the sportsmen's cardiovascular system, estimation of the sportsmen's parameters, adaptation of the model to the sportsmen.

Keywords: eHealth, Service Oriented Architecture, modelling, cardiovascular system.

1 Introduction

In recent years wearable smart systems that measure and process a human vital sings, [26, 35], have become very popular, mainly due to widespread use of wireless sensors and smartphones [39]. Many commercial and academic teams develop solutions for support of training (for professional and recreational purposes). In general, the role of wearable smart systems are: to gather measurement data, to deliver these data to destination places, to process these data and to support decision making (for sportsmen and for trainer) by delivering feedback information about the training process, [14, 24, 41].

Polar[1] belongs to the most important companies offering wearable devices and data processing tools. A wrist worn watch and a chest worn heart rate sensor are used to measure and present data to the user. Polar provides equipment for fitness improvement and for maximisation of performance. These devices are used in motivational feedback, for example generate beeps every time when certain amount of calories is burnt. Suunto[2] provides devices that generate a personalised training plan. Based on the results from monitoring of the user, these devices are capable of making recommendations for

[1] Polar, www.polar.fi, accesed: 4 April 2012.
[2] Suunto, www.suunto.com, accesed: 4 April 2012.

J.W. Tweedale and L.C. Jain (Eds.): KES 2012, CCIS 246, pp. 291–304, 2013.

training volume, for example its frequency, duration and intensity. The *miCoach* application[3] made by Adidas company is an advanced training tool that optimizes a training plan for endurance training, strength and flexibility. Data is measured for stride and the heart rate. The website[4] allows the user to manage the training process. Important feature of the system is *digital coaching*. It motivates the user by voice notifications, such as *"speed up"* or *"slow down"*. The Mobile Personal Trainer (MOPET) is an example of enhancement project [6] under development. It uses a mathematical model of the user to predict their performance and to support training.

Typically, all functionalities of commercial eHealth applications are unique to specific systems and cannot be directly accessed by others (the lack of interoperability). The key idea behind the proposed system design is that the system is proposed to deliver its functionalities in the form of services in the network, taking advantage of the Service Oriented Architecture (SOA) [15–17, 37, 40].

The aim of research was to develop a system that supports recreational physical activity. The system is composed of wireless measurement devices, smartphones that receive measurement data and distributed computer system that delivers communication and computational services in unified framework [5, 31]. The basic system setting includes Bluetooth devices to measure: acceleration, heart rate, traveled distance and Global Positioning System (GPS) location. An advanced configuration is also comprised of an: Electromyogram (EMG), Eelectrocardiogram (ECG) and respiration rate monitors.

In general, data is acquired to perform modelling, optimization, pattern recognition and control tasks. The novelty of the proposed system is that all algorithms are provided in the form of computer services [3, 10, 19, 30]. This aids the user in composing the data processing flow that meet his or her requirements [13].

In Section 2 details about the system architecture and services available are given. Inn Section 3, the custom application composed to support training intensity is described. Problems solved by the system are posed in detail in Section 4. Typical application of the system is described in the form of use-case scenario in Section 5.

2 System Architecture

2.1 General View

The main non-functional requirement to be met by the system is the operate „anywhere and anytime" and provided services availability [18, 22, 32]. The system is scalable with respect to the number of users and number of services [23, 33, 38]. It allows for flexible composition of new use-case scenarios on the basis of available components and services [36, 37]. A three-tier architecture is composed from the following types of devices:

Tier 1: Bluetooth sensors and mobile phones,

[3] Adidas miCoach, The Interactive Personal Coaching and Training System, `www.micoach.com`, accesed: 4 April 2012.

[4] `www.micoach.com`, accesed: 4 April 2012.

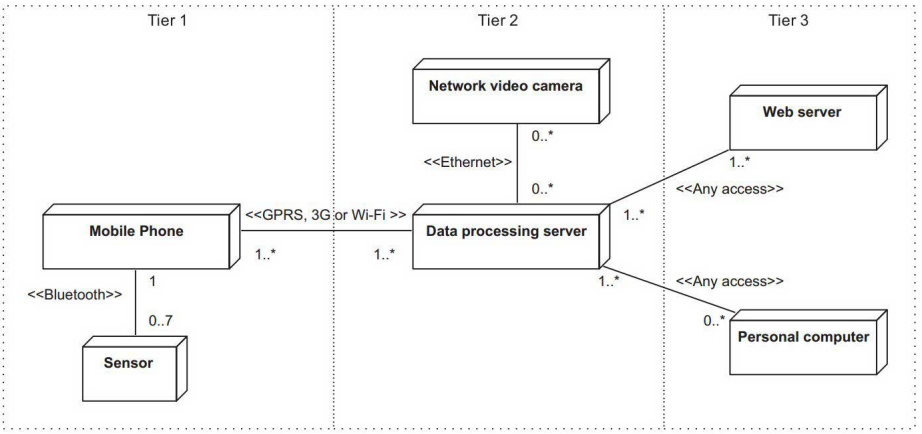


Fig. 1. Deployment diagram of the system

Tier 2: network video cameras and data processing services,
Tier 3: web server and end-user terminals.

Any device may be used either by a sportsman, the system administrator or a trainer. Complete architecture is presented on the deployment diagram (Fig. 1).

2.2 Use-Case Perspective

The use-case perspective describes how user activity affects network communication and task scheduling for the system. To set up a training process, the sportsman wears the sensors (Shimmer[5], Zephyr HxM or Zephyr BT[6]) and launches a mobile application, which is implemented in JAVA for Android. The main purpose of a software installed on the smartphone is to provide functionalities of data acquisition and visualization. The mobile application establishes Bluetooth connections with all sensors worn by the sportsman. Then it collects data, presents them to the user and transmits via General Packet Radio Service (GPRS), 3^{rd} generation of mobile telecommunications technology (3G) or Wireless Fidelity (Wi-Fi) to data processing servers using Extensible Messaging and Presence Protocol (XMPP). Data processing servers hosts computational services. For example, in the case when the sportsman trains in a gym or any court, services will connect to network video cameras and employ computer vision algorithms to analyze the sportsman's performance. Results of data processing are sent to the web site or to a desktop application using XMPP.

2.3 The System Tiers

As it was mentioned, the system is composed of three tiers. The first one consists of Bluetooth sensors, mobile devices (phones, tablets) and JAVA application for sportsman

[5] `www.shimmer-research.com`, accesed: 4 April 2012.
[6] `www.zephyr-technology.com/products/`, accesed: 4 April 2012.

(see architecture in Fig. 2). There are three significant types of components in mobile phone application: network controllers, sensor (device) drivers and scenarios. All of them are JAVA classes, which provide application scalability using polymorphism. The set of device drivers is easily extensible simplifying the code maintenance and allowing to be up-to-date with market novelties. The scenario is JAVA package that allows sportsman to monitor his activity. It may provide: data visualization on a smartphone, data preprocessing calculation of simple statistics, notifications, etc.

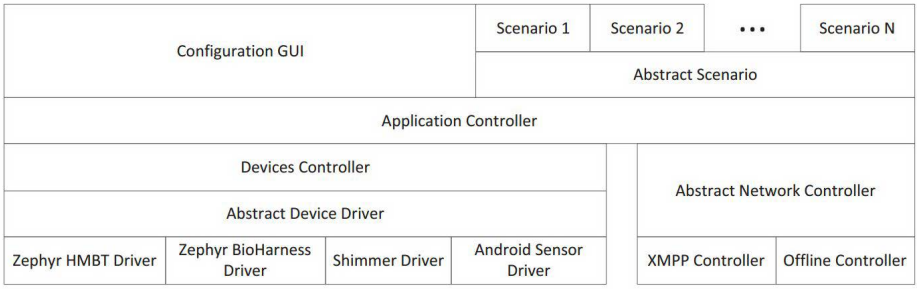


Fig. 2. Architecture of JAVA application for the sportsmen's smartphone

3 Support of Endurance Training Task

An illustrative example of the system describes a custom application that serves to generate the schedule of a single training session (the so called training protocol). Typically, for such applications, the user data undergo the following processing stages:

- parameters of the exerciser's model estimation;
- model based optimization of the training protocol (planning); and
- training support by the control algorithm.

The following computational services should be available in order to support physical activity of the user:

- the system identifier that makes use of some machine learning techniques to adjust the user model to measurement data;
- the system simulator that may predict the user state for different training protocols;
- training protocol optimizer that works out the best training protocol;
- the controller that supports the user in performing optimal training protocol.

This can be demonstrated by simply focusing on a footrace, where all stages of data processing mentioned before are present. The goal of a footrace is to run a given distance in the shortest possible time. Some limitations on the training intensity may be introduced if the user has any health problems. Typical health problems include: cardiovascular diseases, obesity and diabetics. In case of any cardiovascular problems or obesity, the heart rate must be monitored to prevent such incidents as fainting, heart

attack and brain stroke. In case of diabetic, it is crucial to keep blood glucose level within the normal range [11]. This task requires the use of glucometer to measure blood glucose level.

The most important issue is to spread the effort of a runner in such a way that the time of the run is the shortest. Similar applications are being considered by others [16, 17, 25]. In this research the treadmill in a laboratory or fitness club is used. The treadmill controls speed with high accuracy. Our application takes into account that the user may perform physical activity at many different places (in the forest or along city streets). Therefore, the user speed is controlled with low accuracy by notifications generated by the smartphone.

Two variables are measured by wireless sensors in a real-time: the user speed (u) and the Heart Rate (HR). All decisions are made on the basis of mathematical model (Eqs. 1-3) which describe the cardiovascular system during exercise [9]:

$$x_1'(t) = -a_1 x_1(t) + a_2 x_2(t) + a_2 u^2(t), \tag{1}$$

$$x_2'(t) = -a_3 x_2 + \phi(x_1(t)). \tag{2}$$

The nonlinear part is:

$$\phi(x_1) = \frac{a_4 x_1}{1 + exp(a_5 - x_1)}, \tag{3}$$

where x_1 is heart rate change from the rest (resting heart rate), u denotes current speed of the user, x_2 may be considered as fatigue, caused by such a factors as: vasodilation in the active muscles leading to low arterial blood pressure, accumulations of metabolic byproducts (for example lactic acid), sweeting and hyperventilation. Fatigue cannot be directly measured, it may only be worked out on the basis of the HR. Parameters $a_1, ..., a_5$ take nonnegative values. Values of these parameters are obtained by the estimation procedure, with use of data measured from few experiments. Each user may be characterized by different values of the model parameters. The user has their own speed limit (u^{max} and fatigue limit x_2^{max}). After the user reaches the fatigue limit, he/she terminates the footrace.

An example of the model response for simple training protocol is given in Fig. 3. The model responses are obtained by solving the system of differential Eqs. 1-3 with the 4-th order multistep Runge-Kutty method [8].

Note that during exercise fatigue increases faster than it decreases during recovery. HR is illustrated as a difference between the current heart rate and the basal heart rate HR_0. The basal HR is the user heart rate value during rest.

The training protocol is a function $u(t)$ (see Fig. 4) composed of step-like functions describing the recovery (zero speed), the exercise (high speed) and the resting (low speed) periods. Each period may be specified by two numbers: duration and speed. The distance D made by the user is equivalent to the area under the $u(t)$:

$$D = \int_0^T u(t)dt, \tag{4}$$

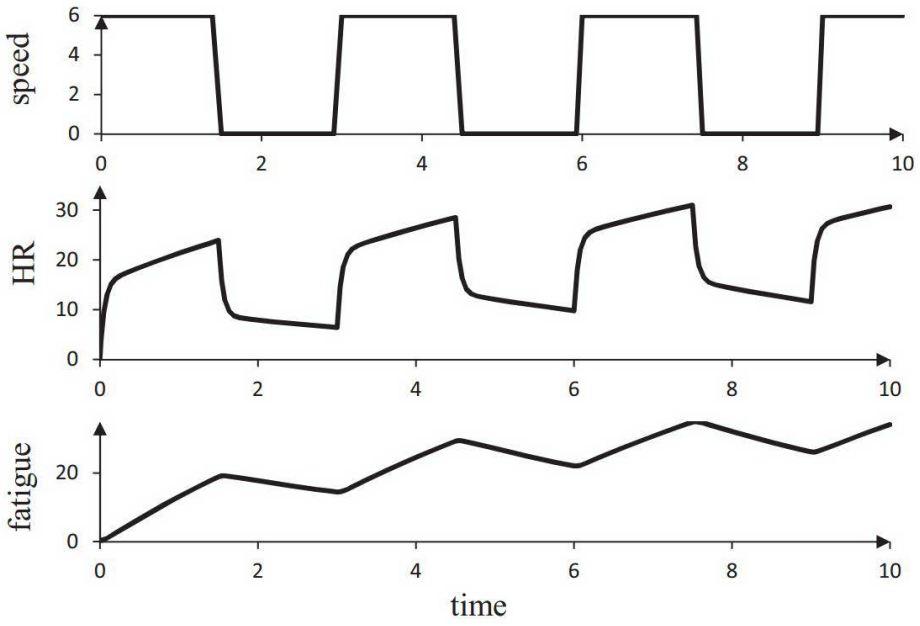


Fig. 3. The model response for the step like input signal

where T is the total footrace time.

In order to learn the user parameter, many heart rate profiles are registered during different training protocols. Then, an identification algorithm is applied in order to estimate values of its parameters. Two main problems with identification performance concern: the model nonlinearity, immeasurable variable x_2 in the model. To cope with those problems, nonlinear identification algorithm is performed. In the work [9] authors employed the Levenberg-Marquardt method [8]. We decided to use simulated annealing [8].

On the basis of the model, the search for optimal training protocol is conducted. The optimization task is: for a **given**: distance D to be made, the user model, fatigue limit $x_2^{\max}$, speed limit $u^{\max}$, the task is **to find** such a training protocol $u(t)$, for which a desired distance is completed in a shortest time T^*. Thus, the performance index is the time T that is solution of the Eq. 4 with respect T, for given D and $u(t)$.

There is also a constraint for fatigue:

$$\max_{0\le t\le T} x_2(t) \le x_2^{\max},$$

where x_2 is related to the model Eqs. 1-3. It seems reasonable to force the user to do his best, by introducing requirement, that at the end of the race the fatigue reaches its maximum value: $x_2(T) = x_2^{max}$.

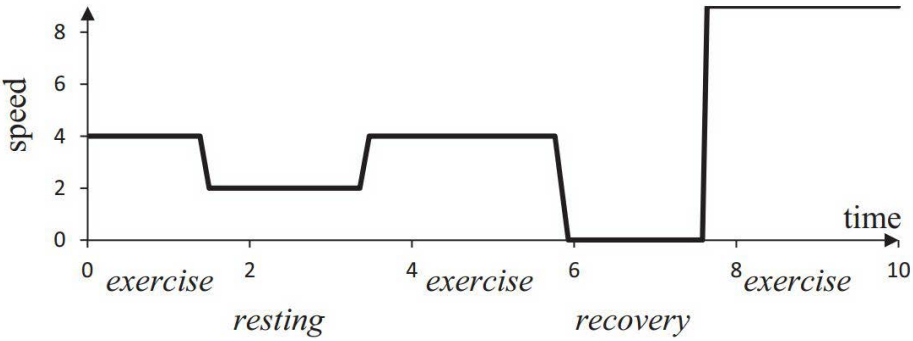


Fig. 4. Typical training protocol

Moreover, due to fitness level, the user cannot run faster than $u^{\max}$:

$$\max_{0 \le t \le T} u(t) \le u_{\max},$$

In order to solve the optimization problem posed above, it is necessary to search the space of feasible functions $u(t)$ witht $t \in < 0, T >$. To reduce the problem to real-valued optimization we parameterized the function $u(t)$. It is represented as a sequence of pairs of numbers, describing the duration and the speed of successive exercise periods. Parameterizations makes optimization easier, but introduces another problem. Since the footrace is completed after the distance D is made, the number of parameters describing the solution may vary during the optimization process. Therefore, we applied simulated annealing as optimization routine. Sometimes training protocols generated by simulated annealing are similar to those obtained by the well known, in physiological literature, Interval Method [28], but the most common type of protocol is different. It looks very interesting and an example is depicted in Fig. 5. The optimal training protocol is just a reference signal. The next step is to actuate it. The user is expected to follow it as close as possible. The user should be supported in switching between exercise, resting and recovery periods as well as in maintaining the correct speed during these periods. This may be done using voice commands uttered by the smartphone (for example "run faster", "slow down please") or by modifying music genre. In the former case it suffices to employ on/off controller with hysteresis.

4 Tracking Execution of Exercise Plan

In the previous section, endurance scenario was discussed – a type of training, which is composed of only one exercise. Nonetheless, in majority of cases, both professional and amateur training plan is consisted of a sequence of strictly defined exercises, which serve to improve different sportsman's characteristics such as endurance, strength and technique. In order to extend mentioned approach, and make the system more flexible, a method of recognizing the sportsman activity in real-time is required.

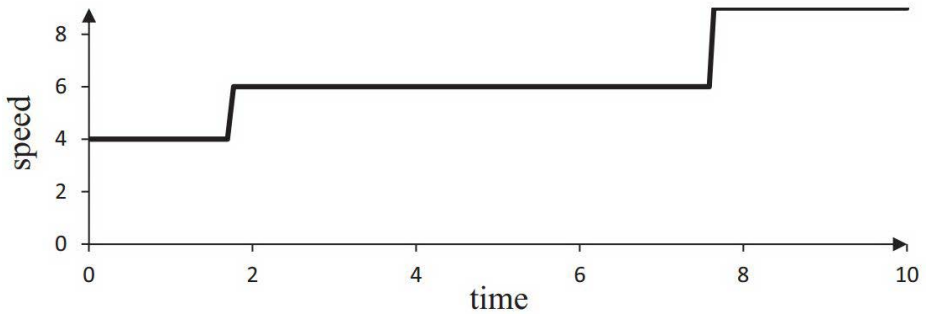


Fig. 5. The training protocol commonly generated by simulated annealing

Results of such detection might be utilized in the few different ways. First of all, we could switch between mathematical models in order to conduct decision support, which is adjusted to one particular exercise. Secondly, we could take advantage of well-established statistics such as the metabolic equivalent described by Plowman [29] to provide a feedback, which corresponds to the whole training period.

Because of mentioned reasons, the authors were motivated to investigate a machine learning approach to track an exercise plan execution.

4.1 Problem Formulation

The research relies on two three-axis accelerometers, which are mounted on sportsman thigh and wrist. This was an arbitrary decision based on a literature review or scientific papers related to sporting disciplines, fall detection and Activities of Daily Living (ADL) recognition. In the recent studies, usually two or three sensors were used [20, 21] and they were located on either the waist, wrist or chest [1, 7, 12, 25, 27]. Such approaches guaranteed a satisfying performance. Results of the literature review are presented in the Table 1.

After experimental environment is defined, it is possible to introduce two new research questions. These include:
RQ1: What feature extraction method is the most appropriate for tracking execution of exercise plan?
RQ2: Which classifier achieve the best performance in this context?

4.2 The Experiment – Selection of "Time Window" Size

An experiment was conducted in order to answer to the research questions. At first, the authors recorded volunteer data, who were doing the following exercises in varying order: running, sit-ups, push-ups, squat, stretching and weight lifting. After that, all recordings were labeled and divided into two sets, one for learning (60% of samples) and the second as test set (40%).

Secondly, variables were identified in the experiment: size of "time window" (Fig. 6), feature vector and type of classifier (Naive Bayes, k-NN, decision tree or decision table).

Table 1. Activity detection based on accelerometer signals - results of literature review [25]-[12]

Type of activity	Classification accuracy	Sensor location
walking	92.85% - 95.91%	thigh
walking, posture	80.00% - 90.00%	left hip, right hip
walking, posture	89.30%	chest, thigh
Kung Fu strike	96.67%	wrist
walking, posture, writing, bicycling	95.80%	chest, hip, wrist, forearm

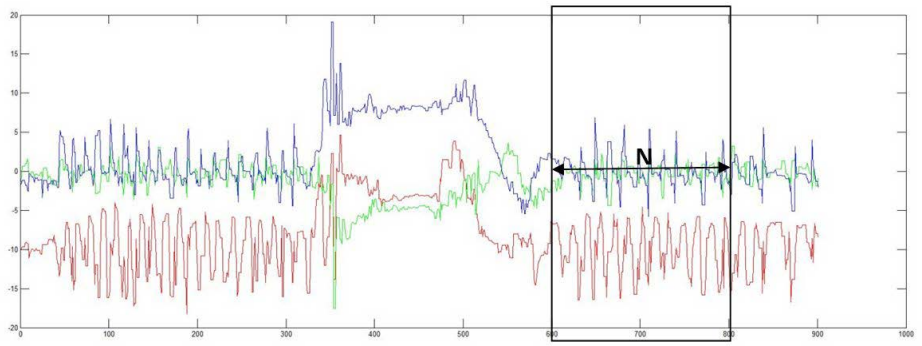


Fig. 6. Exemplary signal from accelerometer. In the picture a "time window" of size N is presented. In the experiment we assumed 7 possible values of N - $16, 32, 64, 128, 256, 512$ and 1024.

Thirdly, given maximal feature vector (feature vector, that is composed of all considered features) as a constant, we classified test set using all possible classifiers and "time window" sizes. The best results were obtained for Naive Bayes (Table 2).

According to the simulations, regardless for classifier, the best results were achieved, when "time window" size was set to 128, what corresponds to 2.56 s. of exercise (50Hz sampling rate).

4.3 The Experiment - Selection of Classifier and Feature Vector

In the second phase of experiment, we set "time window" size as a constant (128 measurements), and conducted simulation using different combinations of classifiers and feature vectors.

At given time t, accelerometers signals might be represented as six elements vector $m(t)$, where each element correspond to one axis (in the simulation, we are using two three-axis sensors). Based on that data representation, the authors started their simulations with the following features:

$$x^{(1)} = \frac{1}{N}\sum_{n=0}^{N} m(t-n) \tag{5}$$

Table 2. Accuracy metric calculated for each time window size. Results correspond to performance of the best classifier (Naive Bayes)

Time window size	Running	Sit-ups	Weight lifting	Push-ups	Squats	Stretching
16	89.58%	57.29%	78.13%	99.22%	19.01%	100.00%
32	100.00%	80.11%	79.06%	97.92%	62.83%	100.00%
64	100.00%	100.00%	79.17%	98.96%	100.00%	100.00%
128	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
256	100.00%	100.00%	79.17%	100.00%	100.00%	100.00%
512	100.00%	100.00%	75.00%	100.00%	100.00%	100.00%
1024	100.00%	100.00%	33.33%	100.00%	100.00%	100.00%

$$x^{(2)} = \frac{1}{N-1}\sum_{n=0}^{N}(m(t-n)-x^{(1)})^2 \tag{6}$$

$$x^{(3)} = \frac{1}{N}\sum_{n=0}^{N}|m(t-n)|^2 \tag{7}$$

The introduced features (Eqs. 5-7) corresponds to mean value of the signal, standard deviation and energy respectively. The parameter N represents a size of "time window". As is was mentioned before, a simulation was conducted for each classifier and feature vector, given N=128. The results are presented in Table 3.

Table 3. Accuracy of classification

Feature(s)	Naive Bayes	Nearest Neighbors	Decision Tree
$x^{(1)}$	84.03%	86.46%	52.78%
$x^{(2)}$	100.00%	98.26%	79.17%
$x^{(3)}$	100.00%	99.65%	77.78%
$x^{(1)},x^{(2)}$	98.96%	92.02%	80.56%
$x^{(1)},x^{(3)}$	100.00%	99.31%	77.78%
$x^{(2)},x^{(3)}$	100.00%	97.57%	81.95%
$x^{(1)},x^{(2)},x^{(3)}$	100.00%	98.26%	81.95%

To a great surprise of the authors, introduction of three simple features is enough to solve this particular problem. Nonetheless, the authors expect, that if number of analyzed exercises grow up, the number of features should be increased.

Presented method of tracking exercise execution plan is one of the elements of the wearable system for support decision in sport training. This element allows to choose, in real-time, a mathematical model, which is suitable for sportsman exercise. Such solution, aimed at automated model selection and adjustment, allows us to build comprehensive system, that could serve in both amateur and professional sport training purposes.

5 Use-Case Scenario

The detailed list of available computational services includes:

- Differential Equations Solver (DES);
- the System Identification Service (SIS);
- the Model Validation Service (MVS);
- the Training Protocol Optimizer (TPO) and
- the Speed Control Service (SCS).

MVS returns the result of binary classification: whether parameters of the user model are up to date or not. The service is fed up by the latest measurement data and the current user model. If the user model responses become significantly different than the measured ones, DES should be run to update the parameters. Actuation of the training protocol is supported by SCS. Since DES and SCS are simple procedures, their executable files may be sent to the smartphone and then run. On the other hand, MVS, SIS and TPO need high computational power, therefore they need to be run in a computer center. The smartphone only receives results of computations.

The user story is the following. Before the application is used to support physical activity, the user is asked to perform few predefined training protocols. This is the first adjustment of the model to the user and SIS is executed. Typical usage scenario starts from typing:

- the footrace distance and
- maximum speed then
- the limit of fatigue (maximum speed and fatigue limit are learned by the system later on).

TPO works out optimal training protocol on the basis of the user model simulations made by DES. The training protocol is actuated by SCS. In the background, MVS compares the model responses to the measured signals. If the difference becomes significant, SIS is called to update the user model and then TPO is run once again, with the new user model. This may happen mainly due to worse or better (compared to average) user condition (which takes place very often after the party) and problems with tracking the training protocol (for example after the fall). Each time MVS detects that the user does not follow the reference trajectory, TPO recalculates optimal training protocol for the remaining part of the route. With time, the user improves (or decreases) his/her performance and the model updates are necessary. Therefore, from time to time, MVS checks the user model validity and, if necessary, SIS is called to update the user model parameters. All the mentioned services provide adaptive control system supporting the user in a real-time (see Fig. 7). Adaptation takes place in response to events occurring during the single training as well as reaction to long term effects caused by systematic training.

6 Final Remarks

The proposed application is just an example of personalization scheme which relies on composition of new services on the basis of predefined services. It is possible to deliver

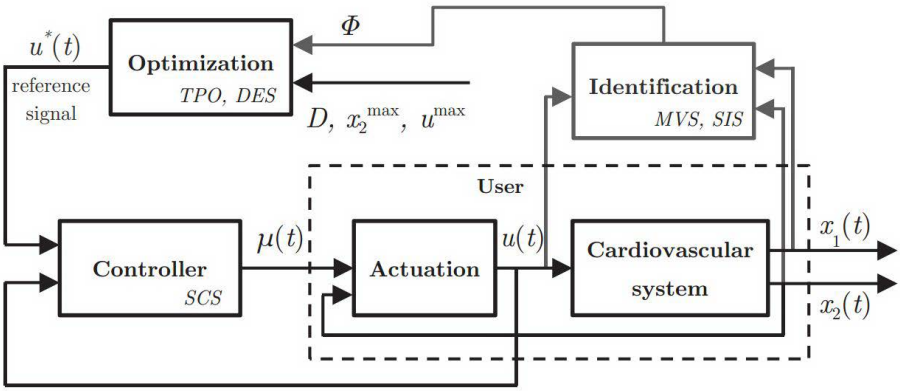


Fig. 7. Adaptive control system

more advanced applications, making use of services derived from other applications. For instance, when one wants to deliver the same applications for a diabetic user, a new service may be composed. This service may support physical training, taking blood glucose level into account. In such a case, additional service to keep blood glucose level within normal range is executed. Additional constraint for TPO is defined to make sure that the training scenario generated by TPO will be safe for the user. Further works focus on development of advanced functionalities that make use of pattern recognition methods.

Acknowledgement. The research presented in this paper has been partially supported by the European Union within the European Regional Development Fund program no. POIG.01.01.02-00-045/09.

References

1. Aminian, K., Robert, P., Buchser, E.E., Rutschmann, B., Hayoz, D., Depairon, M.: Physical activity monitoring based on accelerometry: validation and comparison with video observation. Medical & Biological Engineering & Computing 37(3), 304–308 (1999)
2. Brzostowski, K., Drapała, J., Grzech, A., Świątek, P.: Adaptive decision support system for automatic physical effort plan generation: data-driven approach. Cybernetics and Systems 44(2/3), 204–221 (2013)
3. Brzostowski, K., Drapała, J., Świątek, J.: System analysis techniques in eHealth systems: A case study. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ACIIDS 2012, Part I. LNCS (LNAI), vol. 7196, pp. 74–85. Springer, Heidelberg (2012)
4. Bourke, A.K., O'Brien, J.V., Lyons, G.M.: Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm. Gait & Posture 26(2), 194–199 (2007)
5. Burakowski, W., Bęben, A., Tarasiuk, H., Śliwiński, J., Janowski, R., Batalla, J.M., Krawiec, P.: Provision of End-to-End QoS in Heterogeneous Multi-Domain Networks. Annals of Telecommunications 63, 559–577 (2008)
6. Buttussi, F., Chittaro, L.: MOPET: A Context-Aware and User-Adaptive Wearable System for Fitness Training. Artificial Intelligence in Medicine 42, 153–163 (2008)

7. Chambers, G.S., Venkatesh, S., West, G.A.W., Bui, H.H.: Hierarchical recognition of intentional human gestures for sports video annotation. In: Proc. of the 16th International Conference on Pattern Recognition, vol. 2, pp. 1082–1085 (2002)
8. Cheney, W., Kincaid, D.: Numerical Mathematics and Computing, 6th edn. Thomson Brooks/Cole (2008)
9. Cheng, T.M., Savkin, A.V., Celler, B.G., Su, S.W., Wang, L.: Nonlinear Modeling and Control of Human Heart Rate Response During Exercise With Various Work Load Intensities. IEEE Trans. on Biomedical Engineering 55(11), 2499–2508 (2008)
10. Cheng, T.M., Savkin, A.V., Celler, B.G., Su, S.W., Wang, L.: Heart Rate Regulation During Exercise with Various Loads: Identification and Nonlinear H_inf Control. In: Proc. of the 17th World Congress of The International Federation of Automatic Control IFAC, Seoul, Korea, pp. 11618–11623 (2008)
11. Dalla, M.C., Breton, M.D., Cobelli, C.: Physical Activity into the Meal Glucose-Insulin Model of Type 1 Diabetes: In Silico Studies. Journal of Diabetes Science and Technology 3, 56–67 (2009)
12. Foerster, F., Smeja, M., Fahrenberg, J.: Detection of posture and motion by accelerometry: a validation in ambulatory monitoring. Computers in Human Behaviour 15, 571–583 (1999)
13. Fraś, M., Grzech, A., Juszczyszyn, K., Kołaczek, G., Kwiatkowski, J., Prusiewicz, A., Sobecki, J., Świątek, P., Wasilewski, A.: Smart Work Workbench: integrated tool for IT services planning, management, execution and evaluation. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS (LNAI), vol. 6922, pp. 557–571. Springer, Heidelberg (2011)
14. Greene, B.R., McGrath, D., O'Neill, R., O'Donovan, K.J., Burns, A., Caulfield, B.: An adaptive gyroscope-based algorithm for temporal gait analysis. Journal of Medical and Biological Engineering and Computing 48, 1251–1260 (2010)
15. Grzech, A., Świątek, P., Rygielski, P.: Dynamic Resources Allocation for Delivery of Personalized Services. In: Cellary, W., Estevez, E. (eds.) Software Services for e-World. IFIP AICT, vol. 341, pp. 17–28. Springer, Heidelberg (2010)
16. Grzech, A., Rygielski, P.: Translations of service level agreement in systems based on service oriented architecture. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010, Part II. LNCS (LNAI), vol. 6277, pp. 523–532. Springer, Heidelberg (2010)
17. Grzech, A., Świątek, P.: Modeling and optimization of complex services in service-based systems. Cybernetics and Systems 40(8), 706–723 (2009)
18. Grzech, A., Świątek, P.: Complex Services Availability in Service Oriented Systems. In: Proc. of 21st International Conference on Systems Engineering, Las Vegas, Nevada, August 16-18, pp. 227–232. IEEE Computer Society Conference Publishing Services (2011)
19. Janiak, A., Kozik, A., Lichtenstein, M.: New perspectives in VLSI design automation: deterministic packing by sequence pair. Annals of Operations Research 179(1), 35–56 (2010)
20. Kangas, M., Konttila, A., Lindgren, P., Winblad, I., Jämsä, T.: Comparison of low-complexity fall detection algorithms for body attached accelerometers. Gait Posture 28(2), 285–291 (2008)
21. Kangas, M., Vikman, I., Nyberg, L., Korpelainen, R., Lindblom, J., Jämsä, T.: Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects. Gait Posture 35(3), 500–505 (2012)
22. Kołaczek, G.: Multiagent security evaluation framework for service oriented architecture systems. In: Velásquez, J.D., Ríos, S.A., Howlett, R.J., Jain, L.C. (eds.) KES 2009, Part I. LNCS (LNAI), vol. 5711, pp. 30–37. Springer, Heidelberg (2009)
23. Kosek, K., Natkaniec, M., Vollero, L., Pach, A.R.: An Analysis of Star Topology IEEE 802.11e Networks in the Presence of Hidden Nodes. In: The Internation Conference on Information Networking, ICOIN, Busan, Korea, pp. 1–5 (2008)

24. Liu, S., Gao, R.X., John, D., Staudenmayer, J.W., Freedson, P.S.: Multisensor Data Fusion for Physical Activity Assessment. IEEE Trans. on Biomedical Engineering 59(3), 687–696 (2012)
25. Lee, S.W., Mase, K.: Activity and location recognition using wearable sensors. IEEE Pervasive Computing 1(3), 24–32 (2002)
26. Lornicz, K., Chen, B., Challen, G.W.: Mercury: A Wearable Sensor Network Platform for High-Fidelity Motion Analysis. In: Conference on Embedded Networked Sensor Systems, SenSys 2009, pp. 183–196 (2009)
27. Mantyjarvi, J., Himberg, J., Seppanen, T.: Recognizing human motion with multiple acceleration sensors. In: IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, pp. 747–752 (2001)
28. Michailov, M.: Evolvement and Experimentation of a New Interval Method For Strength Endurance Development. The Engineering of Sport 6, 291–296 (2006)
29. Plowman, S.A., Smith, D.L.: Exercise Physiology for Health, Fitness and Performance. Lippincott Williams & Wilkins (2010)
30. Prusiewicz, A., Zięba, M.: The Proposal of Service Oriented Data Mining System for Solving Real-Life Classification and Regression Problems. In: Camarinha-Matos, L.M. (ed.) Technological Innovation for Sustainability. IFIP AICT, vol. 349, pp. 83–90. Springer, Heidelberg (2011)
31. Rygielski, P., Gonczarek, A.: Migration-aware Optimization of Virtualized Computational Resources Allocation in Complex Systems. In: Proc. of 21st International Conference on Systems Engineering, Las Vegas, Nevada, August 16-18, pp. 212–216. IEEE Computer Society Conference Publishing Services (2011)
32. Rygielski, P., Świątek, P.: Graph-fold: an efficient method for complex services execution plan optimization. Systems Science 36(3), 25–32 (2010)
33. Rygielski, P., Tomczak, J.M.: Context change detection for resource allocation in service-oriented systems. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS (LNAI), vol. 6882, pp. 591–600. Springer, Heidelberg (2011)
34. Scalzi, S., Tomei, P., Verrelli, C.M.: Nonlinear Control Techniques for the Heart Rate Regulation in Treadmill Exercises. IEEE Trans. on Biomedical Engineering 59(3), 599–603 (2012)
35. Scully, C.G., Lee, J., Meyer, J., Gorbach, A.M., Granquist-Fraser, D., Mendelson, Y., Chon, K.H.: Physiological Parameter Monitoring from Optical Recordings With a Mobile Phone. IEEE Trans. on Biomedical Engineering 59(2), 303–306 (2012)
36. Stelmach, P., Grzech, A., Juszczyszyn, K.: A model for automated service composition system in SOA environment. In: Camarinha-Matos, L.M. (ed.) Technological Innovation for Sustainability. IFIP AICT, vol. 349, pp. 75–82. Springer, Heidelberg (2011)
37. Świątek, P., Juszczyszyn, K., Brzostowski, K., Drapała, J., Grzech, A.: Supporting content, context and user awareness in Future Internetapplications. In: Álvarez, F., et al. (eds.) FIA 2012. LNCS, vol. 7281, pp. 154–165. Springer, Heidelberg (2012)
38. Tomczak, J.M.: On-line change detection for resource allocation in service-oriented systems. In: Camarinha-Matos, L.M., Shahamatnia, E., Nunes, G. (eds.) DoCEIS 2012. IFIP AICT, vol. 372, pp. 51–58. Springer, Heidelberg (2012)
39. Świątek, P., Rygielski, P., Juszczyszyn, K., Grzech, A.: User assignment and movement predistion in wireless networks. Cybernetics and Systems 43(4), 340–353 (2012)
40. Świątek, P., Stelmach, P., Prusiewicz, A., Juszczyszyn, K.: Service Composition in Knowledge-Based SOA Systems. New Generation Computing 30(2-3), 165–188 (2012)
41. Twomey, N., Faul, S., Marnane, W.P.: Comparison of accelerometer-based energy expenditure estimation algorithms. In: Proc. of the 4th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), pp. 1–8. IEEE Press, Munich (2010)

Author Index