

LNAI 6422

Jeng-Shyang Pan  
Shyi-Ming Chen  
Ngoc Thanh Nguyen (Eds.)

# Computational Collective Intelligence Technologies and Applications

Second International Conference, ICCCI 2010  
Kaohsiung, Taiwan, November 2010  
Proceedings, Part II

2  
Part II

 Springer



Lecture Notes in Artificial Intelligence 6422

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Jeng-Shyang Pan Shyi-Ming Chen  
Ngoc Thanh Nguyen (Eds.)

# Computational Collective Intelligence

## Technologies and Applications

Second International Conference, IICCI 2010  
Kaohsiung, Taiwan, November 10-12, 2010  
Proceedings, Part II

**Series Editors**

Randy Goebel, University of Alberta, Edmonton, Canada

Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Wolfgang Wahlster, DFKI and University of Saarland, Saarbrücken, Germany

**Volume Editors**

Jeng-Shyang Pan

National Kaohsiung University of Applied Sciences

Department of Electronic Engineering

415 Chien-Kung Road, Kaohsiung 807, Taiwan

E-mail: jspan@cc.kuas.edu.tw

Shyi-Ming Chen

National Taiwan University of Science and Technology

Department of Computer Science and Information Engineering #43, Sec.4

Keelung Rd., Taipei, 106,Taiwan

E-mail: smchen@mail.ntust.edu.tw

Ngoc Thanh Nguyen

Wroclaw University of Technology, Institute of Informatics

Str. Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland

E-mail: ngoc-thanh.nguyen@pwr.wroc.pl

Library of Congress Control Number: 2010937276

CR Subject Classification (1998): I.2, I.2.11, H.3-4, C.2, D, H.5

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-642-16731-4 Springer Berlin Heidelberg New York

ISBN-13 978-3-642-16731-7 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

[springer.com](http://springer.com)

© Springer-Verlag Berlin Heidelberg 2010

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India  
Printed on acid-free paper 06/3180

## Preface

This volume composes the proceedings of the Second International Conference on Computational Collective Intelligence—Technologies and Applications (ICCCI 2010), which was hosted by National Kaohsiung University of Applied Sciences and Wroclaw University of Technology, and was held in Kaohsiung City on November 10-12, 2010. ICCCI 2010 was technically co-sponsored by Shenzhen Graduate School of Harbin Institute of Technology, the Tainan Chapter of the IEEE Signal Processing Society, the Taiwan Association for Web Intelligence Consortium and the Taiwanese Association for Consumer Electronics. It aimed to bring together researchers, engineers and policymakers to discuss the related techniques, to exchange research ideas, and to make friends. ICCCI 2010 focused on the following themes:

- Agent Theory and Application
- Cognitive Modeling of Agent Systems
- Computational Collective Intelligence
- Computer Vision
- Computational Intelligence
- Hybrid Systems
- Intelligent Image Processing
- Information Hiding
- Machine Learning
- Social Networks
- Web Intelligence and Interaction

Around 500 papers were submitted to ICCCI 2010 and each paper was reviewed by at least two referees. The referees were from universities and industrial organizations. 155 papers were accepted for the final technical program. Four plenary talks were kindly offered by: Gary G. Yen (Oklahoma State University, USA), on “Population Control in Evolutionary Multi-objective Optimization Algorithm,” Chin-Chen Chang (Feng Chia University, Taiwan), on “Applying De-clustering Concept to Information Hiding,” Qinyu Zhang (Harbin Institute of Technology, China), on “Cognitive Radio Networks and Its Applications,” and Lakhmi C. Jain (University of South Australia, Australia), on “Intelligent System Design in Security.”

We would like to thank the authors for their tremendous contributions. We would also express our sincere appreciation to the reviewers, Program Committee members and the Local Committee members for making this conference successful. Finally,

we would like to express special thanks for the financial support from the National Kaohsiung University of Applied Sciences, Kaohsiung City Government, National Science Council and Education Ministry, Taiwan, in making ICCCI 2010 possible.

Novermber 2010

Ngoc Thanh Nguyen  
Jeng-Shyang Pan  
Shyi-Ming Chen  
Ryszard Kowalczyk

# **ICCCI 2010 Conference Organization**

## **Honorary Chair**

Chun-Hsiung Fang	National Kaohsiung University of Applied Sciences, Taiwan
Jui-Chang Kung	Cheng Shiu University, Taiwan

## **General Chair**

Ngoc Thanh Nguyen	Wroclaw University of Technology, Poland
-------------------	--

## **Program Committee Chair**

Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Ryszard Kowalczyk	Swinburne University of Technology, Australia

## **Special Session Chairs**

Bao-Rong Chang	National University of Kaohsiung, Taiwan
Chang-Shing Lee	National University of Tainan, Taiwan
Radoslaw Katarzyniak	Wroclaw University of Technology, Poland

## **International Advisory Chair**

Bin-Yih Liao	National Kaohsiung University of Applied Sciences, Taiwan
--------------	--

## **International Publication Chair**

Chin-Shin Shieh	National Kaohsiung University of Applied Sciences, Taiwan
Bing-Hong Liu	National Kaohsiung University of Applied Sciences, Taiwan

## Local Organizing Committee Chair

Mong-Fong Horng

National Kaohsiung University of Applied Sciences,  
Taiwan

## ICCCI 2010 Steering Committee

### Chair

Ngoc Thanh Nguyen

Wroclaw University of Technology, Poland

### Co-chair

Ryszard Kowalczyk

Swinburne University of Technology, Australia  
National Taiwan University of Science and  
Technology, Taiwan

Shyi-Ming Chen

Wroclaw University of Technology, Poland  
University of South Australia, Australia  
Inha University, South Korea  
Polish Academy of Sciences, Poland  
AGH-UST, Poland  
Kyoto University, Japan

Adam Grzech

Lakhmi C. Jain

Geun-Sik Jo

Janusz Kacprzyk

Ryszard Tadeusiewicz

Toyoaki Nishida

## ICCCI 2010 Technical Program Committee

Jeng Albert B.

Jinwen University of Science and Technology, Taiwan

Gomez-Skarmeta Antonio F.

Murcia University, Spain

Shih An-Zen

Jinwen University of Science and Technology, Taiwan

Andres Cesar

Universidad Complutense de Madrid, Spain

Hsieh Cheng-Hsiung

Chaoyang University of Technology, Taiwan

Lee Chin-Feng

Chaoyang University of Technology, Taiwan

Badica Costin

University of Craiova, Romania

Godoy Daniela

Unicen University, Argentina

Barbucha Dariusz

Gdynia Maritime University, Poland

Greenwood Dominic

Whitestein Technologies, Switzerland

CAPKOVIC Frantisek

Slovak Academy of Sciences, Slovakia

Yang Fuw-Yi

Chaoyang University of Technology, Taiwan

Huang Hsiang-Cheh

National University of Kaohsiung, Taiwan

Chang Hsuan-Ting

National Yunlin University of Science and Technology,  
Taiwan

Lee Huey-Ming

Chinese Culture University, Taiwan

Deng Hui-Fang

South China University of Technology, China

Czarnowski Ireneusz

Gdynia Maritime University, Poland

Lu James J.

Emory University, USA

Kacprzyk Janusz

Polish Academy of Sciences, Poland

Marecki Janusz	IBM T.J. Watson Research, USA
Sobecki Janusz	Wroclaw University of Technology, Poland
Jung Jason J.	Yeungnam University, South Korea
Nebel Jean-Christophe	Kingston University, USA
Dang Jiangbo	Siemens Corporate Research, USA
Huang Jingshan	University of South Alabama, USA
Chang Jui-fang	National Kaohsiung University of Applied Sciences, Taiwan
Nunez Manuel	Universidad Complutense de Madrid, Spain
Gaspari Mauro	University of Bologna, Italy
Khurram Khan Muhammad	King Saud University, Saudi Arabia
Sheng Quan Z.	University of Adelaide, Australia
Katarzyniak Radoslaw	Wroclaw University of Technology, Poland
Unland Rainer	University of Duisburg-Essen, Germany
Ching Chen Rung	Chaoyang University of Technology, Taiwan
Shen Rung-Lin	National Taipei University, Taiwan
Yang Sheng-Yuan	St. John's University, Taiwan
Yen Shu-Chin	Wenzao Ursuline College of Languages, Taiwan
Chen Shyi-Ming	National Taiwan University of Science and Technology, Taiwan
Zadrozny Slawomir	Polish Academy of Sciences, Poland
Hammoudi Slimane	ESEO, France
Hong Tzung-Pei	National University of Kaohsiung, Taiwan
Hsu Wen-Lian	Academia Sinica, Taiwan
Pedrycz Witold	University of Alberta, Canada
Baghdadi Youcef	Sultan Qaboos University, Oman
Lo Yu-lung	Chaoyang University of Technology, Taiwan
Cheng Yuh Ming	Shu-Te University, Taiwan
Huang Yung-Fa	Chaoyang University of Technology, Taiwan
Ye Yunming	Harbin Institute of Technology, China

## Keynote Speakers

Gary G. Yen	Oklahoma State University, USA
Lakhmi C. Jain	University of South Australia, Australia
Chin-Chen Chang	Feng Chia University, Taiwan
Qinyu Zhang	Harbin Institute of Technology Shenzhen Graduate School, China

## Program Committee of Special Sessions

Dariusz Barbucha	Gdynia Maritime University, Poland
Bao-Rong Chang	National University of Kaohsiung, Taiwan
Hsuan-Ting Chang	National Yunlin University of Science and Technology, Taiwan

Chuan-Yu Chang	National Yunlin University of Science and Technology, Taiwan
Rung-Ching Chen	Chaoyang University of Technology, Taiwan
Shyi-Ming Chen	National Taiwan University of Science and Technology, Taiwan
Kazimierz Choroś	Wrocław University of Technology, Poland
Mohamed Hassoun	ENSSIB Villeurbanne, France
Mong-Fong Horng	National Kaohsiung University of Applied Sciences, Taiwan
Chien-Chang Hsu	Fu-Jen Catholic University, Taiwan
Wu-Chih Hu	National Penghu University of Science and Technology, Taiwan
Chien-Feng Huang	National University of Kaohsiung, Taiwan
Tien-Tsai Huang	Lunghwa University of Science and Technology, Taiwan
Huey-Ming Lee	Chinese Culture University, Taiwan
Che-Hung Lin	Cheng Shiu University, Taiwan
Lily Lin	China University of Technology, Taiwan
Piotr Jędrzejowicz	Gdynia Maritime University, Poland
Jeng-Shyang Pan	National Kaohsiung University of Applied Sciences, Taiwan
Chia-Nan Wang	National Kaohsiung University of Applied Sciences, Taiwan

## Table of Contents – Part II

### Social Networks

Influence Clubs in Social Networks .....	1
<i>Chin-Ping Yang, Chen-Yi Liu, and Bang Ye Wu</i>	
A Method of Label-Dependent Feature Extraction in Social Networks .....	11
<i>Tomasz Kajdanowicz, Przemysław Kazienko, and Piotr Doskocz</i>	
Identifying Representative Reviewers in Internet Social Media .....	22
<i>Sang-Min Choi, Jeong-Won Cha, and Yo-Sub Han</i>	
Fitcolab Experimental Online Social Networking System .....	31
<i>Haris Memic</i>	
General Network Properties of Friendship Online Social Network .....	41
<i>Haris Memic</i>	

### Innovations in Computation and Application

Fuzzy Decision Making for IJV Performance Based on Statistical Confidence-Interval Estimates .....	51
<i>Huey-Ming Lee, Teng-San Shih, Jin-Shieh Su, and Lily Lin</i>	
Designing a Learning System Based on Voice Mail – A Case Study of English Oral Training .....	61
<i>Huey-Ming Lee and Chien-Hsien Huang</i>	
A Gene Selection Method for Microarray Data Based on Sampling .....	68
<i>Yungho Leu, Chien-Pang Lee, and Hui-Yi Tsai</i>	
The Similarity of Video Based on the Association Graph Construction of Video Objects .....	75
<i>Ping Yu</i>	
Using SOA Concept to Construct an e-Learning System for College Information Management Courses .....	85
<i>Chung C. Chang and Kou-Chan Hsiao</i>	
Cryptanalysis on Sun-Yeh's Password-Based Authentication and Key Distribution Protocols with Perfect Forward Secrecy .....	95
<i>Wen-Gong Shieh and Wen-Bing Horng</i>	

Using Genetic Algorithms for Personalized Recommendation .....	104
<i>Chein-Shung Hwang, Yi-Ching Su, and Kuo-Cheng Tseng</i>	
 <b>Intellinet Signal Processing for Human-Machine Interaction (I)</b>	
Pretreatment for Speech Machine Translation .....	113
<i>Xiaofei Zhang, Chong Feng, and Heyan Huang</i>	
Emotion Aware Mobile Application .....	122
<i>Radoslaw Nielek and Adam Wierzbicki</i>	
Boosting-Based Ensemble Learning with Penalty Setting Profiles for Automatic Thai Unknown Word Recognition .....	132
<i>Jakkrit TeCho, Cholwich Nattee, and Thanaruk Theeramunkong</i>	
AONT Encryption Based Application Data Management in Mobile RFID Environment .....	142
<i>Namje Park and Youjin Song</i>	
A Query Answering Greedy Algorithm for Selecting Materialized Views .....	153
<i>T.V. Vijay Kumar and Mohammad Haider</i>	
Capturing Users' Buying Activity at Akihabara Electric Town from Twitter .....	163
<i>The-Minh Nguyen, Takahiro Kawamura, Yasuyuki Tahara, and Akihiko Ohsuga</i>	
 <b>Novel Approaches to Intelligent Applications</b>	
Private Small-Cloud Computing in Connection with WinCE Thin Client .....	172
<i>Bao Rong Chang, Hsiu Fen Tsai, Chien-Feng Huang, and His-Chung Huang</i>	
Honey Bee Mating Optimization Algorithm for Approximation of Digital Curves with Line Segments and Circular Arcs .....	183
<i>Shu-Chien Huang</i>	
Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment .....	193
<i>Namje Park and Youngsoo Kim</i>	
Efficient GHA-Based Hardware Architecture for Texture Classification .....	203
<i>Shiow-Jyu Lin, Yi-Tsan Hung, and Wen-Jyi Hwang</i>	

Stability by Feedback of the Second Order Singular Distributed Parameter Systems Containing Infinite Many Poles . . . . .	213
<i>Feng Liu, Guodong Shi, and Jianchun Wu</i>	
Mining Generalized Association Rules with Quantitative Data under Multiple Support Constraints . . . . .	224
<i>Yeong-Chyi Lee, Tzung-Pei Hong, and Chun-Hao Chen</i>	
<b>Intelligent Technologies for Medical Related Applications</b>	
Comparison of Edge Segmentation Methods to Tongue Diagnosis in Traditional Chinese Medicine . . . . .	232
<i>Chieh-Hsuan Wang, Ching-Chuan Wei, and Che-Hao Li</i>	
The Grading of Prostatic Cancer in Biopsy Image Based on Two Stages Approach . . . . .	239
<i>Shao-Kuo Tai, Cheng-Yi Li, Yee-Jee Jan, and Shu-Chuan Lin</i>	
Automatic Drug Image Identification System Based on Multiple Image Features . . . . .	249
<i>Rung-Ching Chen, Cho-Tsan Pao, Ying-Hao Chen, and Jeng-Chih Jian</i>	
On the Development of a Brain Simulator . . . . .	258
<i>Wen-Hsien Tseng, Song-Yun Lu, and Hsing Mei</i>	
Obstructive Sleep Apnea Diagnosis from Electroencephalography Frequency Variation by Radial Basis Function Neural Network . . . . .	268
<i>Chien-Chang Hsu and Jie Yu</i>	
Authentication and Protection for Medical Image . . . . .	278
<i>Chih-Hung Lin, Ching-Yu Yang, and Chia-Wei Chang</i>	
<b>Intellinet Signal Processing for Human-Machine Interaction (II)</b>	
FLC-Regulated Speaker Adaptation Mechanisms for Speech Recognition . . . . .	288
<i>Ing-Jr Ding</i>	
Monitoring of the Multichannel Audio Signal . . . . .	298
<i>Eugeniusz Kornatowski</i>	
Watermark Synchronization Based on Locally Most Stable Feature Points . . . . .	307
<i>Jiansheng Qian, Leida Li, and Zhaolin Lu</i>	

Audio Watermarking with HOS-Based Cepstrum Feature .....	316
<i>Bo-Lin Kuo, Chih-Cheng Lo, Chi-Hua Liu, Bin-Yih Liao, and Jeng-Shyang Pan</i>	
Processing Certificate of Authorization with Watermark Based on Grid Environment .....	324
<i>Heng-Sheng Chen, Tsang-Yean Lee, and Huey-Ming Lee</i>	
<b>Novel Approaches to Collective Computations and Systems</b>	
Probability Collectives Multi-Agent Systems: A Study of Robustness in Search .....	334
<i>Chien-Feng Huang and Bao Rong Chang</i>	
An Improved Ant Algorithm for Fuzzy Data Mining .....	344
<i>Min-Thai Wu, Tzung-Pei Hong, and Chung-Nan Lee</i>	
Information-Driven Collective Intelligences .....	352
<i>Francesca Arcelli Fontana, Ferrante Formato, and Remo Pareschi</i>	
Instructional Design for Remedial English e-Learning .....	363
<i>Chia-ling Hsu, Ai-ling Wang, and Yuh-chang Lin</i>	
A Web-Based E-Teaching System under Low Bit-Rate Constraint .....	373
<i>Wei-Chih Hsu, Cheng-Hsiu Li, and Tzu-Hung Chuang</i>	
Using Freeware to Construct a Game-Based Learning System .....	381
<i>Yuh-Ming Cheng and Li-Hsiang Lu</i>	
<b>Intelligent Systems</b>	
Three Kinds of Negations in Fuzzy Knowledge and Their Applications to Decision Making in Financial Investment .....	391
<i>Zhenghua Pan, Cen Wang, and Lijuan Zhang</i>	
Using Fuzzy Neural Network to Explore the Effect of Internet on Quality of Life .....	402
<i>Jui-Chen Huang</i>	
The Power Load Forecasting by Kernel PCA .....	411
<i>Fang-Tsung Liu, Chiung-Hsing Chen, Shang-Jen Chuang, and Ting-Chia Ou</i>	
A Study of USB 3 in Perspective Aspect .....	425
<i>Yeh Wei-Ming</i>	

A Study of CAPTCHA and Its Application to User Authentication . . . . .	433
<i>Albert B. Jeng, Chien-Chen Tseng, Der-Feng Tseng, and     Jiunn-Chin Wang</i>	
TAIEX Forecasting Based on Fuzzy Time Series and the Automatically Generated Weights of Defuzzified Forecasted Fuzzy Variations of Multiple-Factors . . . . .	441
<i>Shyi-Ming Chen and Huai-Ping Chu</i>	
<b>Advanced Knowledge Management (I)</b>	
Concept Document Repository to Support Research of the Coal Industry Development Forecasting . . . . .	451
<i>Liudmila Takayashvili</i>	
Forecasting Coal and Rock Dynamic Disaster Based on Adaptive Neuro-Fuzzy Inference System . . . . .	461
<i>Jianying Zhang, Jian Cheng, and Leida Li</i>	
Context-Aware Workflow Management Based on Formal Knowledge Representation Models . . . . .	470
<i>Fu-Shiung Hsieh</i>	
A Consensus-Based Method for Fuzzy Ontology Integration . . . . .	480
<i>Ngoc Thanh Nguyen and Hai Bang Truong</i>	
Contextual Information Search Based on Ontological User Profile . . . . .	490
<i>Nazim uddin Mohammed, Trong Hai Duong, and Geun Sik Jo</i>	
Rough Sets Based Association Rules Application for Knowledge-Based System Design . . . . .	501
<i>Shu-Hsien Liao and Yin-Ju Chen</i>	
<b>Author Index . . . . .</b>	511

# Influence Clubs in Social Networks

Chin-Ping Yang, Chen-Yi Liu, and Bang Ye Wu

National Chung Cheng University, ChiaYi, Taiwan 621, R.O.C.  
[bangye@cs.ccu.edu.tw](mailto:bangye@cs.ccu.edu.tw)

**Abstract.** A new model “influence club” for cohesion group in a social network is proposed. It generalizes the definition of  $k$ -club and has two advantages. First, the influence between two nodes does not only depend on the their distance but also on the numbers of pathways of different lengths. Second, the new model is more flexible than  $k$ -club and can provide middle results between  $k$ -club and  $(k+1)$ -club. We propose a branch-and-bound algorithm for finding the maximum influence club. For an  $n$ -node graph, the worst-case time complexity is  $O(n^3 1.6^n)$ , and it is much more efficient in practical: a graph of 200 nodes can be processed within 2 minutes. The performance compared to  $k$ -clubs are tested on random graphs and real data. The experimental results also show the advantages of the influence clubs.

**Keywords:** Social network analysis, algorithm, cohesion group, influence,  $k$ -club.

## 1 Introduction

A social network is a graph in which the nodes are actors and an edge represents some kind of social relation between two actors. Finding cohesion groups, or *community detection*, is an important and interesting problem in social network analysis [9,19]. Intuitively we want to find a sufficiently large node subset such that the actors in the group are closely related. The most widely known in this type of research is to find out *clique*, defined as a set of nodes all directly linked to each other by an edge, i.e. a complete subgraph. Although clique is a natural and simple definition for cohesion group, it is not practical at all. Finding the maximum clique in a graph is a well-known NP-hard problem [8], so it is very unlikely to develop a worst case efficient algorithm theoretically. But in the practical perspective, there are several algorithms which are sufficiently efficient for many of applications [11]. However, the main drawback comes from the imperfect data of social networks. Because of the too restrict definition of clique, only few lost edges collapse a clique very much. To overcome this problem, several relaxations have been proposed, such as  $k$ -cliques,  $k$ -clubs,  $k$ -clan,  $k$ -plex, and so on [1,9,13,19].

A  $k$ -*clique* relaxes the distance constraint such that the distance from each node to any other is at most  $k$ . However, the shortest path between two nodes may pass through a node outside the  $k$ -clique. This situation may be troublesome

in some applications. Hence *k*-club was proposed. A *k*-club is defined as a node subset, by which the induced subgraph is of diameter at most *k*. That is, for any pair of nodes in the group, there is a path of length at most *k* and passing only nodes in the group. Club problems seem even more difficult than clique problems although both problems are NP-hard. UCINet is a popular software for social network analysis [18]. It provides a function to compute *k*-clique but no function for *k*-club. Jean-Marie, Gilbert, Gilles [3] showed that the maximum *k*-club problem in an undirected graph is NP-hard and gave an exact branch-and-bound algorithm. They also proposed three heuristics in an earlier paper [4].

The *k*-clique and *k*-club models consider only longest shortest paths among the actors in a group. They implicitly assume that the centralities (powers) of actor are based on geodesic paths but ignore the number of paths which are not shortest. But in many applications it is not suitable. For example, the spread of rumors and diseases does not only depend on the distances but also the number of pathways and their lengths. Centralities on different models have been studied, e.g., network-flow [7] and current-flow [5,15]. In [2], centralities on different types of network flows were studied. The author categorize network flows into several types according to the kinds of trajectories that traffic may follow (geodesics, paths, trails, or walks) and the method of spread (broadcast, serial replication, or transfer).

Let  $\gamma_H$  be a predefined binary relation on pair of nodes in a graph (subgraph)  $H$ . For a given graph  $G$ , finding a maximum group  $U$  such that  $\gamma_G(u, v) = 1$  for any  $u$  and  $v$  in  $U$  can be reduced to the Clique problem. But if the constraint is replaced with  $\gamma_{G[U]}(u, v) = 1$ , i.e., the relation is defined on the subgraph induced by the found group, there is no such reduction, and the problem becomes more complicated. In this paper we propose a generalized club problem, especially the *influence club* problem. *Influence* is a measurement of an actor on the other in a social network. An *influence club* is a group of actors in which any actor has a sufficiently large influence on any other in the group. In the terminology of graph theory, given an undirected and unweighted graph  $G = (V, E)$ , an influence club is a node subset  $C \subseteq V$  such that  $\Phi_{G[C]}(s, t) \geq \theta$  for any nodes  $u, v \in C$ , in which  $\Phi_{G[C]}(s, t)$  is the influence function between  $u$  and  $v$  in the induced subgraph  $G[C]$  and  $\theta$  is a threshold. Our influence function is based on [10,12] and defined by

$$\Phi_H(s, t) = \sum_{P \in \mathcal{P}(H, s, t)} \lambda^{|P|-1}, \quad (1)$$

in which  $\mathcal{P}(H, s, t)$  is the set of all *simple paths* between  $s$  and  $t$  in  $H$ ,  $|P|$  is the length (number of edges) of a path  $P$ , and  $\lambda < 1$  is an attenuation factor (usually set to 0.5).

Another reason to propose the generalized club definition is that it is more flexible. For many social networks, the 1-club (clique) is too restrict but the 2-club is too loose. The influence club proposed in this paper can provide middle results between the two definitions. In this paper, we give a branch and bound algorithm

to find the influence club of maximum cardinality and compare the influence club with  $k$ -club. Experimental results on random graphs and real data are shown. The random graphs includes two models: Poisson [6,17] and Price's model for network growth [14,16]; and the real data comes from PTT (the most popular BBS in Taiwan). The experimental results show the new model is effective and our algorithm is efficient.

The rest of the paper is organized as follows. In section 2, we give some preliminary discussion on the generalized clique and club problems. The algorithms for computing the maximum influence club are in Section 3. Then we show the experimental results and discussions in Section 4. Finally, conclusions are given in Section 5.

## 2 Generalized Club Problem

Let  $G = (V, E, w)$  be a network, in which  $w$  is an edge weight function. For any  $u, v \in V(G)$ , let  $\gamma_G(u, v)$  be a binary function defining the reachability from  $u$  to  $v$  on  $G$ . A  $\gamma$ -clique is a subset  $C$  of  $V$  such that  $\gamma_G(u, v) = 1$  for any  $u, v \in C$ . A  $\gamma$ -club is a subset  $C$  of  $V$  such that  $\gamma_{G[C]}(u, v) = 1$  for any  $u, v \in C$ , in which  $G[C]$  denotes the subgraph of  $G$  induced by  $C$ . For example, the traditional  $k$ -clique and  $k$ -club problems are defined by  $\gamma_H(u, v) = 1$  iff  $d_H(u, v) \leq k$ , where  $d_H(u, v)$  is the distance (the length of the shortest path), and  $H = G$  for clique problem and  $H = G[C]$  for club problem. Once the reachability function is defined, the  $\gamma$ -clique is equivalent to the clique in the graph  $G' = (V, F)$ , in which  $(u, v) \in F$  iff  $\gamma(u, v) = 1$ . But there is no such reduction for the  $\gamma$ -club problem.

Suppose that the weight of edge  $(u, v)$  is the an attenuation factor of the influence of  $u$  to  $v$  or the probability of a message flow from  $u$  to  $v$ . For a simple path  $P$  from  $s$  to  $t$ , the contribution of the influence along path is  $\delta(P) = \prod_{e_i \in P} w(e_i)$ . One way to define the reachability is that  $\gamma_H(s, t) = 1$  iff  $\Phi_H(s, t) = \sum_{P \in \mathcal{P}} \delta(P) \geq \theta$ , in which  $\mathcal{P}$  is the set of all simple paths, or walks as required, from  $s$  to  $t$  in graph  $H$  and  $\theta$  is a threshold. When the edge weights are uniform, it is equivalent to the conventional exponential attenuation shown in (1).

In the following we shall consider only undirected graph with uniform edge weight. First we consider that  $\mathcal{P}$  is the set of all walks. A walk allows repeated node but a path doesn't. Let  $A$  be the adjacency matrix, i.e.,  $A[i, j] = 1$  if  $(i, j) \in E$  and is 0 otherwise. A well-known result is that the number of walks from  $i$  to  $j$  of length  $l$  can be computed by

$$\text{nwalk}^{(l)}(i, j) = A^l[s, t] \quad (2)$$

Thus

$$\Phi_G(s, t) = \sum_{i=1}^{\infty} (\lambda^{i-1}) A^i[s, t] \quad (3)$$

In practical, we don't need to compute walks of all possible lengths. By the exponential attenuation, a long walk has very few contributions to the total

influence. We can usually limit the length walk (or path) by a small integer  $p$ . It turns out the following formula.

$$\Phi_G(s, t) = \sum_{i=1}^p (\lambda^{i-1}) A^i[s, t] \quad (4)$$

For example, using  $p = 2$ ,  $\lambda = 1/2$  and  $\theta = 1$ ,  $\gamma_G(s, t) = 1$  iff  $(s, t) \in E(G)$  or they have at least two common neighbors. When  $\mathcal{P}$  is the set of all paths but not walks, there is no simple general form for calculating the number of paths of length  $i$ . In the next section, we shall show how to compute it for  $p \leq 4$ .

### 3 Algorithms for Maximum Influence Clubs

In this section, we show a branch and bound algorithm for finding the maximum influence club in an undirected unweighted graph  $G = (V, E)$ . The influence function is defined by

$$\Phi_H(s, t) = \sum_{P \in \mathcal{P}(H, s, t)} \lambda^{|P|-1},$$

in which  $\mathcal{P}(H, s, t)$  is the set of all simple paths between  $s$  and  $t$  in  $H$  and  $\lambda < 1$  is an attenuation factor set to  $1/2$ . Let  $\sigma_H^{(i)}(s, t)$  denote the number of paths of length  $i$  between  $s$  and  $t$  in  $H$ . We can rewrite the formula

$$\Phi_H^{(p)}(s, t) = \sum_{i=1}^p 2^{1-i} \sigma_H^{(i)}(s, t) \quad (5)$$

The superscript “ $(p)$ ” indicates we limit the path length to  $p$ . We shall use  $p = 4$  in our experiments. The problem can be formally defined as follows.

**PROBLEM:** Maximum Influence Club Problem.

**INSTANCE:** An undirected graph  $G = (V, E)$  and a threshold  $\theta > 0$ .

**GOAL:** Find a node subset  $C$  of maximum cardinality such that  $\Phi_{G[C]}^{(p)}(s, t) \geq \theta$  for any  $s, t \in C$ .

We remind that when  $p = 2$ ,  $\lambda = 1/2$  and  $\theta = 1/2$ , the  $\gamma$ -club problem is equivalent to the 2-club problem and is therefore NP-hard.

#### 3.1 Influence Function

In this subsection, we show how to compute the influence function. The most important thing is how to find the number of simple paths between two nodes. Apparently  $\sigma_H^{(1)}(s, t) = A[s, t]$ , in which  $A$  is the adjacency matrix. For  $l = 2$ , any walk is a path unless  $s = t$ . Therefore, by (2),

$$\sigma_H^{(2)}(s, t) = A^2[s, t]$$

When  $l = 3$ , if  $A[s, t] = 0$ , a walk is a path. If  $A[s, t] = 1$ , for any neighbor  $x$  of  $s$ , there is a non-path walk  $(s, x, s, t)$ . Therefore we have

$$\sigma_H^{(3)}(s, t) = \begin{cases} A^3[s, t] & \text{if } A[s, t] = 0 \\ A^3[s, t] - (\Delta_s + \Delta_t - 1) & \text{if } A[s, t] = 1 \end{cases} \quad (6)$$

in which  $\Delta_s$  is the degree of  $s$ . The reason of the additional “-1” in the second case of (6) is that the walk  $(s, t, s, t)$  is counted twice. The case of length 4 is more complicated. We shall divide it into 3 sub-cases according to the distance between  $s$  and  $t$ .

- distance > 2: A walk is a path, and  $\sigma_H^{(4)}(s, t) = A^4[s, t]$ .
- distance = 2: The number of paths is  $A^4[s, t] - |N_{st}|(\Delta_s + \Delta_t) - \sum_{x \in N_{st}} (\Delta_x - 2)$ , in which  $N_{st}$  denotes the set of common neighbors of  $s$  and  $t$ .
- distance = 1: The total number of paths is

$$A^4[s, t] - |N_{st}|(\Delta_s + \Delta_t - 5) - \sum_{x \in N_{st}} \Delta_x - A^3[s, s] - A^3[t, t]$$

**Lemma 1.** *The influences, defined by (5), between all pairs of nodes in a graph  $H$  can be computed in  $O(n^3)$  time, in which  $n$  is the number of nodes.*

### 3.2 A Branch and Bound Algorithm

In this subsection, we show a branch and bound algorithm for the maximum influence club problem. A configuration  $(S, \bar{S}, U)$  is a partition of  $V$  into three subsets, in which  $S$ , and  $\bar{S}$  resp., contain the nodes in, and not in resp., the club, and  $U$  is the set of the undetermined nodes. The algorithm uses a stack to contain all configurations to be explored. So it is a depth-first-search tree-searching algorithm. The algorithm is as follows.

#### Algorithm. ICLUB

Input: A graph  $G = (V, E)$  and a threshold  $\theta$ .

Output: An influence club of maximum cardinality.

- 1: initialize a stack  $T$ ;
- 2:  $BestClub \leftarrow \emptyset$ ;
- 3: push  $(\emptyset, \emptyset, V)$  into  $T$ ;
- 4: while  $T$  is not empty do
  - 5: pop  $(S, \bar{S}, U)$  from stack  $T$ ;
  - 6: let  $H = G[S \cup U]$ ;
  - 7: compute  $\Phi_H(u, v)$  for each  $u, v \in V(H)$ ;
  - 8: if there exist  $u, v \in S$  such that  $\Phi_H(u, v) < \theta$  then
    - 9: goto step 4;
  - 10: if  $|S| + |\{u \in U \mid \min_{s \in S} \Phi_H(u, s) \geq \theta\}| \leq |BestClub|$  then
    - 11: goto step 4;
  - 12: compute  $f(v) = |\{u \in V(H) \mid \Phi_H(u, v) < \theta\}|$  for each  $v \in V(H)$ ;

```

13: if  $f(v) = 0$  for each  $v \in V(H)$  then /* $S \cup U$  is a better club*/
14:   replace BestClub by  $S \cup U$ ;
15:   goto step 4;
16:  $R_1 = \{v \in U \mid \exists s \in S, \Phi_H(s, v) < \theta\}$ ; /* nodes conflicting with  $S^*$ /
17: if  $R_1 \neq \emptyset$  then
18:   push  $(S, \bar{S} \cup R_1, U - R_1)$  into  $T$ ;
19:   goto step 4;
20:   /*  $\exists u, v \in U, \Phi_H(u, v) < \theta$  */
21:    $x \leftarrow \arg \max_{v \in U} f(v)$ ;
22:    $R_2 = \{v \in U \mid \Phi_H(x, v) < \theta\}$ ; /* nodes conflicting with  $x^*$ /
23:   push  $(S, \bar{S} \cup \{x\}, U - \{x\})$  into  $T$ ;
24:   push  $(S \cup \{x\}, \bar{S} \cup R_2, U - \{x\} - R_2)$  into  $T$ ;
25: endwhile;
25: output BestClub.

```

In each iteration, we pop a configuration  $(S, \bar{S}, U)$  from the stack and do the following. If there exist  $u, v \in S$  such that  $\Phi_H(u, v) < \theta$ ,  $S$  cannot be contained in any feasible subset of  $H$ , and we simply discard the configuration (Steps 8–9). At Step 10 we compute an upper bound, if this bound is not better than the currently best, we also discard the configuration (Step 11). If the condition in Step 13 holds,  $V(H)$  is a feasible club. Since it is not discarded at Step 11, it must be a better club. At Step 16, we compute  $R_1$  as the set of nodes which cannot be in a feasible club with  $S$ . If there exists such a node, we remove them to generate a configuration which will be explored later. For otherwise, since there exists conflicting pair but no node conflicts with node in  $S$ , there must be conflicting pair of nodes in  $U$ , seeing Lemma 2. We then generate two configurations: one with  $x$  in  $S$  and the other with  $x$  in  $\bar{S}$ . If  $x$  is put into  $S$ , all nodes conflicting with  $x$  should be put into  $\bar{S}$  (Step 23). The way we choose  $x$  is to maximize the conflicting nodes ( $R_2$ ) so that the undetermined nodes ( $|U|$ ) of the generated configuration is minimized.

**Lemma 2.** *The  $R_2$  found at Step 21 is not empty.*

**Theorem 1.** *The algorithm ICLUB finds the maximum influence club of a undirected graph in  $O(n^3(1.618)^n)$  time and  $O(n^2)$  space, in which  $n$  is the number of nodes.*

## 4 Experimental Results

### 4.1 Test Data

**Random Graphs.** Two kinds of random graphs are used in our experiments: Poisson networks [6,17] and Price's model for network growth (Price networks) [14,16]. For a Poisson random graph, the existence of each edge has the same probability  $d$  and independent to any other. The expected density is  $d$ . However, it is believed that typical social networks has two features that doesn't happen

in Poisson networks: power-law distribution and fat-tail [14]. So we include the other kind of random networks. Price network has has the two features and it is generated by simulating the growth of a citation network in a “riches-get-richer” manner.

**Real Data.** PTT is the most popular BBS in Taiwan. The author-board relation is a two-mode social network, i.e., a bipartition graph  $G = (X \cup Y, E)$ , in which  $X$  and  $Y$  are the sets of nodes representing boards and authors respectively, and  $(x, y) \in E$  if author  $y$  has post a paper on board  $x$ . By the one-mode projection, we obtain a weighted undirected graph  $G_X = (X, E_X, w)$ , in which, for each edge  $(x_1, x_2) \in E_X$ ,  $w(x_1, x_2)$  is the number of common authors of boards  $x_1$  and  $x_2$ . Then by a threshold  $t$ , we transform  $G_X$  into an unweighted graph  $G^{(t)} = (X, E')$  such that  $(x_1, x_2) \in E'$  iff  $w(x_1, x_2) \geq t$ , i.e., boards  $x_1$  and  $x_2$  have at least  $t$  common authors. In our experiments,  $|X| = 50$ .

## 4.2 Experimental Results

**Random Graphs.** The random graphs generated for experiments are categorized by three parameters: model, number of nodes ( $n$ ), and density ( $d$ ), in which the density of a graph with  $n$  nodes and  $e$  edges is defined by  $e/(n^2)$ . The model is either Poisson or Price. For  $n = 50$  and  $100$ , several densities are tested, but we only show some typical densities in the table. The statics are computed over  $100$  graphs for each type.

For each generated random graphs, we find the maximum size 1-club, 2-club, and influence club of different thresholds  $\theta$ . We compute the average size, average density, and average cluster coefficient for the found clubs. The cluster coefficient of a club  $H$  is defined by  $cc(H) = \Pr\{(a, c) \in E | (a, b) \in E \wedge (b, c) \in E\}$ . Tables 1 and 2 show the experimental results. For the influence clubs, it happens for some types of inputs that the found club has size 1 or very large. To avoid these extreme cases, we only consider “effective” club, defined by the size is at least 3 and at most 90% of the whole graphs. The column labeled by  $n^*$  is the number of graphs in which the found club is effective.

By definition the influence club has the advantage on the pairwise influence. The influence scores are not shown in the tables. For examples, when  $n = 50$ ,  $d = 0.1$  and  $\theta = 1$ , the average sizes of maximum 2-club, 3-club and influence club are 11.2, 44.8, and 32.0, respectively, and the influences are 45.8, 83.3 and 97.2, respectively. The size of influence club is between the ones of 2-club and 3-club, and the influence score is larger than theirs. When we use density or CC as the index of cohesion, it is natural that the density decreases as the size increases. In some of the cases, our method can find influence clubs between 1-clubs and 2-clubs. In some cases, such as  $n = 50$  and  $d \geq 0.2$ , it is hard to find influence clubs between 1-clubs and 2-clubs. The reason is that, for relative high density, the influence score increases rapidly as the size increases because the number of paths increases exponentially. In the meantime, for a high threshold  $\theta$ , no club of small size can exist because there are not sufficiently many paths.

**Table 1.** Results for random graphs of  $n = 50$ 

Poisson							Price							
	Size	Dens	CC	$n^*$		Size	Dens	CC	$n^*$		Size	Dens	CC	$n^*$
d = 0.1	$\theta = 1$	32.0	0.16	0.12	97	$\theta = 1$	33.6	0.16	0.18	100	$\theta = 1$	33.6	0.16	0.18
	$\theta = 2$	15.3	0.37	0.28	87	$\theta = 3$	20.5	0.30	0.29	87	$\theta = 5$	11.7	0.60	0.55
	$\theta = 3$	8.5	0.58	0.48	6	$\theta = 5$	11.7	0.60	0.55	20	1-club	4.0	1.00	1.00
	1-club	3.2	1.00	1.00	100	1-club	4.0	1.00	1.00	100	2-club	17.8	0.27	0.34
	2-club	11.2	0.26	0.20	100	2-club	17.8	0.27	0.34	100				
d = 0.2	$\theta = 6$	41.5	0.20	0.20	2	$\theta = 10$	43.5	0.23	0.27	26	$\theta = 20$	33.8	0.33	0.37
	$\theta = 9$	40.0	0.23	0.22	22	$\theta = 30$	26.4	0.41	0.44	9	$\theta = 30$	26.4	0.41	0.44
	$\theta = 23$	39.0	0.28	0.27	6	1-club	4.2	1.00	1.00	100	1-club	5.4	1.00	1.00
	1-club	4.2	1.00	1.00	100	1-club	6.2	1.00	1.00	100	2-club	30.4	0.36	0.39
	2-club	23.7	0.32	0.26	100	2-club	40.8	0.33	0.38	100				
d = 0.3	$\theta = 55$	43.5	0.33	0.32	6	$\theta = 30$	43.7	0.30	0.35	12	$\theta = 70$	41.5	0.36	0.35
	$\theta = 70$	41.5	0.36	0.35	11	$\theta = 65$	34.8	0.40	0.43	45	$\theta = 80$	39.0	0.37	0.35
	$\theta = 80$	39.0	0.37	0.35	4	$\theta = 70$	25.6	0.52	0.54	16	1-club	5.2	1.00	1.00
	1-club	5.2	1.00	1.00	100	1-club	6.2	1.00	1.00	100	2-club	44.0	0.32	0.31
	2-club	44.0	0.32	0.31	100	2-club	40.8	0.33	0.38	100				

**Table 2.** Results for random graphs of  $n = 100$ 

Poisson							Price							
	Size	Dens	CC	$n^*$		Size	Dens	CC	$n^*$		Size	Dens	CC	$n^*$
d = 0.05	$\theta = 0.55$	41.1	0.12	0.08	100	$\theta = 1.05$	66.1	0.10	0.13	100	$\theta = 3.05$	27.9	0.24	0.24
	$\theta = 1.05$	18.2	0.25	0.15	100	$\theta = 4.05$	23.6	0.30	0.31	75	$\theta = 5.05$	19.6	0.34	0.35
	$\theta = 1.55$	9.6	0.44	0.29	77	$\theta = 7.6$	0.54	0.40	19	1-club	5.4	1.00	1.00	100
	$\theta = 2.05$	7.6	0.54	0.40	19	1-club	5.4	1.00	1.00	100	2-club	30.4	0.14	0.17
	1-club	3.5	1.00	1.00	100	2-club	30.4	0.14	0.17	100				
d = 0.08	$\theta = 2.5$	68.9	0.13	0.10	96	$\theta = 5$	67.1	0.13	0.17	100	$\theta = 9$	35.9	0.25	0.28
	$\theta = 3.5$	44.1	0.24	0.20	40	$\theta = 15$	20.3	0.43	0.48	5	1-club	4.3	1.00	1.00
	$\theta = 4.8$	16.5	0.51	0.50	6	1-club	4.3	1.00	1.00	100	2-club	30.3	0.19	0.27
	1-club	3.6	1.00	1.00	100	2-club	30.3	0.19	0.27	100				
d = 0.1	$\theta = 5$	85.5	0.12	0.11	84	$\theta = 12$	71.5	0.15	0.19	95	$\theta = 17$	45.9	0.25	0.28
	$\theta = 7$	77.4	0.13	0.12	54	$\theta = 22$	34.3	0.32	0.34	27	1-club	5.0	1.00	1.00
	$\theta = 9$	70.5	0.15	0.14	4	1-club	5.0	1.00	1.00	100	2-club	34.5	0.20	0.29
	1-club	4.0	1.00	1.00	100	2-club	34.5	0.20	0.29	100				

As a result, for such situations, the influence clubs are either relative large or contains only one node. However, the influence club is still useful. It in fact finds clubs between 2-clubs and 3-clubs. That is, it still has the flexibility between  $k$ -club and  $(k + 1)$ -club but not necessarily for  $k = 1$ .

**Real Data.** For a threshold  $t$ ,  $G^{(t)}$  is a graph such that  $(x_1, x_2)$  is an edge iff boards  $x_1$  and  $x_2$  have at least  $t$  common authors. We construct  $G^{(t)}$  for different  $t$  and find the maximum clique in  $G^{(t)}$ . Apparently, if  $t_1 < t_2$ , the edge set of  $G^{(t_2)}$  is a subset of the one of  $G^{(t_1)}$ , and a clique in  $G^{(t_2)}$  remains a clique in  $G^{(t_1)}$ . In Table 3, the rows with  $t=3200, 2500, 2200, 2000$ , and  $1500$  are the maximum clique in the graphs  $G^{(t)}$  for corresponding  $t$ . There are three different rows. Two of them are marked by  $\theta = 18$  and  $\theta = 17$  which represents the maximum influence club in the graph  $G^{(3200)}$  with the corresponding  $\theta$ -values. The other one marked by “2-club” is the maximum 2-club in  $G^{(3200)}$ . For each found club, we find their common authors in the original bipartite graph and record the number in the last column. We explain the meaning of the results as follows.

**Table 3.** PTT data

$t$	Club Size	Club members	Common authors
3200	7	1 3 10 11 14 17 18	182
2500	8	1 3 11 14 16 17 18 32	79
$\theta = 18$	8	1 3 10 11 14 16 17 18	67
2200	9	1 3 6 11 14 16 17 18 32	36
2000	10	1 3 6 9 11 14 16 17 18 32	21
$\theta = 17$	10	1 3 6 10 11 14 16 17 18 32	18
1500	11	1 3 6 9 10 11 14 16 17 18 32	11
2-club	14	1 3 6 9 10 11 14 16 17 18 32 36 38 41	3

Suppose we have only the graph  $G^{(3200)}$  and want to find cohesion groups. By the traditional methods, we can only find the clique or the 2-club, or  $k$ -club for  $k > 2$ . In our case, the clique has 7 nodes and the 2-club has 14 nodes. The influence club give us other alternatives such as 8 nodes for  $\theta = 18$  and 10 nodes for  $\theta = 17$ . Because it is an experiment, we may have  $G^{(t)}$  for different  $t$ , and we can verify the results. The number of common authors of each group is also an evidence of the correctness. The result shows that the influence club method can really help us to find the group in the middle of  $k$ -club and  $(k + 1)$ -club.

### 4.3 Running Time

For  $n = 50$ , the average running time is  $< 10^{-3}$  seconds. For larger  $n$ , the running times are diverse for different input densities and output club sizes. When the club size is very large or very small, the running time is small. For  $n = 100$ , it is about 0.8 seconds in average. And for  $n = 150$  and  $n = 200$ , the average times are approximately 15 and 120 seconds, respectively. In our test we find that the time to compute influence club is much less than that of computing 2-clubs. Especially, as shown in [4], the case of  $d = 0.15$  and  $n = 100, 150, 200$  for 2-club is very time consuming. It takes hundreds of seconds, or even more than ten thousands of seconds, for one graph. But such hard cases do not happen in our tests for influence clubs. The reason may be that the influence score increases rapidly as the club size increases. In summary the algorithm is much more efficient than the brute force method.

## 5 Concluding Remarks

In this paper, we propose a new model, the influence club, for cohesion groups in a social network. For nodes within an influence club, their influence scores are large enough. We give a branch-and-bound algorithm for finding the influence club of maximum size. The experimental results show that the algorithm is efficient and the model is effective.

For large graphs, it is interesting to develop heuristic algorithm which runs fast but does not ensure the optimal solution. Interesting future works also include club problem under different network flow models defined in [2].

## Acknowledgments

This work was supported in part by NSC 97-2221-E-194-064-MY3 and NSC 98-2221-E-194-027-MY3 from the National Science Council, Taiwan.

## References

1. Alba, R.D.: A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology* 3, 113–126 (1973)
2. Borgatti, S.P.: Centrality and network flow. *Social Networks* 27, 55–71 (2005)
3. Bourjolly, J.M., Laporte, G., Pesant, G.: Heuristics for finding k-clubs in an undirected graph. *Computers Operations Research* 27, 559–569 (2000)
4. Bourjolly, J.M., Laporte, G., Pesant, G.: An exact algorithm for the maximum k-club problem in an undirected graph. *European Journal of Operational Research* 138, 21–28 (2002)
5. Brandes, U., Fleischer, D.: Centrality measures based on current flow. In: Diekert, V., Durand, B. (eds.) STACS 2005. LNCS, vol. 3404, pp. 533–544. Springer, Heidelberg (2005)
6. Erdős, P., Rényi, A.: On random graphs. *Publicationes Mathematicae* 6, 290–297 (1959)
7. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks* 13(2), 141–154 (1991)
8. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to The Theory of NP-Completeness. Freeman, New York (1979)
9. Hanneman, R.A., Riddle, M.: Introduction to social network methods (2005), <http://www.faculty.ucr.edu/hanneman/nettext/>
10. Hubbell, C.H.: An input-output approach to clique detection. *Sociometry* 28, 277–299 (1965)
11. Johnson, D.S., Trick, M.A. (eds.): Cliques, coloring, and Satisfiability: Second DIMACS Implementation Challenge. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society, Providence (1996)
12. Katz, L.: A new status index derived from sociometric analysis. *Psychometrika* 18, 39–43 (1953)
13. Mokken, R.J.: Cliques, clubs, and clans. *Quality and Quantity* 13, 161–173 (1979)
14. Newman, M.E.J.: The structure and function of complex networks. *SIAM Review* 45, 167–256 (2003)
15. Newman, M.E.J.: A measure of betweenness centrality based on random walks. *Social Networks* 27, 39–54 (2005)
16. de Solla Price, D.J.: Networks of scientific papers. *Science* 149, 510–515 (1965)
17. Solomonoff, R., Rapoport, A.: Connectivity of random nets. *Bulletin of Mathematical Biophysics* 13, 107–117 (1951)
18. UCINet, Release 6.0, Mathematical Social Science Group, Social School of Science, University of California at Irvine
19. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)

# A Method of Label-Dependent Feature Extraction in Social Networks

Tomasz Kajdanowicz, Przemysław Kazienko, and Piotr Doskocz

Wrocław University of Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland  
[{tomasz.kajdanowicz,kazienko,piotr.doskocz}@pwr.wroc.pl](mailto:{tomasz.kajdanowicz,kazienko,piotr.doskocz}@pwr.wroc.pl)

**Abstract.** The paper is referring to within-network classification problem. A new method for feature extraction is proposed in the paper. New features are obtained by combination of network structure information and class labels assigned to nodes. The influence of various features on classification performance has also been studied in the paper. The experiments based on real-world data have shown that the proposed method provides with features leading to significant improvement of classification accuracy.

**Keywords:** Label-dependent features, within-network classification, feature extraction, social network analysis.

## 1 Introduction

Classification task is one of most important concepts in Machine Learning. It is usually based on the data that represents relationships between a fixed set of attributes and one target class. These relations describe each object independently that means no direct correlations between objects in the classification phase are taken into account. An exception may be additional input features, which aggregate information about the entire group a given object belongs to. However, it requires any clustering process be launched before. There are some applications and research methods, especially related to social networks, which are able to produce data with dependencies between labels of interconnected objects, referred as relational autocorrelation [14]. Based on these connections additional input information should be added to the classification process. If the considered objects are humans and the classification is utilized on their profiles then the social network can be extracted from complementary data (different from people's profiles) about common activities and mutual communication [9, 10, 15]. Overall, a social network is a set of nodes (human entities, objects) and node-node relationship between pairs of nodes [18]. According to [17], all network objects may be described by three distinct types of information that can be easily used in label classification: correlation between the object's label (class) and its attributes, correlation between the object's label and the observed labels of other objects in its neighborhood and, consequently, correlation between the object's label and unobserved labels of other objects in its neighborhood.

Basic task of within-network classification [1, 12] is to assign the correct labels to the unlabeled nodes from a set of the possible class labels. For example, based on the

network of communication interactions, it could be determined whether a given company's employee is either an executive or a performer. Willing to obtain the best possible results of classification, all three types of information should be evaluated: nodes attributes (profiles), node-node network relations to the known labels in the neighborhood (labeled neighbors) and relations to the neighboring objects with unknown labels. Main difficulty here is to extract the set of most discriminative features from the network nodes and their connections to achieve the best classification model.

A new approach for network feature extraction is proposed in further sections. Some of these structural features have discriminative distribution, which may directly influence classification performance.

Section 2 covers related work while in Section 3 a new method for network feature extraction is presented. Sections 4 and 5, contain descriptions of the experimental setup and the obtained results, respectively. The paper is concluded in Section 6.

## 2 Related Work

In general, network classification problems, may be solved using two main approaches: by within-network and across-network inference. Within-network classification, for which training entities are connected directly to entities, whose labels are to be classified, stays in contrast to across-network classification, where models learnt from one network are applied to another similar network [11]. Overall, the networked data have several unique characteristics that simultaneously complicate and provide leverage to learning and classification. More generally, network data allow the use of the features of the node's neighbors to label them, although it must be performed with care to avoid increase of variance estimation [7].

There have been developed many algorithms and models for classification in the network. Among others, statistical relational learning (SRL) techniques were introduced, including probabilistic relational models, relational Markov networks, and probabilistic entity-relationship models [2, 6, 13, 16]. Two distinct types of classification in networks may be distinguished: based on collection of local conditional classifiers and based on the classification stated as one global objective function. The most known implementations of the first approach are iterative classification (ICA) and Gibbs sampling algorithm (GS), whereas example of the latter are loopy belief propagation (LBP) and mean-field relaxation labeling (MF) [17]. Generally speaking, there exist many pretty effective algorithms of collective classification as well as graph-based semi-supervised learning methods. It refers, especially logForest, a logistic model based on links, wvRN, a relational neighbor model, SSL Gaussian random field model, ghostEdge, combination of statistical relational learning and semi-supervised learning for sparse networks and theirs collective classification supplements [5].

One of the most crucial problems in the network classification is feature extraction. According to [4] the derived features are divided into two categories: label-dependent (LD) and label-independent (LI). Features LD use both structure of the network as well as information about labels of the neighboring nodes labels, e.g. number of

neighbors with given class label. Features LI, in turn, are calculated using the network structure only, e.g. betweenness of a node. The LI like features, therefore, are independent from the distribution of labels in the network and might not be informative. However, they can be perfectly calculated regardless of the availability of labels. What is worth mentioning, most of the proposed network classification methods were usually applied to the data sets with very limited access to labels. Their authors assumed that their applications need to deal even with only 1% labeled nodes. This problem is known as classification in sparsely labeled networks [4, 5].

It appears that the majority of network-based structural measures used as features in network classification may be useful and may potentially improve classification performance.

Social networks, being a network representation of interactions between people is a subject of research in terms of classification in networks as well [4].

### 3 Extraction of the Features from the Social Network

#### 3.1 Problem Statement

Let us suppose that a social network is a graph  $G = (V, E, X, L, Y, A)$ , where  $V$  is a set of nodes (objects, social entities);  $E$  is a set of edges (connections)  $e_{ij}$  between two nodes  $v_i$  and  $v_j$ ,  $E=\{e_{ij}: v_i, v_j \in V, i \neq j\}$ ;  $X$  is a set of attribute vectors  $x_i$ , a separate one for each node  $v_i$  (a profile of  $v_i$ ),  $X=\{x_i: v_i \in V \Leftrightarrow x_i \in X\}$ ;  $L$  is the set of distinct labels (classes) possible to be assigned to nodes;  $Y$  is a list of actual labels assignments to nodes,  $Y=\{<v_i, y_i>: v_i \in V \wedge y_i \in L\}$ ;  $A$  is a set of edge weights,  $\forall a_{ij} \in A \ a_{ij} \geq 0$  and  $a_{ij}$  indicates the strength of edge  $e_{ij}$ .

Having known the values of  $y_i$  for a given subset of nodes  $V^K \subset V$ , classification may be described as the process of inferring the values of  $y_i$  for the remaining set of nodes  $V^U$ ,  $V^U = V \setminus V^K$ .

The first step in the process of node classification is a translation of network data into a set of unified vectors, one for each node. A single vector corresponding to node  $v_i$  contains all information from  $x_i$  as well as some additional information (new attributes) derived by feature extraction methods based on the network profile. Next, the obtained set of vectors is used in classical, supervised classification.

#### 3.2 Features Extraction

Feature extraction from social networks is a general term for methods of constructing variables from the connectivity graph, expressing the position and importance of each node with respect to the others. As mentioned in Section 1, the generated features may be label-independent or label-dependent. For commodity and clarity, while describing label-dependent features, it is made a basic assumption in the paper that feature extraction is based only on correlation between the object's label and the observed labels of other objects in its neighborhood.

Three basic label-independent and three label-dependent features are presented in the following sub-sections, as well as generalization for label-dependent features extraction.

### 3.2.1 Label-Independent Features

#### *Betweenness Centrality*

Betweenness centrality of node  $v_i$  pinpoints to what extent  $v_i$  is between other nodes. Nodes with high betweenness are very important in the network as other nodes are connected with each other mainly through them. Betweenness centrality  $B(G, v_i)$  of node  $v_i$  in graph  $G$  can be calculated according to the following equation:

$$B(G, v_i) = \sum_{v_j, v_k, v_i \in G(V); j \neq k \neq i} \frac{P(G, v_j, v_k, v_i)}{P(G, v_j, v_k)}, \quad (1)$$

where:

$P(G, v_i, v_j)$  - a function returning the number of shortest paths between  $v_i$  and  $v_j$  in graph  $G$ ;

$P(G, v_j, v_k, v_i)$  - a function that returns the number of shortest paths between  $v_k$  and  $v_j$  that pass through  $v_i$  in graph  $G$ .

Obviously, Equation 1 is calculated only for pairs  $v_j, v_k$ , for which there exists a path from  $v_j$  to  $v_k$  to prevent the denominator from equaling 0.

#### *Degree Centrality*

Degree centrality is defined as the number of connections (edges) incident upon a given node. It is the simplest and most intuitive measures that can be used in the network analysis. Nodes with the high degree centrality are recognized as a crucial cog that occupies a central location in the network. Degree centrality  $D(G, v_i)$  of node  $v_i$  in graph  $G$  can be computed using Equation 2:

$$D(G, v_i) = \frac{\text{card}(n(G, v_i))}{\text{card}(V)-1}, \quad (2)$$

where:

$n(G, v_i)$  - a set of neighboring nodes of node  $v_i$  in graph  $G$ .

#### *Local Clustering Coefficient*

The local clustering coefficient  $CC(G, v_i)$  of a node  $v_i$  in graph  $G$  quantifies how close  $v_i$ 's neighborhood is to a complete graph, see Equation 3.

$$CC(G, v_i) = \frac{\text{card}(R(n(G, v_i)))}{\text{card}(n(G, v_i))(\text{card}(n(G, v_i))-1)}, \quad (3)$$

where:

$R(V)$  - an operator returning the number of all connections between nodes from set  $V$ .

### 3.2.2 Label-Dependent Features

While introducing label-dependent features there are proposed two custom manners of their composition. Both of them relays on the idea of selective definition of sub-networks based on the labels assigned to each node. It means that a sub-network for a given label  $l$  consists of only those nodes that share label (class)  $l$  together with all edges connecting these selected nodes. For that purpose, a new selection operator  $O(G,l)$  for graph  $G$  and label  $l$  is defined. It returns a sub-network  $G_l$  labeled with  $l$ :  
 $G_l = (V_l, E_l, X_l, \{l\}, Y_l, A_l)$  such that  $V_l = \{v_i : \langle v_i, l \rangle \in Y_l\}$ ,  $Y_l = \{\langle v_i, y_i \rangle : v_i \in V_l \wedge y_i = l\}$ ,  
 $E_l = \{e_{ij} : v_i, v_j \in V_l \wedge e_{ij} \in E\}$ ,  $X_l = \{x_l : v_l \in V_l \Leftrightarrow x_l \in X\}$ .

Afterwards, for each sub-network  $G_l$  (each label  $l$ ), new features are computed.

First group of label-dependent features composition is based on new custom measures derived from the interaction between a given node and its neighboring nodes only. The measures respect either the number of connections or their strengths.

#### *Normalized Number of Connections to the Labeled neighbors*

The measure for the normalized number of connections to the labeled neighbors  $NCN(G,l,v_i)$  represents the proportion of the number of connections to the neighboring nodes in the sub-network with label  $l$  ( $G_l$ ) by the number of connections to the labeled neighbors in the whole primary graph  $G$  (with all labels).

The measure  $NCN(G,l,v_i)$  is defined as follows:

$$NCN(G,l,v_i) = \frac{card(n(O(G,l),v_i))}{card(n_L(G,v_i))}, \quad (4)$$

where:

$n(O(G,l),v_i)$  - a set of the neighboring nodes for node  $v_i$  in sub-network  $O(G,l)$ ,

$n_L(G,v_i)$  - a set of  $v_i$ 's labeled neighbors in graph  $G$ , each neighbor must be labeled with any label  $l \in L$ .

Note that the value of  $card(n(O(G,l),v_i))$  is the same as the number of connections between  $v_i$  and  $v_i$ 's neighbors (each  $v_i$ 's neighbor has one connection with  $v_i$ ). Similarly, the value of  $card(n_L(G,v_i))$  equals the number of connections between  $v_i$  and all  $v_i$ 's labeled (and only labeled) neighbors.

The measure  $NCN(G,l,v_i)$  is computed separately for each label  $l$  and in general, for two labels  $l_k$  and  $l_m$ , the value of  $NCN(G,l_k,v_i)$  may differ from  $NCN(G,l_m,v_i)$ .

For the example network from Fig. 1, and the measure  $NCN(G,'red',v_1)$  calculated for node 1 in the sub-network with nodes labeled with the 'red' class, the value of  $NCN(G,'red',v_1)$  is 4 divided by 8 (total number of nodes in graph  $G$ ).

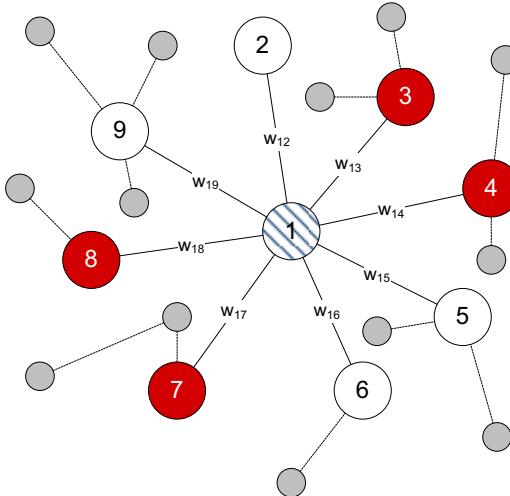
#### *Normalized Sum of Connection Strengths to the Labeled Neighbors*

The value of the normalized sum of connection strengths to the labeled neighbors  $NCS(G,l,v_i)$  is the proportion of node  $v_i$ 's activity towards  $v_i$ 's neighbors (measured by the aggregated connection strengths) in the sub-network with label  $l$  ( $G_l$ ) normalized by the equivalent value of strengths to the neighbors with any label in the whole graph  $G$ . The value of  $NCS(G,l,v_i)$  for graph  $G$  and label  $l$  is expressed in the following way:

$$NCS(G, l, v_i) = \frac{\sum_{v_j \in n(O(G, l), v_i)} a_{ij}}{\sum_{v_j \in n_L(G, v_i)} a_{ij}}, \quad (5)$$

Similarly to  $NCN(G, l, v_i)$ , the measure  $NCS(G, l, v_i)$  is evaluated separately for each label  $l$  and differs for different labels  $l$ .

For the network from Fig. 1, the measure  $NCS(G, 'red', v_1)$  is computed for node 1 and label (class) 'red', as the sum of  $w_{13}$ ,  $w_{14}$ ,  $w_{17}$ , and  $w_{18}$  normalized by sum of all eight connection strengths.



**Fig. 1.** Feature calculation based on label dependent neighborhood. For each of label class {white, red}  $w$  is calculated.

### 3.2.3 General Method for Label-Dependent Features Extraction

In the domain of social network analysis (SNA), a number of measures characterizing network nodes have been introduced in the literature. Majority of them is label-independent and it is possible to define many methods that will extract label-dependent features based on them. A general concept of creation of any label-dependent feature  $M_l(G, l, v_i)$  for label  $l$  and node  $v_i$  in the social network  $G$  applies label-independent feature  $M$  to the appropriate labeled sub-network  $G_l=O(G, l)$ , as follows:

$$M_l(G, l, v_i) = M(G_l, v_i), \quad (6)$$

where:

$M_l(G_l, v_i)$  - denotes any structural network measure for node  $v_i$  applied to sub-network  $G_l=O(G, l)$ , e.g degree, betweenness or clustering coefficient;  
Obviously,  $M_l(G, l, v_i)$  is computed separately for each label  $l$  using the appropriate sub-network  $G_l=O(G, l)$ .

As an example, the label-dependent clustering coefficient ( $CC_l$ ) is defined in accordance with Equation 3 as:

$$CC_l(v_i) = \frac{card(R(n(G_l, v_i)))}{card(n(G_l, v_i))(card(n(G_l, v_i))-1)} \quad (7)$$

## 4 Experimental Setup

### 4.1 Data Set

The data set used for experiments, "Attendee Meta-Data" (AMD), was downloaded from UCI Network Data Repository (<http://networkdata.ics.uci.edu/data.php?d=amdhope>). The AMD data set was an output of a project, which used RFID (Radio Frequency Identification) technology to help connect conference participants at "The Last HOPE" Conference held in July 18-20, 2008, New York City, USA. All attendees received an RFID badges that uniquely identified and tracked them across the conference space. The data set contains descriptions of interests of participants, their interactions via instant messages, as well as their location over the course of the conference. Conference attendees were asked to "tag" themselves based on a diverse set of interests. Thanks to location tracking, a list of attendances was extracted for each conference talk. Additionally, participants could email or send a text message to "ping" the people who had similar interests.

In general, the data set contains information about conference participants, conference talks and presence on talks. Initial import contained 767 different persons, 99 talks, 10,110 presences reported during talks. In the cleaning process, these contributors who did not give any information about their interests were excluded from further studies. As a result, 334 persons with 99 lectures and 3,141 presences have left after cleaning.

Afterwards, the social network was build. Ties in the network were constructed based on the fact that participants were present on the same talks. Moreover, strengths of the connections between each pair of contributors were calculated as the proportion of number of talks attended by both participants by the total number of talk presences of the first participant. It provided 68,770 directed, weighted connections. The raw data contained 4 attributes: 3 nominal (sex, cell phone provider, country) and 1 numerical (age). Additionally, each participant was described by unordered set of interests that in our experiments was chosen as the classification target. Since each network node (participant) could have multiple interests assigned, it was decided to construct 20 separate experimental data sets that formed a binary assignment of each interest. For the clarity of the experiment, the binary classification problem was established as it did not contrive a loss of generality of the proposed feature extraction approach.

## 4.2 Extracted Features

According to methodology presented in Section 3, 17 attributes were calculated in the experiments. These features were grouped in 4 sets. The first contained raw data attributes. In the second there were label-independent network based features. In the third group label-dependent features obtained from proposed method were introduced. The last, fourth group attach all previously introduced features. Finally, the obtained 20 data sets, used in the experiment, may be downloaded from <http://www.zsi.pwr.wroc.pl/~kazienko/datasets/amd/amd.zip> in the arff format.

## 4.3 Classification

Experiments were conducted for 20 data sets using 3 classification algorithms, AdaBoost, Multilayered Perceptron, SVM. Classification was performed in 10% - 90% proportion of labeled and unlabeled nodes, respectively, using 10-cross fold validation.

## 5 Results

The obtained results have revealed that the average accuracy of classification using various feature sets really differs. The average accuracy is greater by about 23% for feature set 3 and 4 compared to set 1 and 2. Simultaneously, F-Measure and precision improves by usage of label-dependent feature sets (set 3 and 4) by 33% and 35%, respectively, see Table 1.

Irrespectively of the used feature data set, all utilized classification algorithms: AdaBoost, Multilayered Perceptron, SVM, provide similar results.

Classification based on feature set 3 and 4 seems to be more stable than for feature set 1 and 2. In particular, standard deviation of accuracy for 20 data sets in first case equals 1% and in the second 12%.

Additionally, experiments have revealed that classification based on feature set 4 returns in average worse accuracy than classification based on feature set 3 (see Table 1). Let remind that feature set 4 contains all features from sets 1, 2 and 3. Worse classification performance might be an effect of too many relative poor input features, from which some weaken classification and have contrary discriminative distributions. It refers features from set 1 and 2 that degrade high correlation between output and label-dependent features from set 3. It means that the features extracted from the social network are so good that regular profiles of the tested cases only decrease classification performance and should not be even taken into account.

Owing to the carried out experiments, it is visible that the proposed label-dependent features used in classification undoubtedly provide the best results.

**Table 1.** Average results of experiments for 20 data sets

Algorithm	Feature Set				Measure
	1	2	3	4	
AdaBoost	0.76	0.76	0.99	0.98	Accuracy
	0.62	0.63	0.99	0.99	Precision
	0.67	0.68	0.99	0.98	F-measure
Multilayer Perceptron	0.74	0.76	0.99	0.98	Accuracy
	0.67	0.63	0.99	0.98	Precision
	0.69	0.68	0.99	0.98	F-measure
SVM	0.76	0.76	0.98	0.98	Accuracy
	0.64	0.61	0.98	0.98	Precision
	0.69	0.67	0.98	0.98	F-measure

## 6 Conclusions and Future Work

A new method for label-dependent feature extraction from the social network was proposed in the paper. The main principle behind the method is based the selective definitions of sub-graphs for which new features are defined and computed. These new features provide additional quantitative information about the network context of the case being classified.

According to collected experimental evidences, the proposed label-dependent feature extraction appears to be significantly more effective and improves classification performance in high extent. Obtained, so good, results were even surprising to authors. These results have shown that the new approach to classification extended with features derived from the social network may return very satisfactory and promising outcomes.

Feature work will focus on further experimentations on the method, especially in terms of its validity for variety of local network measures. Additionally, the proposed feature extraction method will also be examined against the usage of global objective functions for classification. Yet another direction of future studies will be development of new ensemble algorithms, which would have network measures already incorporated, especially based on boosting concept [8].

**Acknowledgments.** This work was supported by The Polish Ministry of Science and Higher Education, the development project, 2009-11.

## References

- [1] Desrosiers, C., Karypis, G.: Within-network classification using local structure similarity. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS, vol. 5781, pp. 260–275. Springer, Heidelberg (2009)
- [2] Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: Proceedings of the International Joint Conference on Artificial Intelligence IJCAI 1999, pp. 1300–1309 (1999)
- [3] Fruchterman, T., Reingold, E.: Graph Drawing by Force-directed Placement. Software – Practice and Experience 21, 1129–1164 (1991)
- [4] Gallagher, B., Eliassi-Rad, T.: Leveraging Label-Independent Features for Classification in Sparsely Labeled Networks: An Empirical Study. In: Proceedings of the Second ACM SIGKDD Workshop on Social Network Mining and Analysis (SNA-KDD 2008), Las Vegas, NV (2008)
- [5] Gallagher, B., Tong, H., Eliassi-Rad, T., Faloutsos, C.: Using ghost edges for classification in sparsely labeled networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 256–264 (2008)
- [6] Getoor, L., Friedman, N., Koller, D., Taskar, B.: Learning probabilistic models of link structure. Journal of Machine Learning Research 3, 679–707 (2002)
- [7] Jensen, D., Neville, J., Gallagher, B.: Why collective inference improves relational classification. In: The Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 593–598 (2004)
- [8] Kajdanowicz, T., Kazienko, P., Kraszewski, J.: Boosting Algorithm with Sequence-loss Cost Function for Structured Prediction. In: Graña Romay, M., Corchado, E., Garcia Sebastian, M.T. (eds.) HAIS 2010. LNCS (LNAI), vol. 6076, pp. 573–580. Springer, Heidelberg (2010)
- [9] Kazienko, P., Musiał, K., Kajdanowicz, T.: Profile of the Social Network in Photo Sharing Systems. In: 14th Americas Conference on Information Systems, AMCIS 2008, Minitrack: Social Network Analysis in IS Research, Association for Information Systems, AIS (2008) ISBN: 978-0-615-23693-3
- [10] Kazienko, P., Musiał, K., Kajdanowicz, T.: Multidimensional Social Network in the Social Recommender System. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans (accepted) (in press)
- [11] Lu, Q., Getoor, L.: Link-based classification. In: Proceedings of the 20th International Conference on Machine Learning ICML 2003, pp. 496–503 (2003)
- [12] Macskassy, S., Provost, F.: A brief survey of machine learning methods for classification in networked data and an application to suspicion scoring. In: Airoldi, E.M., Blei, D.M., Fienberg, S.E., Goldenberg, A., Xing, E.P., Zheng, A.X. (eds.) ICML 2006. LNCS, vol. 4503, pp. 172–175. Springer, Heidelberg (2007)
- [13] Macskassy, S., Provost, F.: Classification in networked data: A toolkit and a univariate case study. Journal of Machine Learning Research 8, 935–983 (2007)
- [14] McPherson, M., Smith-Lovin, L., Cook, J.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444 (2007)
- [15] Musiał, K., Bródka, P., Kazienko, P., Gaworecki, J.: Extraction of Multi-layered Social Networks from Activity Data. Journal of Global Information Management (to appear)
- [16] Neville, J., Jensen, D.: Relational dependency networks. Journal of Machine Learning Research 8, 653–692 (2010)

- [17] Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective classification in network data. *Artificial Intelligence Magazine* 29(3), 93–106 (2008)
- [18] Wasserman, S., Faust, K.: *Social network analysis: Methods and applications*. Cambridge University Press, New York (1994)

# Identifying Representative Reviewers in Internet Social Media\*

Sang-Min Choi<sup>1</sup>, Jeong-Won Cha<sup>2</sup>, and Yo-Sub Han<sup>1, \*\*</sup>

<sup>1</sup> Department of Computer Science, Yonsei University, Seoul, Republic of Korea  
*{jerassi, emmous}@cs.yonsei.ac.kr*

<sup>2</sup> Department of Computer Engineering, Changwon National University, Changwon,  
Republic of Korea  
*jcha@changwon.ac.kr*

**Abstract.** In a society, we have many forms of relations with other people from home, work or school. These relationships give rise to a social network. People in a social network receive, provide and pass lots of information. We often observe that there are a group of people who have high influence to other people. We call these high influence people opinion leaders. Thus, it is important and useful to identify opinion leaders in a social network. In Web 2.0, there are many user participations and we can create a social network from the user activities. We propose a simple yet reliable algorithm that finds opinion leaders in a cyber social network. We consider a social network of users who rate musics and identify representative users of the social network. Then, we verify the correctness of the proposed algorithm by the T-test.

**Keywords:** Social network, opinion leaders, representative reviewers, music.

## 1 Introduction

There are lots of information from mass media and the information affects us in various ways. However, one interesting thing is that most people often receive information not by mass media directly but by opinions of some other persons. We call these persons who influence other people *opinion leaders*. Note that opinion leaders affect people more than mass media [8]. The general public does not accept information by mass media uncritically whereas they accept information from opinion leaders easily. This implies that opinion leaders are representatives of a social network. Most people use the Internet and, thus, the Web becomes a place to share information for the general public. Since there are many users

\* Choi and Han were supported by the Basic Science Research Program through NRF funded by MEST and Cha was supported by the IT R&D program of MKE/IITA 2008-S-024-01.

\*\* Corresponding author.

and lots of user participations, the Web become a cyber society. The current Web shows that lots of information creates a community among peoples regardless of age, gender or nationality. Moreover, there are some influential people in the community. There are many researches to find opinion leaders in a Web social network [2,9]. We design an algorithm that finds opinion leaders in a social network of users for musics. We apply the proposed algorithm in Yahoo! music data [1] and verify the method using the T-test.

In Section 2, we revisit the previous approaches to identify opinion leaders in Twitter<sup>1</sup> and blogs. Then, we propose a new algorithm that finds opinion leaders in an Internet social network in Section 3. Then, we verify the algorithm using the T-test in Section 4. We conclude the paper with future directions of this research in Section 5.

## 2 Related Work

We briefly describe previous researches that find opinion leaders in Twitter and blogs.

### 2.1 Twitter

Twitter is one of the most popular social networking applications for Internet users [4]. Twitter composes a social network using a function called follow that allows users to add another users in the user lists. It also provides other functions such as `reply` and `post`. The research to find opinion leaders in Twitter is based on functions in Twitter such as `follower` or `retweet`.

**The Number of Followers:** Follower means the users following a particular user called followee. Followers can confirm uploaded posts by followee, and reply or retransmit (called `retweet`) it immediately. Namely, a follower is a user who can immediately response behavior of followee. The right side of Table 1 shows the followee who has the most followees in top 15.

**The Number of Retweets:** Tweet is a post of Twitter users. Followers can confirm tweet of followee and also deliver tweet to follower of their own. In this delivery case, followers can add their message.

Twitter users response their opinion for others tweets using different styles and ways [5]. In Fig. 1, users who have incoming edges are followees, and users who have outgoing edges are followers. For instance, 1 denotes a followee of users 2, 3, . . . , 8. When 1 uploads a message (`tweet`) to its own twitter, all users from 2 to 8 can read the message immediately. Furthermore, user 8 can retweet the message to users 9, 10 and 11 or response the message by adding opinions. This function is called `retweet`. Namely, a tweet that has many retweets can be considered as an influential opinion in the Twitter network and a followee

---

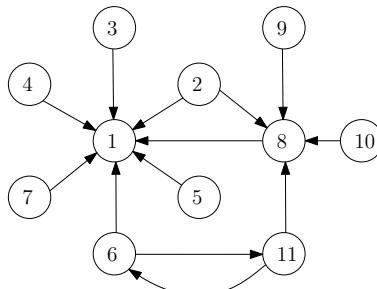
<sup>1</sup> <http://twitter.com>

**Table 1.** Number of Followers VS. Number of Retweets in Twitter [9]

Rank	most number of retweets		most number of followers	
	Name	Remark	Name	Remark
1	Pete Cashmore	News on social media	Ashton Kutcher	Actor
2	BNO News	News	Britney Spears	Musician
3	TweetMeme	News on Twitter	Ellen DeGeneres	Show host
4	Oxfordgirl	Journalist	CNN Breaking News	New
5	CNN Breaking News	News	Oprah Winfrey	Show host
6	Michael Arrington	News on technology	Twitter	Twitter
7	Fabolous	Musician	Barack Obama	President of U.S.
8	The New York Times	News	Ryan Seacrest	Show host
9	Lil duval	Comedian	THE REAL SHAQ	Sport star
10	Iran	about Iran	Kim Kardashian	Model
11	ESPN Sports News	News	John Mayer	Musician
12	Persiankiwi	about Iran	Demi Moore	Actress
13	Ashton Kutcher	Actor	Iamddiddy	Musician
14	Raymond Jahan	about Iran	Jimmy Fallon	Actor
15	Alyssa Milano	actress	Lance Armstrong	Sport star

who has many retweets can be considered as an opinion leader. The left side of Table 1 shows the top 15 users who have the most number of retweets.

The left side of the top 15 are usually celebrities in Table 1. However, we cannot infer whether followers of the celebrities react as faithful fans or having other rational reasons. Because of these reasons, we cannot conclude that the celebrities become the opinion leaders in Tweeter. Whereas, the mass media appear in the left side of Table 1. This difference show that followers react more sensitively in tweet from mass media than celebrities. In conclusion, we can consider the group of followers in the left side of Table 1 opinion leaders.

**Fig. 1.** An example of social network construction by the function `following` in Twitter [13]. The directed edge shows the `following` relation between users.

## 2.2 Blogs

Blog is an application that supplies various information and functions in the Web. One of the important functions of blog is to post various media contents and to tag the relevant information among each other. The blog influences the user in the Web. For example, the preferences of some influential bloggers affect the consume of visitors [7,10]. Thus, many advertisers can make big profits by advertising in the influential blog [12]. Hence, the influential bloggers can be opinion leaders. Some experiments examine essential issues of identifying influential bloggers, evaluate the effects of various collectible statistics from a blog site on determining blog-post influence, develop unique experiments using Digg<sup>2</sup> and conduct experiments by using the whole history of blog posts.

The research to find influential bloggers classifies the characteristics of bloggers into active, inactive, influential and non-influential bloggers based on intuition which active bloggers not equal to influential bloggers. The active bloggers mean users who often list their posts, and influential bloggers mean users who have influential posts. This research determines the influential posts using social gestures such as comments, incoming links, outgoing links and length of posts, and find influential bloggers based on characters of bloggers.

This research shows us the clear conclusion for the existence of influential bloggers, and how much it relates with each other between the influential bloggers and other common visitors [2].

## 3 Proposed Method

We identify the representative user of evaluative group of music using Yahoo! music data [1]. And we test the extracted users to know that users can be representative like a opinion leader. The average rating of music means collective opinion. This value can consider with the representative figure from collective opinion. We use the T-test to verify whether or not identified users are propose representative of the group.

In the case of adding new song, because rating of opinion leaders for new songs has representativeness, the opinion leaders can process these songs. For example, the searching or recommendation systems exclude the new songs because of a lack of information. In that case, ratings of opinion leaders have representative like a average rating. Rating of opinion leader become one of the way to solve cold start problem in information search and recommendation system [3,11].

### 3.1 Data Set

We use Yahoo! Webscope music data. The data is as follows:

1. User id: There exist 15,400 users in total, and given by integer number form 1 to 15,400.

---

<sup>2</sup> <http://www.digg.com>

2. Song id: There exist 1,000 songs in total, and given by integer number form 1 to 1,000.
3. Rating: As integer number from 1 to 5 there exist approximately 300,000 ratings.

In the dataset, each user gives rating to at least 10 songs. The rating presents the number from 1 to 5. The higher number means the higher rating.

### 3.2 Identifying Representative Reviewer

We use the following equation for identifying representative users from in the music rater group:

$$U_s = \frac{\sum_{i \in A} |(R_S(i) - R_\mu(i))|}{|A|}, \quad (1)$$

where

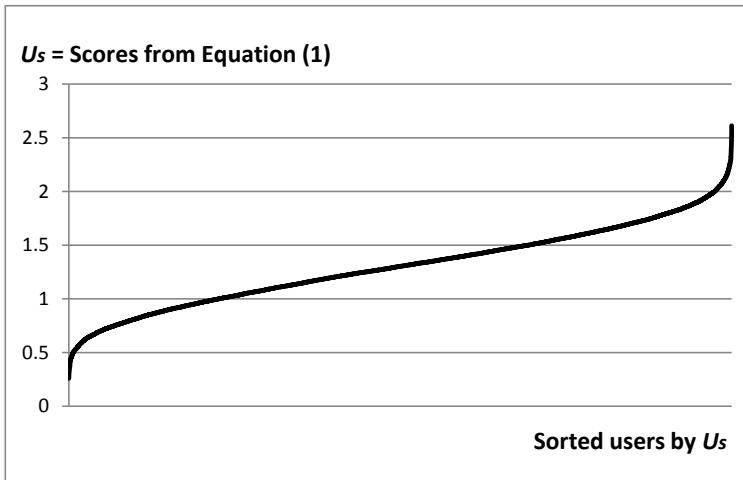
- $A$  is a set of songs evaluated by user  $S$ .
- $|A|$  is the cardinality of  $A$ .
- $R_S(i)$  is the rating of song  $i$  by user  $S$ .
- $R_\mu(i)$  is the average rating of song  $i$  in the database.

Remark that the result of Equation (1) shows how close each rating of user compared to the average rating. We select those who have low score from Equation (1) as opinion leaders.

**Table 2.** Top 10 scores and bottom 10 scores out of Equation (1)

Rank	score	Rank	score
1	0.2598	1	2.6126
2	0.2989	2	2.5453
3	0.2996	3	2.5056
4	0.3047	4	2.5015
5	0.3086	5	2.4922
6	0.3126	6	2.4856
7	0.3140	7	2.4660
8	0.3211	8	2.4087
9	0.3219	9	2.3977
10	0.3277	10	2.3888

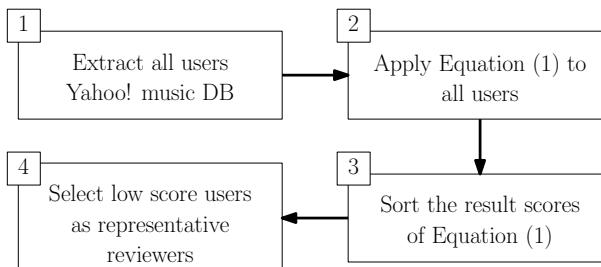
Table 2 shows the top 10 users and the bottom 10 users by Equation (1). For instance, the top ranked user has a gap of 0.2598 between its own rating and the average rating in Table 2 whereas the bottom ranked user has a gap of 2.6126.



**Fig. 2.** The distribution of scores ( $U_s$ ) from Equation (1) in ascending order

In Fig. 2, the y-axis is the value of Equation (1) and the x-axis is the set of users that are sorted by the score of Equation (1) in ascending order.

Fig. 3 illustrates the procedure of identifying representative reviewer in music rater group. First, we extract all users who evaluate songs at least 10 times. Second, applying Equation (1) to the user group. We sort the results of Equation (1) in ascending order, and select high rank users as representative reviewers in music rater group.

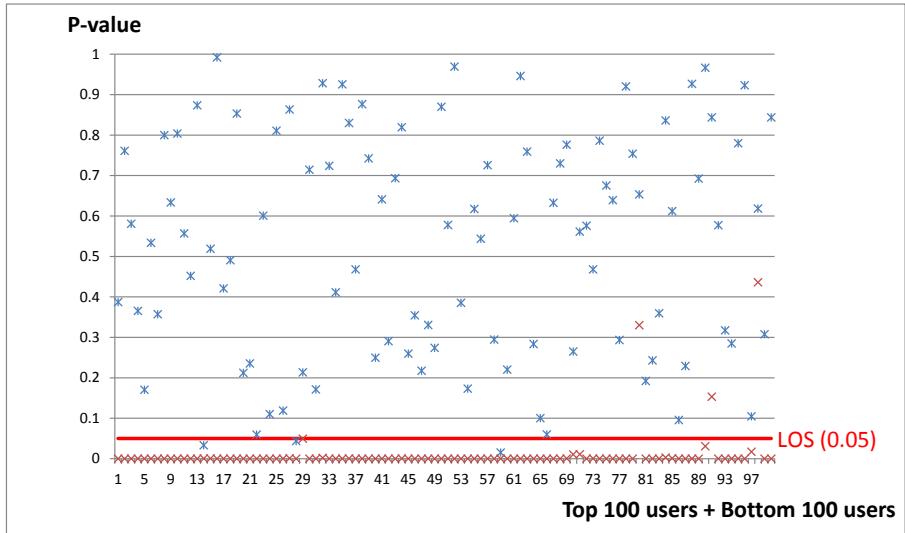


**Fig. 3.** The procedure of identifying representative users from Yahoo! music dataset

## 4 Experiments and Analysis

Now we demonstrate the validness of our algorithm for identifying representative users by the T-test. The T-test is a statistical hypothesis test in which the test statistic follows a certain distribution if the null hypothesis is supported. For

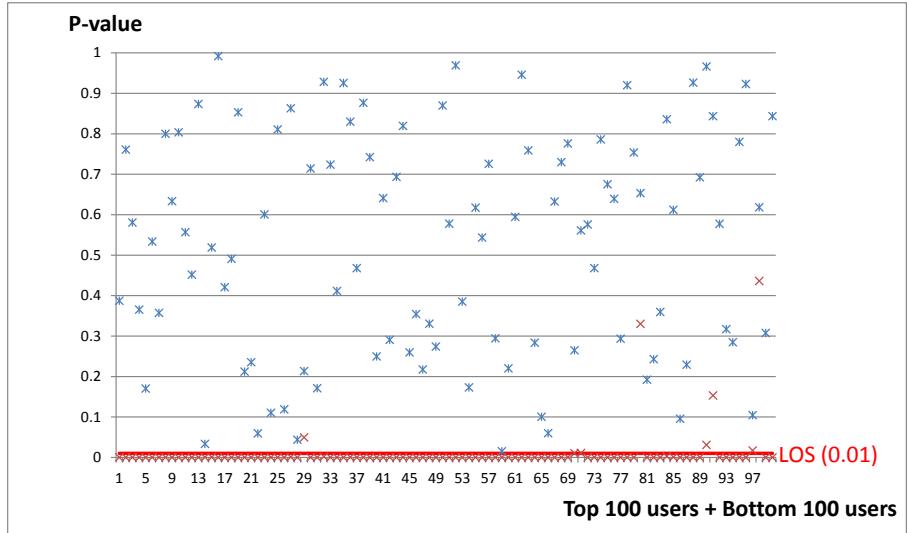
details on the T-test, refer to the text [6]. We compare two sets using the T-test: The first is a set of ratings and the second is a set of average ratings. We know that rating of user can represent an average rating. Note that our null hypothesis is valid when the result of T-test (called *P-value*) is more than significance level, and therefore, those users are representativeness. Alternative hypothesis is the opposite of the null hypothesis.



**Fig. 4.** Result of the T-test with level of significance (LOS) 0.05 depicted by the red line. The \* denotes the top 100 users and the x denotes the bottom 100 users.

Fig. 4 shows a result of T-test applying to the 100 high ranking users and the 100 low ranking users from Equation (1). The X-axis is users and the y-axis is p-value from the T-test. In Fig. 4, for example, the P-value of the first user from the top 100 is 0.388 and the P-value of the first user from the bottom 100 is very close to 0. Since the level of significance is 0.05, we can say that the first user from the top 100 has representativeness. On the other hand, the first user from the bottom 100 does not have representativeness because the P-value is smaller than the significance level. In Fig. 4, when the significance level is 0.05, all users from the top 100 except for three users are above the significance level and, thus, these users are representative reviewers. On the other hand, we say that the bottom 100 users are not representative reviewers since their P-values are less than 0.05.

Fig. 5 shows the top 100 and the bottom 100 users when the significance level is 0.01. The significance level 0.01 contains more users than 0.05. In the significance level 0.01 and 0.05, we cannot reject null hypothesis for most top 100 users. We can conclude that most top 100 users have representativeness in



**Fig. 5.** Result of the T-test with level of significance (LOS) 0.01 depicted by the red line. The \* denotes the top 100 users and the x denotes the bottom 100 users.

their music rater group. Most of the bottom 100 users are, on the other hand, in reject position of null hypothesis. Thus, we can say that the extracted users through our approach have representativeness.

## 5 Conclusions and Future Work

There are many different types of user participations in Web 2.0 and we can construct a social network of users based on these participations. For example, in Twitter, users use a special function called follow to other users. This function gives a network of users. Similarly, in blogs, people post an article and other users response it by comments or trackback. Again, we can make a social network of blog users based on these activities. In a real world society, people often make an opinion based on their trustful opinion leaders rather than mass media directly. An opinion leader is a person who has high influence to other people in a community. Since, we now have a social network in Web, it is natural to identify opinion leaders from the network. As a special opinion leader, we consider a user who has an accurate rating to items among many users; we call such user *representative user*. We have designed an algorithm that identifies representative users among many users based on the history of their ratings. We, then, have applied the proposed algorithm to the Yahoo! music dataset and demonstrated the usefulness of the algorithm by the T-test: We have formulated a hypothesis that opinions of influential users can represent all users and verified the statement.

Given overwhelming amount of information in Web, it is not easy to search good information solely by machines. Moreover, there are several user participation in Web 2.0. Researchers investigate how to use user participations for finding good information. One of the most well-known examples is collaborative filtering. There are many applications of collaborative filtering such as recommendation systems in Amazon<sup>3</sup> or Simania<sup>4</sup>, information retrieval or decision support systems in Google News<sup>5</sup>. However, we must have enough data to find good information. Namely, we cannot use the collaborative filtering without sufficient user data [3,11]. For resolving this problem, a method of finding representative users or opinion leaders is very useful. We plan to apply the proposed algorithm to a recommendation system with small size of users for movies or songs.

## References

1. Webscope from yahoo! labs, <http://webscope.sandbox.yahoo.com>
2. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM, pp. 207–218 (2008)
3. Billsus, D., Pazzani, M.J.: Learning collaborative information filters. In: ICML, pp. 46–54 (1998)
4. Blake, B.P., Agarwal, N., Wigand, R.T., Wood, J.D.: Twitter quo vadis: Is twitter bitter or are tweets sweet. In: ITNG, pp. 1257–1260 (2010)
5. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In: HICSS, pp. 1–10 (2010)
6. Bulmer, M.: Principle of Statistics. Dover Publications, New York (1979)
7. Gruhl, D., Guha, R.V., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: KDD, pp. 78–87 (2005)
8. Katz, E., Lazarsfeld, P.: Personal influence: the part of played by people in the flow of mass communications. Free Press, New York (1955)
9. Kwak, H., Lee, C., Park, H., Moon, S.B.: What is twitter, a social network or a news media? In: WWW, pp. 591–600 (2010)
10. Mishne, G., de Rijke, M.: Deriving wishlists from blogs show us your blog, and we'll tell you what books to buy. In: WWW, pp. 925–926 (2006)
11. Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: ACM Conference on Electronic Commerce, pp. 158–167 (2000)
12. Elkin, T.: Just an online minute online forecast,  
<http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticle&artid=29803>
13. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)

---

<sup>3</sup> <http://www.amazon.com>

<sup>4</sup> <http://simania.co.il>

<sup>5</sup> <http://news.google.com>

# Fitcolab Experimental Online Social Networking System

Haris Memic

Dzemal Bijedic University,  
Mostar, Bosnia and Herzegovina  
[memich@gmail.com](mailto:memich@gmail.com)

**Abstract.** Scholars have recently started to explore specific characteristics of increasingly popular online social networks. This paper presents Fitcolab online social network (OSN). This real life, modern OSN was created as an experimental research network which should allow for examination of various phenomena pertaining to network structure of online social networks. The main goal of the paper is to thoroughly describe and document Fitcolab OSN so that it can be easier referenced in future research work, as our research results that follow will be based on the wealth of collected data from this OSN. Additionally, anonymised network data will be gradually released to research community. The paper first presents users' wider environment and describes development process of the OSN. It then focuses on description of the final version of the system which was used to collect research data. Lastly, network boundary specification is carried out.

**Keywords:** Online social networks, social networks, social network analysis.

## 1 Introduction

Online social networks have become permanent habitats for many computer users. They have also started to attract growing number of researchers. Researchers from different disciplines attacked different phenomena pertaining to online social networks. Most of these studies are of a limited value, as they explore only very specific aspects of these networks. The same can be said about the research on network structure of online social networks which is the central theme of our research project. For example, as it has almost exclusively been done in the past, to study network of one relation (e.g. network of friendships) in isolation of networks of other relations (e.g. networks of blog comments, private messages, etc.) on same online social networking systems is a limited scholarly endeavor.

Our main research goal is a more encompassing one as we want to study the network and its creation from multiple angles. In order to be able to accomplish a more comprehensive study of structure in online social networks, created was an experimental OSN which will be presented in this paper. Beside the description of the OSN, the paper presents users' wider setting environment and elaborates the development of the system. At last network boundary specification of the experimental network was carried out.

## 2 Setting

Fitcolab was educational OSN in a sense that its users were students of FIT educational institution, as described below.

Faculty of Information Technology (FIT) is a higher-educational institution which acts as a very independent department of Dzemal Bijedic University in Mostar, Bosnia. Students of FIT come from various parts of the country. Academic year at FIT consists of two semesters, with no summer sessions. The most interesting experimentally collected data stems from 2008/09 academic year, and the highest quality data from the first semester of this year.

User population consisted of regular students (about one third) and distance learning (DL) students (about two thirds). Regular students attended their classes on-campus. DL students were receiving their study materials electronically and were usually only required to attend FIT for their final examinations.

## 3 Development of Fitcolab OSN

### 3.1 Evaluation of Users' Needs and Competing Systems

First phase of the development of Fitcolab consisted of evaluation of users' needs and evaluation of competing online systems. One of the reasons for selecting the chosen educational institution and their students as users of the system was the finding that DL students of this institution were neither connected nor did they have efficient means to network themselves.

Before the Fitcolab was launched almost all of the students of FIT had used "CS" online community. CS was an institutionalized online platform for FIT students and its educators and contained forum, blog and photo-gallery functionalities. The purpose of that system was almost exclusively dedicated for discussions related to individual courses. CS was pretty popular but was not a social networking system. It was built with sophisticated "Community Server" software.

Some of FIT students also used general social networking systems such as Facebook and MySpace, but at the time of the data collection for this research were not well connected with the help of those.

With regards to the above mentioned, it was estimated that a modern, sophisticated, institutionalized online social networking system, not closely related to students coursework, could be an attractive platform for the targeted users.

### 3.2 Selection of Underlying Software Solution

A requirement for setting up a sophisticated online social network, which would be able to compete with CS, ruled out, because of very expensive time costs, option of programming new software solution from the scratch. Because of that, for the purposes of this research project, an analysis of existing software products appropriate for the creation of the online social network was accomplished. Focus of the analysis was on open-source software products, which could be modified in order to satisfy the research purposes. Preliminary analysis narrowed down the possible solutions to

following products, all of whom were regarded as promising at time of their evaluation (from 03/2007 to 06/2007):

- AROUNDme (<http://www.barnraiser.org/aroundme>),
- Drupal (<http://drupal.org>),
- Elgg (<http://elgg.org>),
- PHPizabi ([www.phpizabi.net](http://www.phpizabi.net)).

AROUNDme was at the time of evaluation pretty new and untested platform and was used only from several websites. Primarily goal of the application was online collaboration but it also allowed online social networking. One of its major drawbacks was unsophisticated user interface.

Drupal was, from the preselected four, by far the most stable and most widely used software. As a generic software solution, Drupal was used for building simple personal webpages but was also extendable for the purposes of this research, if configured with a number of additional modules.

Elgg and PHPizabi were, as opposed to Drupal and AROUNDme, specifically designed for the purpose of building online social networking websites. At the moment of their evaluation both were in their beta software development phases, contained certain errors and were relatively rough software products. Despite that, both were characterized as very promising platforms for online social networking. However, PHPizabi was, from security aspects, characterized as vulnerable software, as relatively high number of PHPizabi sites was cracked. On the other side, Elgg was at the time successfully used from several educational institutions (e.g. STOA, Claremont Graduate University, Athabasca University, University of Brighton, etc.).

Security problems of PHPizabi and the fact that AROUNDme was an untested product resulted in elimination of these two. Although Drupal was by far the most stable software and had very large number of additional modules, Elgg was chosen as the software for the new OSN because of the following reasons: a) Elgg was already successfully deployed in few other educational settings; b) it was specifically designed as social networking software.

Few months after it was chosen for the purposes of this research, Elgg was together with Drupal voted as the second best open-source platform for social networking [1] (after WordPress which was not relevant for this research since it didn't have some of the standard functionalities of online social networking systems).

### **3.3 Initial Adjustments and Improvements of the Chosen Platform**

Official "launch" of the Fitcolab OSN was scheduled for the first semester of 2007/08 academic year. It was, however, pushed back because of waiting for upcoming Elgg 0.8 release and because the researcher wanted to accomplish initial adjustments and improvements of the software. After Elgg 0.8 was released, next three months were spent for a) understanding the source code of the software, b) translation of the user interface, c) adjustment of system settings, and d) some fundamental modifications and improvements. After initial system modifications and adjustments, 15 trial users were registered for the purposes of testing. Based on their feedback some of the system errors were found and corrected.

### 3.4 Official Start and Registration

Fitcolab OSN was officially launched on 29.12.2007. Registration of the users, students who were enrolled at the time, was accomplished during January 2008. Manual registration included setting the following data to be visible to all registered users: real name, place of residence, personal picture, year of study and way of study (on-campus or DL).

### 3.5 Further Improvements of the System

After the initial registration of most of the enrolled students was accomplished, it was continued with further improvements of the system. Improvements could be classified in those related to usability of the system and those related to individual functionalities of the system.

Usability improvements addressed some of the major users' complaints about the system, related mainly to very large number of options and sub-options of the OSN which in some cases were confusing to some of the new users. On the other hand, researcher also extensively worked on many hundreds of changes to the system, related to improvements of existing and addition of new functionalities.

## 4 Description of Fitcolab OSN

The final version of Fitcolab which was used for the purposes of data collection possessed all standard functionalities of popular modern OSN systems.

### 4.1 Main Menu Functionalities

The main menu of Fitcolab contained links to most of the main functionalities of Fitcolab OSN: Activity, Blog, Files, Wiki, My Network, Messages, Profile, Photogallery, Forum and Chat. Fig. 1 illustrates outlook of the home page of Fitcolab OSN. Short description of each of main menu functionalities follows.

*Activity.* Functionality that enabled users to view activities related to blogs, files and wikis. Activities could be filtered according to the type and source of activity.

*Blog.* Functionality that enabled users to create and publish blogs. Users could opt for the public blogs which were then shown on the home page or for the restricted blogs that were shown on home pages of selected users.

*Files.* Each user was allocated personal file space with the help of which he could share his files with other users.

*Wiki.* Wiki was a simplified Wikipedia-like personal feature. By the default all users and groups have been allocated their personal wikis.

*My Network.* The place from where users managed their connections with other users/groups. Main part of this feature represented friendship links with other users. As Fitcolab implemented directed friendship links users had two options for viewing their friendships, depending on whether the links were incoming or outgoing.

The screenshot shows the top portion of the Fitcolab website. At the top, there is a navigation bar with links like 'File', 'Edit', 'View', 'History', 'Bookmarks', 'Tools', and 'Help'. Below the navigation bar, the main header reads 'FITcolab' and 'Fakultet Informacijskih Tehnologija'. A search bar with placeholder text 'Pretraga' and 'Trazi' is located at the top right. Below the header, there is a menu bar with items such as 'Pocetna', 'Aktivnosti', 'Blog', 'Fajlovi', 'Wiki', 'Moja Mreza', 'Poruke (4)', 'Profil', 'Fitogalerija', 'Forum', 'eChat', 'Prestavke', and 'Ostvari se'. A sub-menu for 'eChat' is open, showing 'Postjedne slike' (Recent photos) featuring a photo of a person sitting on a bench, and 'Fitcolab, online zajednica za druženje i učenje' (Fitcolab, online community for networking and learning). To the right of the sub-menu, there is a sidebar titled 'Fitcolab Upstuvo' (Fitcolab News) with several news items from February 2009. Below the sub-menu, there are sections for 'Trenutno logovani' (Currently logged in) showing 'Maris Memic', and 'Vijesti sa News.com' (News from News.com) with a link to 'Yahoo vents frustration over App Store process'.

(a)

Fig. 1a. Outlook of Fitcolab's home page (upper part of the screen)

The screenshot shows the lower portion of the Fitcolab website. It features a news feed with a prominent red box containing the headline 'Are gamers really overweight and depressed?' dated 'Aug 19, 2009 7:51:00 PM'. Below the news feed, there are sections for 'Fitoci' (Friends) and 'Fitube' (Videos), each displaying four user profiles. A 'Desavanja koja slijede' (Upcoming events) section lists several events with small thumbnail images. On the right side, there is a 'Postjedni Blogovi' (Recent blogs) section showing posts by users like Sanjin Custovic, Leid Zejinovic, and Mesud Krupic, each with a brief description and a thumbnail image.

(b)

Fig. 1b. Outlook of Fitcolab's home page (lower part of the screen)

*Profile.* The standard profile functionality of OSN websites that served primarily as a place for publishing basic personal information about the owner of the user account. Profile information could be saved as public for all or public for only certain users.

*Photogallery.* It allowed users to publish their personal pictures as well as to view pictures of other users. Additionally, one of the five most recently published pictures from all users was randomly displayed on the home page of Fitcolab.

*Forum.* The forum was an integrated vBulletin forum software. It came with variety of useful options along with the possibilities of subscriptions to individual topics and automatic email notifications for new postings. Researcher also programmed the possibility for anonymous postings.

*Chat.* A standard Chat package was integrated. However, it was very rarely used.

## 4.2 Other Important Functionalities

Some of the functionalities that were not associated with the main menu, but were significant, are described below.

*Neretva.* On the front page of every user a customized flow of information was displayed, pertaining to his and his friends' actions, but also some of the actions of friends of his friends. This was analog and similar to Facebook's "News Feed".

*Fitube.* Functionality which allowed for embedding videos from popular video websites such as YouTube.

*Calendar.* All of the users were assigned their personal calendar, to which could add private or public events. Upcoming public events were shown to everyone on the home page of the OSN.

*Comment Wall.* It allowed users to post publicly visible comments on profiles of other users.

*Buddies.* Link 'Jarani' (eng. buddies), placed in the header of the OSN, was associated with rang list of users with most incoming friend connections (i.e. most popular users).

## 4.3 Content of the Home Page

The home page of Fitcolab OSN contained overview of the most relevant information. It contained: for each user personally customized "Neretva", last published photos, list of currently logged users, most recently active forum topics, last blogs, current news from the world, pictures of 8 randomly selected user profiles, one of the last embedded videos, and upcoming events taken from users calendars.

## 4.4 Creation and Dissolution of Online Friendships

Users created online friendships with other users by visiting their profiles and selecting option to add them as friends. Users could have also restricted all other users to add them without their previous consent, but most didn't opt to do so.

Deletion of a user from a friendship lists was accomplished by visiting the profile of a specific user and then selecting the appropriate option for friend deletion.

#### 4.5 Automatic Message Notifications

Important role for the interactivity and popularization of the OSN, like in most other social networking websites, had automatic private message notifications accompanied by automatic email messages to users. Private messages/emails were automatically sent to users in the case of new forum posts inside forum topics where they previously participated, in the case of comments on their pictures/files/profiles/blogs, and when they received private messages.

#### 4.6 Additional Activities on Fitcolab OSN

Beside the role of the classic online social network, Fitcolab was also place for few education activities. Two courses at FIT required students to use Fitcolab for some of their coursework activities. FIT also offered its students to do group-work on Fitcolab. Not many students, however, opted for this possibility.

All of these “side” activities were completed by the end of 2007/08 academic year, and were therefore not influencing research data that started to be collected at a later time and from users that didn’t participate in any of these ‘side’ activities.

### 5 Data Collection

The final version of the Fitcolab which was used for the research purposes and data collection possessed all standard functionalities of popular modern OSN systems.

#### 5.1 Data Collection Period

Start of data collection was the beginning of first semester of 2008/09 academic year. This was an appropriate starting time because of the following:

- During this semester there were no non-traditional (educational) activities on Fitcolab which could disturb the original purpose of this system as an online social networking system.
- First year students that enrolled for 2008/09 academic year were not influenced in any shape or form from any previously held academic activities on Fitcolab. They solely used the system for the purpose of online social networking. As it will be later elaborated, this was the main reason that the network boundary specification took into account this population of users.

Data for analysis for the research purposes was collected between 14.10.2008. and 22.02.2009. (131 days). Almost all of the users that were encompassed with the research experiment were registered on Fitcolab between 14.10.2008. and 17.10.2008.

#### 5.2 Data Collection Methods

Multiple data collection methods were used, among other things for the reasons of data redundancy. Data was collected and stored in: database of the OSN system, textual log files, and Apache web server log files. Google Analytics was also used.

Collected data included all resource requests (users' clicks) together with corresponding times. Extracted are friendship networks, messaging networks, different kinds of comment networks (blog/forum/picture comments), and other statistical data. These and other anonymised data will be gradually published at [www.memich.org/fitcolab/](http://www.memich.org/fitcolab/).

## 6 Network Boundary Specification

Laumann et al. in [2] identified three ways of network boundary specification for social networks: approach based on position, approach based on event, and relational approach. These approaches are not mutually exclusive. Some of research efforts used combination of the above suggested approaches (see [3]) for defining the boundaries of their networks. This research project used a combination of the approach based on position and the approach based on event.

### 6.1 Boundary Specification Based on Position

Boundary specification based on position identified the group of users that were first year students for the first time (enrolled in 2008/09). For the purposes of this research project it was decided that this would be the user population to be studied. This was decided because of the following reasons:

- a) As opposed to other students, first year students that enrolled in 2008/09 were not "burdened" with previous educational activities connected with this OSN. This generation of students used the Fitcolab as a classical online social network.
- b) Earlier research ([4], [5]) showed that students in online social networks tend to strongly segregate by their year/level of study. It was assumed that this would also apply to our freshmen generation of 2008/09, and that these users could thus be regarded as a distinct subsystem and therefore as a distinct online social network. Indeed, after the data collection was accomplished it was found that this population of 257 students (20.6% from all registered users) created 1165 online friendship links toward all registered users (not counting administrator). Out of 1165, 884 or 76% of links was directed towards users of the same freshmen group of 257 users. By chance it would be expected that less than 21% from the links be directed toward the "257" group. This indeed confirmed very strong homogeneity of the selected user population.
- c) Vast majority of these users was registered at Fitcolab at about the same time (middle of October 2008.). Almost all of the other users were registered much earlier (January 2008.). Research that would encompass both the selected and the remaining users would be significantly influenced by this large difference.

Because of the above points, administrator of the system (researcher) created on the OSN a distinct group in which all of the users of "257" population were placed.

## 6.2 Boundary Specification Based on Event

An analysis of OSN usage of the population “257” showed that certain number of these users never or almost never used the system during the time of data collection or spent very little time on the system. In order to be able to better reflect the reality, an analysis of individual usage of the OSN was accomplished, so as to find those users that could and should not be considered real users in later research analyses.

Criteria for this type of the network boundary specification were set to be: a) minimum number of visits to the OSN, b) minimum amount of total individual time spent in the OSN.

It is worth to remark that recent research, as opposed to earlier ones, considers “lurker” users (users that are present on the system but do not participate actively) as members of online communities (e.g. [6], [7], [8], [9]). Moreover, active users of online systems regard lurkers as users that do belong to online communities [7]. Because of the mentioned, lurkers will not be exempt from our user population.

On the other hand it is very difficult to justify inclusion of users that never or almost never visit online communities as real members. By analyzing user log data it was found that 30 users of “257” population never logged in. These were the students that were manually registered from administrator but because of certain reasons never logged to the OSN. Additional 9 users logged on the OSN only during a single calendar day and never after that. Users that never logged as well users that logged only during a single day without ever returning back were not treated as members of the OSN. After the exemption of these there remained 218 users.

In order to find the true specification of user population, analyzed was the system usage of remaining 218 users, taking into account criteria b) of boundary specification based on event. More specifically, following two sub-criteria were used:

- amount of total individual time spent using the OSN,
- time passed between the first and the last visit of every user.

To the best of our knowledge, literature doesn't provide specific minimal thresholds for these two sub-criteria. It would certainly be preferential to draw the boundary by removing those users that were barely present on the system, as well as those users that were only present during a very short time period.

Because of important link dependencies in social networks in general, we decided to set relatively loose elimination criteria. Users that didn't satisfy the following were not considered the members of the OSN:

- Total time spent on the system > 10 minutes (Average time of a single visit from the population “218” was calculated to be  $\approx 10$  minutes).
- Time passed from the first to the last visit  $\geq 7$  days. This criteria is in the line with the results obtained by Leskovec et al. (2008), where it was observed that it was problematic to model nodes that were active less than a week (Average time between the first and last visits for “218” population was  $\approx 107$  days).

Out of 218 users, first of the above conditions didn't satisfy 5 of them. Total times spent on the system for these users were  $\approx 3, 4, 5, 6$  and 8 minutes.

Additional four users didn't satisfy the second condition. Time differences between the first and the last visit for these users were 1, 2, 2 and 3 days.

By removing users that didn't satisfy above two criteria obtained was the "real" member population of Fitcolab OSN consisting of 209 users. A verification check showed that out of 48 (257 – 209) users that were excluded by network boundary specification neither had outgoing links (online friendships), which further justified removal of those users. The "209" population will be the user population that, together with the relations between them, will represent the online social network to be researched.

## 7 Conclusion

The paper presented and documented Fitcolab experimental online social network. Rich datasets from this OSN have been collected and are yet to be analyzed and modeled as a part of a larger research project whose goal is to study network structural features of online social networks more comprehensively than it has been done to date. Described were also the environment setting of the system and its users, development process of the system, and data collection methods. At the last, network boundary specification was accomplished in order to disclose and remove users that, although registered, in reality did not belong to the social network and thus should not be included in statistical analyses that will follow.

## References

1. Open Source CMS Awards 2007, Pact Publishing (29.10.2007),  
<http://www.packtpub.com/article/wordpress-wins-best-open-source-social-networking-cms>
2. Laumann, E., Marsden, P., Prensky, D.: The Boundary Specification Problem in Network Analysis. In: Burt, R., Minor, M. (eds.) Applied Network Analysis, pp. 18–34. Sage, Beverly Hills (1983)
3. Marin, A., Wellman, B.: Social Network Analysis: An Introduction. In: Carrington, P., Scott, J. (eds.) Forthcoming in Handbook of Social Network Analysis. Sage, London (2010)
4. Traud, A., Kelsic, E., Mucha, P., Porter, M.: Community Structure in Online Collegiate Social Networks, arXiv:0809.0960 (2008)
5. Mayer, A., Puller, S.: The Old Boy (and Girl) Network: Social Network Formation on University Campuses. Journal of Public Economics 92(1-2), 329–347 (2008)
6. Preece, J., Maloney-Krichmar, D.: Online Communities: Focusing on Sociability and Usability. In: Jacko, J., Sears, A. (eds.) Handbook of Human-Computer Interaction. Lawrence Erlbaum, Mahwah (2003)
7. Nonnecke, B., Andrews, D., Preece, J.: Non-public and Public Online Community Participation: Needs, Attitudes and Behavior. Electronic Commerce Research 6(1), 7–20 (2006)
8. Preece, J.: Online Communities: Designing Usability, Supporting Sociability. Wiley & Sons, New York (2000)
9. Schoberth, T., Preece, J., Heinzl, A.: Online Communities: A Longitudinal Analysis of Communication Activities. In: Proceedings of the Hawaii International Conference on Systems Sciences, Big Island, Hawaii (2003)

# General Network Properties of Friendship Online Social Network

Haris Memic

Dzemal Bijedic University,  
Mostar, Bosnia and Herzegovina  
memich@gmail.com

**Abstract.** This paper explores visiting metrics and some of the more important general network properties of Fitcolab online social network (OSN). The wide array of statistics was explored in order to obtain general insight that will not only be useful by itself but would also serve as the starting platform for more focused research endeavors that are to be based on the same experimental network. Longitudinal OSN usage patterns are studied for number of visits, average duration of visits and number of active users. Network structural characteristics analyzed included reachability, average distance, diameter, clustering coefficient, in/out degree distributions. Longitudinal analyses of a number of structural characteristics were also carried out. Partial interpretation of the results followed.

**Keywords:** Online social networks, social networks, social network analysis.

## 1 Introduction

This paper analyses more important web usage patterns as well as some of the more significant general network properties of Fitcolab online social network (OSN). Fitcolab was created as an experimental platform for a larger project (see [1]) whose goal is to study network structural patterns in online social networks more comprehensively than it has been done to date. The analysis presented in this paper is a starting point for the work that will follow. Here we explore the wide array of statistics in order to obtain the general insight that will not only be useful by itself but should also serve as a platform for more focused future research endeavors.

The paper starts with analysis of OSN usage which encompasses longitudinal analyses of relevant visiting metrics. It then explores basic structural characteristics such as reachability, average distance, diameter, clustering coefficient and degree distribution. Longitudinal analyses of a number of structural characteristics were also carried out. Partial interpretation of the results is included.

## 2 OSN Usage Analysis

A visit, in vocabulary of web analytics, denotes an interaction between a user and a certain web system and consists of browsing one or more web pages of that web

system. Browsing time for individual web pages is usually calculated by taking the difference between the moment a user accesses current page and the moment he accesses the next page. This way of calculating works well, except for the last visited page on web systems, because its time cannot often be calculated as it is not possible to log neither time (moment) of visit to web pages on other systems nor time when users close their browsers.

Because of this, most of professional programs for measurement of individual visits to web systems denote a specific visit/session to be ended when a user is idle for 30 or more minutes ([2], [3], [4]). In order to be able to measure visiting patterns of a subset of entire user population, as it was necessary for this work because of network boundary specifications (see [1]), researcher implemented a computer program to analyze visits of selected 209 users. Because of the above, it was specified that an individual visits/session ended provided that user was idle for 30 or more minutes.

Web usage statistics of the online social network are summarized in Table 1.

**Table 1.** Cumulative usage patterns of Fitcolab OSN

<b>Traffic</b>	
Total number of visits	18,172
Avg. number of visits per user	86.9
Total time spent from all users (minutes)	191,519
Avg. time spent per one visit (minutes)	10.54
Avg. difference between first and last visit (days)	110.6

During the period of data collection, which encompassed 131 days of observation, there were 18,172 unique visits from 209 users. On average, users visited the network 87 times. Altogether users spent on the system 191,519 minutes or about 3,192 hours. Average duration of visits was 10.5 minutes. Average time differences between first and last visits of individual users amounted to more than 110 days.

Longitudinal movements of the following statistics were analyzed:

- number of visits,
- average duration of visits, and
- number of active users.

These longitudinal measurements were done in intervals of 7 days, with the exception of the first and last intervals, as it was impossible to nicely divide the period of observation by number seven. First interval was specified to be shorter, as almost all of the selected user population was registered during that time frame. In order to allow for more accurate measurements, borders for time intervals were set to 4 a.m., a time when the system usage was about the lowest. This almost entirely eliminated the problem that would be present if interval borders would be placed during busier hours, as then some of visits would be counted twice (since they would be crossing interval borders).

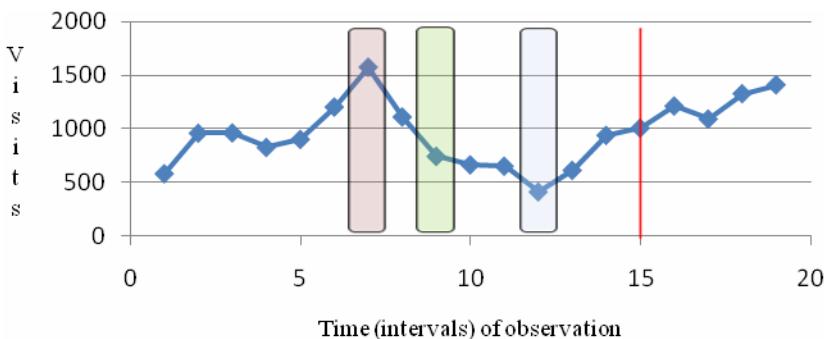
Considering above points, following intervals were specified for the longitudinal analysis of the OSN usage.

**Table 2.** Specified intervals for longitudinal analysis

Interval	Period
Interval 1	14-10 (15:00:00h) to 19-10 (04:00:00h)
Interval 2	19-10 (04:00:01h) to 26-10 (04:00:00h)
Interval 3	26-10 (04:00:01h) to 02-11 (04:00:00h)
...	...
...	...
Interval 16	25-01 (04:00:01h) to 01-02 (04:00:00h)
Interval 17	01-02 (04:00:01h) to 08-02 (04:00:00h)
Interval 18	08-02 (04:00:01h) to 15-02 (04:00:00h)
Interval 19	15-02 (04:00:01h) to 22-02 (15:16:25h)

*Number of Visits per Week*

Longitudinal movement of number of visits per defined intervals is shown in Fig. 1.

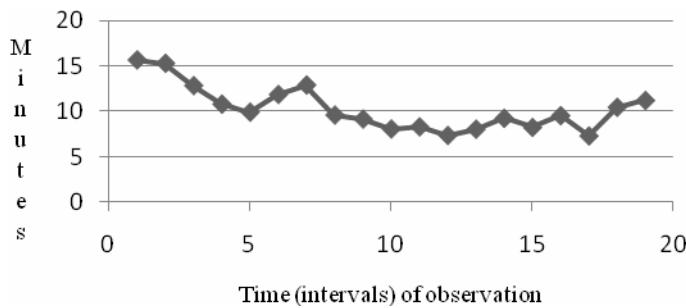
**Fig. 1.** Longitudinal display of the total number of visits

In Fig. 1 pink color represents week of midterm exams (no classes during that period), green color represents the period of holidays, blue color represents the second period of holidays, and red line represents the last day of classes. Evident are periods of the mild “saturation” (interval 4), rapid increase of the popularity of the OSN (intervals 6-7), followed by decreasing popularity (intervals 9-12) and lastly the new period of “popularization”.

Interesting to find was that extreme points of the Fig. 1 overlap with the important events in students/users lives. Interval with the largest number of visits (interval 7) overlaps with the period of midterms. Period of final examinations was also characterized by high number of visits. Intervals of low system usage correspond to periods of holydays.

*Average Time of Visits*

Analysis of average times users spent per a visit during each of the intervals produced data shown in Fig. 2.



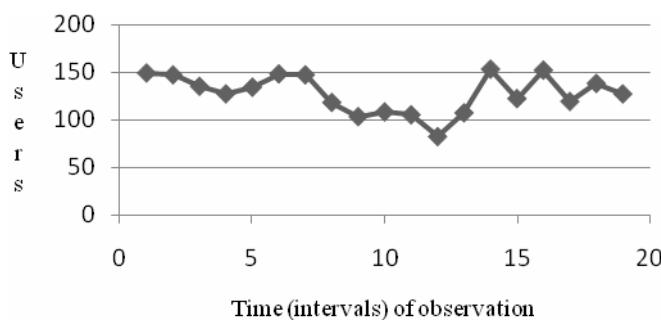
**Fig. 2.** Longitudinal display of average duration of visits

This illustration, in combination with Fig. 1 shows that users tend to spend more time per visit:

- during first weeks of system usage,
- during intervals which have more visits (when network is more active).

#### *Active Users*

Longitudinal movement of the number of users that were present (visited at least once) during certain intervals is illustrated with the next figure.



**Fig. 3.** Longitudinal movement of number of active users

Obvious is again the same period of the low visiting rate (intervals 9-13).

### **3 Basic Structural Characteristics of the Network**

Some of the more important characteristics of the final online friendship network are shown in Table 3.

**Table 3.** General structural characteristics of the online friendship social network

General network characteristics	
Number of nodes	209
Number of links	845
Avg. number of in/out links	4.04
Density	0.0194
Reachability	0.332
Avg. shortest path between two nodes	3.32
Diameter	8
Percentage of reciprocal pairs	49.3
Clustering coefficient	0.262
Largest weakly connected component	143 nodes (68.4%)
Largest strongly connected component	101 nodes (48.3%)

The network consists of 209 nodes and 845 directed links. Average number of in/out links per node is 4.04. About one third of all possible pairs are connected either directly or indirectly. Average length of shortest paths between directly or indirectly connected nodes is 3.32. Diameter (longest shortest path between two nodes) is 8. Percentage of reciprocated pairs is 49.3% and was calculated with the formula:

$$\text{NumberOf}(x_{ij} > 0 \text{ AND } x_{ji} > 0) / \text{NumberOf}(x_{ij} > 0 \text{ OR } x_{ji} > 0). \quad (1)$$

#### Degree Distributions

Incoming degree distribution is displayed in Fig. 4, where upper (black) rows show the number of incoming links and lower (white) rows the number of users with the corresponding number of incoming links. 53 users (~25%) of OSN do not have incoming links, and are thus online friends of no one. Out of those users that have incoming links, large majority or ~86% have between 1 and 9 incoming links. None of the users has more than 26 incoming links.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
53	32	19	18	18	13	8	9	9	8	1	4	3	1	6	1	1	0	0	1
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
40	41	42	43	44	45	46	47	48	49	50									
0	0	0	0	0	0	0	0	0	0	0									

**Fig. 4.** Distribution of incoming links (range[0...50])

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
90	24	17	9	12	5	11	1	5	3	5	3	7	1	2	3	0	0	0	1
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
1	2	0	0	1	0	0	1	2	0	1	1	0	0	0	0	0	0	0	0
40	41	42	43	44	45	46	47	48	49	50									
0	0	0	0	0	1	0	0	0	0	0									

**Fig. 5.** Distribution of outcoming links (range[0...50])

Distribution of outcoming links is displayed in Fig. 5, where upper rows show the number of outcoming links and lower the number of users with the corresponding number of outcoming links. Full 90 (~43%) users do not have any outgoing links (don't have any online friends). Distribution of outcoming links is more dispersed then distribution of incoming links. User activity with regards to friendships is more variable then user popularity. The most extreme outlier had 45 outgoing links, 14 more than the second most active user.

## 4 Longitudinal Analysis of Structural Characteristics

Longitudinal movements of the following relevant network characteristics were analyzed:

- density,
- reachability,
- clustering coefficient, and
- reciprocity.

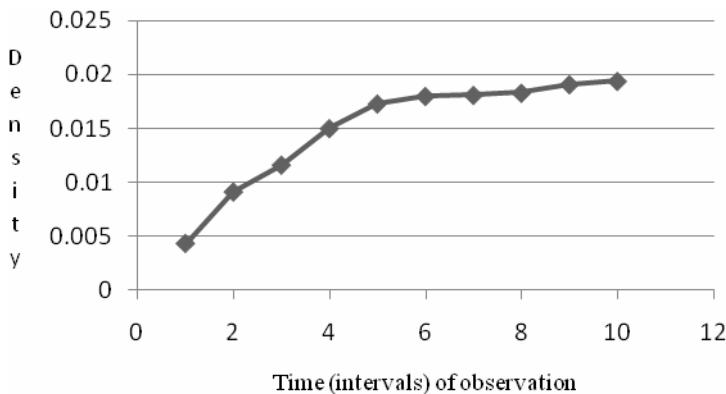
These were measured every second week (14 days). As longitudinal measurements took into account cumulative statistics, and taking into account that the increase in number of total links significantly decreased two months after the data collection started, it didn't make sense to take more frequent observations. First moment of the observation was 19.10.2008. All of the moments are shown in Table 4.

**Table 4.** Observation moments for longitudinal analysis

Observation	Moment
Observation 1	19-10
Observation 2	02-11
Observation 3	16-11
Observation 4	30-11
Observation 5	14-12
Observation 6	28-12
Observation 7	11-01
Observation 8	25-01
Observation 9	08-02
Observation 10	22-02

### Density

Fig. 6 displays the increase of network density against time (observations).

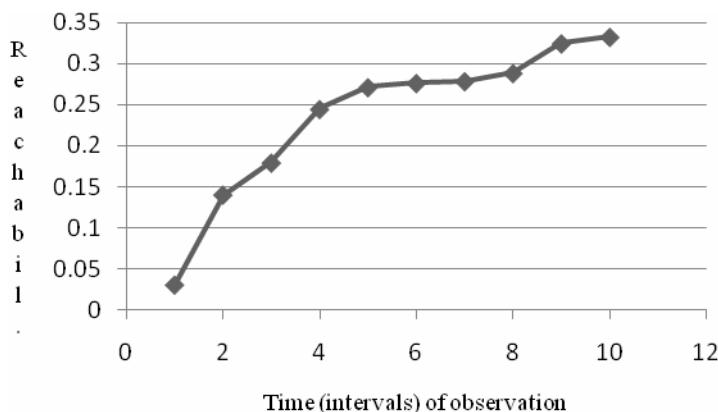


**Fig. 6.** Longitudinal display of network density

It can be seen that the network density increased rapidly during first five moments of observation. After that network continued to grow (number of links increased) albeit much slower.

### Reachability

Fig. 7 shows longitudinal movements of network reachability.

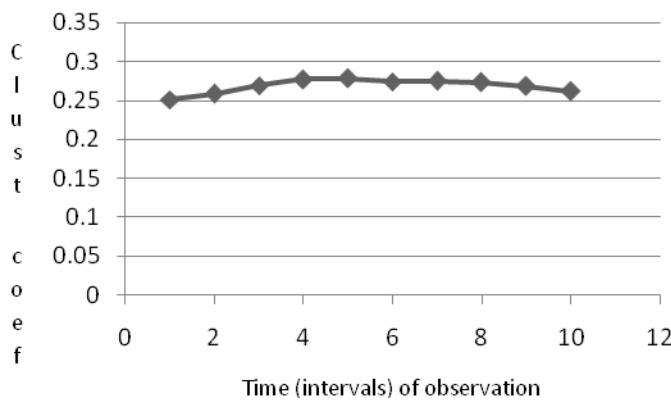


**Fig. 7.** Longitudinal display of network reachability

As expected, increase in number of links results in increased reachability in the network.

### *Clustering Coefficient*

Longitudinal movement of clustering coefficient is illustrated with Fig. 8.

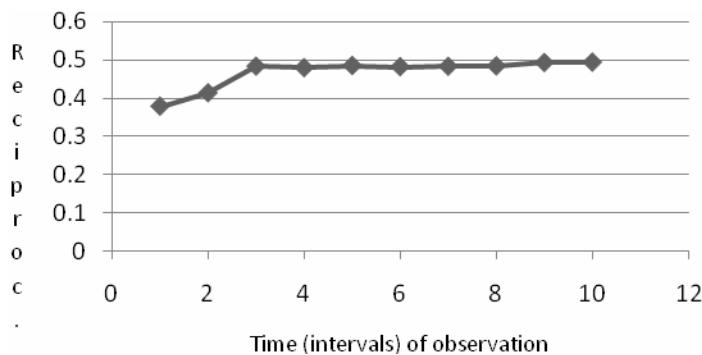


**Fig. 8.** Longitudinal display of clustering coefficient

It is interesting to observe that the value of clustering coefficient first mildly increases and then after some time starts mildly to decrease.

### *Reciprocity*

Longitudinal movement of reciprocity of pairs is shown in Fig. 9.

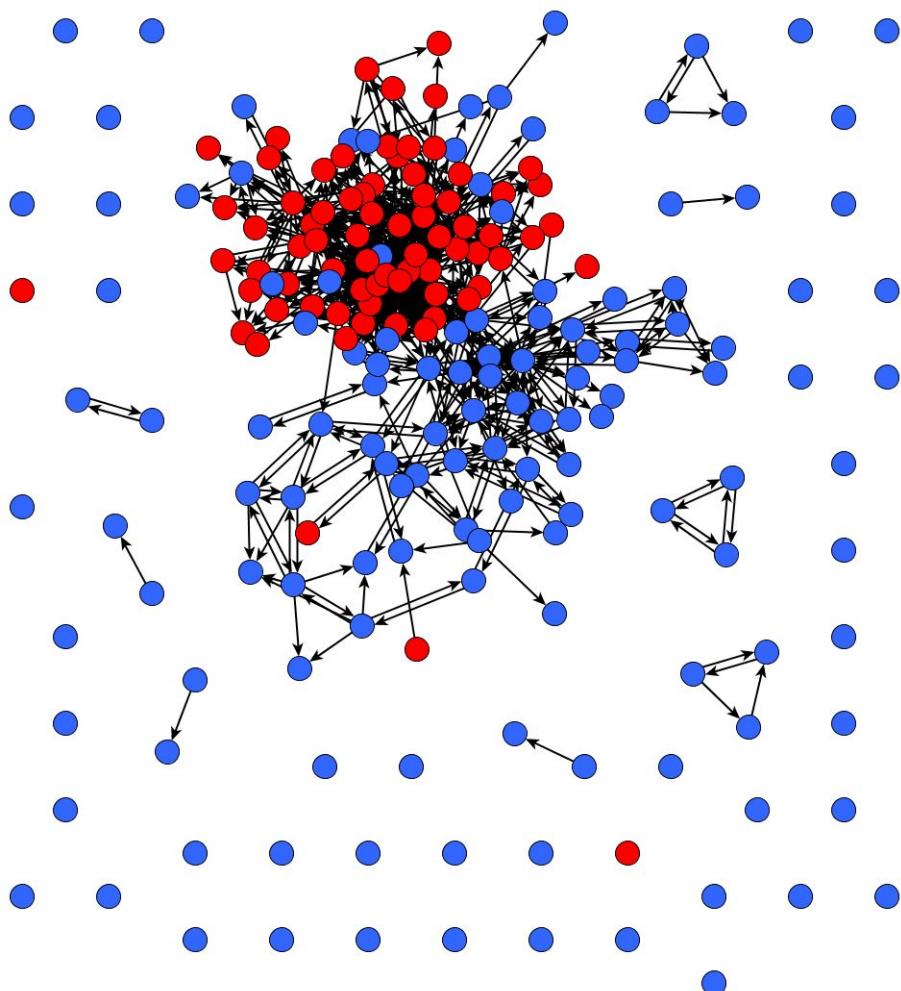


**Fig. 9.** Longitudinal display of percentage of reciprocal pairs

Proportion of reciprocal pairs increases at first and then stays about constant.

## 5 Visualization of the Network

As part of this initial analysis the online social network was visualized. By using stress-minimization algorithm, as elaborated in [5] and implemented in visone<sup>1</sup> program, obtained was the layout of Fitcolab online social network, shown in Fig. 10. Users with the attribute DL (distance learning students, see [1]) are marked blue, whereas on-campus students are marked red.



**Fig. 10.** Graph of realized Fitcolab OSN: displayed with stress-minimization algorithm

---

<sup>1</sup> <http://visone.info>

Vivid is strong segregation of network nodes based on the user attribute. It can also be seen that red nodes (regular students) are more strongly connected than blue nodes (DL students). From displayed network we can also conclude that DL students have greater tendency to be isolated.

## 6 Conclusion

This paper analyzed usage patterns and some of the more significant general network properties of Fitcolab online social network. Exploration of wide variety of descriptive statistics was accomplished in order to obtain general insights about the network and structural properties of its online friendship network.

Among more important findings of this preliminary analysis are: 1. usage patterns of users/students interact heavily with the important moments in their academic lives, 2. there is very high tendency for reciprocation of online friendships, 3. distribution of outgoing friendship links is more dispersed then distribution of incoming links, 4. network density increases fast at the beginning and more slowly as the time passes, 5. strong tendency for segregation based on attribute of way of studying.

## References

1. Memic, H.: Fitcolab Experimental Online Social Networking System. In: Pan, J.-S., Chen, S.-M., Nguyen, N.T. (eds.) ICCCI 2010, Part II. LNCS (LNAI), vol. 6422, pp. 31–40. Springer, Heidelberg (2010)
2. Kaushik, A.: Web Analytics: An Hour a Day. Wiley Publishing, Indiana (2007)
3. Sostre, P., LeClaire, J.: Web Analytics for Dummies. Wiley, Hoboken (2007)
4. Burby, J., Atchison, S.: Actionable Web Analytics: Using Data to Make Smart Business Decisions. Wiley, Chichester (2007)
5. Baur, M., Schank, T.: Dynamic Graph Drawing in Visone. Technical University Karlsruhe, Karlsruhe (2008),  
[http://i11www.iti.uni-karlsruhe.de/extra/publications/  
bs-dgdv-08.pdf](http://i11www.iti.uni-karlsruhe.de/extra/publications/bs-dgdv-08.pdf)

# Fuzzy Decision Making for IJV Performance Based on Statistical Confidence-Interval Estimates

Huey-Ming Lee<sup>1</sup>, Teng-San Shih<sup>2</sup>, Jin-Shieh Su<sup>2</sup>, and Lily Lin<sup>3</sup>

<sup>1</sup> Department of Information Management, Chinese Culture University  
55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 11114, Taiwan  
[hmllee@faculty.pccu.edu.tw](mailto:hmllee@faculty.pccu.edu.tw)

<sup>2</sup> Department of Applied Mathematics, Chinese Culture University  
55, Hwa-Kang Road, Yang-Ming-Shan, Taipei 11114, Taiwan  
[shih@mail.pccu.edu.tw](mailto:shih@mail.pccu.edu.tw), [sjs@faculty.pccu.edu.tw](mailto:sjs@faculty.pccu.edu.tw)

<sup>3</sup> Department of International Business, China University of Technology  
56, Sec. 3, Hsing-Lung Road, Taipei 116, Taiwan  
[lily@cute.edu.tw](mailto:lily@cute.edu.tw)

**Abstract.** This paper considers the International Joint Ventures (IJVs) problem using interval-valued fuzzy sets and the compositional rule of inference in the statistical sense. We consider the performance effects of evaluation criteria facts and weights based on fuzzy set theory to determine the performance ranking among IJVs. Due to the lack of precise information based on some fuzzy language is used in the evaluation criterion. We use the statistical confidence-interval estimates and apply the signed distance method to defuzzify.

**Keywords:** Joint Ventures, Fuzzy decision making, Compositional rule, Confidence-Interval; Signed distance.

## 1 Introduction

The object of International Joint Ventures (IJVs) is better market performance. However many factors have been suggested in the literature as potentially important determinants of IJV performance [4], financial indicators [2, 6], survival or venture liquidation [9], duration [7, 10], and instability or ownership changes [3, 5]. Shieh *et al.* [12] used the interval-value fuzzy sets to evaluate International Joint Ventures. There have been a number of factors identified in the previous studies as having a significant influence on IJV performance. Due to the lack of precise information based on some fuzzy language is used in the evaluation criterion. The fuzzy manual judgment concept considered in a decision making process therefore becomes very important. If we consider the performance effects of these factors together, evaluating IJV performance becomes a very important issue in this uncertain and fuzzy real world.

In this paper, we consider the evaluation criteria facts and weight performance effects. We use the statistical confidence interval concept to fuzzify the weight with the triangular fuzzy numbers and defuzzify by the signed distance method, to determine the performance ranking among the IJVs. This paper is organized as follows. Section 2

presents some definitions and propositions. Section 3 considers fuzzy decision making using the fuzzy number and interference composition rule in the statistical sense. Section 4 provides an example.

## 2 Preliminaries

This section contains some definitions and propositions used in the Section 3.

**Definition 1.** A fuzzy set  $\tilde{A}$  defined on  $R$  is called the level  $\lambda$  triangular fuzzy number if its membership function is

$$\mu_{\tilde{A}}(x) = \begin{cases} \lambda \frac{(x-a)}{b-a}, & a \leq x \leq b \\ \lambda \frac{(c-x)}{c-b}, & b \leq x \leq c \\ 0, & \text{otherwise.} \end{cases}$$

where  $a < b < c$ ,  $0 < \lambda \leq 1$ , then  $\tilde{A}$  is called the level  $\lambda$  fuzzy number and denoted by  $\tilde{A} = (a, b, c; \lambda)$ . When  $\lambda=1$  is called a triangular fuzzy number and denoted by  $\tilde{A} = (a, b, c)$ .

**Definition 2.** ([11]) Suppose that  $\tilde{B}^L = (a, b, c; \lambda)$ ,  $\tilde{B}^U = (p, b, r; \rho)$ , where  $p < a < b < c < r$ ,  $a, b, c, p, r \in R$ ,  $0 < \lambda < \rho \leq 1$ . Let  $\tilde{B} = [\tilde{B}^L, \tilde{B}^U]$  and the membership grade of  $\tilde{B}$  at  $x \in R$  belongs to the interval  $[\mu_{\tilde{B}^L}(x), \mu_{\tilde{B}^U}(x)]$  (see Fig. 1),  $\tilde{B}$  is called a level  $(\lambda, \rho)$  interval-valued fuzzy number, and called a level  $(\lambda, \rho)$  i-v fuzzy number for short.

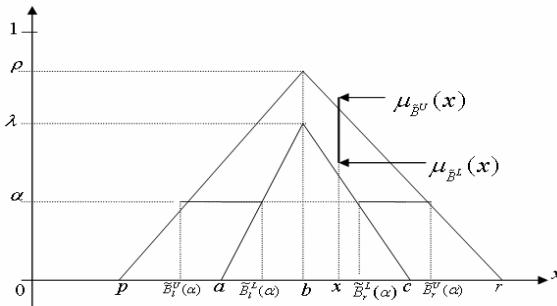


Fig. 1.  $\alpha$ -Cut of level  $(\lambda, \rho)$  i-v fuzzy number  $\tilde{B}$

The  $\alpha$ -level set of  $\tilde{B} = [(a, b, c; \lambda), (p, b, r; \rho)]$  is defined as follows:

If  $0 \leq \alpha \leq \lambda$ ,  $\alpha$ -level set of  $\tilde{B}$  is defined as

$$B(\alpha) = \{x | \mu_{\tilde{B}^U}(x) \geq \alpha\} - \{x | \mu_{\tilde{B}^L}(x) > \alpha\} = [\tilde{B}_l^U(\alpha), \tilde{B}_l^L(\alpha)] \cup [\tilde{B}_r^L(\alpha), \tilde{B}_r^U(\alpha)] \quad (\text{see Fig. 1}) \quad (1)$$

where

$$\begin{aligned}\tilde{B}_l^L(\alpha) &= a + (b - a) \frac{\alpha}{\lambda}, \quad \tilde{B}_r^L(\alpha) = c - (c - b) \frac{\alpha}{\lambda} \quad (\text{see Fig. 1}) \\ \tilde{B}_l^U(\alpha) &= p + (b - p) \frac{\alpha}{\rho}, \quad \tilde{B}_r^U(\alpha) = r - (r - b) \frac{\alpha}{\rho}\end{aligned}\quad (2)$$

For each  $\alpha \in [0, \lambda]$  the mapping

$$[\tilde{B}_l^U(\alpha), \tilde{B}_l^L(\alpha)] \leftrightarrow [\tilde{B}_l^U(\alpha), \tilde{B}_l^L(\alpha); \alpha] \text{ and } [\tilde{B}_r^L(\alpha), \tilde{B}_r^U(\alpha)] \leftrightarrow [\tilde{B}_r^L(\alpha), \tilde{B}_r^U(\alpha); \alpha] \quad (3)$$

are one-to-one mapping.

If  $\lambda \leq \alpha \leq \rho$  then

$$B(\alpha) = [\tilde{B}_l^U(\alpha), \tilde{B}_r^U(\alpha)] \quad (\text{see Fig. 1}) \quad (4)$$

$$\tilde{B}_l^U(\alpha) = p + (b - p) \frac{\alpha}{\rho}, \quad \tilde{B}_r^U(\alpha) = r - (r - b) \frac{\alpha}{\rho} \quad (5)$$

For each  $\alpha \in [\lambda, \rho]$ ,

$$[\tilde{B}_l^U(\alpha), \tilde{B}_r^U(\alpha)] \leftrightarrow [\tilde{B}_l^U(\alpha), \tilde{B}_r^U(\alpha); \alpha] \quad (6)$$

is one-to-one mapping. According to Decomposition Theory, Fig. 1, (1)~(6),  $\tilde{B}$  can be denoted as follows:

$$\tilde{B} = \bigcup_{0 \leq \alpha < \lambda} ([\tilde{B}_l^U(\alpha), \tilde{B}_l^L(\alpha); \alpha] \cup [\tilde{B}_r^L(\alpha), \tilde{B}_r^U(\alpha); \alpha]) \cup \bigcup_{\lambda \leq \alpha \leq \rho} [\tilde{B}_l^U(\alpha), \tilde{B}_r^U(\alpha); \alpha]. \quad (7)$$

Let

$$F_{IV}(\lambda, \rho) = \{(a, b, c; \lambda), (p, b, r; \rho) \mid p < a < b < c < r, a, b, c \in R, 0 < \lambda < \rho \leq 1\}.$$

**Definition 3.** For each  $a, 0 \in R$  we define the signed distance from  $a$  to 0 by  $d_0(a, 0) = a$ . [8]

**Definition 4.** (a) Let  $\tilde{B} = [(a, b, c; \lambda), (p, b, r; \rho) \in F_{IV}(\lambda, \rho)]$ , the signed distance from  $\tilde{B}$  to  $\tilde{0}$  is defined by [8]

$$d(\tilde{B}, \tilde{0}) = \frac{1}{16} [6b + a + c + 4p + 4r + 3(2b - p - r) \frac{\lambda}{\rho}]$$

(b) The signed distance of the  $\rho$  triangular fuzzy number  $\tilde{A} = (p, b, r; \rho)$  is defined by

$$d(\tilde{A}, \tilde{0}) = \frac{1}{2\rho} \int_0^\rho (\tilde{A}_l(\alpha) + \tilde{A}_r(\alpha)) d\alpha = \frac{1}{4} (2b + p + r)$$

From Definition 4, the ranking of level  $(\lambda, \rho)$  i-v fuzzy number of  $F_{IV}(\lambda, \rho)$  is defined as follows:

**Definition 5.** For  $\tilde{B}, \tilde{C}, \tilde{D} \in F_{IV}(\lambda, \rho)$ , the following rankings on  $F_{IV}(\lambda, \rho)$  are defined as

$$\tilde{B} \prec \tilde{C} \Leftrightarrow d(\tilde{B}, \tilde{0}) < d(\tilde{C}, \tilde{0}) \quad \tilde{B} \approx \tilde{C} \Leftrightarrow d(\tilde{B}, \tilde{0}) = d(\tilde{C}, \tilde{0})$$

By relation " $\prec$ " is a linear order on  $R$  and Definition 5, we have the following Proposition.

**Proposition 1.** For  $\tilde{B}, \tilde{C}, \tilde{D} \in F_{IV}(\lambda, \rho)$ , the following properties hold:

(a) There is only one of  $\tilde{B} \prec \tilde{C}, \tilde{B} \approx \tilde{C}, \tilde{C} \prec \tilde{B}$  holds

(b) It satisfies the three ranking axioms:

$$(1) \tilde{B} \prec \tilde{B} \quad (2) \text{If } \tilde{B} \prec \tilde{C}, \tilde{C} \prec \tilde{B} \text{ then } \tilde{B} \approx \tilde{C} \quad (3) \text{If } \tilde{B} \prec \tilde{C}, \tilde{C} \prec \tilde{D}, \text{ then } \tilde{B} \prec \tilde{D}$$

Using Proposition 1, the ordering " $\prec, \approx$ " is a linear order on  $F_{IV}(\lambda, \rho)$

**Definition 6.** For  $\tilde{B}_j \in F_{IV}(\lambda, \rho), j=1,2,\dots,n$ , the following holds:

$$\tilde{B}_q = \underset{j \in \{1,2,\dots,n\}}{\text{Max}} \tilde{B}_j, q \in \{1,2,\dots,n\}$$

if  $\tilde{B}_j \prec \tilde{B}_q$ , for all  $j \in \{1,2,\dots,n\} \Leftrightarrow d(\tilde{B}_j, \tilde{0}) \leq d(\tilde{B}_q, \tilde{0})$  for all  $j \in \{1,2,\dots,n\}$ .

It is easy to know that

$$(a, b, c; \delta) \oplus (e, g, h; \delta) = (a+e, b+g, c+h; \delta), 0 < \delta < 1, \text{ and}$$

$$k(a, b, c; \delta) = \begin{cases} (ka, kb, kc; \delta), & k > 0 \\ (kc, kb, ka; \delta), & k < 0 \end{cases}, k \in R$$

Using Definitions 4 we have the following proposition.

**Proposition 2.** If  $\tilde{B}, \tilde{C} \in F_{IV}(\lambda, \rho), k \in R$ , then

$$d(\tilde{B} \oplus \tilde{C}, \tilde{0}) = d(\tilde{B}, \tilde{0}) + d(\tilde{C}, \tilde{0}) \text{ and } d(k\tilde{B}, \tilde{0}) = k d(\tilde{B}, \tilde{0})$$

### 3 Fuzzy Decision Making with Statistical Confidence Interval Evaluations Approaches

There are  $n$  alternatives  $A = \{A_1, A_2, \dots, A_n\}$  under  $m$  evaluations criteria  $B = \{B_1, B_2, \dots, B_m\}$  are determined by  $r$  experts. Let  $a_{ijq} \in [0,1]$  denote the effectiveness score of the  $j^{th}$  alternative under the  $i^{th}$  evaluation criterion given by the  $q^{th}$  expert, where  $i = 1, 2, \dots, m, j = 1, 2, \dots, n, q = 1, 2, \dots, r$ .

### 3.1 Fuzzy Relation with Membership Grade [12]

Let  $a_{ij} = \frac{1}{r} \sum_{q=1}^r a_{ijq} \in [0,1], i=1,2,\dots,m, j=1,2,\dots,n$ . We have the fuzzy relation with membership grade on  $B \times A$  denoted by  $\tilde{R}$  as follows:

$$\tilde{R} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \quad (8)$$

For each  $i=1,2,\dots,m$ , let  $w_{iq} \in [0,1]$  be the weight given to the  $i^{th}$  criterion by the  $q^{th}$  expert, and for each  $q \in \{1,2,\dots,r\}$ ,  $\sum_{i=1}^m w_{iq} = 1$ , then we have  $w_i = \frac{1}{r} \sum_{q=1}^r w_{iq}, i=1,2,\dots,m, w_i \in [0,1], i=1,2,\dots,m$ ,  $\sum_{i=1}^m w_i = 1$  so,  $w_i$  is the average weight of the  $i^{th}$  evaluation criteria  $B_i$ .

Let  $\tilde{W}$  be the fuzzy vector of weight on  $B$  as follows:

$$\tilde{W} = (w_1, w_2, \dots, w_m) = \sum_{i=1}^m \frac{w_i}{B_i} = \frac{w_1}{B_1} + \frac{w_2}{B_2} + \dots + \frac{w_m}{B_m} \quad (9)$$

From (8), (9) and the compositional rule of inference, we have

$$\tilde{W} \circ \tilde{R} = (C_1, C_2, \dots, C_n) = \sum_{j=1}^n \frac{C_j}{A_j} = \frac{C_1}{A_1} + \frac{C_2}{A_2} + \dots + \frac{C_n}{A_n},$$

where  $C_j = \sum_{i=1}^m a_{ij} w_i, j=1,2,\dots,n$ .

If  $C_{i0} = \max_{j \in \{1,2,\dots,n\}} C_j, i \in \{1,2,\dots,n\}$  then  $A_{i0}$  be the best alternative. (10)

Let  $T_C = \sum_{t=1}^n C_t$ . We obtain the evaluation value ratio

$$\frac{C_j}{T_C}, j = 1, 2, \dots, n. \quad (11)$$

This is the usual compositional inference rule method.

### 3.2 Fuzzy Decision Making on Fuzzy Number and Composition Rule of Inference with Statistical Sense

It is difficult to determine the value of point  $a_{ijq}$  with the interval shown in (12).

$$[a_{ijq} - \Delta_{ijq1}, a_{ijq} + \Delta_{ijq2}], \quad (12)$$

where  $0 < a_{ijq} - \Delta_{ijq1} < a_{ijq} + \Delta_{ijq2} \leq 1, 0 < \Delta_{ijq}, p = 1, 2$ .

In corresponding to the interval (12), we have the triangular fuzzy number  $\tilde{a}_{ijq}^*$ , as follows:

$$\tilde{a}_{ijq}^* = (a_{ijq} - \Delta_{ijq1}, a_{ijq}, a_{ijq} + \Delta_{ijq2}) \quad (13)$$

$$0 < a_{ijq} - \Delta_{ijq1} < a_{ijq} + \Delta_{ijq2} \leq 1, 0 < \Delta_{ijq}, p = 1, 2. \quad (14)$$

Defuzzify  $\tilde{a}_{ijq}^*$  using Definition 4(b) we have

$$d(\tilde{a}_{ijq}^*, \tilde{0}) = a_{ijq} + \frac{1}{4}(\Delta_{ijq2} - \Delta_{ijq1}) \in [a_{ijq} - \Delta_{ijq1}, a_{ijq} + \Delta_{ijq2}]$$

$d(\tilde{a}_{ijq}^*, \tilde{0})$  is the estimate score of  $j^{th}$  alternative under the  $i^{th}$  evaluation criterion given by the  $q^{th}$  expert.

Let  $b_{ijq} \in [0, 100]$  denote an effectiveness score of the  $j^{th}$  alternative  $A_j$  under the  $i^{th}$  evaluation criterion  $B_i$  given by the  $q^{th}$  expert and  $\mu_{ij}$  denote an effectiveness score of the  $j^{th}$  alternative under the  $i^{th}$  evaluation criterion in population  $i=1, 2, \dots, m$ ,  $j=1, 2, \dots, n$ .

Since the probability of error in point estimation  $\bar{b}_{ij} = \frac{1}{r} \sum_{q=1}^r b_{ijq}$  and  $\mu_{ij}$  is unknown,

hence  $(1-\alpha) \times 100\%$  confidence interval of  $\mu_{ij}$  is taken into consideration in statistics, as follows:

$$[\bar{b}_{ij} - t_{r-1}(\alpha_1) \frac{s_{ij}}{\sqrt{r}}, \bar{b}_{ij} + t_{r-1}(\alpha_2) \frac{s_{ij}}{\sqrt{r}}], \quad (15)$$

where  $\alpha_1 + \alpha_2 = \alpha, 0 < \alpha < 1, 0 < \alpha_j < 1, j = 1, 2$  and

$$0 < t_{r-1}(\alpha_1) \frac{s_{ij}}{\sqrt{r}} < \bar{b}_{ij}, s_{ij}^2 = \frac{1}{r} \sum_{q=1}^r (b_{ijq} - \bar{b}_{ij})^2, i = 1, 2, \dots, m, j = 1, 2, \dots, n. \quad (16)$$

Let  $T$  be a random variable of the  $t$  distribution with  $r-1$  degrees of freedom, where  $t_{r-1}(\alpha_i)$  satisfies  $P(T \geq t_{r-1}(\alpha_i)) = \alpha_i, i = 1, 2$ . Since the  $(1-\alpha) \times 100\%$  confidence

interval in (15) is an interval, we shall choose a point within the interval to estimate  $\mu_{ij}$ . We find that if we choose a score of  $\bar{b}_{ij}$ , there is no difference between choosing score  $\bar{b}_{ij}$  and the point estimate  $\bar{b}_{ij}$  and the error is definitely 0. We obtain the maximum of confidence level, let it be  $\rho=1-\alpha$ . we have level  $\rho$  fuzzy number (17) corresponding to (15) as follows:

$$\tilde{b}_{ij}^U = (\bar{b}_{ij} - t_{r-1}(\alpha_1) \frac{s_{ij}}{\sqrt{r}}, \bar{b}_{ij}, \bar{b}_{ij} + t_{r-1}(\alpha_2) \frac{s_{ij}}{\sqrt{r}}; \rho), \quad (17)$$

where  $0 < t_{r-1}(\alpha_1) \frac{s_{ij}}{\sqrt{r}} < \bar{b}_{ij}$ ,  $\alpha_1 + \alpha_2 = \alpha$ ,  $\rho = 1 - \alpha$ ,  $i = 1, 2, \dots, m$ ,  $j = 1, 2, \dots, n$ .

Since the membership grade  $\bar{b}_{ij}$  also has an accuracy problem (this means that it would be better if we consider the membership grade at  $\bar{b}_{ij}$  within an interval). We then take the following level  $(\lambda, \rho)$  interval-valued fuzzy number into consideration.

Let  $0 < \alpha < \gamma < 1$ ,  $0 < \alpha_1 + \alpha_2 = \alpha$ ,  $\gamma_1 + \gamma_2 = \gamma$ , we have  $(1-\gamma) \times 100\%$  confidence interval  $[\bar{b}_{ij} - t_{r-1}(\gamma_1) \frac{s_{ij}}{\sqrt{r}}, \bar{b}_{ij} + t_{r-1}(\gamma_2) \frac{s_{ij}}{\sqrt{r}}]$ .

In corresponding to confidence interval above, we have the level  $\lambda$  triangular fuzzy number as

$$\tilde{b}_{ij}^L = (\bar{b}_{ij} - t_{r-1}(\gamma_1) \frac{s_{ij}}{\sqrt{r}}, \bar{b}_{ij}, \bar{b}_{ij} + t_{r-1}(\gamma_2) \frac{s_{ij}}{\sqrt{r}}; \lambda), \quad (18)$$

where  $0 < t_{r-1}(\gamma_1) \frac{s_{ij}}{\sqrt{r}} < \bar{b}_{ij}$ ,  $\gamma_1 + \gamma_2 = \gamma$ ,  $\lambda = 1 - \gamma$ .

Since  $0 < \lambda < \rho < 1$ , using (17) and (18), we have the level  $(\lambda, \rho)$  interval-valued fuzzy number as follows:

$$\tilde{b}_{ij} = [\tilde{b}_{ij}^L, \tilde{b}_{ij}^U], i = 1, 2, \dots, m, j = 1, 2, \dots, n \quad (19)$$

By (19), we obtain the fuzzy relation  $\tilde{H}$  with level  $(\lambda, \rho)$  interval-valued fuzzy number  $\tilde{b}_{ij}$  on  $B \times A$  as follows:

$$\tilde{H} = \begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} & \cdots & \tilde{b}_{1n} \\ \tilde{b}_{21} & \tilde{b}_{22} & \cdots & \tilde{b}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{b}_{m1} & \tilde{b}_{m2} & \cdots & \tilde{b}_{mn} \end{bmatrix} \quad (20)$$

For each  $i = 1, 2, \dots, m$ , let  $w_{iq} \in [0, 1]$  be the weight given to the  $i^{th}$  criterion by the  $q^{th}$  expert, and for each  $q \in \{1, 2, \dots, r\}$ ,  $\sum_{i=1}^m w_{iq} = 1$ , then we have  $w_i = \frac{1}{r} \sum_{q=1}^r w_{iq}$ ,  $i = 1, 2, \dots, m$ ,

$w_i \in [0, 1]$ ,  $i=1,2,\dots,m$ ,  $\sum_{i=1}^m w_i = 1$  so,  $w_i$  will be the average weight of the  $i^{th}$  evaluation criteria  $B_i$ . Let  $\tilde{W}$  be the fuzzy vector of the weight  $B$  in (9), we obtain Theorem 1 as follows:

**Theorem 1.** By considering  $n$  alternatives  $A=\{A_1, A_2, \dots, A_n\}$  under  $m$  evaluation criteria  $B=\{B_1, B_2, \dots, B_m\}$ , if we consider the weighted value of fuzzy number for criterion  $B$  to get the vector of weights in the fuzzy sense  $\tilde{W}=(w_1, w_2, \dots, w_m)$  in (9), and fuzzy relation with  $(\lambda, \rho)$  interval-valued fuzzy numbers on  $B \times A$  is shown as (20), we then have the following result:

(a) Fuzzy decision making based on fuzzy number and compositional rule of inference  $\tilde{W} \circ \tilde{H} = (\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_n)$  and

$$\tilde{h}_j = (w_1 \tilde{b}_{1j} \oplus w_2 \tilde{b}_{2j} \oplus \dots \oplus w_m \tilde{b}_{mj}) = [\tilde{h}_j^L, \tilde{h}_j^U], j=1,2,\dots,n \quad (21)$$

where

$$\begin{aligned} \tilde{h}_j^L &= \left( \frac{1}{r} \sum_{i=1}^m w_i (\bar{b}_{ij} - t_{r-1}(\gamma_1) \frac{s_{ij}}{\sqrt{r}}), \frac{1}{r} \sum_{i=1}^m w_i \bar{b}_{ij}, \frac{1}{r} \sum_{i=1}^m w_i (\bar{b}_{ij} + t_{r-1}(\gamma_2) \frac{s_{ij}}{\sqrt{r}}); \lambda \right) \\ \tilde{h}_j^U &= \left( \frac{1}{r} \sum_{i=1}^m w_i (\bar{b}_{ij} - t_{r-1}(\alpha_1) \frac{s_{ij}}{\sqrt{r}}), \frac{1}{r} \sum_{i=1}^m w_i \bar{b}_{ij}, \frac{1}{r} \sum_{i=1}^m w_i (\bar{b}_{ij} + t_{r-1}(\alpha_2) \frac{s_{ij}}{\sqrt{r}}); \rho \right) \end{aligned}$$

Defuzzify using signed distance method, we have

$$\begin{aligned} h_j^* &\equiv d(\tilde{h}_j, \tilde{0}) \\ &= \frac{1}{r} \sum_{i=1}^m w_i \sum_{q=1}^r \bar{b}_{iq} + \frac{1}{16r} \sum_{i=1}^m w_i \sum_{q=1}^r (t_{r-1}(\gamma_2) - t_{r-1}(\gamma_1)) \frac{s_{ij}}{\sqrt{r}} \\ &\quad + \frac{1}{4r} \sum_{i=1}^m w_i \sum_{q=1}^r (t_{r-1}(\alpha_2) - t_{r-1}(\alpha_1)) \frac{s_{ij}}{\sqrt{r}} + \frac{3\lambda}{16r\rho} \sum_{i=1}^m w_i \sum_{q=1}^r (t_{r-1}(\alpha_2) - t_{r-1}(\alpha_1)) \frac{s_{ij}}{\sqrt{r}} \end{aligned} \quad (22)$$

(b) The order of  $\tilde{h}_{(j)}$ , and  $A_{(j)}$  could be ranked as follows:

$$\tilde{h}_{(1)} \prec \tilde{h}_{(2)} \prec \dots \prec \tilde{h}_{(n)} \Leftrightarrow h_{(1)}^* < h_{(2)}^* < \dots < h_{(n)}^* \Leftrightarrow A_{(1)} \prec A_{(2)} \prec \dots \prec A_{(n)}$$

where  $(1), (2), \dots, (n)$  is a permutation of  $1, 2, \dots, n$ .

(c) If  $\max_{j \in \{1,2,\dots,n\}} \tilde{h}_j = \tilde{h}_{(n)}$  i.e.  $\max_{j \in \{1,2,\dots,n\}} h_j^* = h_{(n)}^*$  holds, then  $A_{(n)}$  is the best alternative.

(d) Evaluation value ratio.

If  $T_h = \sum_{t=1}^n h_t^*$  then the value ratio of evaluation of  $A_j$  is  $\frac{h_j^*}{T_h}$ ,  $j = 1, 2, \dots, n$ .

## 4 Example [12]

The following cases examine the IJV selection process for operating the decision making with five IJVs, four evaluations criteria and evaluation results provided by the four experts. The process is prepared to select the best performance among these five alternatives (IJVs). Let  $A_j, j=1,2,3,4,5$  represents the five IJVs as follows:  $A_j = j^{\text{th}}$  IJV and  $B_i, i=1,2,3,4$  represents the four evaluation criteria as follows:  $B_1$  = Financial indicators.  $B_2$  = Survival or liquidation of venture.  $B_3$  = Duration.  $B_4$  = Instability or ownership change.

**Case 1:** Fuzzy relation with membership grade on  $B \times A$  (in Section 3.1).

Let  $\tilde{W} = (0.2, 0.3, 0.3, 0.2) \cdot A_5$  is the best performance of all IJVs.

**Case 2: statistical sense.**

Let the averages score  $\bar{b}_{ij}$  and variances  $S_{ij}^2$  of alternatives  $A_j, j=1,2,3,4,5$  under evaluation criteria  $B_i, i=1,2,3,4$  and four experts are shown in Table 1.

**Table 1.** The Averages Score  $\bar{b}_{ij}$  and Variances  $S_{ij}^2$

$i$	$\bar{b}_{i1}(s_{i1}^2)$	$\bar{b}_{i2}(s_{i2}^2)$	$\bar{b}_{i3}(s_{i3}^2)$	$\bar{b}_{i4}(s_{i4}^2)$	$\bar{b}_{i5}(s_{i5}^2)$
1	36 (0.985)	56 (1.392)	80 (0.973)	49 (1.456)	50 (1.968)
2	40 (1.356)	49 (0.998)	90 (1.214)	52 (2.012)	55 (1.598)
3	62 (1.148)	61 (2.341)	78 (2.135)	60 (1.956)	58 (0.976)
4	58 (2.354)	60 (2.654)	69 (1.368)	48 (1.568)	52 (1.967)

Let  $\alpha=0.05, \alpha_1=0.028, \alpha_2=0.022, \gamma=0.08, \gamma_1=0.035, \gamma_2=0.045$ . Thus  $\rho=1-\alpha=1-0.05=0.95$ ,  $\lambda=1-\gamma=1-0.08=0.92$ . Let  $\tilde{W}=(0.2, 0.3, 0.3, 0.2)$ . By  $t$  distribution with degree of freedom  $r-1=3$  we have the following:

$t_3(\alpha_1)=t_3(0.028)=3.0825, t_3(\alpha_2)=t_3(0.022)=3.4538, t_3(\gamma_1)=t_3(0.035)=2.8504, t_3(\gamma_2)=t_3(0.045)=2.5188$ .

**Table 2.** The Estimate Score of  $A_j, j=1,2,3,4,5$  by Signed Distance

IJV	$A_1$	$A_2$	$A_3$	$A_4$	$A_5$
Estimate score $h_j^*$	49.08219	56.59212	79.33144	52.34104	53.83727
rank	5	2	1	4	3
$h_j^*/T_h$	0.168561	0.194352	2.72444	0.179752	0.184891

By Table 1,  $r=4$ , and (16)~(18), we have  $\tilde{b}_{ij} = [\tilde{b}_{ij}^L, \tilde{b}_{ij}^U]$  in (19),  $i=1,2,3,4, j=1,2,3,4$ . By Theorem 1, we have estimate score of  $A_j, j=1,2,3,4,5$  by signed distance as shown in Table 2.

From Table 2, we obtain A3 is the best performance of all IJVs.

## 5 Conclusion

In case 1, applying the usual compositional inference rule method to estimate, we have that  $A_5$  is the best performance of the IJVs. In case 2, by the statistical sense, we have that  $A_3$  is the best performance of the IJVs.

According to the demand in real case, the decision-maker can change the weight and select the suitable confidence-interval to estimate.

## References

1. Kaufmann, A., Gupta, M.M.: *Introduction to Fuzzy Arithmetic Theory and Applications*. Van Nostrand Reinhold, New York (1991)
2. Dang, T.: *Ownership, Control and Performance of the Multinational Corporation. A study of US Wholly-owned Subsidiaries and Joint Ventures in the Philippines and Taiwan*. Unpublished Doctoral Dissertation, University of California at Los Angeles, 1-28 (1977)
3. Franco, L.: *Joint venture Survival in Multinational Corporations*. Praeger, New York (1971)
4. Glaister, K.W., Buckley, P.J.: *Performance Relationships in UK International Alliances*. Management International Review 39(2), 123–147 (1999)
5. Gomes-Casseres, B.: *Joint Venture Instability: Is it a problem?* Columbia Journal of World Business 22(2), 97–102 (1987)
6. Good, L.: *US Joint Ventures and Manufacturing Firms in Monterey, Mexico: Comparative styles of management*. Unpublished doctoral dissertation, Cornell University (1972)
7. Harrigan, K.R.: *Strategic Alliances and Partner Asymmetries*. In: Contractor, F., Lorange, P. (eds.) *Cooperative Strategies in International Business*. Lexington, New York (1988)
8. Yao, J.-S., Wu, K.-M.: *Ranking Fuzzy Numbers Based on Decomposition Principle and Signed Distance*. Fuzzy Sets and Systems 116, 275–288 (2000)
9. Killing, P.: *Strategies for Joint Venture Success*. Praeger, New York (1983)
10. Kogut, B.: *Joint Ventures: Theoretical and Empirical Perspective*. Strategic Management Journal 9, 319–332 (1988)
11. Gorzalezany, M.B.: *A Method of Inference in Approximate Reasoning Based on Interval – Valued Fuzzy Sets*. Fuzzy Sets and Systems 21, 1–17 (1987)
12. Shieh, T.-S., Su, J.-S., Lee, H.-M.: *Fuzzy Decision Making Performance Evaluation for International Joint Venture*. ICIC Express Letters 3(4B), 1197–1202 (2009)
13. Zimmermann, H.-J.: *Fuzzy Set Theory and Its Applications*, 4th edn. Kluwer Academic Publishers, Boston (2001)

# Designing a Learning System Based on Voice Mail – A Case Study of English Oral Training

Huey-Ming Lee and Chien-Hsien Huang

Department of Information Management, Chinese Culture University  
55, Hwa-Kang Road, Yang-Ming-San,  
Taipei, Taiwan  
[hmlee@faculty.pccu.edu.tw](mailto:hmlee@faculty.pccu.edu.tw), [iamkan@gmail.com](mailto:iamkan@gmail.com)

**Abstract.** Voice mail use mainly in mobile phone and PBX, but similar voice record device had already used in assist teaching. Whether the voice mail also suitable to use in the assist teaching? About the voice mail's research, most in the distributed processing or the message's content comparison that never have research in assist teaching. In this study we use voice mail to developed a learning system, Through the system, the learner had chance to spend several minutes to practice English oral ability every day. This study may achieve the effectiveness of decreasing the learner's anxiety about the English oral practice, and further help the learner learning.

**Keywords:** Learning system, voice Mail, oral training.

## 1 Introduction

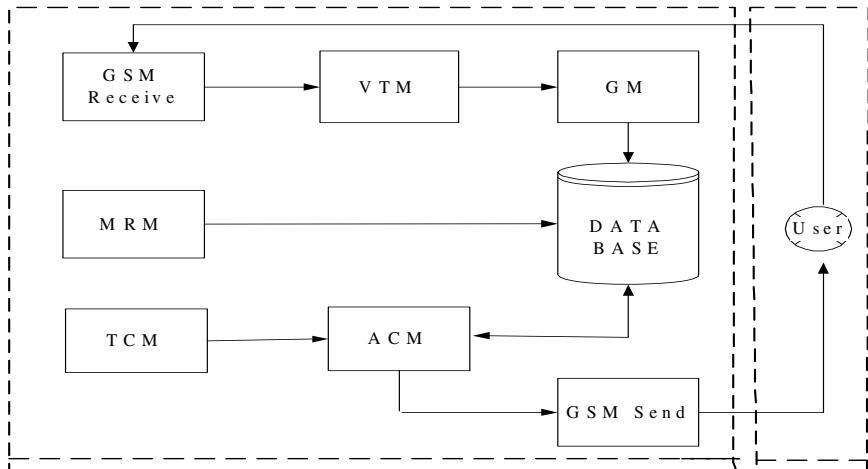
M-learning almost use on PDA or smart phone. But those device are not popular than mobile phone. Therefore how effect is an import issue that learning used voice mail of mobile phone.

There are some studies for voice mail as follows: Hobbs [2] proposed the medium is the message: politeness strategies in men's and women's voice mail messages. Brown *et al* [3] Proposed Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval. Dillman *et al* [4] proposed response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet.

Therefore in this study, we based on the problem-posing approach, action learning and interactive voice response to build up an oral training system. Via this system, we can transfer the oral material into voice mail and store the reply message. When the reply message transferred to text, and then could automatically produce degree via our judgment system. By the training system of the voice mail, our oral training will become more efficient.

## 2 Problem Statement and Preliminaries

In this section, we present an oral reaction training system based on fuzzy inference, as shown in Figure 1.



**Fig. 1.** The system architecture of oral training system

There are five modules in this system, namely, material record module (MRM), access control module (ACM), voice transfer module (VTM), grade module (GM), and time control module (TCM).

The functions of these five modules are as follows:

- (1) Teaching material record module (MRM) - MRM functions recording the teaching material.
- (2) Access control module (ACM) - ACM functions transferring the teaching material record into the voice mail and send a short message to learner.
- (3) Voice transfer module (VTM) - VTM functions transferring the replying voice record to be text. Teacher depends on rules to inference the elementary vocabulary.
- (4) Time control module (TCM) - TCM functions recording time of transfer teaching material and time sent by short message service (SMS).
- (5) Grade module (GM) - GM functions determining the situations of vocabulary by fuzzy inferences, and grades the text through the scoring system and records in the data.

## 3 System Judging

The Oral exam adjusts standards based on fluency, comprehensibility and accuracy [11]. In this study, we focus on the accuracy of content as follows:

(1) Accuracy and appropriateness of content: if talking issues are closely interrelated to the content then it gets the point.

(2) Scoring standard: the closer interrelation between talking issues.

(3) Talking issues relation degree constructing method: In accordance with the database of the elementary talking issues of GEPT (General English Proficiency Test, Taiwan), expert who is professional in this field sets up the relation dimension between talking issues, such as, the relation degree of family with members of family is 1, but it will be different from the perception by different experts.

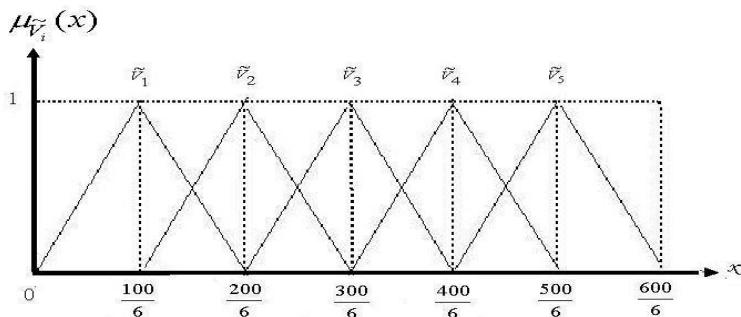
So the system adopts the fuzzy environment to run out the reasonable and more approximate to the actual result. Fuzzy set theory was introduced by Zadeh [12] to deal with problem in which vagueness is present, linguistic value can be used for approximate reasoning within the framework of fuzzy set theory [13] to effectively handle the ambiguity involved in the data evaluation and the vague property of linguistic expression, and normal triangular fuzzy numbers are used to characterize the fuzzy values of quantitative data and linguistic terms used in approximate reasoning.

With regard to fuzzy decision-making problem, Lee [1] applied fuzzy set theory to evaluate the aggregative risk in software development under fuzzy circumstances. Lin and Lee [5-7] presented facility site selection model using fuzzy set theory. Lin and Lee [8] presented a new fuzzy algorithm to evaluate the user satisfaction of software quality. Lin and Lee [9] presented the fuzzy assessment on sampling survey analysis. Lin and Lee [10] presented the two algorithms with the linear fuzzy linguistic for the group assessment.

The criteria ratings of relations between talking issues are linguistic variables with linguistic values  $V_1$ ,  $V_2$ , ...,  $V_5$ , where  $V_1$  = very high,  $V_2$  = high,  $V_3$  = middle,  $V_4$  = low,  $V_5$  = very low. The triangular fuzzy number representations of the linguistic values are shown in Table 1. The membership functions of the set of criteria rating of relations are shown in Figure 2. The system sets up the relation degree from 0 to 1, and establishing five rules of fuzzy inferences are shown in Table 2.

**Table 1.** Triangular fuzzy numbers of the criteria of relations

Rating of relation	Triangular fuzzy number
$V_1$ : very high	$\tilde{V}_1 = (0, \frac{100}{6}, \frac{200}{6})$
$V_2$ : high	$\tilde{V}_2 = (\frac{100}{6}, \frac{200}{6}, \frac{300}{6})$
$V_3$ : middle	$\tilde{V}_3 = (\frac{200}{6}, \frac{300}{6}, \frac{400}{6})$
$V_4$ : low	$\tilde{V}_4 = (\frac{300}{6}, \frac{400}{6}, \frac{500}{6})$
$V_5$ : very low	$\tilde{V}_5 = (\frac{400}{6}, \frac{500}{6}, \frac{600}{6})$



**Fig. 2.** Membership functions of the set of the criteria rating of relations

**Table 2.** Rules of inference

Relation	Relation degree
1 The relation of talking issues is very high	1
2 The relation of talking issues is high	0.75
3 The relation of talking issues is middle	0.5
4 The relation of talking issues is low	0.25
5 The relation of talking issues is very low	0

## 4 System Implementation

To assess the learning performance of the proposed oral reaction training system, this study recruited twenty-one freshmen that were majoring in the Department of physical Education at Chinese Culture University. When they entranced into the university, they had do an English examination to identify the ability of English, so we can make sure that all of they had failed in the elementary of GEPT.

In those thirty experiment days, we change the training material, which just spend thirty seconds, every day, and send a message to inform the statement of judgment to those students and to remind those students to listen the next training material. The teacher can listen the student's answer from his computer. The student can use his mobile phone to receive the SMS of system's judgment. When students receive the SMS, he will know how about his answer of the yesterday's question, and prepare to call the system's assigned phone number to listen the new question today.

The experiment's purpose is to verify whether the system's qualification and satisfaction and the student's anxiety of English speaking have been improved or not. So we will do oral anxiety pretest before the experiment. After we finished the experiment, we did a survey about the system's qualification and satisfaction and student's anxiety of English speaking.

According to the elementary speaking material of GEPT, we recorded thirty seconds voice file for the training material, the reference showed in Table3.

**Table 3.** The example teaching material

Teaching Material	Example
<1>	We all know one week has seven days. There are Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. Please answer below question. What day is today? What day is today?

After we finished the experiment, we did a survey. the reference is shown in Table 4-Table 6.

**Table 4.** The descriptions of question types

Question type	n	Description
System operation	10	Questions related to the user interface and the content of learning materials
Learning attitude	10	To investigate whether the system can enhance learners' learning motivation or Interests and promote their learning achievements or not

(n: The Question of numbers ).

**Table 5.** The satisfaction evaluation results of questionnaire about the operation of system

Question	SA	A	NO	D	SD
1. It is easy for me to learn how to use this system	14.2	57.1	23.8	4.7	0
2. It is easy for me to proficient at use this system	9.52	61.9	23.8	4.7	0
3. I think use this system will increase my learning of speaking English efficiency and score	9.5	42.8	33.3	9.5	4.7
4. I think the system's operation is simple and efficiency	9.5	57.1	28.5	4.7	0
5. I can't understand some parts of the operation in the system	0	9.5	28.5	47.6	14.2
6. I think the system's operation is not so easy	0	4.76	23.8	61.9	9.5
7. I think the system is stability	4.7	38.1	38.1	14.2	4.7
8. I think the system cannot improve my learning of speaking English	4.7	9.5	42.8	33.3	9.5
9. I think the system's reaction is too slow	4.7	19.0	38.1	33.3	4.7
10. When I use the system, I can't use it normally every time	4.7	19.0	33.3	33.3	9.5

(SA: Strongly agreed, A: Agreed, NO: No opinion, D: Disagreed, SD: Strongly Disagreed).

The results by statistics are pointed that the operation of this system is easy for the user.

**Table 6.** The satisfaction evaluation results of questionnaire about the learning attitude of the system

Question	SA	A	NO	D	SD
1. I satisfied with this oral reaction training system	9.5	38.1	42.8	4.7	4.7
2. I satisfied with the environment of this training system	14.2	33.3	42.8	4.7	4.7
3. For whole parts, I satisfied with the learning efficiency that offered by this system	4.7	23.8	47.6	19.0	4.7
4. I dissatisfied with the environment of this training system	4.7	9.5	47.6	28.5	9.5
5. I dissatisfied with this oral reaction training system	4.7	14	38.1	38.1	4.7
6. For whole parts, I have an higher willing to use this system for improvement my speaking English ability	4.2	38.1	38.1	4.7	4.7
7. I will select this system for my oral training in the future	9.5	33.3	42.8	9.5	4.7
8. For whole parts, I have no willing to use this system	4.7	14.2	33.3	38.1	9.5
9. For whole parts, I dissatisfied with the learning efficiency	9.5	23.8	28.5	28.5	9.5
10. Even in the future I will do not to select this system for my oral training	4.7	14.2	38.1	28.5	14.2

(SA: Strongly agreed, A: Agreed, NO: No opinion, D: Disagreed, SD: Strongly Disagreed).

The results by statistics are pointed this system is helpful in the oral training.

## 5 Conclusions

This study presents that the oral training system satisfied the Elementary level student training. The student can do more training after school. It satisfied in English course frequently request of the speaking and listening. So the oral training system can be as auxiliary teaching materials for English course. It proved the voice mail can be as assist teaching materials.

**Acknowledgments.** The authors gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

## References

1. Lee, H.-M.: Applying Fuzzy Set Theory to Evaluate the Rate of Aggregated Risk in Software Development. *Fuzzy Set and Systems* 79, 323–336 (1996)
2. Hobbs, P.: The medium is the message: politeness strategies in men's and women's voice mail messages. *Journal of Pragmatics* 35, 243–262 (2003)

3. Brown, M.G., Foote, J.T., Jones, G.J.F., Jones, K.S., Young, S.J.: Open-Vocabulary Speech Indexing for Voice and Video Mail Retrieval. *Readings in Multimedia Computing and Networking*, 237–246 (2002)
4. Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., Messer, B.L.: Response Rate and Measurement Differences in Mixed-mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet. *Social Science Research* 38, 1–18 (2009)
5. Lee, H.-M., Lin, L.: Fuzzy Facility Site Selection Model Based on Signed Distance Method. *International Journal of Innovative Computing, Information and Control* 5(6), 1505–1514 (2009)
6. Lin, L., Lee, H.-M.: A Fuzzy Decision Support System for Facility Site Selection of Multinational Enterprises. *International Journal of Innovative Computing, Information and Control* 3(1), 151–162 (2007)
7. Lin, L., Lee, H.-M.: A New Assessment Model for Global Facility Site Selection. *International Journal of Innovative Computing, Information and Control* 4(5), 1141–1150 (2008)
8. Lin, L., Lee, H.-M.: A Fuzzy Software Quality Assessment Model to Evaluate User Satisfaction. *International Journal of Innovative Computing, Information and Control* 4(10), 2639–2647 (2008)
9. Lin, L., Lee, H.-M.: Fuzzy Assessment Method on Sampling Survey Analysis. *Expert Systems with Applications* 36(3), 5955–5961 (2009)
10. Lin, L., Lee, H.-M.: Group Assessment Methods Based on Two Algorithms of the Linear Fuzzy Linguistic. *International Journal of Innovative Computing, Information and Control* 6(1), 263–274 (2010)
11. Heaton, J.B.: Writing English Language tests. Longman, New York (1988)
12. Zadeh, L.A.: Fuzzy sets. *Information and Control* 8, 338–353 (1965)
13. Zadeh, L.A.: The Concept of a Linguistic Variable and Its Application to Approximate Reasoning. *Information Sciences* 8, 199–249 (1975); (I), 301–357 (II), 9, 43–58 (III) (1976)
14. Zimmermann, H.-J.: Fuzzy Set Theory and Its Applications. Second Revised edn. Kluwer Academic Publishers, Boston (1991)

# A Gene Selection Method for Microarray Data Based on Sampling

Yungho Leu, Chien-Pang Lee, and Hui-Yi Tsai

Department of Information Management, National Taiwan University of Science and Technology, 43, Keelung Road, Section 4, Taipei, Taiwan 10607, ROC  
yhl@cs.ntust.edu.tw, {D9509302,M9709119}@mail.ntust.edu.tw

**Abstract.** Microarray technology has become an important tool for biologists in recent years. It can obtain the expressions of a large amount of genes in a single experiment. One of the research issues of microarray is to select a set of relevant genes from a large number of genes to assist clinical diagnosis. In this paper, we propose a method for gene selection in microarray data. In the proposed method, we first classify genes into three different groups of genes according to their expressions in the microarray experiment. Then, we use probability sampling method to generate several candidate subsets of genes. Finally, we use  $\chi^2$ -test for homogeneity to select the relevant genes. The experiment results show that the proposed method is better than the other methods in terms of classification accuracy and the number of genes selected.

**Keywords:** Gene selection, Microarray data, Probability sampling,  $\chi^2$ -test for homogeneity.

## 1 Introduction

Recently, microarray technology has become a popular technique for bioinformatics and clinical diagnosis. It allows researchers to measure the expression levels of thousands of genes simultaneously in a microarray experiment. Microarray data usually contain thousands of genes (sometimes more than 10,000 genes) and a small number of samples (usually less than 100 samples). Although thousands of genes are experimented simultaneously, most of them are irrelevant or insignificant to a clinical diagnosis [1]. Therefore, it is important to find a set of genes that are relevant to a diagnosis. Data mining, machine learning and statistical methods have been widely used in finding the relevant genes. For example, de Haan et al. [2] used principal components analysis and Hotelling's T-square test to select genes on microarray data. Wang et al. [3] used a feature selection method to rank the importance of genes and selected a set of top-ranked genes as the relevant genes. Cho et al. [4] used a modified kernel Fisher discriminant analysis (KFDA) to analyze the breast cancer dataset [5]. The authors used the mean square error (MSE) as the gene selection criterion to assist KFDA classifier in gene selection. Li et al. [6] proposed a hybrid method, termed GA/KNN, to analyze the colon dataset [7]. In [6], a genetic algorithm is used to generate large number of genes and then the k-nearest neighbor classifier (KNN) is used

to filter the genes according to classification accuracy. In [8], Lee and Leu used a genetic algorithm and  $\chi^2$ -test for homogeneity to select relevant genes.

This paper proposes a simple method to find relevant genes. While the method is simple, its performance outperforms other complicated methods in the sense that it selects fewer genes with higher classification accuracy. The proposed method consists of three stages. In the first stage, we classify the genes into three different groups according to their levels of expressions in the microarray experiment. In the second stage, we use a probability sampling method to generate many candidate gene sets. Finally, in the third stage, we use  $\chi^2$ -test for homogeneity to determine the set of the relevant genes.

The remainder of this paper is organized as follows. Section 2 describes the microarray dataset. Section 3 introduces the details of the proposed method. Section 4 reports the experimental results and Section 5 concludes this paper.

## 2 Microarray Dataset

A microarray data is commonly represented as an  $N \times M$  matrix, where  $N$  is the number of samples and  $M$  is the number of genes in the experiment. Each cell in the matrix, as shown in Fig. 1, is the level of expression of a specific gene in a specific experiment.

Three different microarray experiments are temporal, duplicate and perturbation [9]. In this paper, we use three duplicate microarray datasets to test the proposed method. The first dataset is the colon dataset [7], which contains 62 tissue samples, including 22 normal tissue samples and 40 tumor tissue samples, with 2,000 genes. The second dataset is the acute lymphoblastic leukemia (ALL)/ acute myeloid leukemia (AML) dataset [10], which contains 72 samples with 7,129 genes. The Third dataset is the Lymphoma dataset [11], which contains 47 samples with 4,026 genes. A description of these datasets is shown in Table 1.

**Table 1.** Microarray datasets

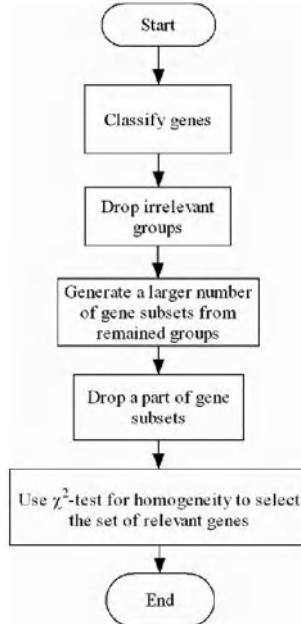
Dataset	Samples	Genes	Classes	Reference
Colon	62	2,000	2	Alon et al. [7]
Leukemia	72	7,129	2	Golub et al. [10]
Lymphoma	47	4,026	2	Alizadeh et al. [11]

$$\begin{matrix}
 & & & \overbrace{M \text{ genes}} \\
 & & g_{11} & g_{12} & \cdots & g_{1M} \\
 N \text{ samples} \left\{ \begin{matrix} g_{21} & g_{22} & \cdots & g_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ g_{N1} & g_{N2} & \cdots & g_{NM} \end{matrix} \right.
 \end{matrix}$$

**Fig. 1.** A matrix representation of a microarray dataset

### 3 The Proposed Method

Our method comprises three stages, which are described in the following. The flowchart of the proposed method is shown in Fig. 2.



**Fig. 2.** The flowchart of our proposed method

#### 3.1 The First Stage

##### *Step 1:* classify genes

Each gene in a 2-class microarray data can be classified into three different groups according to their expression levels as shown in Table2. In Table 2, the symbols “ $\mu_1$ ” and “ $\mu_2$ ” denote the average expression levels of a gene in class 1 samples and class 2 samples, respectively. For each gene in the microarray data, we perform test on the difference of the two means. Then, we classify the gene into Group 1, group 2, or Group 3, if  $\mu_1 > \mu_2$ ,  $\mu_1 < \mu_2$ , or  $\mu_1 = \mu_2$ , respectively.

**Table 2.** Groups of gene expression for a 2-class microarray dataset

Group 1	Group 2	Group 3
$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	$\mu_1 = \mu_2$

##### *Step 2:* drop the useless group

A relevant gene should have good discernment for clinical diagnosis in the sense that  $\mu_1 \neq \mu_2$  [12]. Therefore, we drop Group 3 in further analysis.

### 3.2 The Second Stage

**Step 3:** generate gene subsets by sampling

Having determined the two groups of genes, we then generate many gene subsets, each containing 10 different genes, by sampling the genes in the two groups. To select a gene subset, we associate with each group a probability. Then, we randomly select a group according to its associated probability. That is, the probability a group is selected is in proportion to its associated probability. Having selected a group, we perform sampling without replacement to choose one gene from the group into the gene subset. Repeat the process for 10 times to generate a gene subset with 10 genes. Having generated one subset of genes, the selected genes are put back to their original groups, and we start all over again to generate another subset of genes. The probability associated with each group is determined by the capability of a gene in the group in distinguishing between two the classes of samples in the microarray dataset. Probability  $P_1$  and  $P_2$  of group 1 and group 2, respectively, are defined as following.

$$t_i = \frac{\bar{x}_{i1} - \bar{x}_{i2}}{\sqrt{\frac{s_{i1}^2}{n_1} + \frac{s_{i2}^2}{n_2}}}, \text{ for gene } i \text{ in a group,} \quad (1)$$

$$T_1 = \sum_{i=1}^n \frac{|t_i|}{n}, \text{ for all gene } i \text{ in group 1,} \quad (2)$$

$$T_2 = \sum_{i=1}^m \frac{|t_i|}{m}, \text{ for all gene } i \text{ in group 2,} \quad (3)$$

$$P_1 = \frac{T_1}{T_1 + T_2}, \quad (4)$$

$$P_2 = \frac{T_2}{T_1 + T_2}, \quad (5)$$

where  $\bar{x}_{i1}$  and  $\bar{x}_{i2}$  in (1) denotes the average expression levels of gene  $i$  in the samples belonging to class 1 and class 2, respectively.  $S_{i1}^2$  and  $S_{i2}^2$  denote the variances of expression levels of gene  $i$  in the samples belonging to class 1 and class 2, respectively.  $n_1$  and  $n_2$  are numbers of samples belonging to class 1 and class 2, respectively.  $t_i$  is the  $t$ -value of gene  $i$ . In (2) and (3),  $n$  and  $m$  are the numbers of genes in group 1 and group 2, respectively.  $T_1$  and  $T_2$  are denoted the average absolute  $t$ -value belonging to Group 1 and Group 2, respectively.  $P_1$  and  $P_2$  in (4) and (5) denote the sampling probabilities of Group 1 and Group 2, respectively.

Note that in this stage, we generate 100,000 gene subsets for further analysis.

**Step 4:** Drop less discriminating gene subsets

Having derived a large number of gene subsets, we drop the gene subsets with less discrimination capability from the set of all generated gene subsets. The discrimination capability of a gene subset is determined by its classification accuracy. To test the

classification accuracy of a gene subset, we choose to use the KNN method to perform classification on the microarray dataset using the expression levels of the genes in the subset as input. In this paper, the gene subsets with more than one error in classification are dropped. The remained gene subsets are called candidate gene subsets.

### 3.3 The Third Stage

#### **Step 5:** Select relevant genes

We sort the genes in the candidate gene subsets by their frequencies of occurrence in all the candidate gene subsets. The sorted genes are stored in an ordered list. Then, from the head of the list, we apply  $\chi^2$ -test for homogeneity [8] on two consecutive genes in the list to test their difference in occurrence frequencies. Subsequently, we find a set of relevant genes by choosing the genes from the beginning of the list up to the genes in the list after which the  $\chi^2$ -tests shows insignificant difference in occurrence frequencies. In some microarray dataset, it is possible that some of the irrelevant genes have significant difference in the frequency of occurrence in the set of selected candidate gene subsets. In this case, our proposed method might choose too many genes. To prevent this, before we perform the  $\chi^2$ -test, we drop the genes whose frequencies are below the average frequency of all the genes in the ordered list from the ordered list. Table 3 shows a  $2 \times 2$  contingency table for  $\chi^2$ -test for homogeneity. The  $\chi^2$ -test for homogeneity is computed according to equation (6).

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{n(ad - bc)^2}{(a+c)(b+d)(c+d)(a+b)}, \quad (6)$$

where  $O_{ij}$  denotes the observation of  $i$ th row and  $j$ th column;  $E_{ij}$  denotes the expected value of the  $i$ th row and  $j$ th column;  $a, b, c, d, n$  are defined in Table 3.

**Table 3.** The  $2 \times 2$  contingency table of the  $\chi^2$ -test for homogeneity

	# appeared in candidate gene subsets	# not appeared in candidate gene subsets	Total
$i^{th}$ -ranked gene	$a$	$b$	$a+b$
$(i+1)^{th}$ -ranked gene	$c$	$d$	$c+d$
Total	$a+c$	$b+d$	$n$

## 4 Results and Performance

In this section, we report the results and performance of our proposed method. The significant level of  $\chi^2$ -test for homogeneity is set to 0.05 for all the experiments. Fig. 3 shows the relationship between the classification accuracy and the number of selected genes for the Lymphoma dataset. It shows that, by using our proposed method, the classification accuracy increases as the number of selected gene increases. However, as the number of selected gene greater than 11, the classification accuracy begins to decrease. This shows that to select too many genes has a negative effect on the classification accuracy. Therefore, it is better to select just enough genes for analysis. Table 4

reprots the classification accuracy of our proposed method and the other existing methods. In Table 4, the numbers in parentheses denote the numbers of the selected genes. Note that for our proposed method, we use the SVM for classification.

Table 4 shows that for the Leukemia dataset, the methods of Huerta et al. [13] and Cho et al. [4] both attain 100 percent classification accuracy; however, they both selected too many genes. In contrast, our proposed method attains 97.06 percent accuracy with only 6 selected genes. For the colon dataset, the method of Li et al. [6] achieves 100 percent accuracy with 50 genes. In comparison, our proposed method achieves 97.30 percent accuracy with only 10 genes. For the Lymphoma dataset, our proposed method achieves 100 percent accuracy with only 9 genes, while the other methods chose many more genes to attain less accuracy. In general, our proposed method performs equally well for all the three dataset, while the other methods are not.



**Fig. 3.** The classification accuracy with top  $n$ -ranked selected genes of Lymphoma

**Table 4.** The classification accuracy (%)

Dataset	Method				
	Huerta et al. [13]	Paul et al. [14]	Li et al. [6]	Cho et al. [4]	Proposed method
Leukemia	100 (100)	90 (4)		100 (21)	97.06 (6)
	90.3 (100)	78 (8)	100 (50)	82.08 (10)	97.30 (10)
Colon	93.7 (100)	91 (10)	84.62 (50)		100 (9)
Lymphoma					

## 5 Conclusion

Microarray experiment has become a popular technique for bioinformatics and clinical diagnosis in recent years. This paper proposes a sampling-based gene selection method for gene selection. The experimental results show that the proposed method achieves high classification accuracy with few selected genes. Hence, the proposed method offers a useful alternative for gene selection in microarray data.

## References

- Zhou, X., Tuck, D.P.: MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics* 23, 1106–1114 (2007)
- de Haan, J.R., Wehrens, R., Bauerschmidt, S., Piek, E., Schaik, R.C.v., Buydens, L.M.C.: Interpretation of ANOVA models for microarray data using PCA. *Bioinformatics* 23, 184–190 (2007)
- Wang, L., Chu, F., Xie, W.: Accurate Cancer Classification Using Expressions of Very Few Genes. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 4, 40–53 (2007)
- Cho, J.-H., Lee, D., Park, J.H., Lee, I.-B.: Gene selection and classification from microarray data using kernel machine. *FEBS Lett.* 571, 93–98 (2004)
- Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, L.N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A., Trent, J.: Gene-Expression Profiles in Hereditary Breast Cancer. *N. Engl. J. Med.* 344, 539–548 (2001)
- Li, L., Darden, T.A., Weingberg, C.R., Levine, A.J., Pedersen, L.G.: Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/k-nearest Neighbor Method. *Comb. Chem. High Throughput Screen* 4, 727–739 (2001)
- Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6745–6750 (1999)
- Lee, C.-P., Leu, Y.: A novel hybrid feature selection method for microarray data analysis. *Appl. Soft. Comput.* 11, 208–213 (2011)
- McIntosh, T., Chawla, S.: High Confidence Rule Mining for Microarray Analysis. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 4, 611–623 (2007)
- Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286, 531–537 (1999)
- Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511 (2000)
- Thanyaluk, J.-U., Stuart, A.: Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics* 6, 148 (2005)
- Huerta, E., Duval, B., Hao, J.-K.: Fuzzy Logic for Elimination of Redundant Information of Microarray Data. *Genomics, Proteomics & Bioinformatics* 6, 61–73 (2008)
- Paul, T.K., Iba, H.: Selection of the most useful subset of genes for gene expression-based classification. In: Proc. IEEE Congr. Evolut. Comput., pp. 2076–2083 (2004)

# The Similarity of Video Based on the Association Graph Construction of Video Objects

Ping Yu

Department of Information Management,  
Chinese Culture University, Taipei, Taiwan  
yp@faculty.pccu.edu.tw

**Abstract.** We have proposed 9DST approach to represent the spatial-temporal relations between objects in a symbolic video and the similarity of videos is relevant to the users' requests. In this paper, based on the 9DST approach, we proposed the similarity retrieval algorithm. First, we construct the 9DST index structure, from the 9DST-strings, which contains the spatial-temporal relations for each pair of objects in a video database. Second, we use the similar pairs to define various types of similarity measures and construct the association graph to calculate the similarity between videos. By providing different level types of similarity between videos, our proposed similarity retrieval algorithm has discrimination power about different criteria.

**Keywords:** Video similarity, association graph, 3D Z-string, 9DST.

## 1 Introduction

Recently, the video database is gaining an increasing interest because of its expressive power. The effective video retrieval and summarization method an urgent need [1]. There are many researches in various kinds of database management systems for managing information from videos. For example, KMED [2], QBIC [3], VideoQ [4], etc. A number of techniques for video content modeling involving temporal events also have been proposed. Some of these techniques rely on modeling the interplay among objects over time along with spatial relations between these objects [5]. Hu et al. [6] proposed a cluster-based tracking algorithm to acquire motion trajectories and cluster hierarchically. From the learning activity model, they construct a hierarchical semantic for indexing and retrieving the objects' activities. The PC-FSM model uses finite automata to analysis video and generates personalized highlights of sport events [7]. For the semantic gap between what we can derive automatically from the visual data and the semantic interpretation, are discussed in [8].

From the literature, we can see to retrieve desired videos from a video database; one of the most important methods for discriminating videos is the perception of the objects and the spatial-temporal relations. To represent the spatial and temporal relations between the objects in a video, many iconic indexing approaches, extended from the notion of 2D string [9] have been proposed. For example, 2D B-string, 2D C-Tree, 9DLT strings, 3D-list, 3D C-string, and 3D Z-string [10]. In the 3D Z-string [10],

extended from the 2D Z-string [11], used the projections of objects to represent spatial and temporal relations between the objects in a video, and to keep track of the motions and size changes of the objects in a video. In this paper, we extend the idea behind the similarity retrieval of images in 9D-SPA[13] to 9DST approach. First, we construct the spatial relation sequence and temporal relations for each pair of objects of video database in the 9DST index structure. Second, we define different level of types to measure the similarity between videos and propose the similarity retrieval algorithm. Many criteria can be used to define similarity measures. The different types of similarity can assign different multi-granularity to meet users' need. By providing various types of similarity between videos, our proposed similarity retrieval algorithm has discrimination power about different criteria. It also can be easily applied to an intelligent video database management system to query spatial-temporal relations between the objects and to retrieve the videos similar to a query video from a video database.

The remainder of this paper is organized as follows. We first brief review the 9DST approach of representing symbolic videos in section 2. In Section 3, the 9DST index structure is proposed which can reduce search space. We discuss the similarity retrieval algorithm based on association graph between videos in Section 4. Finally, concluding remarks are made in Section 5.

## 2 9DLT Video Model

In the 9DST approach, we use the projections of objects to represent the spatial-temporal relations between the objects in a video. We propose two structures, VO-string (Video Object-string) and 9DST-string (9 Direction Spatial-Temporal string), to record the spatial-temporal information of objects and relations of object pairs respectively. In the VO-string, for each object, we keep track of the initial location and size of the object by a minimum bounding rectangle (MBR) whose sides are parallel to the x- or y-axes. We also record the information about their motions and size changes. The structure of VO-string is brief as following.

**Definition 1:** The VO-string is a 3-tuple  $(O, A, R_t)$  where

1.  $O$  is a set of objects in a video
2.  $A$  is a set of attributes to describe the objects in  $O$ , including the number of start frame and end frame that the object appears in the video, initial location on the projection on x- (y-) axis, and size (length and width);
3.  $R_t = \{\#, \uparrow, \downarrow\}$ , where operator “#” is the interval operator to denote the interval of a motion/size change. Operator “ $\uparrow$ ” and “ $\downarrow$ ” are the motion operators to denote the direction of the motion of an object. Operator “ $\uparrow$ ” (“ $\downarrow$ ”) denotes that the object moves along the positive (negative) direction of the x- (or y-) axis.

Those metric measures are listed as follows.

1. The size (length and width) or interval of an object: the size or interval of its x- (y-) or time projection is equal to  $s$ , where  $s = \text{End}(P) - \text{Begin}(P)$ , where the  $\text{Begin}(P)$  and  $\text{End}(P)$  are the begin-bound and end-bound or begin-frame and end-frame of the projection of object  $P$  in x- (y-) or time axis, respectively.

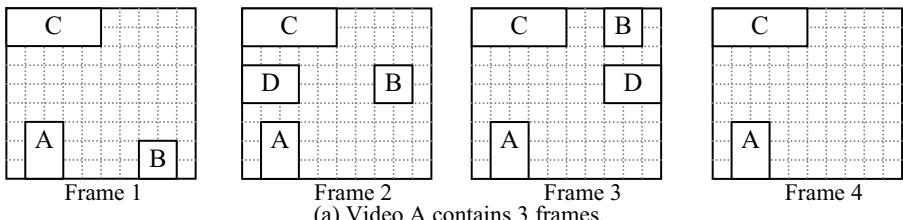
2. The velocity and rate of size change associated with motion operators  $\uparrow_{v,r}$  and  $\downarrow_{v,r}$ : Operators  $\uparrow_{v,r}$  and  $\downarrow_{v,r}$  have two subscripts (fields). “ $v$ ” is the velocity of the motion and “ $r$ ” is the rate of size change of the object. For example, an VO-string:  $\uparrow_{2,1}$  denotes that object with the velocity = 2 and the rate of size change =1. That is, the velocity of object is moving along the positive direction in 2 units/ frame and the size remains unchanged.
3. The interval associated with operator #:  $\#_i$  denotes that the interval length associated with the motion/size change is equal to  $i$ . If an object has not motion or size changing, there is not the interval operator associated with the object.

**Table 1.** The definitions of temporal-relations, codes, and topologies of object pair in the 9DST

Temporal Topology	Operators	Conditions	Binary code/Value
Disjoin	$P < Q$	$E_P < B_Q$	$(00000000)_2=0$
	$P <^* Q$	$E_Q < B_P$	$(00000001)_2=1$
Join	$P \sqcap Q$	$E_P = B_Q$	$(00010010)_2=18$
	$P \sqcap^* Q$	$E_Q = B_P$	$(00010011)_2=19$
Part-Ovlp	$P \sqcap Q$	$B_P < B_Q < E_P < E_Q$	$(00100100)_2=36$
	$P \sqcap^* Q$	$B_Q < B_P < E_Q < E_P$	$(00100101)_2=37$
Belong	$P [ Q$	$B_P = B_Q, E_P > E_Q$	$(01000110)_2=70$
	$P \% Q$	$B_P < B_Q, E_P > E_Q$	$(01000111)_2=71$
	$P ] Q$	$B_P < B_Q, E_P = E_Q$	$(01001000)_2=72$
Inside	$P [* Q$	$B_Q = B_P, E_Q > E_P$	$(10001001)_2=137$
	$P \%* Q$	$B_Q < B_P, E_Q > E_P$	$(10001010)_2=138$
	$P ]* Q$	$B_Q < B_P, E_Q = E_P$	$(10001011)_2=139$
Equal	$P = Q$	$B_P = B_Q, E_P = E_Q$	$(11001100)_2=204$

**Table 2.** The 9 neighborhood areas and codes of Op in 9D-SPA[13]

Area 4: $(00001000)_2=8$	Area 3: $(00000100)_2=4$	Area 2: $(00000010)_2=2$
Area 5: $(00010000)_2=16$	Area 0: MBR of Op :(00000000)=0	Area 1: $(00000001)_2=1$
Area 6: $(00100000)_2=32$	Area 7: $(01000000)_2=64$	Area 8: $(10000000)_2=128$



$(A,1,2,0,6,2,1, \#_2 \uparrow_{5,1} \uparrow_{0,1}), (B,1,3,1,4,2,1, \#_3 \uparrow_{2,1} \uparrow_{1,0}), (C,1,3,3,0,2,1), (D,1,3,3,0,3,3)$

(b) The corresponding VO-string of Video A.

$(A, B ; 134 ; 1|2 ; 192,12,40|32,2,19), (A, C ; 134 ; 1|2 ; 128,8,0|32,2,19), (A, D ; 134 ; 1|2 ; 128,8,0|96,6,41), (B, C ; 198 ; 1|2|3 ; 128,8,18|64,4,198|32,2,19), (B, D ; 198 ; 1|2|3 ; 128,8,18|192,4,134|96,6,4), (C, D ; 198 ; 1-3 ; 7,0,134)$

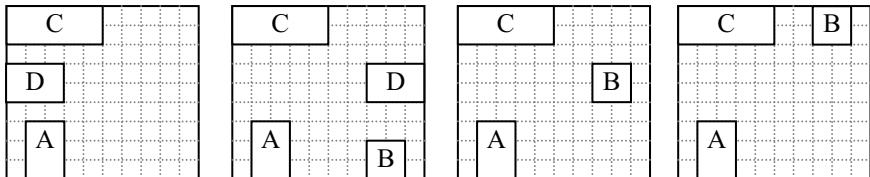
(c) The corresponding 9DST-string of Video A.

**Fig. 1.** Example of video A and the corresponding strings of 9DST approach

In the 9DST-string, there are 13 relations between time projects, those relations can be represented by seven spatial operators whose operators, conditions and corresponding topology and code of value used in the paper are listed in Table 1, where  $B_P$  and  $E_P$  are the beginning point and ending point of time projection of object  $P$ . The direction code of spatial relation between object pairs is  $D_{pq}$  and  $D_{qp}$ ,  $D_{pq}$  represents the value assigned to the directional relation between areas of objects  $O_p$  and  $O_q$  with  $O_p$  as the reference object [13]. Suppose that a video  $V$  contains  $n$  objects ( $O_1, O_2, \dots, O_n$ ), The structure of 9DST-string is defined as follows.

**Definition 2:** The 9DST-string is a 4-tuple  $(O_i, O_j ; TR_{ij} ; Sh_{ij} ; SR_{ij})$  where

1.  $O_i$  and  $O_j$  is a object pair in a video,  $1 \leq i < j \leq n$ , where we also suppose the identification of object is alphabetic order
2.  $TR_{ij}$  is the temporal relations between object pair of  $O_i$  and  $O_j$ , the code as show in the table 1;
3.  $Sh_{ij} = \{“l”\}$  is the set of interval of shots. Each shot represent the change of spatial relation of object pair in the co-exist interval, and split by time operator “l”. If the code of  $TR_{ij}$  is smaller than 36, the  $Sh_{ij}$  is not exist ;
4.  $SR_{ij} = \{“l”\}$  is the set of spatial relation of the shot  $Sh_{ij}$ . For each  $Sh_{ij}$ , exist one spatial relation and also split by time operator “l”. In the  $SR_{ij}$  contains three codes of object pair. There are two direction relation codes  $D_{ij}$  and  $D_{ji}$  , and one spatial topological relations code  $T_{ij}$ , as showing in the Table 1 and Table 2;



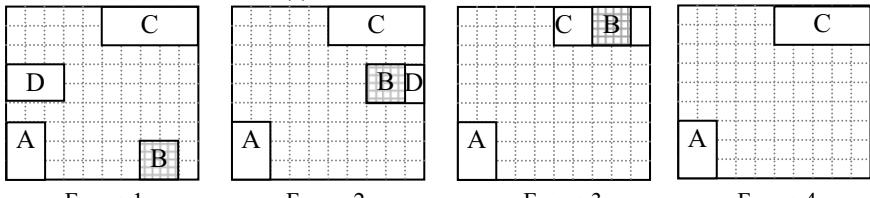
Frame 1

Frame 2

Frame 3

Frame 4

(a) Video B contains 4 frames.



Frame 1

Frame 2

Frame 3

Frame 4

(b) Video C contains 3 frames.

$\{(A, B; 72; 2|3-4; 128, 48, 0|2, 32, 0), (A, C; 204; 1-4; 14, 64, 0), (A, D; 70; 1|2; 12, 64, 0|2, 32, 0), (B, C; 139; 2-3|4; 8, 128, 0; 16, 1, 0), (B, D; 71; 2; 6, 64, 0), (C, D; 70; 1|2; 32, 6, 0|128, 8, 0)\}$

(c) The corresponding 9DST-string of Video B

$\{(A, B; 72; 1|2-3; 1, 24, 0|2, 32, 0), (A, C; 204; 1-4; 2, 32, 0), (A, D; 70; 1|2; 12, 64, 0|2, 32, 0), (B, C; 139; 1-2|3; 14, 64, 0; 17, 0, 0), (B, D; 72; 1|2; 8, 128, 0|1, 0, 137), (C, D; 70; 1|2; 14, 64, 0|17, 0, 71)\}$

(d) The corresponding 9DST-string of Video C

**Fig. 2.** Example of video B and Video C and the corresponding 9DST-strings

The example of VO-string and 9DST-string of video A is shown in Fig. 1. The corresponding VO-string of the video is shown in Fig. 1(c). The corresponding 9DST-string of the video is shown in Fig. 1(d).

### 3 The 9DST Index Structure

To reduce the search space of similarity retrieval, we propose the *9DST index structure* to index the spatial-temporal relation between a pair of objects with video identifications that contain those objects and spatial-temporal relations in the video database. This indexing structure is extended from the 9D-SPA image spatial relation indexing structure to contain spatial-temporal relations. In the index structure has three levels index structure to facilitate the video similarity retrieval. The first level index contains all the object pairs of videos in the video database, the second level indexes those spatial-temporal relation of the object pair, and the third level index contains corresponding video identifications which have the spatial-temporal relation of object pairs in the second level index. In the second level index contains three linked list point, pointing to global temporal relation list “GT”, spatial relation list “SR”, and spatial topological relation “STR”. In the global temporal relation list, each node contains two fields; one is global temporal relation  $T_{ij}$  of object pair in the video, another points to next node. In the spatial relation list, each node presents the spatial-temporal relation of a shot that contains three fields; two fields are direction relation  $D_{ij}$  and  $D_{ji}$ , and the last field points to next node. In the spatial topological relation list, each node contains two fields that are spatial topological relation of shot and point to next node. The 9DST index structure is showed as in the Fig.3. For example, we construct the structure to contain three videos, Video A, B, and C. The corresponding 9DST-strings of Video B and Video C are showing as Fig. 2, and the structure shows as Fig. 4.

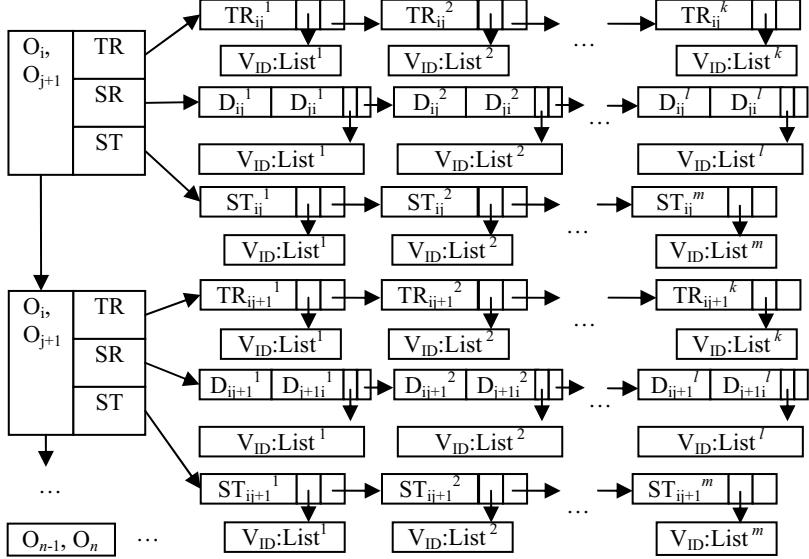
### 4 Similarity Retrieval

In this section, we extend the idea behind the spatial-temporal similarity retrieval in the 3D C-string [12] to 9DST approach. First of all, we define the notations used in the similarity retrieval process, then we describes the similarity between object pairs and propose the similarity retrieval algorithm which construct the association graph approach to find the similarity between videos.

The spatial relation between object  $A$  and that of object  $B$  may change over time in the video. These changing relations will form a spatial relation sequence. Therefore, we need to compute and record the new spatial relation whenever their spatial relation changes. We construct the spatial relation sequence of object pair in their overlapped time interval. In the 9DST index structure, we already record the individual spatial relations. So, we need to organize the spatial relation sequence to construct the similar object pair to the query video.

**Definition 3:** If  $A$  and  $B$  are the object pair in video  $V$ , a spatial relation sequence  $SRS^{DAB} = SR^{DAB}_1, SR^{DAB}_2, \dots, SR^{DAB}_n$  (or  $SRS^{DBA} = SR^{DBA}_1, SR^{DBA}_2, \dots, SR^{DBA}_n$ ) where  $SR^{DAB}_k$  (or  $SR^{DBA}_k$ ) means that the spatial relation between  $A$  and  $B$ . We said the  $SRS^{DAB}_{PQ}$  (or  $SRS^{DBA}_{PQ}$ ) is the spatial relation sequence between  $A$  and  $B$  in the spatial direction relation of video  $V$ .

**Definition 4:** If  $A$  and  $B$  are the object pair in video  $V$ , a temporal relation  $TR_{AB}$  means the global temporal relation between the time-projection of object  $A$  and that of object  $B$  in video  $V$ .



**Fig. 3.** The 9DST index structure

The similarity between videos based on the spatial-temporal relations between objects in the videos. Assume that a pair of objects  $(A, B)$  in a video  $V'$  matches a pair of objects  $(A, B)$  in query video  $V$ . We use the following notations to define the spatial-temporal relations between video  $V$  and  $V'$ .

**Definition 5:** Given two spatial relation sequences  $SRS = SR_1, SR_2, \dots, SR_n$  and  $SRS' = SR'_1, SR'_2, \dots, SR'_{m'}$  where  $n \geq m > 0$ , if  $SR_j = SR'_{j'}, j_1 < j_2 < \dots < j_m$ , for all  $i=1, 2, \dots, m$ , we can say that  $SRS'$  is a sub-sequence of  $SRS$ . The  $(A, B)$  is called a *spatially similar pair* between videos  $V'$  and  $V$ , if  $SRS'^{D_{AB}}_{AB}$  and  $SRS'^{D_{BA}}_{AB}$  both are the sub-sequences of  $SRS^{D_{AB}}_{AB}$  and  $SRS^{D_{BA}}_{AB}$ , respectively.

We also use the spatial topological relation sequence to present the topology relations between  $A$  and  $B$  which can get from the decision tree of spatial topological tree in the Fig. 5.

**Definition 6:** If  $A$  and  $B$  are the object pair in video  $V$ , a spatial topological sequence is a sequence of  $SC_1, SC_2, \dots, SC_n$ , where  $SC_i$  is the topological relation of the  $i$ th spatial relation between  $A$  and  $B$ .  $SCS_{AB}$  is the spatial topological sequence between  $A$  and  $B$  of video  $V$ .

**Definition 7:** Given two spatial topological sequences  $SCS = SC_1, SC_2, \dots, SC_n$  and  $SCS' = SC'_1, SC'_2, \dots, SC'_{m'}$  where  $n \geq m > 0$ , if  $SC_{j_i} = SC'_{j'}, j_1 < j_2 < \dots < j_m$ , for all  $i=1, 2, \dots, m$ , we can say that  $SCS'$  is a sub-sequence of  $SCS$ .  $(A, B)$  is called a *spatially st-similar pair* between videos  $V'$  and  $V$ , if  $SCS'^{SCS}_{AB}$  is the sub-sequences of  $SCS_{AB}$ .

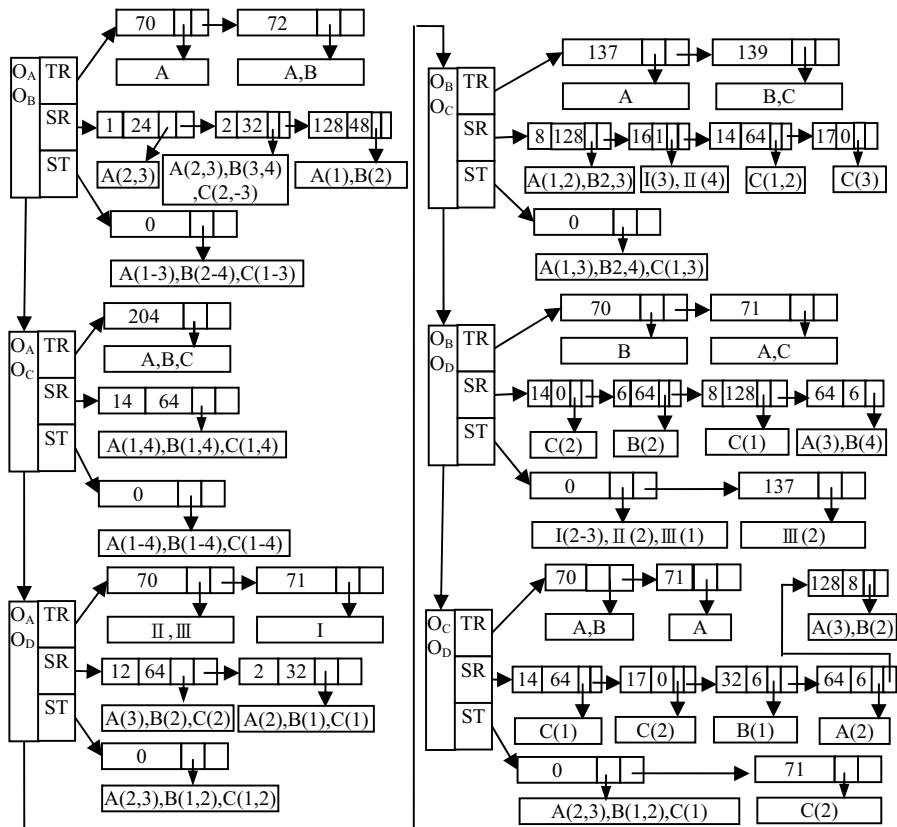


Fig. 4. The 9DST index structure contains Video A, B, and C

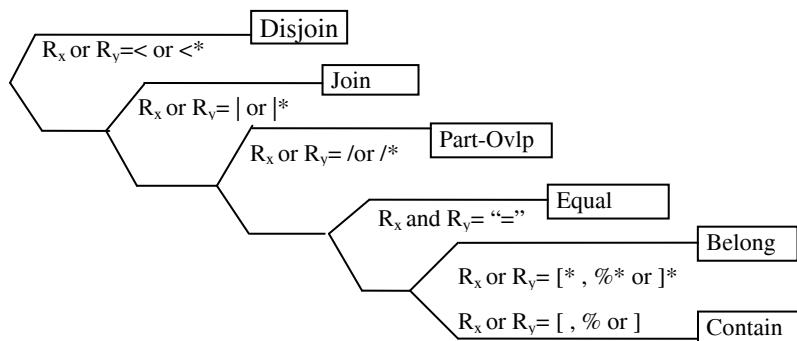


Fig. 5. The decision tree of spatial topological relation

**Definition 8:** Let  $TTR'_{AB}$  and  $TTR_{AB}$  be the temporal topological relations of object  $A$  and that of object  $B$  in video  $V'$  and  $V$ ,  $(A, B)$  is called a *temporally tt-similar pair*

between videos  $V'$  and  $V$ , if the first four bits of  $TR'_{AB}$  equal to the first four bits of  $TR_{AB}$ .

**Definition 9:** Let  $TR'_{AB}$  and  $TR_{AB}$  be the temporal relations of object  $A$  and that of object  $B$  in video  $V'$  and  $V$ ,  $(A, B)$  is called a *temporally similar pair* between videos  $V'$  and  $V$ , if  $TR'_{AB} = TR_{AB}$ .

By defining the spatial-temporal similarity between an object pair, we can define different criteria to measure the similarity degree between the object pair. For each criterion, there are two levels of similarity. The similarity between  $(A, B)$  in video  $V'$  and  $(A, B)$  in video  $V$  can be the combinations of different levels of those criteria.  $(A, B)$  is called a *similar pair*, and objects  $A$  and  $B$  are called *similar objects*. We define four types of different spatial and temporal level as following.

- (1). Spatial type I: If  $(A, B)$  is a *spatially st-similar pair*.
- (2). Spatial type II: If  $(A, B)$  is a *spatially similar pair*.
- (3). Temporal type I: If  $(A, B)$  is a *temporally tt-similar pair*.
- (4).Temporal type II: If  $(A, B)$  is a *temporally similar pair*.

Form assign the different types of similar pairs in the query, users can extract different spatial-temporal levels of information according to their interests.

To find the similarity between videos  $V'$  and  $V$ , we must consider all possible matched object sets from both videos. However, there are a large number of matched object sets, and it seems difficult to find all of them. We solve such a problem by the association graph. We build an association graph in which the vertices are formed by the matched objects between two videos. For every similar pair, the corresponding vertices in the association graph are connected by an edge. Let  $K_j$  be a clique with  $j$  vertices,  $j \geq 2$ . If every pair of vertices in  $K_j$  is similar, we call  $K_j$  a similar pair clique. The number of vertices of the largest similar pair clique in the association graph is the similarity degree between  $V'$  and  $V$ . The *similarity retrieval algorithm* is described as following.

#### Algorithm: similarity retrieval

**Input:** the query videos  $V'$ , the video database of videos  $V$ , and the assign spatial-temporal types.

**Output:** the similarity degree between  $V'$  and  $V$ , the matched objects and interval set associated with  $V$ .

1. Construct the 9DST index structure from the video database.
2. Construct the association graph for videos  $V'$  and  $V$ , where the matched objects between  $V'$  and  $V$  form the set of vertices.
3. Find every similar pair of assign spatial-temporal similar types to find the between videos  $V'$  and  $V$ . For each similar pair, the corresponding vertices in the association graph are connected by an edge which associates with an interval set in the spatial relation. For each interval, the similarity holds.
4. For those edges found in step 5, they form a set of  $K_2$ .
5. Let  $j = 2$ .
6. Repeat steps 9~10 until the largest type-std clique is found.
7. If every sub-clique  $SK_j$  of a clique  $K_{j+1}$  is a type-std  $K_j$ , construct a type-std clique  $K_{j+1}$  with the interval set which is the intersection of the interval sets associated with those  $SK_j$ .
8. Let  $j = j+1$ .

9. Output the number of vertices, matched objects and interval set associated with the largest clique, where the number of vertices is the similarity degree between  $V'$  and  $V$ .

In comparison with the retrieval method of 3D C-string [12], which is used to find the exactly matched object sets and the retrieval method of 9DST can find the different types and partly matched object sets. That is, the proposed retrieval method can provide a more flexible way to retrieve similar videos. We also compare the performance of our *video similarity retrieval algorithm* with that of the 3D C-string [12] and *9DLT approach* [15], we perform a series of experiments, which is made on the synthesized videos. The experiments are made on the synthesized video indexes. There are four cost factors dominating the performance of the *similarity retrieval algorithm*: the number of objects in a query video, the number of database videos, the average number of objects in a database video, and the average number of frames in a database video. We freely set the values of the four cost factors in the synthesized video database. In the 3D C-string approach, we attempt to compare every possible combination of objects. If a combination satisfies a certain type-std similarity, we append it to the similar set. After finding all the satisfying combinations, we use the ones with the largest similarity degree to be the query result. In the 9DST approach, we first compare the objects in the VO-string. For those similar objects, we generate the 9DST-string and 9DST index structure to find the similar pairs. Similar as 3D C-string, after finding all the similar pairs and construct the association graph, we use the ones with the largest similarity degree to be the query result. In all experiments, the execution time of the 9DST approach grows slowly as the number of objects in a query increases respect to the execution time of the 3D C-string approach. Since the 9DLT approach needs to compare every query frame with the frames in its index structure of the database videos, it takes the most execution time among all of the approaches.

## 5 Conclusions

We have proposed a new spatial-temporal knowledge structure called 9DST approach to represent symbolic videos with the VO-string and 9DST-string. In this paper, we extend the idea behind the similarity retrieval of images in 9D-SPA[13] to 9DST approach. Our approach consists of two phases. First, we construct the spatial relation sequence and temporal relations for each pair of objects of video database in the 9DST index structure. Second, we define different level of types to measure the similarity between videos and propose the similarity retrieval algorithm. Many criteria can be used to define similarity measures. The concept of processing spatial relation sequences and temporal relations can also be easily extended to process other criteria such as velocities, rates of size changes, distances, and so on. The different types of similarity would assign different multi-granularity to meet users' need. By providing various types of similarity between videos, our proposed similarity retrieval algorithm has discrimination power about different criteria. This approach also can be easily applied to an intelligent video database management system to query spatial-temporal relations between the objects and to retrieve the videos similar to a query video from a video database.

A video contains rich visual and audio (or sound) information. In the 9DST approach and the similarity retrieval algorithm, we focused on utilizing the visual information. How to integrate the audio information with the visual information to represent a video and perform similarity retrieval is worth further study.

## References

1. Sebe, N., Lew, M.S., Smeulders, A.W.M.: Video retrieval and summarization. *Computer Vision and Image Understanding* 92, 141–146 (2003)
2. Chu, W.W., Cardenas, A.F., Taira, R.K.: A knowledge-based multimedia medical distributed database system, KMED. *Information Systems* 20(2), 75–96 (1995)
3. Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: the QBIC system. *IEEE Computer* 28, 23–32 (1995)
4. Chang, S., Chen, W., Meng, H.J., Sundaram, H., Zhong, D.: VideoQ: an automated content-based video search system using visual cues. In: Proc. of ACM Intl. Conf. on Multimedia Conference, Seattle, WA, pp. 313–324 (1997)
5. Bai, L., Lao, S., Jones, G.J.F., Smeaton, A.F.: A Semantic Content Analysis Model for Sports Video Based on Perception Concepts and Finite State Machines. In: 2007 IEEE International Conference on Multimedia and Expo., pp. 1407–1410 (July 2007)
6. Djordjevic, D., Izquierdo, E.: An Object- and User-Driven System for Semantic-Based Image Annotation and Retrievalm. *IEEE Trans. On Circuits and Systems for Video Technology* 17(3), 313–323 (2007)
7. Hu, W., Xie, D., Fu, Z., Zeng, W., Maybank, S.: Semantic-Based Surveillance Video Retrieval. *IEEE Trans. Image Processing* 16(4), 1168–1181 (2007)
8. Worring, M., Schreiber, G.: Semantic Image and Video Indexing in Broad Domains. *IEEE Trans. on Multimedia* 9(5), 909–911 (2007)
9. Chang, S.K., Shi, Q.Y., Yan, C.W.: Iconic indexing by 2D strings. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 9, 413–429 (1987)
10. Lee, A.J.T., Yu, P., Chiu, H.P.: 3D Z-string: a new knowledge structure to represent spatial-temporal relations between objects in a video. *Pattern Recognition Letter* (to be published)
11. Lee, A.J.T., Chiu, H.P.: 2D Z-string: a new spatial knowledge representation for image databases. *Pattern Recognition Letter* 24, 3015–3026 (2003)
12. Lee, A.J.T., Yu, P., Chiu, H.P.: Similarity Retrieval of Videos by Using 3D C-String Knowledge Representation. *Journal of Visual Communication and Image Representation* 16, 749–773 (2005)
13. Huang, P.W., Lee, C.H.: Image database design based on 9D-SPA representation for spatial relations. *IEEE Trans. on Knowledge and Data Engineering* 16(12), 1486–1496 (2004)
14. Lee, S.Y., Hsu, F.J.: Spatial reasoning and similarity retrieval of images using 2D C-string knowledge representation. *Pattern Recognition* 25, 305–318 (1992)
15. Chan, Y.K., Chang, C.C.: Spatial similarity retrieval in video databases. *Journal of Visual Communication and Image Representation* 12, 107–122 (2001)

# Using SOA Concept to Construct an e-Learning System for College Information Management Courses

Chung C. Chang and Kou-Chan Hsiao

Department of Information Management, Chinese Culture University,  
Yang Ming Shan, Taipei 111, Taiwan  
[zcj@faculty.pccu.edu.tw](mailto:zcj@faculty.pccu.edu.tw)

**Abstract.** Current e-learning websites are built with different programming languages, data storage formats and system architectures, making it hard for data to be interoperable. Therefore, e-learning systems must have a standardized platform to allow teaching resources and system functions to be shared and reused. This issue can be effectively resolved using SOA (service-oriented architecture) concept and web service technology that has open standards and XML (extensible markup language), which allow e-learning systems developed using different programming languages to share resources, and significantly reduces the cost system developers need to spend in constructing and maintaining the system. This study uses web service components developed with Microsoft .NET and XML to construct a teaching platform with standard specifications under SOA, allowing system developers to rapidly construct an e-learning system based on web services.

**Keywords:** e-learning, web services, service-oriented architecture, SOA, XML.

## 1 Introduction

In the conventional education method, students learn from books or teachers, but this method requires them to learn directly or face to face in a specific time and space. Since the rise of the Internet, rapid developments in information technology and the popularization of the Internet have turned the Internet into an indispensable part of life, and have indirectly caused the rise of a new era – the knowledge economy era. The explosion of knowledge in this era accelerated the replacement of old things with new things, and turned learning efficiency and results into a major issue.

The wealth of resources on the Internet and multimedia technology gave birth to a learning method that uses the Internet to transfer and extract learning information and contents [1], which is the concept of e-learning. The e in e-learning represents electronic or web-based, and it is exactly what e-learning is, using electronic teaching technology to engage in learning activities on the Internet. The appearance of e-learning destroyed the conventional teaching method of teachers directly communicating with students, and

gave a role to computers in teacher-student interaction and communication, giving learning the option of being either synchronous or asynchronous [2].

SOA (service-oriented architecture) changed the conventional system development method, and allowed system integration to become more flexible. Due to its loosely coupling characteristic, SOA allows system construction to be completed by combining components [3].

The prevalence of e-learning has caused government agencies, schools and private enterprises to set up e-learning websites one after another. However, these e-learning websites were constructed with different programming languages, data storage formats and system architectures, hence causing an issue with interoperability. Therefore, this study uses SOA concept and web service technology to effectively resolve this issue.

## 2 Research Motivation and Method

Teaching resources of e-learning are reusable, which is why this study employs the concept of SOA and characteristics of web services to implement an e-learning prototype system that can go across platforms, that is simple, has open standards, has a wide range of integration, is highly efficient, and provides highly flexible integrated services. Not only has this study sought to standardize teaching platforms, but also to share e-learning software components the service provider provides.

Web services view the entire Internet as one large platform; it uses web protocols and data formats with open standards, e.g. HTTP, XML and SOAP, to allow systems developed from different programming languages to be used on heterogeneous platforms, and can also easily integrate them.

This study uses web service technology to construct a standard specification teaching platform; system functions are no longer independent, but integrated into a single environment. Its purpose is to allow software components to be reused by other distributed architectures, and not limited to certain operating platforms or programming languages, allowing even more extensive applications of the system developed by this study.

Furthermore, due to its flexibility in loosely coupling, systems under SOA can reuse software components developed by its preceding system during system updates, saving time and cost of program development [4]. In addition, different platforms can use the same software components, reducing time that system developers spend on becoming familiar with different programming languages. Remote service is achieved by writing common components into a service and using XML, increasing the flexibility and scope of future applications [3].

This study adopts an empirical approach and uses web service technology to construct an e-learning prototype system. To achieve data integration under an operating environment with heterogeneous software components, functions of the system are provided by software components developed using different programming languages.

The Microsoft .NET programming platform was released several years ago, and can be used for designing and developing ASP, C#, VB and smart phone applications, and of course web service applications. Therefore, this study uses .NET as its development tool and information management related courses as the contents to develop an e-learning system based on web services.

### 3 Literature Review

#### 3.1 E-Learning

Learning is a way to achieve one's goals; it is the process of enhancing one's performance and gaining new skills and knowledge. E-learning means to send a series of solutions to strengthen one's knowledge and performance via the Internet. E-learning provides us with a broader perspective, and is even more diverse than course software and traditional teaching when it comes to information collection and distribution and directly supporting performance. We are not only introducing new technology for learning via e-learning, but also introducing new concepts of learning. Learning does not necessarily need to rely on training or teaching. People can learn via various ways – by acquiring well designed information, utilizing new tools to enhance performance, through experience, as well as other different methods [5].

E-learning is the utilization of network technology to achieve the function of passing on knowledge at anytime and anywhere. Any type of learning that uses teaching material not in the conventional paper form and requires the use of electronic equipment can be widely referred to as e-learning [6].

e-learning is otherwise known as electronic learning, web learning, on-line learning, and distance education, and refers to interactive teaching and autonomous learning that is not limited to time and space and is completed via the Internet, information technology or media equipment [7].

Carliner [8] divided e-learning into two categories, one is formal learning, which includes online education, online training, and blended learning that combines conventional classrooms and teaching material in written form; the other is informal learning, which includes knowledge management, electronic performance support system, and blended learning that uses other forms of teaching material. Khan [9] believes that e-learning utilizes attributes and resources of the global information network to create a meaningful learning environment, and that its purpose is to train individuals to automatically and continuously learn. A few years ago, some researchers after studying e-learning indicated that the largest advantage of e-learning is that it is not limited by time and space, and that it allowed self-paced learning [10,11,12]. E-learning has removed the limitations on time and space of conventional learning, and enhanced the learning autonomy of learners [11].

When computers started becoming popular, e-learning system developers around the world used different resources and tools. When teachers began to digitize their teaching materials, they started facing problems, such as system functions could not be reused on heterogeneous platforms because the software tools came from different companies, and resources could not be shared because the format of teaching materials developed by different teachers were inconsistent [13,14]. Learning institutions hence proposed the following standards for system developers to follow [15] :

- (1) Interoperability: Teaching materials could be used in different application systems and teaching platforms, and edited using different types of tools.
- (2) Reusability: Materials and functions for learning on different teaching platforms could be easily integrated or reused.
- (3) Manageability: Records of learners and courses should be traceable and transferrable.
- (4) Accessibility: Students could gain suitable learning contents anytime, anywhere using various tools.
- (5) Durability: Applications or learning contents should not require modification after changes or updates of network technology and related standards, this way learning can last for a long time.

### **3.2 Service-Oriented Architecture**

Service-oriented architecture (SOA) is a structural model composed of standard components, such as web service technology. Its purpose is to provide enterprises, schools or web services providers with a flexible, reusable integrated interface that facilitates communication between applications, users and departments, and enhancement of web services [16].

SOA presents applications and resources as reusable services. It uses standard protocols for communication, providing a distributed environment with higher flexibility, efficiency and information integration [3].

Services are a kind of packaging for applications, and a kind of reusable component in business processes. Services generally have the following characteristics [4]:

- (1) Interoperability: Service components are interoperable, saving time on program development.
- (2) Loosely-coupled: Conventional systems cut system functional requirements of applications up into interrelated components, modules or objects; developers spend a large amount of time understanding how each component was designed and is used, so that they do not violate restrictions imposed by the component's connections. However, this makes it hard for systems to use different components in replacement. The approach of SOA is to combine systems by defining a standard interface. Components can be replaced at will as long as the replacement meets interface requirements. This

significantly enhances system flexibility and further provides a cross-platform mechanism, allowing different services to be used in different environments.

(3) Location Transparency: Transparency means that when the client end calls the server end, the system should automatically handle all affairs related to location search, security and reliability for users. Under SOA, the definition and storage location of services should be registered at a specific location for the client end to access. Users do not need to know the actual location of services.

(4) Well-defined: Uses a common definition independent from any technology and can be used by any technology.

(5) Stateless: Services are stateless; any service is not required to know the contents of the previous step, all services are in an independent state.

Some scholars describe the relationship between SOA and web services as SOA being the steel bars in buildings and web services being the cement. Westerman [17] believes that the program development technology and construction method of SOA are used together in alternation, linking several services together via message passing to implement a new application instead of starting from scratch. Chien [18] pointed out that the elements combined together by SOA normally included software components, services and processes; processes define procedures for handling external requirements, services include all program components required for specific procedures, and software components are programs responsible for executing specific tasks.

## 4 System Architecture and Implementation

This study uses web service technology to build an e-learning system, and uses .NET to code partial software components within the system. Teaching contents of the system are mainly information management related courses. In the future, we will add software components coded using JAVA into the system to achieve true integration of heterogeneous systems. The system architecture is shown in Fig. 1and the system functions are described below:

### (1) Bulletin Board

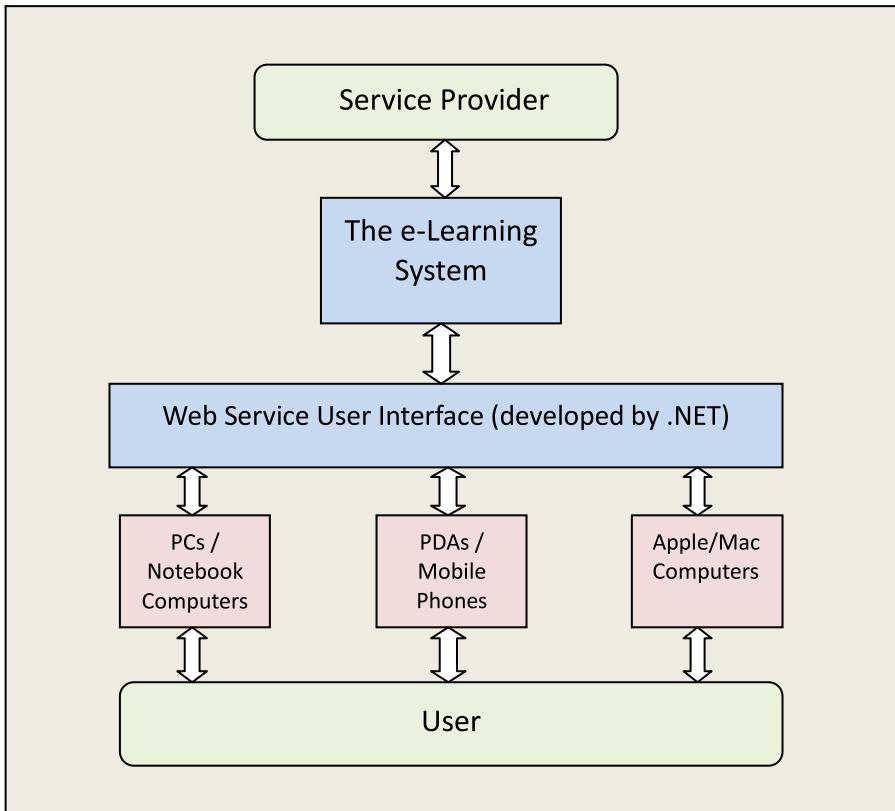
(a) Announcement Information: This function provides summaries and detailed information of course announcements, course information and scores.

(b) Post Announcement: This function allows users to fill in summaries and contents of announcements, and post the announcements on the system.

### (2) Course Materials

(a) Upload Material: This function allows teachers to upload course materials to the system, e.g. power point files, word files and PDF files, for learners to download.

(b) Download Material: This function allows teachers or learners to download course materials from the system to use for teaching or learning.



**Fig. 1.** The overall system architecture

(3) Teaching Videos

(a) Upload Video: This function allows teachers to upload teaching videos to the system for learners to download or play online.

(b) Download Video: This function allows teachers and learners to download teaching videos from the system and use them for teaching or learning.

(c) Play Video: This function allows learners with enough bandwidth to directly play teaching videos online.

(4) Forum

(a) Course Discussion: This function provides an area for teachers and learners to discuss course contents and exchange information.

(5) Online Test

(a) Introduction to Computer Science Assessment: This function provides an online assessment of the course Introduction to Computer Science to understand how well learners understand course contents.

(b) Information Management Assessment: This function provides an online assessment of the course Information Management to understand how well learners understand course contents.

(6) Assignments

(a) Assignment Announcement: This function allows teachers to announce information and notices related to assignments.

(b) Upload Assignment: This function allows learners to upload their assignments to the system for teachers to grade.

(c) Download Assignment: This function allows teachers to download the assignments that learners uploaded.

Users operate system functions using the integrated user interface. When users select a system function, the system component starts to operate. The system component searches for the suitable web service from service providers, service providers return the location of the web service in a WSDL file to the system component, and after the system component reads the information in the WSDL file, it then calls the web service to use and displays processing results on the user interface; the document is stored as a XML file in the system.

Fig. 2 shows the system's user interface and bulletin board function, which is users' initial location after logging in. The bulletin board shows all categories of teaching information for users to choose from. Users may click on the topics that interest them and link to detailed contents. Fig. 3 shows the upload material function. Teachers can use this function to upload teaching materials to the system for others to use.



Fig. 2. A sample screen of the user interface and bulletin board function

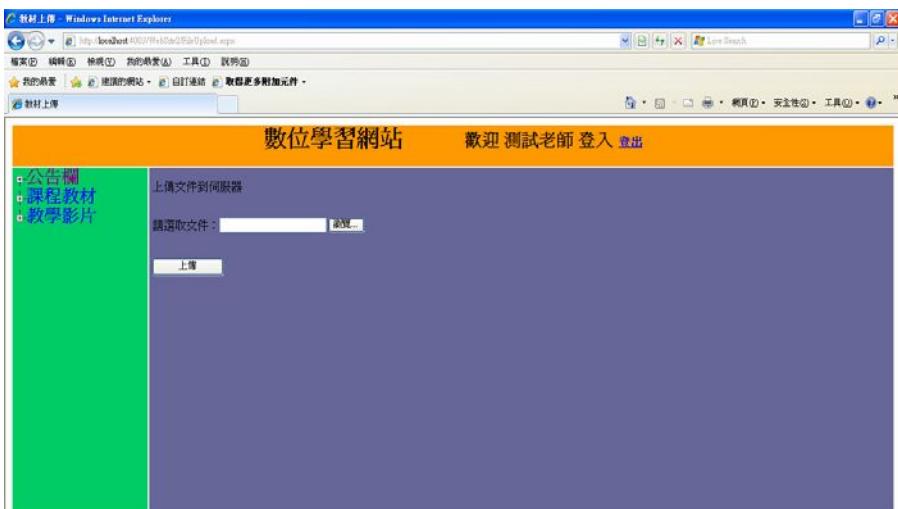


Fig. 3. A sample screen of the upload material function

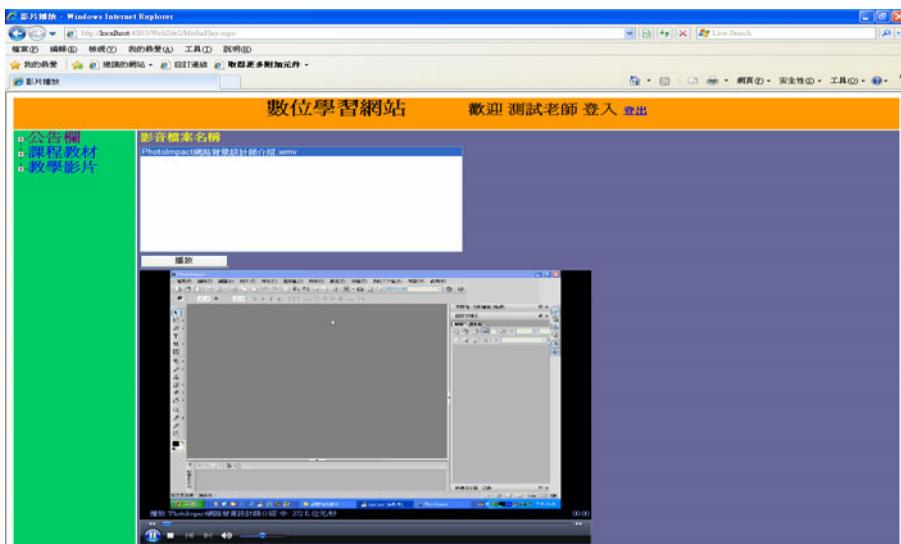


Fig. 4. A sample screen of teaching video function

Fig. 4 shows the teaching video function. When learners have enough bandwidth, they can choose to directly play the teaching video they wish to view online, and are not required to download the video first.

## 5 Conclusion and Future Works

By using SOA concept and web services technology to construct an e-learning system, this study allows users to use the e-learning system via the Internet without imposing any restrictions on time and place, and allows system developers to avoid incompatibility of data formats caused by different development platforms and programming languages, achieving the integration of heterogeneous software components. However, the e-learning system can not be used without a network cable or computer. Therefore, if web services can be integrated with WLAN technology, then users will be able to access the e-learning system at anytime and anywhere using their PDA or smart phone. Furthermore, for e-learning systems based on SOA and web services, besides the .NET development platform, software components can also be developed using JAVA and the Apache AXIS platform; software components developed using different platforms can even be integrated into the same system, and truly achieve the integration of heterogeneous platforms, reducing cost of system development and maintenance.

## References

1. Su, Y.Z.: e-Learning: The Window of Knowledge in the Future World. *Information Pioneers Magazine* 53, 1–2 (2003)
2. Chen, H.Z.: A Study and Practice of Involving Evaluation in a SCROM Learning Environment. Master's Thesis, Department of Information Management, Providence University, Taiwan (2007)
3. Wang, K.F.: Planning and Constructing a High School Collaborative Operation System Based on SOA. Master's Thesis, Department of Information Management, National Kaoshung First University of Science and Technology, Taiwan (2006)
4. Huhns, M.N., Singh, M.P.: Service-Oriented Computing Key Concepts and Principles. *IEEE Internet Computing* 9(1), 75–81 (2005)
5. Rosenberg, M.J.: e-Learning: Strategies for Delivering Knowledge in the Digital Age. The MacGraw-Hill Companies, Inc., New York (2001)
6. Yang, Z.T., Chen, N.H.: Theory and Practice of e-Learning. Dr.Master Publishing Company, Taipei (2006)
7. Chen, Y.Z.: A Framework Design for Knowledge Management Oriented e-Learning System. Master's Thesis, Department of Information Management, Nan-Hua University, Taiwan (2006)
8. Carliner, S.: Designing E-Learning. American Society for Training and Development, Virginia (2002)
9. Khan, B.H.: Web-Based Instruction: An Introduction. *Educational Media International* 35(2), 63–71 (1998)
10. Brown, B.L.: Web-Based Training. *ERIC Digest* 218, 1–8 (2000)
11. Hathorn, L., Ingram, A.: Cooperation and Collaboration Using Computer-Mediated Communication. *Journal of Educational Computing Research* 26(3), 325–347 (2002)
12. Driscoll, M.: Web-Based Training in the Workplace. *Adult Learning* 10(4), 21–25 (1999)
13. Anido, L., Llamas, M.: A Contribution to the e-Learning Standardization. In: Proceedings of the Second IEEE Conference on Standardization and Innovation in Information Technology, pp. 295–309. IEEE Press, New York (2001)

14. Anido, L., Llamas, M., Fernandez, M.J., Caeiro, M., Santos, J., Rodriguez, J.: A Component Model for Standardized Web-Based Education. ACM Journal of Educational Resources in Computing 1(2), 86–95 (2001)
15. The MASIE Center: Making Sense of Learning Specifications & Standards: A Decision Maker's Guide to Their Adoption,  
[http://www.berg-imc.nl/download/masie\\_making\\_sense.pdf](http://www.berg-imc.nl/download/masie_making_sense.pdf)
16. Tseng, B.C.: Introduction to Service-Oriented Architecture,  
[http://www.cc.ntu.edu.tw/chinese/epaper/20070620\\_1008.htm](http://www.cc.ntu.edu.tw/chinese/epaper/20070620_1008.htm)
17. Westerman, J.: Introduction to Service-Oriented Architecture,  
<http://www.information-management.com/news/7992-1.html>
18. Chein, S.T.: Applications of Service-Oriented Architecture,  
[http://www.microsoft.com/taiwan/msdn/columns/soa/SOA\\_overview\\_2004112901.htm](http://www.microsoft.com/taiwan/msdn/columns/soa/SOA_overview_2004112901.htm)

# Cryptanalysis on Sun-Yeh's Password-Based Authentication and Key Distribution Protocols with Perfect Forward Secrecy

Wen-Gong Shieh<sup>1</sup> and Wen-Bing Horng<sup>2</sup>

<sup>1</sup> Department of Information Management,

Chinese Culture University, Taipei 11114, Taiwan, R.O.C.

<sup>2</sup> Department of Computer Science and Information Engineering,

Tamkang University, Taipei 25137, Taiwan, R.O.C.

[wgshieh@faculty.pccu.edu.tw](mailto:wgshieh@faculty.pccu.edu.tw), [horng@mail.tku.edu.tw](mailto:horng@mail.tku.edu.tw)

**Abstract.** Recently, Sun and Yeh proposed three password-based authentication and key distribution protocols with perfect forward secrecy with low, medium, and high security, respectively, in the environment of three principals involved: a client, an application server, and an authentication server. In this paper, we will show that each of their protocols has some security flaw; the protocol with low security is vulnerable to the known-plaintext attack and the other two are subject to the man-in-the-middle attack.

**Keywords:** Diffie-Hellman key exchange, key distribution, password-based authentication, perfect forward secrecy.

## 1 Introduction

In insecure public networks, how to communicate securely becomes an essential problem. In order to provide secure communications, user authentication and session key distribution are two basic services used in network communications. Besides, perfect forward secrecy (PFS) is also an important issue for session key distribution; it means that if a long-term secret (e.g., a user's password in a password-based authentication protocol) is compromised by an attacker, he still cannot derive the session keys of past sessions.

To obtain better security service, a centralized authentication server is constructed to provide the authentication of users to servers as well as servers to users. A well-known protocol of this setting is the Kerberos system [1]. However, Kerberos requires users to use strong passwords for user authentication. Hence, if users use weak passwords, the system will become insecure due to the offline password guessing attack.

To cope with this problem, recently, Sun and Yeh [2] proposed three password-based user authentication protocols with low, medium, and high security, respectively, even if users choose weak passwords. Their authentication protocols provide not only key distribution but also perfect forward secrecy. However, in

this paper, we will show that all of their protocols have security weaknesses. The low security PFS protocol is vulnerable to the known-plaintext attack, while the other two PFS protocols are subject to the man-in-the-middle attack.

The rest of the paper is organized as follows. In Section 2, we briefly review each of Sun-Yeh's protocols and show the security flaw of each protocol. Then, we conclude the paper in the last section.

## 2 Review and Cryptanalysis of Sun-Yeh's Protocols

Sun and Yeh [2] proposed three authentication and key distribution protocols to provide perfect forward secrecy with low, medium, and high security, respectively. All these protocols have three principals involved:

- $A$ : a client who requests services from an application server.
- $B$ : an application server that provides services to clients.
- $S$ : an authentication server that is responsible for authentication and distributes the common session key shared between  $A$  and  $B$ .

It is assumed that (1) the authentication server  $S$ 's public key  $K_S$  is known to all principals, (2) the application server  $B$ 's secret key  $S_B$  is known to  $S$  via a secure channel, and (3) the client  $A$ 's password  $P_A$  is known to  $S$  via a secure channel. The notations used throughout this paper are summarized in Table 1.

**Table 1.** Notation

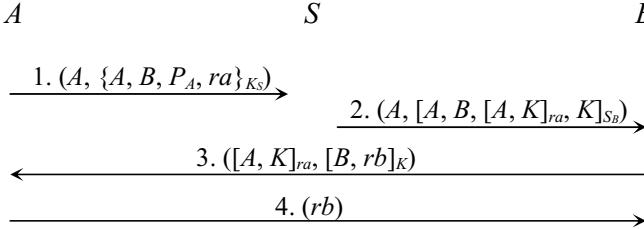
Notation	Meaning
$A$	Client (user)
$B$	Application server
$S$	Authentication server
$P_A$	Password shared between $A$ and $S$
$S_B$	Secret key shared between $B$ and $S$
$K_S$	Public key of the authentication server
$x, y, a, b, ra, rb, rb'$	Random numbers
$h()$	One-way hash function
$A \rightarrow B : M$	$A$ sends the message $M$ to $B$
$g$	Base generator
$P$	Large prime ( $P$ is the modulus of all modular exponentiations)
$[M]_K$	Symmetric encryption of $M$ with key $K$
$\{M\}_K$	Asymmetric encryption of $M$ with public key $K$

### 2.1 Low Security PFS Protocol

**Review.** Fig. 1 shows the protocol providing PFS with low security. The steps are detailed as follows:

- (1)  $A \rightarrow S : (A, \{A, B, P_A, ra\}_{K_S})$

Client  $A$  first chooses a random number  $ra$  and encrypts  $A$ ,  $B$ ,  $P_A$ , and  $ra$  with the authentication server  $S$ 's public key  $K_S$ , where  $P_A$  is  $A$ 's password. Then, he transmits the encrypted message as a request to  $S$ .



**Fig. 1.** Sun-Yeh's low security PFS protocol

(2)  $S \rightarrow B : (A, [A, B, [A, K]_{ra}, K]_{S_B})$

On receiving client  $A$ 's request message,  $S$  decrypts  $\{A, B, P_A, ra\}_{K_S}$  with its private key and checks the legitimacy of  $A$  by verifying the password  $P_A$ . Then,  $S$  chooses a common key  $K$  and computes  $[A, B, [A, K]_{ra}, K]_{S_B}$ , where  $ra$  is used as a one-time key. Finally,  $S$  transmits the encrypted message to  $B$ .

(3)  $B \rightarrow A : ([A, K]_{ra}, [B, rb]_K)$

After receiving the message  $[A, B, [A, K]_{ra}, K]_{S_B}$  from  $S$ , the application server  $B$  first decrypts it with the secret key  $S_B$  to get the common key  $K$ . Then,  $B$  chooses a challenge value  $rb$  and encrypts  $B$  and  $rb$  with the common key  $K$ . Next,  $B$  transmits  $([A, K]_{ra}, [B, rb]_K)$  to  $A$ .

(4)  $A \rightarrow B : (rb)$

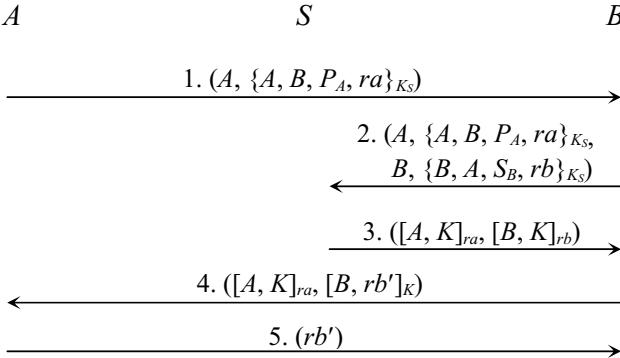
The client  $A$  decrypts the received message  $[A, K]_{ra}$  with  $ra$  to get the common key  $K$ . Then, he decrypts  $[B, rb]_K$ , checks the validity of  $K$ , and sends the response value  $rb$  to  $B$ .

Finally, both  $A$  and  $B$  authenticate each other and compute the common session key  $h(K)$ .

**Cryptanalysis (Known-Plaintext Attack).** This low security PFS protocol is vulnerable to the *known-plaintext attack*. In this protocol, the ciphertext  $[A, B, [A, K]_{ra}, K]_{S_B}$  in the second message transmitted from the authentication server  $S$  to the application server  $B$  is encrypted symmetrically with the secret key  $S_B$  shared between  $B$  and  $S$ . If the length of  $A$  is greater than a block for symmetric encryption, then an attacker might intercept such messages and launch a known-plaintext attack to derive the secret key  $S_B$ . This is because the plaintext  $A$  is also included in the transmitted message in addition to the ciphertext  $[A, B, [A, K]_{ra}, K]_{S_B}$ .

## 2.2 Medium Security PFS Protocol

**Review.** Fig. 2 shows the medium security PFS protocol. The detailed steps are described as follows:

**Fig. 2.** Sun-Yeh's medium security PFS protocol

(1)  $A \rightarrow B : (A, \{A, B, P_A, ra\}_{K_S})$

The client  $A$  first selects a random number  $ra$  and encrypts  $A, B, P_A, ra$  with  $S$ 's public key  $K_S$ , where  $P_A$  is  $A$ 's password. Then,  $A$  transmits the encrypted message as a request to  $B$ .

(2)  $B \rightarrow S : (A, \{A, B, P_A, ra\}_{K_S}, B, \{B, A, S_B, rb\}_{K_S})$

Upon receiving client  $A$ 's request message, the application server  $B$  first chooses a random number  $rb$  and encrypts  $B, A, S_B$ , and  $rb$  with  $S$ 's public key  $K_S$ . Both ciphertexts  $\{A, B, P_A, ra\}_{K_S}$  and  $\{B, A, S_B, rb\}_{K_S}$  together with  $A$  and  $B$  are sent to  $S$ .

(3)  $S \rightarrow B : ([A, K]_{ra}, [B, K]_{rb})$

After receiving the message from  $B$ , the authentication server  $S$  decrypts it with its private key. Then, it checks the legitimacy of  $A$  by verifying his password  $P_A$ , and the legitimacy of  $B$  by verifying its secret key  $S_B$ . The server  $S$  then chooses a common key  $K$ , computes  $([A, K]_{ra}, [B, K]_{rb})$ , and transmits it to  $B$ . Note that both  $ra$  and  $rb$  act as one-time keys.

(4)  $B \rightarrow A : ([A, K]_{ra}, [B, rb']_K)$

The application server  $B$  first decrypts the message  $[B, K]_{rb}$  with  $rb$  to get the common key  $K$ . Then,  $B$  chooses a challenge value  $rb'$ , encrypts  $B$  and  $rb'$  with  $K$ , and sends both  $[A, K]_{ra}$  and  $[B, rb']_K$  to  $A$ .

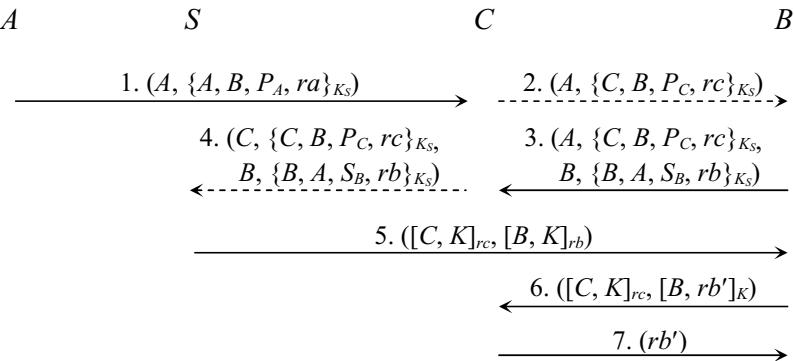
(5)  $A \rightarrow B : (rb')$

On receiving the message from  $B$ , the client  $A$  decrypts  $[A, K]_{ra}$  with  $ra$  to obtain the common key  $K$ . Then,  $A$  decrypts  $[B, rb']_K$ , checks the validity of  $K$ , and sends the response value  $rb'$  to  $B$ .

When the authentication procedure is complete, both  $A$  and  $B$  generate a session key  $h(K)$  for secure communications.

**Cryptanalysis (Man-in-the-Middle Attack).** This medium security PFS protocol is vulnerable to the *man-in-the-middle attack*, as shown in Fig. 3. Suppose that an attacker  $C$ , who is also a legal user in  $S$ , has full control of the network. If  $C$  wants to impersonate the client  $A$  to cheat the application server  $B$ , he might intercept and block  $A$ 's first message  $(A, \{A, B, P_A, ra\}_{K_S})$ , modify

it to  $(A, \{C, B, P_C, rc\}_{K_S})$ , and send it to  $B$ , where  $P_C$  is  $C$ 's password and  $rc$  is a random number chosen by  $C$ . (Note that in Fig. 3, the fake messages are represented in dashed arrows.) After receiving the modified message, the application server  $B$  will send  $(A, \{C, B, P_C, rc\}_{K_S}, B, \{B, A, S_B, rb\}_{K_S})$  to the authentication server  $S$ . At this moment, the attacker  $C$  intercepts and blocks the above  $B$ 's message sending to  $S$ , he then modifies this intercepted message to  $(C, \{C, B, P_C, rc\}_{K_S}, B, \{B, A, S_B, rb\}_{K_S})$  and transmits it to  $S$ . Upon receiving the message from the attacker  $C$ , the authentication server  $S$  will sends  $([C, K]_{rc}, [B, K]_{rb})$  to the application server  $B$ . Then,  $B$  will transmit  $([C, K]_{rc}, [B, rb']_K)$  to  $A$ , where  $rb'$  is a random number chosen by  $B$  acting as a challenge. The attacker  $C$  then intercepts and blocks the above message from  $B$  to  $A$ . Finally,  $C$  will sends  $(rb')$  to  $B$ . After mutual authentication,  $C$  and  $B$  will use the common session key  $h(K)$  to communicate each other.



**Fig. 3.** Man-in-the middle attack on Sun-Yeh's medium security PFS protocol

The steps of the cryptanalysis on this medium security PFS protocol are detailed in the following.

- (1)  $A \rightarrow B : (A, \{A, B, P_A, ra\}_{K_S})$

As before, the client  $A$  transmits the encrypted message as a request to  $B$ .

- (2)  $C \rightarrow B : (A, \{C, B, P_C, rc\}_{K_S})$

The attacker  $C$  first waits for, intercepts, and blocks  $A$ 's first message to  $B$ . Then,  $C$  modifies it to  $(A, \{C, B, P_C, rc\}_{K_S})$  since the authentication server  $S$ 's public key  $K_S$  is known to every one, where  $P_C$  is  $C$ 's password and  $rc$  is a random number chosen by  $C$ . Finally,  $C$  sends the modified message to  $B$ .

- (3)  $B \rightarrow S : (A, \{C, B, P_C, rc\}_{K_S}, B, \{B, A, S_B, rb\}_{K_S})$

Upon receiving the attacker  $C$ 's request message, the application server  $B$  first chooses a random number  $rb$  and encrypts  $B$ ,  $A$ ,  $S_B$ , and  $rb$  with  $S$ 's public key  $K_S$ . Both ciphertexts  $\{C, B, P_C, rc\}_{K_S}$  and  $\{B, A, S_B, rb\}_{K_S}$  together with  $A$  and  $B$  are sent to  $S$  for authentication. Note that at this

moment, the application server  $B$  does not know the inconsistency between  $A$  and  $C$  in the ciphertext  $\{C, B, P_C, rc\}_{K_S}$ .

- (4)  $C \rightarrow S : (C, \{C, B, P_C, rc\}_{K_S}, B, \{B, A, S_B, rb\}_{K_S})$

The attacker  $C$  waits for, intercepts, and blocks the message sent from  $B$  to  $S$ . Then, he modifies the identity of  $A$  into  $C$  in the intercepted message and transmits it to  $S$ .

- (5)  $S \rightarrow B : ([C, K]_{rc}, [B, K]_{rb})$

After receiving the modified message from  $C$ , the authentication server  $S$  decrypts it with its private key. Then, it checks the legitimacy of  $C$  by verifying his password  $P_C$ , and the legitimacy of  $B$  by verifying his secret key  $S_B$ . Surely, the authentication of  $C$  will succeed since  $C$  is a legal user with correct password  $P_C$ . The server  $S$  then chooses a common key  $K$ , computes  $([C, K]_{rc}, [B, K]_{rb})$ , and transmits it to  $B$ .

- (6)  $B \rightarrow A : ([C, K]_{rc}, [B, rb']_K)$

The application server  $B$  first decrypts the message  $[B, K]_{rb}$  with  $rb$  to get the common key  $K$ . Then,  $B$  chooses a challenge value  $rb'$ , encrypts  $B$  and  $rb'$  with  $K$ , and sends both  $[C, K]_{rc}$  and  $[B, rb']_K$  to  $A$ .

- (7)  $C \rightarrow B : (rb')$

The attacker  $C$  waits for, intercepts, and blocks the message sent from  $B$  to  $A$ . Then, the attacker  $C$  decrypts  $[C, K]_{rc}$  with  $rc$  to obtain the common key  $K$ . Then,  $C$  decrypts  $[B, rb']_K$ , checks the validity of  $K$ , and sends the response value  $rb'$  to  $B$ .

After the authentication procedure is complete, both the attacker  $C$  and the application server  $B$  generate a session key  $h(K)$  for subsequent communications.

### 2.3 High Security PFS Protocol

**Review.** Fig. 4 shows the high security PFS protocol. The detailed steps are depicted as follows:

- (1)  $A \rightarrow B : (A, \{A, B, P_A, ra, g^x\}_{K_S})$

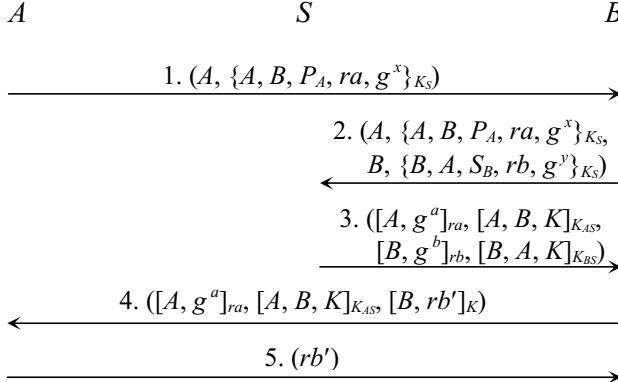
Client  $A$  first chooses two random numbers  $ra$  and  $x$  and computes  $g^x$ . Then,  $A$  encrypts  $A$ ,  $B$ ,  $P_A$ ,  $ra$ , and  $g^x$  with  $K_S$  and transmits the ciphertext to  $B$ , where  $P_A$  is  $A$ 's password.

- (2)  $B \rightarrow S : (A, \{A, B, P_A, ra, g^x\}_{K_S}, B, \{B, A, S_B, rb, g^y\}_{K_S})$

On the receipt of client  $A$ 's request message,  $B$  also selects two random numbers  $rb$  and  $y$  and computes  $g^y$ . Then, he encrypts  $B$ ,  $A$ ,  $S_B$ ,  $rb$ , and  $g^y$  with  $K_S$ . Both ciphertexts  $\{A, B, P_A, ra, g^x\}_{K_S}$  and  $\{B, A, S_B, rb, g^y\}_{K_S}$  together with  $A$  and  $B$  are sent to  $S$ .

- (3)  $S \rightarrow B : ([A, g^a]_{ra}, [A, B, K]_{K_{AS}}, [B, g^b]_{rb}, [B, A, K]_{K_{BS}})$

After receiving the message from  $B$ , the authentication server  $S$  first decrypts it with its private key. Then,  $S$  checks the authenticity of  $A$  by verifying his password  $P_A$  and the authenticity of  $B$  by verifying its secret key  $S_B$ . Next,  $S$  selects two random numbers  $a$  and  $b$  and a common key  $K$ , computes  $([A, g^a]_{ra}, [A, B, K]_{K_{AS}}, [B, g^b]_{rb}, [B, A, K]_{K_{BS}})$ , and transmits it



**Fig. 4.** Sun-Yeh's high security PFS protocol

to  $B$ , where  $K_{AS} = (g^x)^a = (g^a)^x = g^{xa}$  and  $K_{BS} = (g^y)^b = (g^b)^y = g^{yb}$  are used to pass the session key  $K$  securely. Note that both  $ra$  and  $rb$  also act as one-time keys.

- (4)  $B \rightarrow A : ([A, g^a]_{ra}, [A, B, K]_{K_{AS}}, [B, rb']_K)$

Upon receiving the message from  $S$ , the application server  $B$  first decrypts the message  $[B, g^b]_{rb}$  with  $rb$  and computes  $K_{BS} = (g^b)^y = g^{yb}$ .  $B$  then gets the common key  $K$  by decrypting  $[B, A, K]_{K_{BS}}$  with  $K_{BS}$ . Then,  $B$  chooses a challenge value  $rb'$  and encrypts  $B$  and  $rb'$  with the common key  $K$ . Finally,  $B$  sends  $([A, g^a]_{ra}, [A, B, K]_{K_{AS}}, [B, rb']_K)$  to  $A$ .

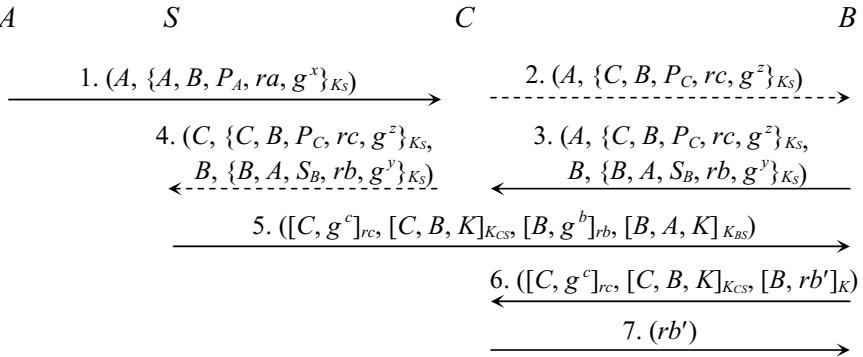
- (5)  $A \rightarrow B : (rb')$

After receiving the message from  $B$ , the client  $A$  decrypts  $[A, g^a]_{ra}$  with  $ra$  and computes  $K_{AS} = (g^a)^x = g^{xa}$ . Then,  $A$  obtains the common key  $K$  by decrypting  $[A, B, K]_{K_{AS}}$  with  $K_{AS}$ . Finally,  $A$  uses  $K$  to decrypt  $[B, rb']_K$ , checks the validity of  $K$ , and sends  $(rb')$  to  $B$ .

After mutually authenticated each other, both  $A$  and  $B$  compute the common session key  $h(K)$ .

**Cryptanalysis (Man-in-the-Middle Attack).** The high security PFS protocol is also vulnerable to the *man-in-the-middle attack*. Suppose that an attacker  $C$ , who is also a legal user in  $S$ , has full control of the network. If the attacker  $C$  wants to impersonate the client  $A$  to login to the application server  $B$ , he might intercept and block  $A$ 's first message  $(A, \{A, B, P_A, ra, g^x\}_{K_S})$ , modify it to  $(A, \{C, B, P_C, rc, g^z\}_{K_S})$ , and send it to  $B$ , where  $P_C$  is  $C$ 's password and  $rc$  and  $z$  are random numbers chosen by  $C$ . After receiving the modified message, the application server  $B$  will send  $(A, \{C, B, P_C, rc, g^z\}_{K_S}, B, \{B, A, S_B, rb, g^y\}_{K_S})$  to the authentication server  $S$ . At this moment, the attacker  $C$  intercepts and blocks the above  $B$ 's message sending to  $S$ , he then modifies this intercepted message to  $(C, \{C, B, P_C, rc, g^z\}_{K_S}, B, \{B, A, S_B, rb, g^y\}_{K_S})$  and transmits it to  $S$ . Upon receiving the message from the attacker  $C$ , the authentication server

$S$  will sends  $([C, g^c]_{rc}, [C, B, K]_{K_{AS}}, [B, g^b]_{rb}, [B, C, K]_{K_{BS}})$  to the application server  $B$ , where  $b$  and  $c$  are two random numbers chosen by  $S$  and  $K$  is a common key. Then,  $B$  will transmit  $([C, g^c]_{rc}, [C, B, K]_{K_{AS}}, [B, rb']_K)$  to the attacker  $A$ , where  $rb'$  is random number used as a challenge. The attacker  $C$  then intercepts and blocks the above message from  $B$  to  $A$ . Finally,  $C$  will sends  $(rb')$  to  $B$ . After mutual authentication,  $C$  and  $B$  will use the common session key  $h(K)$  to communicate each other.



**Fig. 5.** Man-in-the middle attack on Sun-Yeh's high security PFS protocol

The steps of the cryptanalysis on this high security PFS protocol are elaborated in the following.

- (1)  $A \rightarrow B : (A, \{A, B, P_A, ra, g^x\}_{K_S})$

As before, the client  $A$  sends the ciphertext  $\{A, B, P_A, ra, g^x\}_{K_S}$  to  $B$ .

- (2)  $C \rightarrow B : (A, \{A, B, P_C, rc, g^z\}_{K_S})$

The attacker  $C$  first waits for, intercepts, and blocks  $A$ 's first message to  $B$ . Then,  $C$  modifies it to  $(A, \{C, B, P_C, rc, g^z\}_{K_S})$  since the authentication server  $S$ 's public key  $K_S$  is known to every one, where  $P_C$  is  $C$ 's password and  $rc$  and  $z$  are random numbers chosen by  $C$ . Finally,  $C$  sends the modified message to  $B$ .

- (3)  $B \rightarrow S : (A, \{C, B, P_C, rc, g^z\}_{K_S}, B, \{B, A, S_B, rb, g^y\}_{K_S})$

On the receipt of the attacker  $C$ 's request message,  $B$  also selects two random numbers  $rb$  and  $y$  and computes  $g^y$ . Then, he encrypts  $B$ ,  $A$ ,  $S_B$ ,  $rb$ , and  $g^y$  with  $K_S$ . Both ciphertexts  $\{C, B, P_C, rc, g^z\}_{K_S}$  and  $\{B, A, S_B, rb, g^y\}_{K_S}$  together with  $A$  and  $B$  are sent to  $S$ . Note that at this moment, the application server  $B$  does not know the inconsistency between  $A$  and  $C$  in the ciphertext  $\{C, B, P_C, rc, g^z\}_{K_S}$ .

- (4)  $C \rightarrow S : (C, \{C, B, P_C, rc, g^z\}_{K_S}, B, \{B, A, S_B, rb, g^y\}_{K_S})$

The attacker  $C$  waits for, intercepts, and blocks the message sent from  $B$  to  $S$ . Then, he modifies the identity of  $A$  into  $C$  in the intercepted message and transmits it to  $S$ .

- (5)  $S \rightarrow B : ([A, g^a]_{ra}, [A, B, K]_{K_{AS}}, [B, g^b]_{rb}, [B, A, K]_{K_{BS}})$

After receiving the message from the attacker  $C$ , the authentication server  $S$  first decrypts it with its private key. Then,  $S$  checks the authenticity of  $C$  by verifying his password  $P_C$  and the authenticity of  $B$  by verifying its secret key  $S_B$ . Apparently, the attacker  $C$  will pass the authentication since  $C$  is a legal user with correct password  $P_C$  in the authentication server  $S$ . Next,  $S$  selects two random numbers  $c$  and  $b$  and a common key  $K$ , computes  $([C, g^c]_{rc}, [C, B, K]_{K_{CS}}, [B, g^b]_{rb}, [B, A, K]_{K_{BS}})$ , and transmits it to  $B$ , where  $K_{CS} = (g^z)^c = (g^c)^z = g^{zc}$  and  $K_{BS} = (g^y)^b = (g^b)^y = g^{yb}$  are used to pass the session key  $K$ .

- (6)  $B \rightarrow A : ([C, g^c]_{rc}, [C, B, K]_{K_{CS}}, [B, rb']_K)$

Upon receiving the message from  $S$ , the application server  $B$  first decrypts the message  $[B, g^b]_{rb}$  with  $rb$  and computes  $K_{BS} = (g^b)^y = g^{yb}$ .  $B$  then gets the common key  $K$  by decrypting  $[B, A, K]_{K_{BS}}$  with  $K_{BS}$ . Then,  $B$  chooses a challenge value  $rb'$  and encrypts  $B$  and  $rb'$  with the common key  $K$ . Finally,  $B$  sends  $([C, g^c]_{rc}, [C, B, K]_{K_{CS}}, [B, rb']_K)$  to  $A$ .

- (7)  $C \rightarrow B : (rb')$

The attacker  $C$  waits for, intercepts, and blocks the message sent from  $B$  to  $A$ . Then, the attacker  $C$  decrypts  $[C, g^c]_{rc}$  with  $rc$  and computes  $K_{CS} = (g^c)^z = g^{zc}$ . Then,  $C$  obtains the common key  $K$  by decrypting  $[C, B, K]_{K_{CS}}$  with  $K_{CS}$ . Finally,  $C$  uses  $K$  to decrypt  $[B, rb']_K$ , checks the validity of  $K$ , and sends  $(rb')$  to  $B$ .

After mutually authenticated each other, both  $A$  and  $B$  compute the common session key  $h(K)$  for subsequent communications.

### 3 Conclusion

In this paper, we have demonstrated the weaknesses of Sun-Yeh's three password-based authentication and key distribution protocols with perfect forward secrecy. The low security PFS protocol is vulnerable to the known-plaintext attack, while the medium and high security PFS protocols are subject to the man-in-the-middle attack. The security weaknesses of the latter two protocols are owing to the insufficient checking of the legitimacy in their original work.

### References

1. Kohl, J.T., Neuman, B.C., T'so, T.Y.: The evolution of the Kerberos authentication system. In: Distributed Open Systems, pp. 78–94. IEEE Computer Society Press, Los Alamitos (1994)
2. Sun, H.M., Yeh, H.T.: Password-based authentication and key distribution protocols with perfect forward secrecy. Journal of Computer and System Sciences 72, 1002–1011 (2006)

# Using Genetic Algorithms for Personalized Recommendation

Chein-Shung Hwang<sup>1</sup>, Yi-Ching Su<sup>2</sup>, and Kuo-Cheng Tseng<sup>2</sup>

<sup>1</sup> Dept. of Information Management,  
Chinese Culture University, Taipei, Taiwan

<sup>2</sup> Dept. of Information Management,  
Chinmin Institute of Technology, Miaoli, Taiwan  
cshwang@faculty.pccu.edu.tw, poohhh@ms.chinmin.ed.tw,  
tikic@ms.chinmin.edu.tw

**Abstract.** With the high-speed development of customer service orientation, it is essential that the enterprises must find and understand customers' interests and preferences and then provide for suitable products or services. Recommender systems provide one way of circumventing this problem. This paper describes a new recommender system, which employs a genetic algorithm to learn personal preferences of customers and provide tailored suggestions.

**Keywords:** Recommender systems; generic algorithm; collaborative filtering.

## 1 Introduction

The explosive growth of the world-wide-web has led to an influx of users and consequently, a huge increase in the volume of available on-line data. The volume of things is considerably more than any person can possibly filter through to find the ones that he/she will like. Recommender systems have emerged in response to the information overloaded problem. Most Personalized Recommender systems adopt two types of techniques: collaborative filtering approach and content-based approach. Collaborative filtering approach finds other users that have shown similar tastes to the given user and recommends what they have liked to that user. But it is not well-suited to locating information for a specific content information need. On the other hand, content-based approach recommends items based on the item contents that the user has liked in the past. Combining with content-based approach can eliminate the shortcomings of collaborative filtering approach and provide better recommendations [1].

Many hybrid recommender systems have been developed for e-commerce applications. The typical steps of recommender systems can be described as follows. First, customer preference profiles in terms of product features are analyzed and extracted from transaction file and product file. Second, a data mining technique is used to find similar customers who have shown similar interests as on-line customers. Finally, a list of recommendations is provided and can be further adjusted by the subsequent customers' feedbacks. However, for different application strategies, each preference

feature may be associated with different importance. Most of the systems either ignored or used a fixed weight for each feature, which often caused a poor recommendation performance.

Genetic algorithms are adaptive algorithms based on the Darwinian principle of natural selection and are often used to solve optimization problems. In this paper, we propose a hybrid recommender system which uses genetic algorithms for feature weighting. The proposed system consists of three modules. In PGM (Profile Generation Module), the customer's transaction data are analyzed to establish the customers' preference profile candidate table. In NSM (Neighborhood Selection Module), a clustering method is first adopted to segment customers into groups using the profile candidate table. The genetic algorithm is then used to fine-tune profile matching for each active customer. Finally, in the RC Model (Recommendation Module), a list of recommendation is derived and presented. This will enable the recommender system to make more accurate predictions of users' likes and dislikes, and hence will provide better recommendations to users.

## 2 Research Background

Recommender systems have been successfully applied in a number of difference applications such as recommending movies, books, music and products. There are two major techniques used in recommender systems [2-4], content-based approach and collaborative filtering approach.

### 2.1 Collaborative Filtering

The term collaborative filtering was coined by the Goldberg et al. [5] , the developers of the first recommender system – Tapestry. Tapestry allows users to annotate documents that they read. Users can then retrieve a document based on the content of the document or other users' opinions in terms of annotation on that document. However, the recommendations are not automated and require users to explicitly define their collaborative relationships. GroupLens[6][7] provides an automated recommendations using a neighborhood-based algorithm. The system uses the ratings of items to find people who are most similar to you and use their opinions for recommendations. GroundLens provides personalized predictions for Usenet news articles, while other systems use this approach for recommending movies, music, jokes and web pages.

The original collaborative filtering algorithm contains two main steps: *neighborhood formation* and *generation of recommendation*. Neighborhood formation finds a set of users known as neighbors that have similar preference ratings. Common similarity metrics used include Pearson correlation, mean squared difference, and vector similarity. In the generation of recommendation step, the system then computes the predicted ratings on items the active user has not yet seen based on his neighbors' ratings for those items. Finally, the system derives and sorts a set of recommendations by the predicted ratings.

## 2.2 Content-Based Recommendation

Content-based recommendations are based on information on the content of items rather than on other users' opinions. Content-based approach to recommendations has been adopted in information retrieval community and employs similar techniques. In content-based recommendations, every item is represented by a feature vector or an attribute profile. Each feature can be based on a numerical or nominal scale representing different information of items such as color, type, or price. A dissimilarity/similarity measure may be used to measure the similarity of features between items. However, people may place different importance on different attributes when judging the similarity between items. Each feature may be assigned a weight representing the importance of users toward that feature. Accordingly, the overall similarity between items is measured as the weighted sum of individual attributes. One drawbacks of content-based approach is that users need to explicitly specify the weights of features.

## 2.3 Genetic Algorithms

Genetic algorithms (GAs) [8][9] are stochastic search techniques that guide a population of solutions using the principles of evolution and natural genetics. Extensive research has been performed exploiting the robust properties of genetic algorithms and demonstrating their capabilities across a broad spectrum of optimization problems, including feature selection and weighting tasks. GAs are modeled loosely on the principles of the evolution via natural selection, employing a population of individuals that undergo selection in the presence of variation-inducing operators such as mutation and crossover. A fitness function is used to evaluate individuals, and reproductive success varies with fitness.

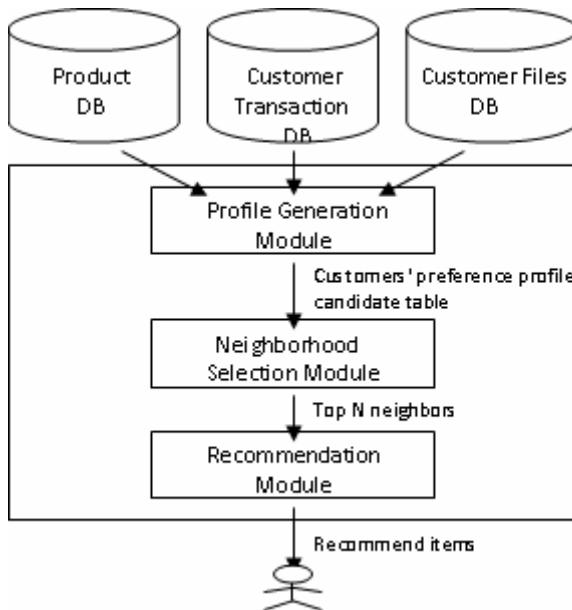
A general algorithm is started with a set of solutions (represented by chromosomes) called population. An initial population is created from a random selection of solutions. At every evolutionary step (generation), the solutions in the current population are evaluated according to some predefined quality criterion, referred to as the *fitness*, or *fitness function*. Solutions from one population are taken and used to form a new population (next generation). Solutions (parents) are selected according to their fitness - the more suitable they are the more chances they have to reproduce. These solutions then "reproduce" to create one or more new solutions (offspring), after which the offspring are produced by *crossover* or *mutation* randomly. The process of fitness-dependent selection and application of genetic operators to generate successive generations of solutions is repeated many times until a satisfactory solution is found. The basic steps of GAs are outlined as follows:

1. [Initialization] Randomly generate an initial population of solutions and evaluate the fitness function
2. [New population] Create a new population by repeating following steps until the new population is complete
  - 2.1[Selection] Select two parent solutions from a population according to their fitness (the better fitness, the bigger chance to be selected)

- 2.2[Crossover] With a crossover probability cross over the parents to form a new offspring. If no crossover was performed, offspring is an exact copy of parents.
- 2.3[Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).
- 2.4[Accepting] Place new offspring in a new population
3. [Evaluation] Compute the fitness values for the new population of N solutions
4. [Test] If the stopping criterion is met, stop, and return the best solution in current population
5. [Loop] Go to step 2.

### 3 System Architecture

In this paper, we propose a hybrid recommender system that uses genetic algorithms for feature weighting to find similar customers who may share the same interests as the active customers, and capture the potential needs of customers. The proposed system consists of three modules, as shown in Fig. 1.



**Fig. 1.** System architecture

#### 3.1 Profile Generation Module ( PGM )

The goal of PGM is to create the preference profile for each customer. The first step in PGM is to build the product profile from the product database. Each product profile is characterized by its feature values and defined as a binary vector as

$$P_j = (f_j^1, f_j^2, \dots, f_j^n) \quad (1)$$

where  $n$  is the number of product features ( $n=12$ , in our experiment.). Each feature is assigned a value of 1 if a product possesses that feature, and 0 otherwise.

To fully understand which product features a customer is of interest, in the second step, we compute the customer product preference profile from the customer transaction data and the product profile. The product preference profile of customer  $k$  is described as

$$CPP_k = (CPP_k^1, CPP_k^2, \dots, CPP_k^n) \quad (2)$$

where  $CPP_k^i$  represents the preference profile of feature  $i$  for customer  $k$  and is obtained by summing up the feature information from the products purchased by customer  $k$ .  $CPP_k^i$  can be defined as

$$CPP_k^i = \frac{\sum_{j \in T_k} f_j^i}{|T_k|} \quad (3)$$

where  $T_k$  represents the set of products purchased by customer  $k$ .

Finally, the customer preference profile is generated by integrating three portions of information: the customer product preference profile, the customer information and customer purchasing behavior. The customer preference profile is defined as

$$\begin{aligned} CP_k &= (CPP_k, CI_k, CT_k) \\ &= (CP_k^1, CP_k^2, \dots, CP_k^{17}) \end{aligned} \quad (4)$$

where  $CI_k = (CI_k^1, CI_k^2)$  contains customer  $k$ 's information including *member level* and *gender* and  $CT_k = (R_k, F_k, M_k)$  contains the information about customer  $k$ 's purchasing behavior measured by the RFM (*recency*, *frequency*, and *monetary*) information. Table 1 shows an example of the customer preference profile.

**Table 1.** The Customer Preference Profile

$C\_no$	$f^1$	$f^2$	...	$f^{12}$	$level$	$sex$	$Recency$	$Frequency$	$Monetary$
002	1	0.33	...	0.33	0	1	0.078	0	0.019

### 3.2 Neighborhood Selection Module ( NSM )

In NSM, firstly, the GAs is used to fine-tune the feature weights for each active customer. Then, the collaborative filtering approach is applied to form the neighborhood of the active customer.

The chromosome in the GA process is represented as a weight vector with 17 genes. Each gene is encoded with 8 bits. The GA begins with random genotypes and an initial population of 100 chromosomes. For each active customer, a randomly selected chromosome is assigned and tested by the fitness function. The fitness function measures the prediction accuracy of products based on the current chromosome.

$$Accuracy(k) = 1 - \frac{\sum_{j=1}^l |P(k, j) - A(k, j)|}{l} \quad (5)$$

Each active customer  $k$  is tested by a random selection of  $l$  products.  $P(k, j)$  and  $A(k, j)$  are the predicted and the actual ratings of customer  $k$  to product  $j$ , respectively. The predicted ratings are calculated by the collaborative filtering algorithm with different neighborhood size.

The algorithm continues to evolve until the termination criteria are met. In our experiment, we set the maximum generation number to 100. For each generation evolution, chromosomes for the next generation are selected using the roulette wheel selection scheme to implement proportionate random selection. All of the chromosomes are then paired up using the single-point crossover strategy with a probability of 0.9. After the crossover, for each of the genes of the chromosomes, the gene is mutated with a probability of 0.05.

After obtaining customer's best feature weights, we can now select the most similar  $n$  neighbors (denoted as  $NB_k$ ) by computing the similarity value.

$$Similarity(k, a) = 1 - \sqrt{\frac{\sum_{i=1}^{17} W_k^i \times (CP_k^i - CP_a^i)^2}{\sum_{i=1}^{17} W_k^i}} \quad (6)$$

where  $W_k = (W_k^1, W_k^2, \dots, W_k^{17})$  is the feature weights of customer  $k$  obtained from GAs.

### 3.3 Recommendation Module ( RC )

RC module recommends products for active customer  $k$  by collecting the information from his/her neighbors. For each product  $j$  we compute its recommendation score as

$$Score(k, j) = \sum_{i \in NB_k} Similarity(k, i) \times purchase(i, j) \quad (7)$$

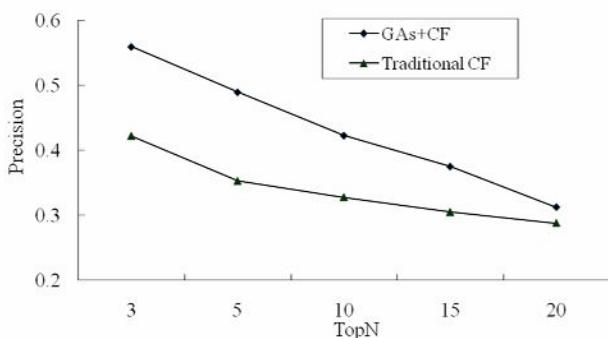
where  $purchase(i, j)$  is set to 1 if customer  $i$  has purchased product  $j$ , 0 otherwise. Finally, all products are sorted in non-increasing order with respective to the recommendation score, and the first  $N$  items are selected as the Top- $N$  recommendation set.

## 4 Experimental Evaluation

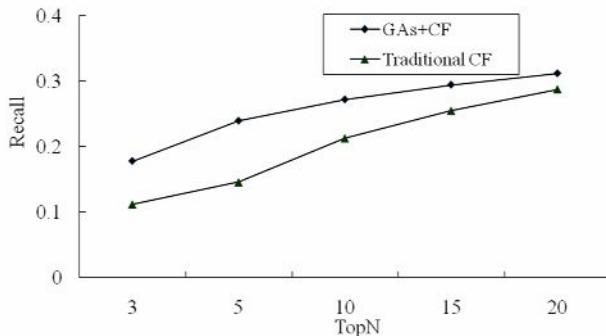
In this study, we use the telemarketing dataset collected by one of the telemarketing company in Taiwan. The dataset was collected over two years from January 2007 to April 2009. After data preprocessing, it contains 15,376 transaction records from 753 users for 239 products. Each user has bought at least 5 transaction records, and each product has been bought at least once.

We employ the 5-fold cross-validation approach and use the precision metric, recall metric and F1-measure metric to evaluate the quality of a recommendation. Precision is the percentage of total number of recommendations that the customer interesting. Recall is the percentage of the customer interesting that we recommend and the customer also interesting. F1-measure is the index that is combination of precision and recall.

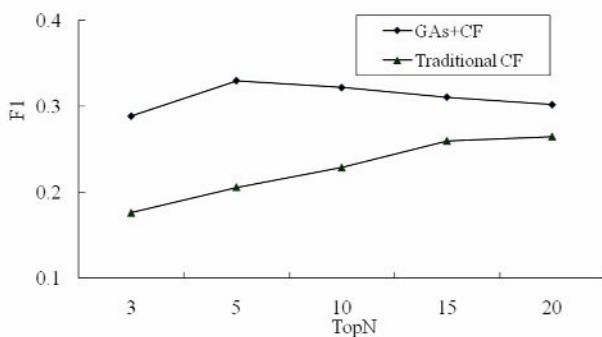
To compare our approach with the collaborative filtering approach, we varied the number of recommendation 3, 6, 10, and 15. For our approach, we set the initial set of population to 100 and the number of generation to 80. Fig. 2 to Fig. 4 show the performance comparisons between our approach and the collaborative filtering approach. As expected, when the number of recommendations increases, the precision drops smoothly but the recall improves gradually. F1-measure also suggests a best recommendation size of 5. However, it can be observed that our approach outperforms the collaborative filtering approach at all values of Top-N.



**Fig. 2.** Comparison of precision between the proposed approach and the traditional CF algorithm



**Fig. 3.** Comparison of recall between the proposed approach and the traditional CF algorithm



**Fig. 4.** Comparison of F1 between the proposed approach and the traditional CF algorithm

## 5 Conclusions

In this paper, we have proposed a hybrid recommender system based on GAs and collaborative filtering technique. The system integrates data from various sources (product, customer, and transaction data) to form the customer preference profile. The GAs are applied to optimize a vector of the feature weights, which are used to measure the similarity among customers. Incorporating weighting information into the collaborative filtering process has proven to be more effective than traditional one.

## References

1. Schafer, J.B., Konstan, J., Riedl, J.: Electronic commerce recommender applications. *Journal of Data Mining and Knowledge Discovery* 5(1/2), 115–152 (2000)
2. Karypis, G.: Evaluation of item-based top-n recommendation algorithms. In: Proceedings of the 10th International Conference on Information and Knowledge Management, pp. 247–254 (2001)

3. Shardanand, U., Maes, P.: Social information filtering: algorithms for automating ‘word of mouth’. In: Proceedings of the Conference on Human Factors in Computing Systems (CHI 1995), pp. 210–217 (1995)
4. Weng, S.-S., Liu, M.-J.: Feature-based recommendations for one-to-one marketing. *Expert Systems with Applications* 26(4), 493–508 (2004)
5. Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information Tapestry. *Communications of the ACM* 35(12), 61–70 (1992)
6. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: Proceedings of ACM Conference on Computer Supported Cooperative World, pp. 175–186 (1994)
7. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: GroupLens: applying collaborative filtering to usenet news. *Communications of the ACM* 40(3), 77–87 (1997)
8. Srinivas, M., Patnaik, L.M.: Genetic algorithms: A survey. *IEEE Computer* 27(6), 17–26 (1994)
9. Davis, L. (ed.): *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York (1991)

# Pretreatment for Speech Machine Translation

Xiaofei Zhang<sup>1</sup>, Chong Feng<sup>2</sup>, and Heyan Huang<sup>2</sup>

<sup>1</sup> Research Center of Computer and Language Information Engineering,  
Chinese Academy of Sciences, Beijing, China, 100097  
[zxlflying@gmail.com](mailto:zxlflying@gmail.com)

<sup>2</sup> School of Computer Science and Technology, Beijing Institute  
of Technology, Beijing, China, 100081  
[{chong.feng,heyan.huang}@bjtu.edu.cn](mailto:{chong.feng,heyan.huang}@bjtu.edu.cn)

**Abstract.** There are many meaningless modal particles and dittographies in natural spoken language, furthermore ASR (automatic speech recognition) often has some recognition errors and the ASR results have no punctuations. And thus the translation would be rather poor if the ASR results are directly translated by MT (machine translation). Therefore, it is necessary to transform the abnormal ASR results into normative texts to fit machine translation. In this paper, a pretreatment approach which based on conditional random field model was introduced to delete the meaningless modal particles and dittographies, correct the recognition errors, and punctuated the ASR results before machine translation. Experiments show that the MT BLEU of 0.2497 is obtained, that improved by 18.4% over the MT baseline without pretreatment.

**Keywords:** Speech machine translation; automatic speech recognition; conditional random field model; pretreatment.

## 1 Introduction

Compared to traditional text machine translation whose input text has normative style of writing, speech machine translation is more difficult. The main challenges are: (1) there are many meaningless modal particles and dittographies in natural spoken language; (2) ASR results have no punctuations; (3) even the state-of-art ASR often has some recognition errors, etc. al. These all badly deteriorate the performance of the machine translation. The following examples extracted from a practical bilingual phone question answering system demonstrate the depressed phenomena.

Example 1:

ASR result: 你 近 店 (Machine translation: Feed you fine I is it ask acoustics nearby hotel of research institutes to want).

Normative sentence: 你 近 店 (Machine translation: Hello! I want to ask the nearby hotel of research institutes of acoustics.)

Example 2:

ASR result:

(Machine translation: You that there is China Merchants Bank of subbranch of Tsing-Hua University of bank of Beijing in the bank near Beijing University want every one).

Normative sentence:

支

? (Machine translation: The banks near Peking University are as follows, Beijing Bank Tsing-Hua University Subbranch and China Merchants Bank, Which one would you like?).

Example 3:

ASR result: 喔 和 我 么 么 过去 (Machine translation: Ah how I want to get there by bus oh in the Summer Palace).

Normative sentence: 和 我 么 过去 ? (Machine translation: Ah, I am in the Summer Palace, how can I get there by bus? ).

From the above examples it can be seen that the translations would be intricate even sometimes very laughable if the ASR results are directly translated by MT. But, after simple correcting and adding punctuations to the ASR results, the translation would be improved greatly. That is especially for RBMT (rule-based machine translation), since RBMT needs to parse the source language. The abnormal phenomena such as having a redundant word or lacking of a word, and having no punctuations all maybe result in failure in parsing and thus generate wrong translation. If the ASR result was transformed into texts with normative style of writing before translation, then the strength of RBMT in parsing would be exerted and good translation could be obtained.

The aim of pretreatment is to transform the abnormal ASR results into normative texts to fit machine translation. Nowadays, there have been many researches focusing on abnormal phenomena of spoken language. To perform the task of dynamic chat language term normalization, reference [1] propose the phonetic mapping models to present mappings between chat terms and standard words via phonetic transcription. Reference [2] described a SDS (Spoken Dialogue Systems) toolkit which deals with Chinese dialogue and makes it easy for not only experts, but also lay people to learn and create their own SDS. Reference [3] found it was important to have acoustic and language models, and statistical pronunciation generation rules adapted to the Northern and Southern varieties of Dutch. Reference [4] presents the overall architecture, the user interface, the design and implementation of the speech recognition grammars, and initial performance results indicating that for sentence level utterance recognition they achieve 60 to 65% of human capability.

## 2 Translation Pretreatment

As mentioned above, the aim of pretreatment is to transform the abnormal ASR results into texts with normative style of writing to fit machine translation. Let's observe the human's action firstly. Manually transforming abnormal ASR result into

normative texts mainly includes several operations such as deletion, insertion, and correction, et al. Here is an example given to demonstrate the transforming process.

ASR result (after segment): 喔 和 要 么 么 过去呢

## 2.1 Deletion Operation

Delete the meaningless modal particles and ditto graph: In this example, delete the second 和 the second 么, and then the sentence was transformed into “ 喔 和 要 么 过去呢 ”.

## 2.2 Insertion Operation

Insert punctuations into the sentence: In this example, insert commas after 和 and 分别, and insert a question mark after 呢, and then the sentence was transformed into “ 喔 和 要 么 过去呢 ”.

## 2.3 Correction Operation

Correct the recognition errors in ASR results: In this example, transform 喔 into 呀, and then the sentence was transformed into “ 呀 和 要 么 过去呢 ”.

Thus, the sentence finally becomes a normative sentence suitable for machine translation.

Transforming the abnormal ASR results into normative texts involve many complicated contextual information such as word partnerships, word position, length of sentence, and the type of the sentence, and even involve some acoustic features such as interval times of the voice and the tone of voice, et al. How to considerate overall those multiple factors is a key problem need to resolve. The conditional random fields (CRFs) models don't force to adhere to the independence assumption such as in Hidden Markov generative models and thus can depend on arbitrary, non-independent features, which are benefit to the specific task, without accounting for the distribution of those dependencies. Second, since CRFs model is able to flexibly utilize a wide variety of features, the sparse problem of training data can also be resolved efficiently. Furthermore, the parameter estimation for CRFs is global, which effectively resolve the label bias problem as in maximal entropy model. Therefore, CRFs model is a feasible choice for our task.

Actually, early in 2001 Lafferty et al. had utilized CRFs models to carry out POS tagging experiments [5]. Especially in recent years, the CRFs have been widely applied in the area of natural language processing such as sentence parsing [6], Chinese word segmentation [7], automatic speech recognition [8], word alignment [9], and named entity recognition [10], and so on.

Based on conditional random fields model, the above pretreatment process can be converted into a sequence labeling task. The pretreatment labeling result was shown as follows:

/I\_ 喔/C\_ \_ NIL /NIL 和 / I\_ /D /NIL 么/NIL 么/D  
 / NIL 过去/NIL 呢/I\_

Where the label  $I_$  indicates inserting a comma after the current word, the label  $C_$  \_ NIL indicates transforming the current word into \_ but without punctuation need to insert, the label D indicates deleting the current word, and the label NIL means nothing to do.

### 3 Pretreatment Algorithm

Give a sentence, which consist of word sequence  $w_1 w_2 \cdots w_i \cdots w_L$ , assume its corresponding pretreatment labeling sequence is  $y_1 y_2 \cdots y_i \cdots y_L$ . The task of CRFs - based pretreatment is to find a label sequence  $y_1 y_2 \cdots y_i \cdots y_L$  to make the probability  $p(y_1 y_2 \cdots y_i \cdots y_L | w_1 w_2 \cdots w_i \cdots w_L)$  maximal. This procedure can be formulated as

$$Y^* = \arg \max_{y_1 y_2 \cdots y_L} p(y_1 y_2 \cdots y_i \cdots y_L | w_1 w_2 \cdots w_i \cdots w_L) \quad (1)$$

To determine the pretreatment label of a word in a given sentence involves many factors. Assume that random variable  $x$  represents these factors and the random variable  $y$  represents the pretreatment label, and then  $p(y|x)$  represents the probability of a word with label  $y$  in the given sentence. This probability can be evaluated based on maximum entropy principle which requires that the probability  $p(y|x)$  should make the entropy defined as following formula (2) get maximum under the condition of some constrains.

$$H(p) = - \sum_{x,y} p(y|x) \log p(y|x) \quad (2)$$

Here constrains are all the known cases, and according to first-order linear-chain-structured CRFs the constrains can be represented by the following formula (3) and (4)

$$f_k(x_i, y_{i-1}, y_i) = \begin{cases} 1, & \text{if } \{x_i, y_{i-1}, y_i\} \text{ satisfying contrains} \\ 0, & \text{else} \end{cases} \quad (3)$$

$$g_k(x_i, y_i) = \begin{cases} 1, & \text{if } \{x_i, y_i\} \text{ satisfying contrains} \\ 0, & \text{else} \end{cases} \quad (4)$$

Where,  $y_{i-1}$  represent the previous word's pretreatment label.  $f_i(x, y)$  and  $g_k(x_i, y_i)$  are all called the features of the CRFs model. These features described the relations between the contextual factors  $x_i$ , previous word's pretreatment label  $y_{i-1}$  and the current word pretreatment label  $y_i$ . It is has been proved in mathematics theory that

the probability  $p(y|x)$  must adhere to the following exponential distribution to accord with the requirement of the maximum entropy principle.

$$p_\theta(y_i | x_i) \propto \exp\left(\sum_k \lambda_k f_k(x_i, y_{i-1}, y_i) + \sum_k u_k g_k(x_i, y_i)\right) \quad (5)$$

The model parameters  $\lambda_i$  and  $\mu_k$  can be trained by L-BFGS, GIS or IIS algorithm [5] [11]. And to avoid overfit learning, Gaussian prior smoothing as in [12] [13] should be applied too.

So, according to first-order linear-chain-structured CRFs theory, the task of pretreatment labeling described in formula (1) can be represented by

$$\begin{aligned} Y^* &= \arg \max_{y_1, y_2, \dots, y_L} p(y_1, y_2, \dots, y_L | w_1, w_2, \dots, w_L) \\ &= \arg \max_{y_1, y_2, \dots, y_L} \prod_{i=1}^L p(y_i | x_i) \\ &= \arg \max_{y_1, y_2, \dots, y_L} \left\{ \prod_{i=1}^L \exp\left(\sum_k \lambda_k f_k(x_i, y_{i-1}, y_i) + \sum_k u_k g_k(x_i, y_i)\right) \right\} \end{aligned} \quad (6)$$

In formula (6), we can consider computing the logarithm of  $p(y_i | x_i)$  in stead of computing  $p(y_i | x_i)$  directly. This is why the CRFs is also called a kind of linear-logarithm model. Thereafter, the CRFs-based pretreatment labeling can be further simply represented by

$$Y^* = \arg \max_{y_1, y_2, \dots, y_L} \sum_{i=1}^L \left( \sum_k \lambda_k f_k(x_i, y_{i-1}, y_i) + \sum_k u_k g_k(x_i, y_i) \right) \quad (7)$$

## 4 Decoding Algorithm

For decoding formula (7), if labeling the current word need to consider the previous words' labeling results, i.e. the state transfer of adjacent elements follows Markov Process, then it can be decoded by Viterbi algorithm to get the globally optimal solution.

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the Viterbi path – that results in a sequence of observed events, especially in the context of Markov information sources, and more generally, hidden Markov models.

Given the model parameters  $\lambda, \mu$  and hidden state set  $S = \{s_1, s_2, \dots, s_N\}$ , a sequence of observed events  $X = x_1 x_2 \dots x_L$  (where  $L$  is the length of the sentence). Then the probability of obtaining the sequence of hidden states  $Y = y_1 y_2 \dots y_L$  (where  $y_i \in S$ ) can be represented by  $P(Y | X, \lambda, \mu)$ .

Define the following local optimal function:

$$\delta_t(s_i) = \max_{y_1 y_2 \cdots y_{t-1}} P(y_1 y_2 \cdots y_{t-1}, y_t = s_i | x_1 x_2 \cdots x_t, \lambda, \mu) \quad (8)$$

The function  $\delta_t(s_i)$  is to find the most likely sequence  $y_1, y_2, \dots, y_{t-1}$  which make the probability  $P(y_1 y_2 \cdots y_{t-1}, y_t = s_i | x_1 x_2 \cdots x_t, \lambda, \mu)$  maximize under the condition of giving the model parameters  $\lambda, \mu$  and some parts of observed sequence  $x_1 x_2 \cdots x_t$ .

In addition, to backward find the optimal state sequence, the local optimal path to  $s_i$  in position t should be recorded. We use an array  $\Psi_t(s_i)$  to record the path.

The Viterbi algorithm detail as follow:

Step 1: Calculate the initial values of the local optimal function.

$$\delta_1(s_i) = \sum_k u_k g_k(x_1, y_i = s_i), \quad 1 \leq i \leq N \quad (9)$$

$$\Psi_1(s_i) = 0 \quad (10)$$

Step 2: recursive calculation

$$\delta_t(s_i) = \max_{1 \leq j \leq N} \left\{ \delta_{t-1}(s_j) + \left( \sum_k \lambda_k f_k(x_t, y_{t-1}, y_i = s_i) + \sum_k u_k g_k(x_t, y_i = s_i) \right) \right\} \quad (11)$$

$$\Psi_t(s_i) = \arg \max_{1 \leq j \leq N} \left\{ \delta_{t-1}(s_j) + \sum_k \lambda_k f_k(x_t, y_{t-1} = s_j, y_i = s_i) \right\} \quad (12)$$

Where  $2 \leq t \leq L, 1 \leq j \leq N$

Step 3: calculate the last optimal state

$$s_T^* = \arg \max_{1 \leq j \leq N} [\delta_T(s_j)] \quad (13)$$

Step 4: backward find the global optimal sequence of states

$$s_t^* = \Psi_{t+1}(s_{t+1}^*), \quad t = L-1, L-2, \dots, 1 \quad (14)$$

Thus, according to the formula (14) we can step by step get the previous optimal state  $s_t^*$  from the current optimal state  $s_{t+1}^*$ , and finally the global optimal sequence of states can be obtained.

## 5 Feature Template

Pretreatment for speech machine translation involve many factors. So we should reasonably select features which benefit to pretreatment. Feature templates used in this paper are presented in table 1.

**Table 1.** Feature templates

ID	Feature	Remark
1	current word $w_i$	The word itself
2	Word position $p_{i-1}$	The position of current word in the sentence
3	Interval time $t_i$	interval times of the voice between two words
4	Label restrict	For example, current label must match previous label
5	previous word $w_{i-1}$	Constrains on context
6	next word $w_{i+1}$	Constrains on context
7	$w_{i-1} \& w_{i+1}$	Constrains on context

## 6 Experiments

To validate the effect of our CRFs-based pretreatment algorithm, we have designed two group contrast tests.

### 6.1 Experimental Corpora

We totally have extracted 10000 ASR sentences from a practical bilingual phone question answering system which mainly provide the service of asking the way in Beijing region. Thereinto, 8000 sentences serve as training set, while the other 2000 sentences as test set.

### 6.2 Test for Sentence Pretreatment Effect

The tests for sentence pretreatment effect divide into close test and open test. In close test, the 8000 sentences training set is also served as the test set. While in the open test, 8000 sentences serve as training set, while the other 2000 sentences as test set. The experimental results are shown in the table 2.

**Table 2.** Sentence pretreatment effect

Test type	Rate of normative sentence (%)	
	Original ASR	After pretreatment
Close test	9.0	59.7
Open test	8.4	42.5

It can be seen from the table that the normative sentence rate increased to 59.7% from 9.0% in the close test that increased by 50.7% compared with original ASR result baseline. Even in the open test, the normative sentence rate also increased to 42.5% from 8.4% that increased by 34.1% compared with original ASR result

baseline. Indeed, the normative sentence rate is still relatively low, but it is improved considerably from original ASR result. The tests also show that the ASR technology needs to further develop.

### 6.3 Test for Machine Translation Effect

In test for machine translation effect, mteval-v13 which is the tool of NIST Open Machine Translation 2009 Evaluation is used to score BLEU. The experimental results are shown in the table 3. Note: our MT system is RBMT.

**Table 3.** Machine translation effect

MT result	Test type	
	Translate directly	After pretreatment
BLEU	0.2038	0.2497

It can be seen from the table that the MT BLEU increased to 0.2497 from 0.2038 that increased by 18.4% compared with baseline MT without pretreatment. The positive experimental results confirm the effect of our pretreatment algorithm.

## 7 Conclusion and Further Study

In this paper, we convert the pretreatment problem into a task of sequence labeling. Pretreatment for speech machine translation involve many complicated contextual information such as word partnerships, word position, length of sentence, and the type of the sentence, and even involve some acoustic features such as voice interval times and the tone of voice, et al. Since the conditional random fields models don't force to adhere to the independence assumption such as in Hidden Markov generative models and thus can depend on arbitrary, non-independent features, which are benefit to the specific task, without accounting for the distribution of those dependencies. Second, since CRFs model is able to flexibly utilize a wide variety of features, the sparse problem of training data can also be resolved efficiently. Furthermore, the parameter estimation for CRFs is global, which effectively resolve the label bias problem as in maximal entropy model. The positive experimental results confirm that CRFs model is a feasible choice for our task.

Further study can focus on the design of feature template and improvement of feature selection algorithm. After all pretreatment for speech machine translation involves many factors, while more factors are considered, the accuracy should be improved further. Secondly, feature selection is also very important. Good feature selection algorithm can not only decrease the number of the features to improve the decoding speed, but also improve the accuracy through reserving the features with high discriminative degree and discarding the ones with low discriminative degree.

**Acknowledgments.** This research is supported by the National Science Foundation of China under Grant No. 60672149 and national 863 program under grant No. 2006AA010109.

## References

1. Xia, Y.Q., Wong, K.F., Li, W.J.: A Phonetic-based Approach to Chinese Chat Test Normalization. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 993–1000 (2006)
2. Liu, Z., Brasser, M., Zheng, T.F., Xu, M.: A New Implementation Approach of Grammar Generation for Text-based SDS. Computer science 11, 205–209 (2006)
3. Despres, J., et al.: Modeling northern and southern varieties of Dutch for STT. In: Inter-speech 2009, Brighton, UK, pp. 96–99 (September 2009)
4. Acero, A., Bernstein, N., Chambers, R., Jui, Y.C., Li, X., Odell, J., Nguyen, P., Scholz, O., Zweig, G.: Live search for mobile: Web services by voice on the cellphone. In: Proc. of ICASSP (2008)
5. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of ICML (2001)
6. Finkel, J.R., Kleeman, A., Manning, C.D.: Efficient, feature-based, conditional random field parsing. In: Proc. ACL/HLT (2008)
7. Zhao, H., Huang, C.N., Li, M.: An improved Chinese word segmentation system with conditional random field. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language (2006)
8. Hifny, Y., Renals, S.: Speech Recognition using Augmented Conditional Random Fields. IEEE Transactions on Audio, Speech, and Language Processing 17(2), 354–365 (2009)
9. Blunsom, P., Cohn, T.: Discriminative word alignment with conditional random fields. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (2006)
10. Watanabe, Y., Asahara, M., Matsumoto, Y.: A Graph-based Approach to Named Entity Categorization in Wikipedia Using Conditional Random Fields. In: Proc. of EMNLP-CoNLL (2007)
11. Berger, A.L., Della Pietra, V.J., et al.: A maximum entropy approach to natural language processing. In: Computational linguistics (1996)
12. Chen, S.F., Rosenfeld, R.: A Gaussian prior for smoothing maximum entropy models. In Technical Report CMUCS (1999)
13. Zhang, X.-f., Zhan, L., Huang, H.-y.: The Application of CRFs in Part-of-Speech Tagging. In: Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC 2009), Hangzhou, China, August 26-27, vol. 2, pp. 347–350 (2009)

# Emotion Aware Mobile Application\*

Radoslaw Nielek<sup>1</sup> and Adam Wierzbicki<sup>2</sup>

<sup>1</sup> Polish Japanese Institute of Information Technology, ul. Koszykowa 86, 02-008  
Warszawa, Poland

[nielek@pjwstk.edu.pl](mailto:nielek@pjwstk.edu.pl)

<sup>2</sup> Polish Japanese Institute of Information Technology, ul. Koszykowa 86, 02-008  
Warszawa, Poland  
[adamw@pjwstk.edu.pl](mailto:adamw@pjwstk.edu.pl)

**Abstract.** In recent years content aware mobile application are boomed. Both, research activities and business focus are kept on the information about physical space understand as proximity (Bluetooth), geographical location (GPS) and movement (accelerometer). Growing computational power of mobile devices allow researchers to make a step further and design application which “feel what their user feels”. The paper promotes an idea of the emotional aware mobile application as a natural next step in content awareness. A novel framework for developing emotion aware mobile applications is presented and possible sources of emotion rich information in mobile environment are identified and pointed out. Information about user’s mood can be used in a variety of application. Some of them like filtering information or tracking stressing situation are described. As a proof-of-concept two application were created.

## 1 Introduction

Multi-user operating systems for mobile phones which allow easy user switching doesn’t exist. It is probably the most striking proof of the fact that mobile phones are exclusively personal devices. In opposite to personal computers we don’t share one device but, still, we don’t go with personalization beyond theme (skin) selection, accounts’ settings and some additional features like five most frequently called numbers.

Supporting and extending personalization is one of main goals of context-aware systems and application but most researches have been focused on a physical or social context[6]. The true personalization can be achieved only by a device which “feels what user feels”. Emotion-aware mobile device followed by emotion-aware services and application are a natural next step in context aware researches. It is worth to notice that already in 1998 Dey [1] listed emotional state of a user as an example of context information but since then not much progress towards sensing mobile device has been done. It is so mostly because the internal state of a person who use a mobile device was usually omitted as difficult to measure (compare to localization or temperature) and commercially

---

\* This research has been supported by the Polish Ministry of Science grant 69/N-SINGAPUR/2007/0.

not very useful. The title of the workshop held during MobileHCI '09 conference "Measuring Mobile Emotions: Measuring the Impossible?" [8] perfectly describes concerns shared by the majority of the research community. Lack of good objective measures dedicated to emotional states (it is questionable if they can exist at all) and subjectivity component make results difficult to compare and to justify particular, chosen approach or algorithm.

Problems with an evaluation of the proposed solution are not only limited to measuring emotion. Mizzaro et. al. [2] claim that even for a localization based services we don't have good benchmarks which allows us to compare the quality of action-event mapping. Interesting approach was proposed by Oulasvirta [3]. Instead of looking for objective benchmark to compare different solution author pinpointed the dominance of the human needs and expectation over technology-driven innovation. In the present-day world innovation is driven neither by technology existing in laboratories nor by pure human needs. In applied computer science innovation are motivated by business. Thus a persuasive use-cases and a proof that proposed solution can be developed with use of existing technology are needed to cause a large-scale deployment.

The approach used in this paper is closer to a business people viewpoint. Developed application work on ordinary smartphone and are dedicated for Android – a publicly available mobile operating system. EmotionML , markup language proposed by W3C group, is used to tagging emotion and communication between processing and storing layer in framework. Developed solution is validated with use of SMS corpus collected at the Department of Computer Science at the National University of Singapore by volunteers.

Measuring, understanding and, even, influencing emotional state of end-users (especial in mobile environment) can be a basis for a variety of m-commerce services and products ranged from m-healthcare to matchmaking. Thus, in this paper a framework for emotional aware mobile applications is sketched. As a proof-of-concept two application are created. SMSemoAlerter that analyze the content of incoming SMS and based on them modify information about user's mood. The second application changes the wallpaper of the mobile phone following by the extracted emotional information. In section 2, collecting, processing and storing of emotional information is discussed. Section 3 is devoted to the presentation of the two proof-of-concepts. In the last section conclusion are drafted and some of the ideas for future are presented.

## 2 Collecting, Processing and Storing Emotional State

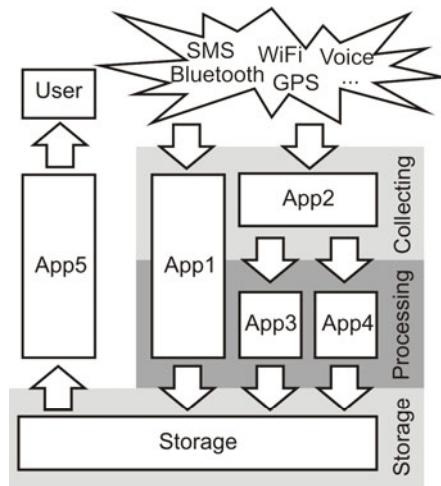
### 2.1 Introduction

The framework sketched in this section is composed of three separate layers. Each layer is focused around some important functions and, thus, can work at least semi-independent. Moreover, some algorithms and solutions are platform-independent and can be used universally on almost all present mobile devices (the only limitation is computational power). Another functions have to be developed for only one operating system or even a particular mobile phone.

Modern smartphones can (at least theoretically) hear what you hear, see what you see and even read what you read. So, the first step in developing a system feeling their users' emotion is collecting and preprocessing data (e.g. SMSs, background noises etc.) containing information about emotion. Such functionality is pursued by top layer in proposed framework and is strongly device-dependent.

Most promising piece of information are identified, filtered out and forwarded to the next layer for further in-depth processing. Algorithms used on this layer are universal and can be deployed on almost every device regardless of operating system. The bottom layer is responsible for calculating present user's mood based on information incoming from higher layer. Additional functionality of the last layer is affording of collected and computed information for different application.

The proposed framework is presented on fig. 1. Application developed within this framework are not limited to only one layer and can go across two (quite often) or even three (rare) layers. Services which only use information about user's mood but don't modify them can easily connect to the bottom layer (like App5 on fig. 1.).



**Fig. 1.** Sketch of the framework with three separate layers

## 2.2 Collecting Facts

A variety of data sources which can deliver some information about user's mood make collecting and preprocessing such data extremely difficult. Necessity to prepare special piece of software not only for every data source but often for many types of mobile devices make the task even more complex. In tab.1. a

selection of possible sources of emotion-rich information in mobile environment is presented. For each data source most promising tools and techniques are given (accompanied by an estimation of required computational power).

Even if for some approaches an efficient and computable algorithms exist it is not always an easy task to gather and processing right data in a real-time. To trace the background noises on a regular basis a microphone embedded into mobile phone has to be active all the time. It will open a lot of question about privacy and decrease the time the device will work on battery to an unacceptable small value. The same problem can be seen more vivid for video stream. On the other hand some emotional rich data like SMSes, email or a history of web browsing are accessible without much effort. Most mobile operating systems allow developers to register their own classes as a handler of particular event. Then an application is activated by the OS right after such event appeared.

Note that many application can be simultaneous registered to the same event, thus a way to integrate the results of emotion extraction algorithms is required to avoid a situation in which the same emotion could be counted twice. Therefore, the framework presented in this paper assume that on the top level only one module is responsible for collecting the data of particular type (let it be video, voice, emails or SMSes) and then, those collected facts are accessible for processing (extracting emotional attitude) on the lower layer.

### 2.3 Processing Information

Extracting or reasoning emotion from intercepted texts, voice, pictures or other behavioral data is an hard problem which require not only a lot of computational power but also many additional information about context. Constantly growing computational power embedded into mobile devices and intelligent algorithms which help reconstructing proper context are useful but, even then, many unsolvable problems still exist (e.g. subjectivity, ).

## 3 Proof-of-Concept

### 3.1 SMS EmoAlerter

**Dataset.** As a basis for validation the proposed solution SMS corpus collected at the Department of Computer Science at the National University of Singapore by volunteers. The corpus, which is distributed under Open Project Library License, contains 10,094 short messages written in English by 133 authors. Average message is composed of 58 characters (standard deviation for the whole dataset is rather big and equal to 39.5 characters). More in-depth analyses and statistics about used dataset can be found in [6].

**Text processing.** The idea to develop a light-weight version of sentiment extraction tool running on mobile phones was driven by the will to verify concepts

**Table 1.** Sources of emotion rich information in mobile environment

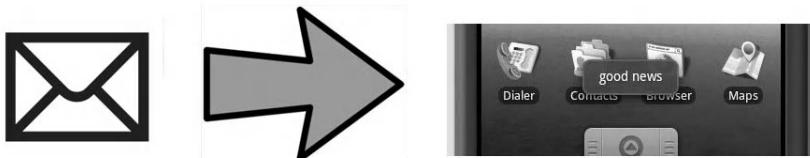
Sources of information	Tools and techniques	Required computational power
SMS, MMS, emails, Natural Language Processing tools, dictionaries, Hidden Markov Model and artificial neural networks, spell checker, etc.	Low to medium for well structured languages; high for languages with complex grammar,	Low to medium for well structured languages; high for languages with complex grammar,
Voice tone, background noises	Frequency analyzes, artificial intelligence, etc.	High (probably dedicated DSP required)
History of visited web pages	Classification and clustering, predefined categories, artificial intelligence, etc.	Low to very low if based on predefined lists of portals,
Proximity	Bluetooth, AI for automatic building of databases,	Low (energy hungry because of Bluetooth)
Localization	GPS and GSM cells, access points (WiFi)	Low but extensive databases are needed
Video, pictures	Embedded camera, AI (face detection)	Very high
Explicitly declaration	Survey	Very low
Implicitly declaration	Playlist selection etc.	Low

sketched in previous chapters. Our aim was to prove that mobile devices existing on the market have enough computational power to support NLP-based emotion extraction tool. The usage of special recompiled standalone version of the General Inquirer[7] was considered but because of size and computational requirements was rejected. Thus, a simple yet robust model was developed based on the Harvard IV-4 dictionary which contains 1915 positive and 2291 negative words.

In the first step every incoming message is intercepted and tokenized. Because English has very limited flexion stemming is actually unnecessary (except genitivus and plural forms). Thus, a tokenized form can be directly compared with dictionary containing positive and negative words. Typically, as for all samples of user generated content [], also in this dataset a lot of misspellings, mistakes, slangs and abbreviation exist. Because of limited number of characters in which can be sent in one message and lacks of full keyboard in most of used devices[] many specific abbreviations have evolved and are in common use.

All extracted tokens were compared with a list of English words contains 58 thousands entity . For over 120 thousands tokens found in dataset slightly more than 81 thousands were recognized as an English word. Thus, a need for correcting and decoding mechanism is obvious. Closer look on the 42,646 unrecognized tokens shows that only 3663 are unique. The most common used unrecognized combination of characters appear in dataset more than 4000 times and some of them (e.g. "tnk" -> "thank") are rather easy to decode. Therefore for most commonly used abbreviation 31 rules were manually crafted. Selection of prepared rules is presented below: "u" → "you", "ur" → "your", "4" → "for", "n" → "and", "r" → "are", "coz" → "because".

Applying rules described in previous paragraph allows to reduce number of unrecognized tokens to 25 thousands which is less than 20 percent of all tokens in the dataset. It's still higher than in typical texts collected from Internet but as can be expected SMSs contain more words such as the names of persons, organizations, locations or expressions of times because of their informative function.



**Fig. 2.** Screenshot of the SMS EmoAlerter application

**Implementation.** To verify an applicability of the proposed framework a proof-of-concept was developed. As a platform for deployment a platform composed of T-mobile G1 smart phone and Android v.1.2 operating system was chosen. Currently, it is at best a medium range platform and much more sophisticated devices and operation systems are already on the market (e.g. iPhone 3GS, HTC HD2, Motorola Droid or Google Nexus). Thus, a similar deployment would be also possible, with very limited problems, on cutting the edge mobile phones.

The SMS EmoAlerter remains the architecture of the App1 presented on fig. 1. Module responsible for collecting facts (in this case it means intercepting incoming short messages) and NLP processing engine (emotion discovering from text) are placed within one application. To intercept incoming SMSes a standard solution delivered by operating system was used. An application was registered as an event handler for the standard event exiting in Android OS which is called *RECEIVE\_SMS*. The security policy of the operating system requires that users will explicitly grant this privilege to the application during installation process. When the scheduled event appears, operating system activates application and passes control to it.

**Tests.** A subset of the dataset described more detailed in the subsection 2.1 was manually tagged by two persons. Every single SMS was evaluated either as positive or neutral or negative. The meaning of the SMS was discovered only by looking on their text content. Therefore all additional information like details about sender or history of communication were hidden. The motivation to make the task more complicated for people were twofold. Firstly, we wanted to eliminate as many collateral effects<sup>1</sup>, which could influence the human evaluation , as possible. Secondly, to make results of automatic classification and human tagging comparable we had to ensure that this classification will be based on the same dimensions.

The inter-agreement between two persons was relative high and exceed 90% what indicate that negative or positive emotion appeared in SMSes, at least at some level, can be treated as objective feature. The remaining 10% of SMSes in which evaluation of two persons do not match was removed from further processing. In tab. 2 the percentage of positive, negative and neutral SMS, according to human evaluation, is presented. The most striking conclusion from that results is that huge majority of messages belong to the neutral category. In overwhelming number of cases a content of SMS was very hard (or even impossible) to understand without precedent messages. Sender of an message, motivated by the limitation of SMS technology and clunky keyboard, usually tries to predict a level of knowledge of receiver and, thus, restricts an content of their message to a new facts which can be placed on the top of the knowledge they share.

The limited precision of the NLP algorithm (in comparison to people) in detecting positivity in messages is can be attributed to the limitation of the dictionaries. Standard Harvard IV-4 dictionary contain neither emoticons like ":@" or ";)" nor short combination of letters mimicking the sound of laugh. This

**Table 2.** Manually vs. automatic tagged SMSes

	<b>positive</b>	<b>neutral</b>	<b>negative</b>
<b>manually tagged</b>	15%	81%	4%
<b>automatic tagged</b>	12%	84%	4%

two drawbacks can be easily eliminated but the advantage will be very limited as long as we do not start to deal with the history of communication rather than with a single message.

---

<sup>1</sup> It is a well-known phenomena that human usually cannot ignore side-information during evaluation process. Such information like nick or even a combination of numbers presented next to the content of SMS could influence the emotional state of the evaluator.

### 3.2 Sensing Wallpaper

Second example of an application, which use an information about emotional state of user, is the Sensing Wallpaper. Going back to the schema presented on fig. 1 this application has a design similar to the "App5". It uses information from the storage layer to adapt current wallpaper to the mood of the user. In general two strategies can be identified either the application can improve user's mood by positive wallpapers every time when some bad events (information) affect or the application can follow user's mood and change wallpaper to positive only when the user is happy. The event-action mapping in a context-aware mobile systems is a general and known problem and it is not the aim of this paper to address it. Therefore, the Sensing Wallpaper application was designed to be agnostic in terms of event-action mapping. Every user can freely change picture-emotional state matching by themselves.

## 4 Conclusion

In previous section the framework for developing emotional aware mobile application was presented. Two developed application proof that, firstly, this framework is practically usable and, secondly, that mobile device which is aware of emotional state of their user can be developed with use of software and hardware already present on the market. Sufficient precision of emotion extraction from short messages (SMSes) can be ensured with relative simple algorithm which, at least for English, can run smoothly on middle range phone with Android OS (G1). On the other hand more complex algorithms and another sources of emotion-rich information are still too computational power hungry to be successfully deployed on mobile devices.

Quite surprising was a very low level of emotionality of SMSes in using dataset. The most probably explanation of this fact is the way the dataset was constructed. Volunteers contributing to this dataset could either change their typical communication habits or filter messages that were too emotional. Disregard of the reason, in our research SMSes have appeared to be a tool for exchanging information and agreeing common activity rather than channel for expressing emotion.

Popularization of precise and efficient algorithms for sensing user's emotion and installing them by default on mobile devices can raise some serious privacy and security concerns. The threat that some application will abuse information about our emotional state cannot be fully dismissed. A vivid example of such abuse can be a situation in which service provider (let it be bank or mobile network operator) shapes their offer after our mood or presents it only when we are happy (it's well known psychological effect that we are more willingly to accept proposition/spend money when we are happy). This problem can be partly solved by a mechanism which is present in almost all

operating systems for mobile devices and requires explicit declaration of privileges for each application but, firstly, most users are unaware of threats and don't understand the consequences of choices (in matter of security) and, secondly, for devices which are permanently on-line and can be freely managed, inspected and even switched off remotely the true security and privacy actually do not exist. Future research, among security and privacy should be focused on new algorithms, which allow extracting emotion from variety of sources (e.g. video, voice or proximity), and will efficiently work with limited computational resources accessible on mobile devices. Yet another interesting topic is looking for practical use-cases for emotion aware application (e.g. monitoring stress factors of user), which will probably raise many additional research questions.

## References

1. Dey, A.K.: Context-aware computing: The CyberDuck project. In: AAAI 1998 Spring Symposium on Intelligent Environments, Palo Alto, pp. 51–54. AAAI Press, Menlo Park (1998)
2. Mizzaro, S., Nazzi, E., Vassena, L.: Retrieval of context-aware applications on mobile devices: How to evaluate? In: IliX 2008, Information in Context. Information in Context (2008)
3. Oulasvirta, A.: Finding Meaningful Uses for Context-Aware Technologies: The Humanistic Research Strategy. In: CHI 2004 (2004)
4. Sanchez, J.M., Cano, J.C., Calafate, C.T., Manzoni, P.: BlueMall: A Bluetooth-based Advertisement System for Commercial Areas. In: PM2HW2N 2008, Vancouver (2008)
5. Nakanishi, Y., Kitaoka, N., Ohyama, M., Hakozaki, K.: Estimating Communication Context through Location Information and Schedule Information - A Study with Home Office Workers. In: CHI 2002 (2002)
6. Korpiapaa, P., Hakkila, J., Kela, J., Ronkainen, S., Kansala, I.: Utilising context ontology in mobile device application personalisation. In: MUM 2004: Proceedings of the 3rd International Conference on Mobile and Ubiquitous Multimedia, pp. 133–140. ACM, New York (2004)
7. Philip, J., et al.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
8. Geven, A., Tscheligi, M., Noldus, L.: Measuring Mobile Emotions: Measuring the Impossible? In: MobileHCI 2009: Proc. of the 11th Int. Conference on Human-Computer Interaction with Mobile Devices and Services (2009)
9. Mody, R.N., Willis, K.S., Kerstein, R.: WiMo: location-based emotion tagging. In: MUM 2009: Proceedings of the 8th International Conference on Mobile and Ubiquitous Multimedia, pp. 1–4. ACM, New York (2009)
10. Tsai, T., Chen, J., Lo, W.: Design and Implementation of Mobile Personal Emotion Monitoring System. In: MDM 2009: Proceedings of the 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware, pp. 430–435. IEEE Computer Society, Los Alamitos (2009)

11. Isomursu, M., Tähti, M., Väinämö, S., Kuutti, K.: Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *Int. J. Human-Computer Studies* 65, 404–418 (2007)
12. Yang, B., Lugger, M.: Emotion recognition from speech signals using new harmony features. *Signal Processing* 90, 1415–1423 (2010)

# Boosting-Based Ensemble Learning with Penalty Setting Profiles for Automatic Thai Unknown Word Recognition\*

Jakkrit TeCho, Cholwich Nattee, and Thanaruk Theeramunkong

School of Information, Computer and Communication Technology,  
Sirindhorn International Institute of Technology, Thammasat University, Thailand  
`{jakkrit, cholwich, thanaruk}@siit.tu.ac.th`

**Abstract.** A boosting-based ensemble learning can be used to improve classification accuracy by using multiple classification models constructing to cope with errors obtained from preceding steps. This paper presents an application of the boosting-based ensemble learning with penalty setting profiles on automatic unknown word recognition in Thai. Treating a sequential task as a non-sequential problem requires us to rank a set of generated candidates for a potential unknown word position. Since the correct candidate might not located at the highest rank among those candidates in the set, the proposed method provides penalties, in the form of a penalty setting profile, to improper ranking in order to reconstruct the succeeding classification model. In addition a number of alternative penalty setting profiles are introduced and their performances are compared on the task of extracting unknown words from a large Thai medical text. Using the naïve Bayes as the base classifier for ensemble learning, the proposed method achieves the accuracy of 89.24%, which is an improvement of 9.91%, 7.54%, 5.25% over conventional naïve Bayes, non-ensemble version, and flat penalty setting profile.

**Keywords:** Ensemble learning; Boosting Technique; Data mining; Unknown word recognition; Word boundary detection.

## 1 Introduction

Unknown word recognition plays an important role in natural language processing since words, fundamental units of a language, may be newly developed and invented. It is necessary to develop techniques to handle words not occurred in the lexicon, so-called unknown words recognition. In the languages with explicit word boundary, it is straightforward to identify an unknown word and its boundary. This simplicity is not conformed to the languages without word

---

\* This work was partially funded by NECTEC of Thailand via research grant for Automatic Tagger for Named Entity in Thai News Corpus Project (NT-B-22-KE-38-52-01).

boundary (later called unsegmented language) such as Thai, Chinese, Japanese, where words are running without any explicit space or punctuation mark. In Thai, our target language, Kawtrakul et al. have defined that the major sources of unknown words [5] are (1) Thai transliteration of foreign words, (2) invention of Thai new technical words, and (3) emerging of Thai proper names. For example, Thai medical texts often abound in transliterated words/terms or technical words/terms which may not be in any dictionary. In Thai news articles, it is common to find a lot of proper names related to persons, organizations, etc.

In the past decade, many researches on the unknown word recognition in Thai language have been proposed to detect unknown words using features extracted from a large corpus of Thai texts with various machine learning techniques, later called Machine Learning-based (ML-based) approach. In the ML-based approach, unknown word recognition can be viewed a process to detect new compound words in a text without using a dictionary to segment the text into words. Charoenpornsawat et al. considered unknown word recognition as a classification problem [1] and proposed a feature-based approach to identify Thai unknown word boundaries. Haruechaiyasak et al. have proposed a semi-automated framework [4] that utilizes statistical and corpus-based concepts for detecting unknown words and then introduced a collaborative framework among a group of corpus builders to refine the obtained results. In the automated process, unknown word boundaries are identified using frequencies of strings. Although several works have been done in this approach, they however used only local information to learn a set of rules for word segmentation/unknown word detection by a single-level learning process (a single classifier).

In practice, the unknown word candidates actually have relationship with adjacent words. Although several works have been done in both approaches, most works did not consider about such issue. As a more recent work, TeCho et al. proposed a framework [7] to combine word segmentation process with learning process that utilizes long-distance context in learning a set of rules for unknown word detection in word segmentation process, where no manual rules are required. Moreover, they proposed a technique to treat the unknown word candidate as a group. However, as the ensemble learning process, the technique also makes a new dataset based on misclassified group more important by updating the weight for a group of candidates. Without considering the correctness of ranking in a group, this technique may not make specialization for a dataset on the next iteration. However, a large set of candidates may be generated, inducing the problem of unbalanced class sizes where the number of positive unknown word candidates is dominantly smaller than that of negative candidates. To solve the problem, this paper presents a technique called “GRE-based boosting with Penalty Setting Profile” is incorporated into ensemble learning in order to generate a sequence of classification models that later collaborate to select the most probable unknown word from multiple candidates. Moreover, in the boosting step, our technique is applied in order to build a dataset for training the succeeding model, by applying different weight updating to each of its candidates

according to their ranking and correctness when the candidates of an unknown word are considered as a group.

## 2 Thai Unknown Word Recognition Framework

Most researches based on ML-based approach for Thai unknown word recognition [4,7] have defined the framework consisting of three processes: (1) unregistered portion detection, (2) unknown word candidate generation and reduction, and (3) unknown word identification. The details of these processes are given in sequence.

*Unregistered Portion Detection.* Normally when we apply word segmentation on a Thai running text with some unknown words, we may face with a number of unrecognizable units due to out-of-vocabulary words. This work has applied a useful concept, namely a Thai Character Cluster (TCC) [8]. In addition, this paper also employed a capability approach [7] in order to detect the unregistered portions from a running text.

*Unknown Word Candidate Generation and Reduction.* Similar to the candidate generation technique proposed by [1]. The  $\pm h$  TCCs surrounding an unregistered portion are merged to form an unknown word candidate. By this setting,  $(h+1)^2$  possible candidates can be generated for each unregistered portion.

*Unknown Word Identification.* Various ML techniques are applied to identify the unknown words using a feature extraction set. Unlike the most previous Thai unknown word recognition treated unknown word candidates independently. In practice, a set of candidates generated from an unregistered portion, should be considered dependently and treated as a group. In this paper, we efficiently apply a technique, namely GRE [7], into an ensemble learning scheme [2], i.e., boosting, in order to generate a sequence of classification models that later collaborate to identify the most probable unknown word from a number of candidates (within a group).

As stated in the previous section, TCCs are used as processing units. We therefore use a sequence of TCCs instead of a sequence of characters to denote an unknown word candidate. To specify whether a candidate is the most probable unknown word or not, a set of suitable features need to be considered. In this work, we employed the same feature set [7] collected from context around an unknown word can be considered as features by applying an algorithm, originally proposed [6] that utilized the sorted sistrings concept. For each sistring (i.e., unknown word candidate), nine types of features are extracted such as Number of TCC, Number of Characters, Number of known words, Sistring Frequency, Left and Right TCCs variety, Probability of a special character on left and right, DF, IDF, and TFIDF score.

### 3 GRE-Based Boosting with Penalty Setting Profile

A technique called “GRE-based Boosting” [7] applied the AdaBoost technique [3] to the unknown word data with treating the candidates as a group. As a famous boosting technique, AdaBoost is a technique to repeatedly construct a sequence of classifiers based on a base learning method. In this technique, each instance in the training set is attached with a weight (initially set to 1.0). On each iteration, the base learning method constructs a classifier using all instances in the training set, and with their weights showing the importance. After evaluation the obtained classifier, the weights of the misclassified examples are increased to make the learning method focus more on the misclassified examples. Originally, AdaBoost evaluates each instance and updates its weight individually. This technique is not suitable for the unknown word data that we treat them as groups of unknown word candidates. Thus, the technique to update the weights for the whole instances in the misclassified groups has been proposed. However, the correctness of ranking in a group are not considered when boosting is performed, this technique may not give an expression of specialization for a next iteration dataset.

#### 3.1 Penalty Setting Profile

This paper efficiently propose different weighing approaches into a group instead of using the same weight for the whole group. We assign a weight to an unknown word candidate according to the correctness of its rank in the group. The candidate which has the potential for an unknown word should be ranked at the top while the rests are ranked at the bottom. The classifier is considered to misclassify a group when the top ranked candidate in the group is not a correct unknown word. The weight of that group is then increased to make the group be more focused in the next iteration. From the concept of boosting process, a ratio of success to unsuccess rate ( $\beta$ ) is calculated from the misclassifying rate. This ratio can be used as the new weight of the misclassified groups in the next iteration. In order to reassign the weight to such a group, this paper proposes a technique namely “penalty setting profile” by applying several weight updating approaches to each of its candidates. In addition, the penalty setting profile is characterized by two parameters i.e.,  $\zeta$  and  $\eta$ , defined as follows.

$$\eta = \gamma\beta \quad (1)$$

$$\zeta = \delta\beta \quad (2)$$

where  $\eta < \zeta$  and  $\gamma, \delta$  are the multiplier for  $\beta$ . Basically, this  $\beta$  is larger than 1. Thus, the classifier constructed in the next iteration will be specialized to the previously misclassified instances. To explain these penalty setting profiles, the following description is first given. Let  $r_{c_i}$  be a ranking position of the potential for an unknown word in a group,  $r_{|G_i|}$  is the last ranking of the  $i$ -th group, and  $r_j$  is a ranking position of the  $j$ -th candidate where  $1 \leq j \leq r_{|G_i|}$ . With the above description, we presents seven potential penalty setting profiles i.e., Profile 1 - Profile 7, can be formally defined in sequence as follows.

Profile 1:	$w_{ij} = \zeta$	if $1 \leq r_j \leq r_{ G_i }$
Profile 2:	$w_{ij} = \begin{cases} \zeta \\ \eta \end{cases}$	if $1 \leq r_j \leq r_{c_i}$ if $r_{c_i} < r_j \leq r_{ G_i }$
Profile 3:	$w_{ij} = \begin{cases} \zeta \\ \eta \end{cases}$	if $r_j = r_{c_i}$ if $1 < r_j \leq r_{ G_i }, r_j \neq r_{c_i}$
Profile 4:	$w_{ij} = \begin{cases} \zeta \\ \eta + \left\{ \frac{(\zeta - \eta)}{(r_{c_i} - 1)} \times (r_{c_i} - r_j) \right\} \\ \eta \end{cases}$	if $r_j = 1, r_j = r_{c_i}$ if $1 < r_j < r_{c_i}$ if $r_{c_i} < r_j \leq r_{ G_i }$
Profile 5:	$w_{ij} = \begin{cases} \zeta \\ \eta + \left\{ \frac{(\zeta - \eta)}{(r_{c_i} - 1)} \times (r_j - 1) \right\} \\ \eta \end{cases}$	if $r_j = r_{c_i}$ if $1 < r_j < r_{c_i}$ if $r_j = 1, r_{c_i} < r_j \leq r_{ G_i }$
Profile 6:	$w_{ij} = \begin{cases} \zeta \\ \eta + \left\{ \frac{(\zeta - \eta)}{(0.5r_{c_i} - 1)} \times (0.5r_{c_i} - r_j) \right\} \\ \eta + \left\{ \frac{(\zeta - \eta)}{(0.5r_{c_i} - 1)} \times (0.5r_j - 1) \right\} \\ \eta \end{cases}$	if $r_j = 1$ if $1 < r_j < 0.5r_{c_i}$ if $0.5r_{c_i} < r_j < r_{c_i}$ if $r_j = 0.5, r_{c_i} < r_j \leq r_{ G_i }$
Profile 7:	$w_{ij} = \begin{cases} \zeta \\ \eta + \left\{ \frac{(\zeta - \eta)}{(0.5r_{c_i} - 1)} \times (0.5r_{c_i} - 1) \right\} \\ \eta + \left\{ \frac{(\zeta - \eta)}{(0.5r_{c_i} - 1)} \times (0.5r_{c_i} - r_j) \right\} \\ \eta \end{cases}$	if $r_j = 0.5r_{c_i}, r_j = r_{c_i}$ if $1 < r_j < 0.5r_{c_i}$ if $0.5r_{c_i} < r_j < r_{c_i}$ if $r_j = 1, r_{c_i} < r_j \leq r_{ G_i }$

Algorithm 1 shows the GRE-based boosting with penalty setting profile technique in details. The algorithm starts with the initial training set  $T_1 = \{G_1, \dots, G_n\}$  with  $G_i = \{(c_{i1}, w_{i1}), \dots, (c_{ir_i}, w_{ir_i})\}$  and  $c_{ij} = (X_{ij}, y_{ij})$ , where  $G_i$  is the group of unknown word candidates generated for the  $i$ -th unregistered portion,  $c_{ij}$  is the  $j$ -th candidate of the  $i$ -th unregistered portion,  $w_{i1}$  is an initial weight (set to 1 at the first iteration) given to  $c_{i1}$ ,  $n$  is the number of unregistered portions,  $r_i$  is the number of unknown word candidates generated for the  $i$ -th unregistered portion,  $X_{ij}$  is the set of feature values representing  $c_{ij}$ , and  $y_{ij} \in \{-1, 1\}$  is the target attribute of  $c_{ij}$  (designated as the class label), stating whether  $c_{ij}$  is the correct unknown word (1) or not (-1).  $K$  iterations are conducted to construct a sequence of base classifiers. At the  $k$ -th iteration, a training set  $T_k$  is fed to INDUCER to construct a base classifier  $m_k$ . The classifier is then evaluated by GRE-INCOR yielding  $E_k \subseteq T_k$ , a set of misclassified groups.  $e_k$ , the error rate of the classifier  $m_k$ , can be calculated from  $E_k$ . It is used to calculate  $\alpha_k$ , and  $\beta_k$  which are the parameters showing the confidence level of the classifier, and the weight for the iteration. Finally, the weight of the candidates in a group  $\in E_k$  are calculated by WEIGHING according to the penalty setting profile approach (a) when the weight for correctly predicted candidate ( $\eta$ ) and the weight for incorrectly predicted candidates ( $\zeta$ ) are given. Otherwise, they are all set to 1.

*Voting Group-based Ranking Evaluation.* From the previous step, we obtain a sequence of base classifiers. Each classifier is attached with its confidence weight ( $\alpha_k$ ). In this section, we employed a technique called “Voting Group-based Ranking

**Algorithm 1.** GRE-based Boosting with penalty setting profile

**Data:**  $T_1$ : an initial training set (with all weights set to 1.0),  $K$ : the number of iterations,  $\gamma$ ,  $\delta$ : the multipliers in order to calculate the  $\eta$  and  $\zeta$ ,  $a$ : an penalty setting profile.

**Result:**  $M$ : a set of base classifiers

```

1  $M = \phi;$ 
2  $T_1 = \{G_1, \dots, G_n\};$ 
3 for  $k=1$  to  $K$  do
4    $m_k = \text{INDUCER}(T_k);$ 
5    $E_k = \text{GRE-INCOR}(m_k, T_k);$ 
6    $e_k = |E_k|/n;$ 
7    $\alpha_k = \ln(e_k/(1 - e_k));$ 
8    $\beta_k = \lceil (1 - e_k) / e_k \rceil;$ 
9    $T_{k+1} = \phi;$ 
10  forall  $G_i \in T_k$  do
11    if  $(G_i \in E_k)$  then
12       $G'_i = \phi;$ 
13      foreach  $(c_{ij}, w_{ij}) \in G_i$  do
14         $w_{ij} = \text{WEIGHING}(c_{ij}, a, \gamma, \delta, \beta);$ 
15         $G'_i = G'_i \cup (c_{ij}, w_{ij});$ 
16      end
17       $T_{k+1} = T_{k+1} \cup \{G'_i\};$ 
18    else
19       $T_{k+1} = T_{k+1} \cup \{G_i\};$ 
20    end
21  end
22   $M = M \cup \{(m_k, \alpha_k)\};$ 
23 end
```

Evaluation” [7] to evaluate a group of unknown word candidates, and predict the unknown word by combining votes from all produced classifiers.

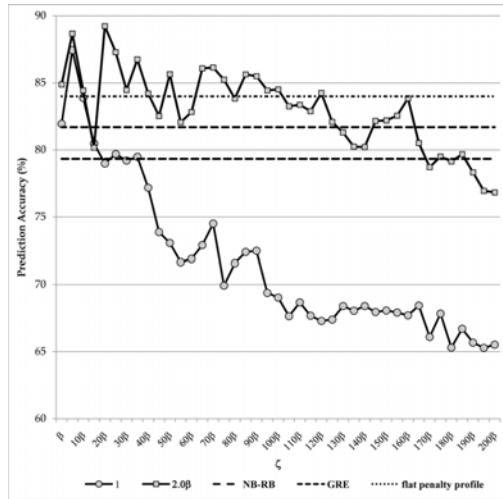
## 4 Experimental Results

In the experiment, we used a corpus of 16,703 medical-related documents gathered from WWW [9] with the size of 8.4 MB for evaluation. The corpus is first preprocessed by removing HTML tags and all undesirable punctuations. To construct a set of features, we apply TCCs and the sorted sistring technique. After applying word segmentation on the running text, we have detected 46,352 unregistered portions. Based on these unregistered portions, 3,321,703 unknown word candidates are generated according to the process shown in Sect. 2. Moreover, these 46,352 unregistered portions came from only 4,170 distinct words. In practice, each group of candidates may contain one or two positive labels. Therefore, 53,089 unknown candidates were assigned as positive and 2,710,909

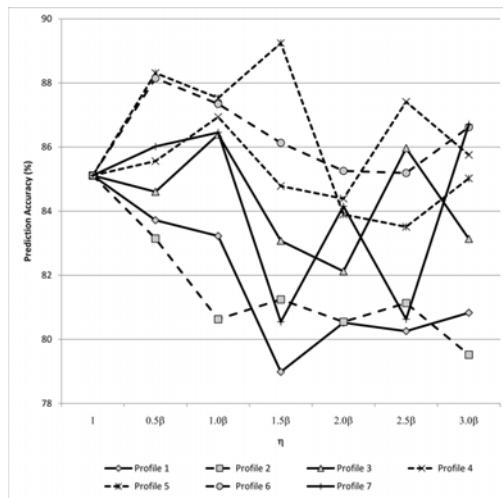
unknown candidates were assigned as negative. The average number of unknown candidates in a group is around 58. Based on preliminarily statistical analysis of the Thai lexicon, we found that the average number of TCCs in a word is around 4.5. In this work, to limit the number of generated unknown word candidates, the maximum number of TCCs surrounding an unregistered portion ( $h$ ) is set to nine. This number is twice of the average number of TCCs in a word. With  $h=9$ , the number of generated unknown word candidates becomes 100. Moreover, it is possible to use two sets of separation markers in Sect. 2 to reduce the number of candidates. For learning process, a naïve Bayes classifier is used as base classifier for the proposed methods, GRE-based boosting with penalty setting profiles, in order to learn ensemble classifiers and V-GRE is also used to identify an unknown word.

For the boosting iteration,  $K$  is set to 10. Here, sequentially ten classifiers are generated and used as Classification committees. Moreover, to evaluate our proposed method in details, we have conducted the experiments to examine the effect of nine features, on the classification result by comparing performance of each possible feature combination with the others. As stated in Sect. 3, we tested all the seven penalty setting profiles with either varying weights for  $\eta$  and  $\zeta$ . In addition, the multiplier for the weight of correctly predicted candidates ( $\gamma$ ) are varied ( $1/\beta$ , 0.5, 1.0, ..., and 3.0). Moreover, the multiplier for the weight of incorrectly predicted candidates ( $\delta$ ) are also varied (1, 5, ..., 95, and 100). In order to investigate the efficiency performance of our proposed method, this experiments were conducted using 10-fold cross validation. Furthermore, a number of experiments are performed with several naïve Bayes classifiers in order to investigate our proposed method such as the conventional naïve Bayes classifier and naïve Bayes with GRE [7]. Table 1 displays the performance of seven penalty setting profiles with seven varying of  $\eta$  as well as 21 varying of  $\zeta$ . However, only 867 settings are conducted since the Profile 1 does not concern about the  $\eta$ . According to the result, the experimental results showed that the classifier achieved the highest accuracy of 89.24% when the Profile 4 is employed when  $\eta$  and  $\zeta$  are set to  $2.0\beta$  and  $20\beta$ , respectively. This is intuitive since the Profile 4 sets high penalty to the candidates that are strongly misclassified, and also focuses on the correctly predicted candidates. The table also presents the highest accuracies yielded from each penalty setting profile using boldface font.

As a deep investigate to the Profile 4 performance, we have conducted more experiments with a number of setting by increasing the multiplier  $\delta$  up to 200 as shown in Fig. 1. As the result, our proposed method (with  $\eta=2.0\beta$ ,  $\zeta=20\beta$ ) achieves the accuracy over the several setting of naïve Bayes classifiers. Moreover, the figure also presents that the accuracy of trends to continue decreasing while increasing the number of  $\zeta$ . In addition, we also plotted a graph that indicated the accuracy of all penalty setting profiles among the several varying of  $\eta$  when  $\zeta$  is set to  $20\beta$  as shown in Fig. 2. From the result, most of penalty setting profile achieved the highest accuracy when a classifier is given the small number of  $\eta$  i.e., 1.0 up to  $2.0\beta$ . According to such phenomenons, The result indicates that the adequately number of  $\zeta$  can achieves the higher accuracy. In other words, we may



**Fig. 1.** Comparison Accuracy among  $\eta$  when the Profile 4 is employed with varied  $\zeta$

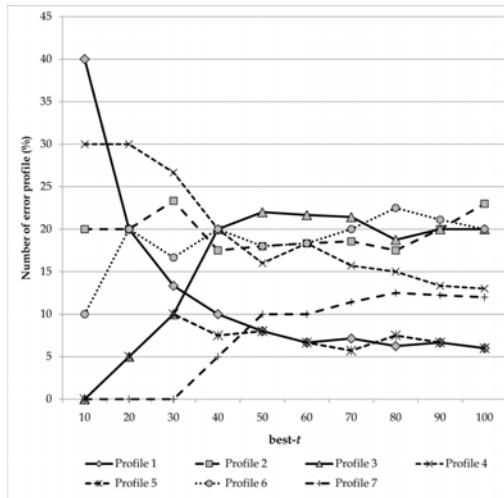


**Fig. 2.** Comparison Accuracy among seven penalty setting profiles when  $\zeta$  is set to  $20\beta$

not make the correctly predicted candidate more important for next iteration on boosting-based learning. As another exploration, we sorted all accuracies in Table 1 and then plotted the number of penalty setting profile against the best- $t$  penalty settings as shown in Fig. 3 when  $t$  is number of penalty setting profile. The figure presents the number of third, fifth, sixth and seventh penalty profile settings trend to increase while the number of Profile 1 and Profile 4 trend to decrease.

**Table 1.** Accuracy comparison among a conventional naïve Bayes classifier, GRE, GRE-based Boosting with penalty setting profiles with varied  $\eta$  and  $\zeta$ . Here  $\gamma=0.0$  to 3.0 step 1.0 and  $\delta=1$  to 100 step 10.

Naïve Bayes with GRE-based Boosting with Penalty setting profiles												
Profile	$\eta$	$\zeta$										
		$\beta$	$10\beta$	$20\beta$	$30\beta$	$40\beta$	$50\beta$	$60\beta$	$70\beta$	$80\beta$	$90\beta$	$100\beta$
1	N/A	83.99	<b>89.23</b>	85.11	86.57	88.82	86.65	85.98	85.99	85.19	84.65	86.04
	1	82.32	83.70	83.72	79.70	78.70	82.36	76.70	73.62	72.72	70.21	76.11
	$1.0\beta$	-	79.99	84.61	87.11	82.28	81.32	77.88	78.85	77.60	73.76	81.85
	$2.0\beta$	-	84.63	<b>88.31</b>	86.59	86.46	87.37	86.56	84.25	81.50	80.64	79.93
2	$3.0\beta$	-	84.37	86.02	82.07	84.37	86.07	86.91	87.38	86.83	86.23	84.43
	1	81.05	78.37	83.22	86.46	85.16	83.50	84.81	83.36	83.36	82.69	81.15
	$1.0\beta$	-	83.18	86.39	86.34	86.20	86.09	86.09	86.00	86.20	86.59	86.15
	$2.0\beta$	-	87.88	87.53	87.57	86.08	86.20	86.21	85.69	85.58	86.35	85.72
3	$3.0\beta$	-	<b>88.48</b>	86.45	86.32	85.38	85.32	84.30	83.50	84.40	84.69	83.41
	1	81.94	83.89	78.99	79.20	77.18	73.06	71.90	74.51	71.57	72.48	69.04
	$1.0\beta$	-	80.55	83.08	86.40	85.42	82.74	80.95	80.20	78.90	78.13	77.08
	$2.0\beta$	-	84.44	<b>89.24*</b>	84.46	84.20	85.64	82.83	86.14	83.84	85.49	84.52
4	$3.0\beta$	-	84.96	80.55	86.76	85.02	85.57	85.26	82.80	81.66	85.28	82.51
	1	80.97	86.75	80.52	81.96	83.51	81.76	80.51	81.08	79.18	78.34	79.17
	$1.0\beta$	-	87.25	82.12	78.86	84.60	81.47	82.55	86.34	84.93	83.54	82.80
	$2.0\beta$	-	88.13	83.90	83.73	82.13	81.93	81.81	83.56	82.64	84.94	83.69
5	$3.0\beta$	-	<b>88.33</b>	84.15	85.06	81.01	79.29	80.58	78.84	82.68	83.21	84.25
	1	81.26	83.28	80.26	80.12	77.42	79.27	77.91	77.00	75.29	73.94	71.37
	$1.0\beta$	-	83.35	85.97	83.70	86.50	<b>87.69</b>	86.38	85.97	87.19	85.97	83.39
	$2.0\beta$	-	84.88	83.51	85.88	86.79	85.60	83.50	87.17	85.99	85.44	85.48
6	$3.0\beta$	-	85.39	80.63	83.42	85.25	85.62	85.18	86.16	82.94	84.05	81.54
	1	81.05	85.67	80.83	80.95	78.78	78.95	78.50	81.11	79.82	80.10	80.23
	$1.0\beta$	-	86.37	83.14	77.24	81.10	84.32	83.43	84.31	81.30	81.21	81.51
	$2.0\beta$	-	86.55	85.02	82.74	80.86	77.66	80.49	83.92	85.01	82.92	82.39
7	$3.0\beta$	-	<b>87.69</b>	86.71	86.35	81.42	82.38	81.02	80.24	81.40	83.52	84.04
	Naïve Bayes with GRE											
	81.70											
	Conventional Naïve Bayes											



**Fig. 3.** Number of penalty setting profile respects to best- $t$  penalty setting profiles

## 5 Conclusion

In this paper, we presented an automated method to recognize unknown words from a Thai running text. We described how to map the problem to a classification task. The naïve Bayes with a smoothing technique classifier is investigated using nine features for evaluating the model. In practice, the unknown word candidates actually have relationship among them. To reduce the complexity in unknown word boundary identification, reduction approaches are employed to decrease a number of generated unknown word candidates. This paper also proposed the GRE-based boosting with penalty setting profiles. This technique considered the unknown word candidates as groups that can be solved the unbalanced datasets problem. In addition, this technique applied different weight updating approaches into a group in order to increase the focusing on next iteration as penalty setting profile. A number of experiments were performed by seven penalty setting profiles with varied weighing given to the candidates when naïve Bayes is applied as base classifier. In this work, we set 867 settings by applying such penalty setting profiles with varying the multiplier. Moreover, all settings are conducted with 10-fold cross validation. As the experimental results, our proposed technique achieves on the highest accuracy 89.24% which is an improvement of 9.91%, 7.54%, 5.25% over simple naïve Bayes, non-ensemble, and flat penalty setting profile.

## References

1. Charoenpornsawat, P.: et al.: Feature-based thai unknown word boundary identification using winnow. In: Proc. of APCCAS 1998, Chiang Mai, Thailand, pp. 547–550 (November 1998)
2. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittle, J., Roli, F. (eds.) Multiple Classifiers Systems, pp. 1–15. Springer, Heidelberg (2000)
3. Freund, Y., Schapire, R.E.: A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* 14(5), 771–780 (1999)
4. Haruechaiyasak, C., et al.: A collaborative framework for collecting thai unknown words from the web. In: Proc. of the COLING/ACL-2006, Sydney, Australia, pp. 345–352 (July 2006)
5. Kawtrakul, A., et al.: Automatic thai unknown word recognition. In: Proc. of NLPRS 1997, Phuket, Thailand, pp. 341–346 (October 1997)
6. Sornlertlamvanich, V., Tanaka, H.: The automatic extraction of open compounds from text. In: Proc. of COLING 1996, Copenhagen, Denmark, pp. 1143–1146 (August 1996)
7. TeCho, J., et al.: A corpus-based approach for automatic thai unknown word recognition using boosting techniques. *IEICE Transactions on Information and Systems* E92-D(12), 2321–2333 (2009)
8. Theeramunkong, T., et al.: Pattern-based features vs. statistical-based features in decision trees for word segmentation. *IEICE Transactions on Information and Systems* E87-D(5), 1254–1260 (2004)
9. Theeramunkong, T., et al.: A framework for constructing a thai medical knowledge base. In: Proc. of KICSS 2007, JAIST, Ishikawa, Japan, pp. 45–50 (November 2007)

# AONT Encryption Based Application Data Management in Mobile RFID Environment\*

Namje Park<sup>1</sup> and Youjin Song<sup>2, \*\*</sup>

<sup>1</sup> Department of Computer Education, Teachers College, Jeju National University,

61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, Korea  
namjepark@jejunu.ac.kr, namjepark@gmail.com

<sup>2</sup> Department of Information Management, Dongguk University,  
707 Seokjang-dong, Gyeongju, Gyeongsangbuk-do, 780-714, Korea  
song@dongguk.ac.kr

**Abstract.** Mobile RFID (radio frequency identification) is a new application that allows the use of a mobile phone as a wireless RFID reader and provides new services to users by integrating RFID and the ubiquitous sensor network infrastructure with mobile communication and wireless internet services. Ensuring the security of mobile RFID's large-capacity database system by depending only on existing encryption schemes is unrealistic. In this regard, data sharing for security management has drawn attention as an extremely secure scheme. However, applying the existing secret sharing scheme to this method makes the size of the share equal to that of the original data, making it unsuitable for application to a large-scale database. To address this problem, this paper proposes secret sharing algorithms that enable efficient data security management through the use of the characteristics of the all-or-nothing transform (AONT) encryption in RFID middleware.

## 1 Introduction

Radio Frequency Identification (RFID) technology is being actively developed to exploit its global market potential. At the same time, it has raised fears among those who believe that it could facilitate a ‘Big Brother’ society. Thus, technology development efforts in areas such as tags, readers, and middleware should address not only information and market needs but also privacy and security concerns. The excessive limitations of RFID tags and readers have made it impossible to apply present codes and protocols. Technologies for information and privacy protection should address the general interconnection among elements, and their RFID characteristics should closely reflect the RFID environment.

Common RFID technologies have been used in B2B (business to business) models (e.g., supply channel, distribution, and logistics management), whereas mobile RFID technologies have been used in the RFID reader attached to an individual owner's mobile phone, through which the owner can collect and use information on objects by

---

\* This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(No. 2009-0087849).

\*\* Corresponding author.

reading their RFID tags. Corporations have applied mobile RFID technologies mainly to B2C (business to customer) models for marketing. Although RFID services have been limited largely to fields requiring less security (e.g., searching for movies and providing information in galleries), such services are expected to be applied more extensively to fields that necessitate privacy and security (e.g., purchasing, healthcare, and electrical drafts).

The secure storage and management of RFID's personal data has become an important issue, leading to the encryption of data to address security threats. However, the encryption scheme requires a lot of time and memory for encoding and decoding. Because encryption schemes such as the Advanced Encryption Standard (AES) encode confidential information in its entirety, there is a risk of the outflow of entire information once the secret key is decoded; hence, the distribution and management of the key is difficult. Unlike the encryption scheme, there is no such thing as the life of a key under the secret sharing scheme; thus, secret sharing scheme presents no problems in terms of public key certificate-based authentication and can substantially reduce operational costs. By contrast, under the existing secret sharing scheme[1], if original data are separately stored, the size of share (i.e., distributed information) equals that of the original data, increasing the data volume to be stored. Recently, it has been found that transforming the plaintext (in the encryption mode) before encoding improves data security while maintaining the efficiency of the existing encryption scheme[2].

This paper proposes the secret sharing algorithms that incorporate the characteristics of the all-or-nothing transform (AONT) encryption mode for the long-term, secure, and efficient sharing, storage, and reconstruction of large-capacity data. That is, the security of the algorithms is improved using the AONT encryption mode, and XOR operations are applied to the algorithms to boost their efficiency. The proposed algorithms are expected to be used for efficiently managing the distribution of large-capacity data (including highly confidential data on customers' personal information) even when there is data leakage from the RFID tags.

The rest of this paper is organized as follows: Chapter 2 describes a secure data management in networked mobile RFID, the AONT encryption mode related to this research and XOR secret sharing scheme; Chapter 3 reviews and designs the proposed algorithms, and Chapter 4 analyzes the proposed scheme; finally, Chapter 5 presents the conclusion.

## 2 Related Research

### 2.1 Networked Mobile RFID Technology

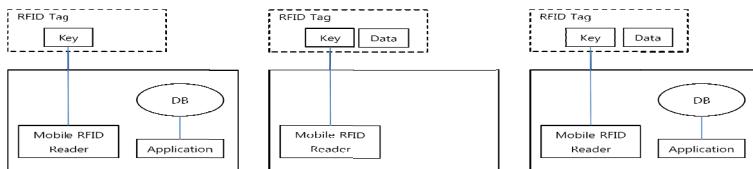
RFID is expected to be the base technology for ubiquitous network or computing, and to be associated with other technology such as telemetric, and sensors. The mobile phone integrated with RFID can activate new markets and end-user services, and can be considered as an exemplary technology fusion. Furthermore, it may evolve its functions as end-user terminal device, or 'u-device (ubiquitous device)', in the world of ubiquitous information technology.

Mobile RFID means an expanded RFID network and communication scope to communicate with a series of networks, inter-networks and globally distributed application systems. So it makes global communication relationships triggered by RFID, for such applications as B2B, B2C, B2B2C, G2C, etc. Networked RFID loads a compact RFID reader in a cellular phone, providing diverse services through mobile telecommunications networks when reading RFID tags through a cellular phone. Internet-enabled mobile phone which equips RFID reader will bring new service concepts to mobile telecommunication.

Mobile RFID technology is focusing on the UHF range (860~960MHz), since UHF range may enable longer reading range and moderate data rates as well as relatively small tag size and cost. Then, as a kind of handheld RFID reader, in the selected service domain the UHF RFID phone device can be used for providing object information directly to the end-user using the same UHF RFID tags which have widely spread. The service area of networked RFID is expected to be unlimited, and its services, diverse; currently, however, the service scenarios using RFID tags chiefly as offline hypertext owing to the constraints of cellular phone performance and business models are still at the proposal stage.

## 2.2 Secure Data Management in Networked Mobile RFID

Networked mobile RFID middleware applications may access application tag data in three ways as Figure 1. The left below is a diagram of a tag being read by a mobile RFID reader to get a key. This is used by an application to access database storage to get additional data. Note that key and storage contain application data, such as personal data, etc. This is at risk of theft and abuse. In the middle, tag key and data are both read by a mobile RFID reader to get the personal data, etc. In this case, no additional data is required. Note that both key and data may contain sensitive data that is at risk. On the right, RFID tag key and data are read by a mobile RFID reader to get the personal data, etc. Key also used by an application to access storage to get additional data, such as whether this person is on a specific watch list. This is the most challenging combination because key, data, and storage are all at risk.



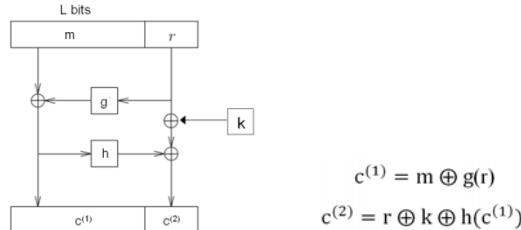
**Fig. 1.** Secure Data Management in Mobile RFID Environment

## 2.3 Encryption Mode

### 2.3.1 AONT Encryption Mode

The AONT consists of the scramble part that transforms the plaintext without using a secret key and the encryption part that encodes data by using a secret key [2]. The bit connection is indicated as  $\parallel$ , and exclusive OR(XOR) between bits, as  $\oplus$ . Here,  $h : \{0,1\}^* \rightarrow \{0,1\}^\ell$  means the hash function,  $g : \{0,1\}^\ell \rightarrow \{0,1\}^*$  refers to the

generator,  $m$  is the L-bit plaintext, and  $k$  represents the secret key shared by the transmitter and the receiver. The random number  $r$  and the plaintext  $m$  are encrypted as follows, at which time the ciphertext is  $c = c^{(1)} \parallel c^{(2)}$ :



**Fig. 2.** Efficient Non-Separable Encryption Model

The non-separable encryption mode is similar to the OAEP(Optimal Asymmetric Encryption Padding). The existing OAEP does not use a secret key in the scramble part, but this encryption mode uses a secret key. As such, the processing speed can be enhanced by adding an encryption function, which eliminates the encryption part. The AONT encryption mode proposed in this paper is OAEP-based; thus, this mode enables the fast processing and flexible setting of the number and size of distributed data and allows the sum of distributed data sizes to be equal to the original data size.

### 2.3.2 Exclusive OR(XOR) Value Secret Sharing Scheme

The secret sharing scheme was proposed by Shamir in 1979[1]. Shamir's  $(k, n)$  threshold secret sharing scheme draws on polynomial interpolation; it is a scheme according to which confidential information is segmented into  $n$  pieces of distributed information. If random  $k$  pieces of information are collected out of the  $n$  pieces, the original confidential information can be reconstructed. At this time of information sharing and reconstruction,  $k-1$ st order polynomial should be solved. However, polynomial operation presents problems to actual applications owing to calculated load. In reference [3], a fast threshold secret sharing scheme that is capable of distribution and reconstruction of confidential information only through Exclusive OR (XOR) operation is proposed. This paper utilizes XOR. Meanwhile, Kurihara, et al proposed an XOR-based  $(k, n)$  threshold secret sharing scheme in 2008[7]. This scheme consists of algorithms that can extend the threshold  $k$  value, generating the random numbers needed for plaintext and secret sharing operations for which it employs XOR; thus creating shares.

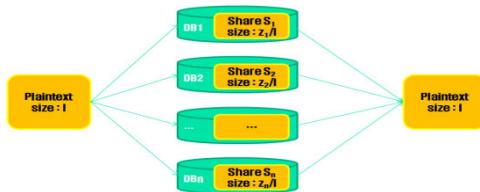
## 3 AONT Based Data Security Management Scheme

In this section, we propose a new scheme based on AONT [2] and Exclusive OR (XOR) operations. The overview of proposed scheme is shown in Figure 3. The proposed scheme generates the share  $s_i (i = 1, \dots, n)$  of the size  $\frac{\ell}{n}$  against the

plaintext of size  $\ell$ . The sum of all share sizes is equal to the size of the plaintext, which can be described as follows:

$$z_1 + z_2 + \dots + z_n = \ell$$

The scheme is more advantageous than the existing Shamir scheme [1] or the secret sharing scheme proposed by Kurihara [7]. Specifically, the existing secret sharing schemes are such that the sum of share sizes is equal to the plaintext size  $x n$ , requiring the same storage space as the plaintext for each database at the time of database storage. By contrast, the proposed algorithm is such that the sum of share sizes is equal to that of the plaintext, requiring storage space averaging as much as  $\frac{\ell}{n}$  per database. Such characteristics are quite suitable for large-capacity databases.



**Fig. 3.** Proposed scheme

### 3.1 Overview

The proposed data sharing system is as follows:

#### 1) Data distribution stage

- Segment the plaintext  $M$  by generating random numbers.
- Align the segmented plaintext in ascending or descending order based on size and perform XOR operations for each segmented plaintext.
- In the process, invert the resulting value from the descending order in even-numbered rounds.
- In terms of the values generated from repeating 16 times, turn the segmented fragments into shares and store each share separately on DB.

#### 2) Data reconstruction stage

- Aggregate each share stored separately on DB.
- For each share, repeat the XOR operations and the ascending/descending-order alignment with reverse direction of the data distribution stage.
- In this process, the plaintext can be reconstructed in odd-numbered rounds by reverting 16 times.

### 3.2 Details of the Proposed Scheme

The terms used in this paper are as follows:

: Exclusive-OR (XOR) operation , : Bit connection

$M$  : Plaintext,  $\ell$  : Size of plaintext,  $t$  : Number of blocks in the plaintext

$S$  : Segmented plaintext,  $r_i$  : Random number generated by BBS generator

### 3.2.1 Data Distribution Stage

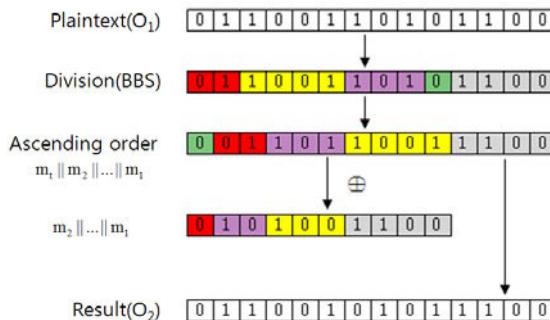
In terms of the  $t$ -sized plaintext  $m$ , generate as many uniform random numbers  $r_i (i = 1, \dots, t - 1)$  less than  $t$  as appropriate by using the BBS a pseudo random number generator[8].

Based on the random number  $r_i$ , segment the plaintext  $m$  into  $m_1, m_2, \dots, m_t$  and rearrange them in ascending order in proportion to the segmented size. At this time, the size is equal to that of the plaintext, despite the rearrangement:

$$m_1 \parallel m_2 \parallel \dots \parallel m_t \rightarrow m_t \parallel m_2 \parallel \dots \parallel m_1, m_1 > \dots > m_t$$

The plaintext  $m_t \parallel m_{t-1} \parallel \dots \parallel m_1$  arranged in ascending order generates the resulting value ( $O_2$ ) through the XOR operation. The XOR operation involves obtaining the segmented value  $m_1$  at the very last as it is (i.e.,  $m_1$  at the very last) and the segmented value of the plaintext immediately before obtaining  $m_1$ . The resulting values from the XOR operation are placed before  $m_1$ . If the calculation includes the segmented value of the plaintext  $m_t$  at the very front, the value ( $O_2$ ) can be derived as follows:

$$m_t \parallel m_2 \parallel \dots \parallel m_1 \oplus m_2 \parallel \dots \parallel m_1 = O_2$$



**Fig. 4.** Example 1

At this time, if the size between the segmented plaintext at the back and that at the front are different, the XOR operation is performed by adjusting the size of the segmented plaintext value at the back starting from the front and subsequently erasing the last part because the operation is possible only if the segmented plaintext size at the back is equal to that at the front.

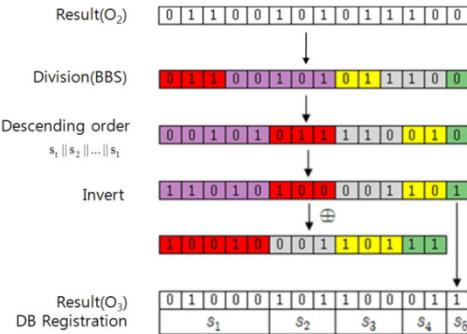
Next, from the resulting value  $O_2$ , segment  $O_2$  into  $s_1, s_2, \dots, s_t$  based on the random number  $r_i$  and rearrange them in descending order as follows:

$$s_1 \parallel s_2 \parallel \dots \parallel s_t \rightarrow s_t \parallel s_2 \parallel \dots \parallel s_1$$

Afterward, invert the value rearranged in descending order and create the resulting value  $O_3$  by performing the same XOR operation as in as follows:

$$\text{Invert}[s_t \parallel s_2 \parallel \dots \parallel s_1] \oplus s_2 \parallel \dots \parallel s_1 = O_3$$

Repeat the aforementioned procedure 16 times until the resulting value  $O_{16}$ , is created and compose the shares based on the fragmented shares of the finally created value  $O_{16}$ , storing them on k databases.



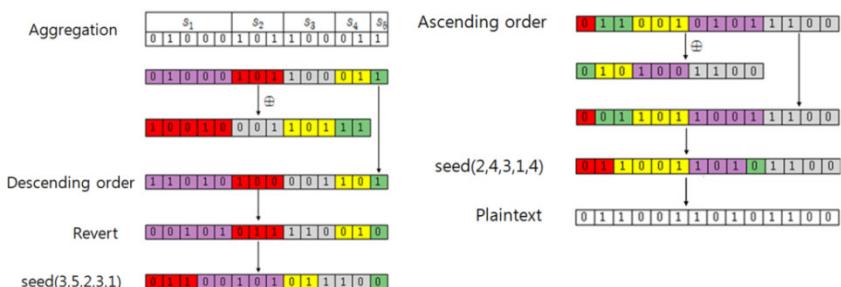
**Fig. 5.** Example 2

### 3.2.2 Data Reconstruction Stage

First, aggregate the data separately stored in k databases. Afterward, perform XOR operations on the segmented  $O_{16}$  values. Here, the XOR operation is performed in the same manner as in the distribution stage in . For the XOR operation, in the case in which it is different from the preceding size, the last part need not be erased to adjust the size of the segmented value for the operation as in the distribution stage; here, in the process of lowering the segmented value because it is at the very last and performing an XOR operation of the value that is lowered as it is, the XOR operation needs to be performed again for the immediately preceding part and for the resulting value derived from the XOR operation, not for the immediately preceding or succeeding part. At this time, if the preceding part is bigger, the succeeding small part should be used repeatedly.

Segment again the resulting values into random numbers and revert them. Here, the random value within the same seed range is known.

The final plaintext value can be obtained by repeating , above.



**Fig. 6.** Example 3

The proposed algorithm is summarized in Table 1. And, data reconstruction algorithm is summarized as follows (see Table 2):

**Table 1.** The proposed algorithm

```

INPUT : P {0,1}^t
OUTPUT : s1, ..., st
1. Generate ri(i=1,...,t) by BBS Generator
2. Calculate set s ⊂ {s0,...,st-1} which
   LENGTH(si)=ri andd si=SUBSTRING(P)riri+1
3. for R←0 to 16 do
4.   for i←0 to n-1 do
5.     if R mod 1 = 0 then
6.       b←MAX(s)
7.       ai←b
8.       discard MAX(s)
9.     else
10.      b←MIN(s)
11.      ai←Invert(b)
12.      discard MIN(s)
13.    end if
14.  end for i
15.  for j←0 to n-1 do
16.    if j=n-1 then
17.      sj←sj||(aj)
18.    else
19.      sj←(aj⊕PADD(aj+1)aj)
20.    end if
21.  end for j
22. end for R
23. return (s1, ..., st)

```

**Table 2.** Data reconstruction algorithm

```

INPUT : s1,...,st
OUTPUT : Plaintext P
1. Generate ri(i=1,...,t) by BBS Generator
2. Calculate set s ⊂ {s0,...,st-1} which
   LENGTH(si)=ri andd si=SUBSTRING
   (P)ri+1
3. for R←0 to 16 do
4.   for i←0 to n-1 do
5.     if R mod 1 = 1 then
6.       b←MAX(s)
7.       ai←b
8.       discard MAX(s)
9.     else
10.      b←MIN(s)
11.      ai←Invert(b)
12.      discard MIN(s)
13.    end if
14.  end for i
15.  for j←n-1 to 0 do
16.    if j=n-1 then
17.      sj←sj||(aj)
18.    else
19.      sj←(aj⊕PADD(Sj+1)aj)
20.    end if
21.  end for j
22. end for R
23. return (s1,...,st)

```

## 4 Analysis

### 4.1 Advantages of the Proposed Scheme

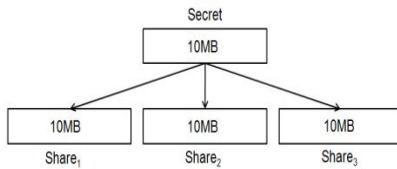
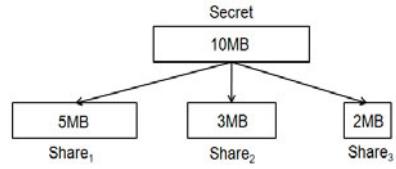
Compared with existing secret sharing algorithms, the proposed algorithms have the following advantages:

#### 1) Improved speed

The Shamir scheme uses the polynomial of degree (k-1) for secret sharing. This scheme incurs a heavy computational cost for making shares and recovering the secret. However, the proposed method, which uses the XOR operation and the pseudo-random number process, is faster than Shamir scheme.

#### 2) Minimization of data storage

The Shamir scheme increases the aggregate of distributed shares. However, the proposed scheme can minimize data storage required. For example, if 10 MB secret is distributed, the Shamir scheme requires 30 MB of storage space (Fig 7). But proposed scheme needs only 10 MB, the size of the secret (Fig 8).

**Fig. 7.** Shamir Scheme**Fig. 8.** Proposed Scheme

### 3) Allocation of storage space according to variable of share size

The  $(k,n)$  threshold secret sharing cannot variably process the size of share because the size of share distributed is the same size as secret. In this respect, there realistically exists the limitation that cannot be selectively stored in a database with various capacity according to size of share. That is, in case of storing secret information to distributed database, it has happened the efficient management problem of storage space by storing shares with same size regardless of the capacity of distributed database. Sharing the secret information, the proposed scheme is able to store share according to size of database because of variable selection of the share size is possible proportion to the capacity of each database. In case of size of share distributed is large, it is stored in a large capacity database, and other case (in case of size of share distributed is small) has a merit that can be stored in a portable storage device like USB.

### 4) Adjustment of the number of shares by share aggregation

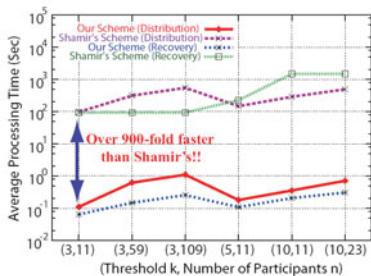
In case of existing  $(k,n)$  threshold secret sharing, a secret is divided into  $n$  pieces of share in the distribution stage. And it is impossible to adjust the number of shares to be less than  $n$  pieces once distributed. The characteristics of share aggregation enable the proposed scheme to adjust the number of share randomly. This facilitate the share management.

### 5) Suitability for large capacity database

In case of a large capacity database, it is desirable that data size is small as soon as possible for high speed processing. And data should be stored selectively according database size. The proposed scheme is suitable for a large capacity database in the point of high speed processing based on XOR and random number, the data minimization of storage space, and the variable selection of share size according to database capacity.

## 4.2 Comparison of Algorithms

Kurihara [7] carried out a simulation to compare the efficiency with the Shamir method based on  $k - 1$  degree polynomial  $f(x)$ . The result of this simulation is as the following picture shown below (see Figure 9).



**Fig. 9.** Computer Simulation: Comparing the processing time with Shamir's scheme (Left)

**Table 3.** Comparison of algorithms (Right)

Factor \ Scheme	Shamir	Kurihara	Proposed Algorithm
Average share size	$\ell$	$\ell$	$\frac{\ell}{n}$
Capability of fast processing	×	○	○
Suitability for large-capacity DBMS	×	△	○
Variability in share size	×	×	○
Variability in share number	×	×	○

(○ : Superior △ : Moderate × : Unsatisfactory)

Here, Shamir method is based on ARITHMETIC ( $GF(q)$ ); therefore, it is heavy computational cost for both distribution and recovery. However, the calculation speed was proved to be improved when the suggested method was simulated (Under the simulation by Kurihara, our scheme realizes much faster operations than Shamir's) on the basis of Kurihara's EXCLUSIVE-OR (XOR) calculation. In short, the proposed schemes are suitable for large-capacity database.

## 5 Conclusion

Ensuring the security of mobile RFID's large-capacity database system by depending only on existing encryption schemes is unrealistic. By adapting secure data management to mobile RFID middleware, we identified the security requirements for the safe storage and management of large-capacity data and AONT-based secret sharing/reconstruction scheme. We proposed the secret sharing algorithms that enable efficient data security management based on the characteristics of AONT encryption mode in RFID middleware.

The proposed scheme has realized the variability of share size and fast processing at the same time. The results of this paper are expected to be used in mobile RFID environment since they are structured to ensure the safe, efficient distribution management of large-capacity data such as highly confidential medical data and trade secrets including customers' personal information. As a further study, we will discuss the scheme to enable privilege management in case of the weight is given to share.

## References

1. Shamir, A.: How to Share a Secret. Communication of the ACM 22(11), 612–613 (1979)
2. Rivest, R.L.: All-or-nothing encryption and the package transform. In: Biham, E. (ed.) FSE 1997. LNCS, vol. 1267, pp. 210–218. Springer, Heidelberg (1997)
3. Kurihara, J., Kiyomoto, S., Fukushima, K., Tanaka, T.: A fast (3, n)-threshold secret sharingscheme using exclusive-or operations. IEICE Trans. Fundamentals E91-A(1), 127–138 (2008)

4. Fujii, Y., Tada, M., Hosaka, N., Tochikubo, K., Kato, T.: Fast (2, n)-threshold scheme and its application. In: Proc. CSS 2005, pp. 631–636 (2005)
5. Tada, M., Fujii, Y., Hosaka, N., Tochikubo, K., Kato, T.: A secret sharing scheme with threshold 3. In: Proc. CSS 2005, pp. 637–642 (2005)
6. Kuwakado, H., Tanaka, H.: Strongly non-separable encryption mode for throwing a media away. Technical Report of IEICE 103(417), 15–18 (2003)
7. Kurihara, J., Kiyomoto, S., Fukushima, K., Tanaka, T.: A New (k, n)-Threshold Secret Sharing Scheme and Its Extension. In: Wu, T.-C., Lei, C.-L., Rijmen, V., Lee, D.-T. (eds.) ISC 2008. LNCS, vol. 5222, pp. 455–470. Springer, Heidelberg (2008)
8. Blum, L., Blum, M., Shub, M.: A Simple Unpredictable Pseudo-Random Number Generator. SIAM Journal on Computing 15, 364–383 (1986)
9. Park, N., Song, Y.: Secure RFID Application Data Management Using All-Or-Nothing Transform Encryption. In: Li, Y. (ed.) WASA 2010. LNCS, vol. 6221, pp. 245–252. Springer, Heidelberg (2010)
10. Park, N., Kim, S., Won, D., Kim, H.: Security Analysis and Implementation leveraging Globally Networked Mobile RFIDs. In: Cuenca, P., Orozco-Barbosa, L. (eds.) PWC 2006. LNCS, vol. 4217, pp. 494–505. Springer, Heidelberg (2006)
11. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)

# A Query Answering Greedy Algorithm for Selecting Materialized Views

T.V. Vijay Kumar and Mohammad Haider

School of Computer and Systems Sciences,  
Jawaharlal Nehru University,  
New Delhi-110067,  
India

**Abstract.** Materialized views aim to improve the response time of analytical queries posed on a data warehouse. This entails that they contain information that provides answers to most future queries. The selection of such information from the data warehouse is referred to as view selection. View selection deals with selection of appropriate sets of views to improve the query response time. Several view selection algorithms exist in literature, most of them being greedy based. The greedy algorithm HRUA, which selects top-k views from a multidimensional lattice, is considered the most fundamental greedy based algorithm. It selects views having the highest benefit, computed in terms of size, for materialization. Though the views selected using HRUA are beneficial with respect to size, they may not account for a large number of future queries and may hence become an unnecessary overhead. This problem is addressed by the Query Answering Greedy Algorithm (QAGA) proposed in this paper. QAGA uses both the size of the view, and the frequency of previously posed queries answered by each view, to compute the profits of all views in each iteration. Thereafter it selects, from among them, the most profitable view for materialization. QAGA is able to select views which are beneficial with respect to size and have a greater likelihood of answering future queries. Further, experimental results show that QAGA, as compared to HRUA, is able to select views capable of answering greater number of queries. Though HRUA incurs a lower total cost of evaluating all the views, QAGA has a lower total cost of answering all the queries leading to an improvement in the average query response time. This in turn facilitates decision making.

**Keywords:** Data Warehouse, Materialized View Selection, Greedy Algorithm, Query Response Time.

## 1 Introduction

Large amounts of data are continuously being generated by disparate data sources spread across the globe. This data needs to be accessed and exploited by organizations in order to have an edge over their competitors. One way to access this data is by gathering data from disparate data sources, integrating and storing it in a repository and then posing queries against the repository. This approach, referred to as the eager

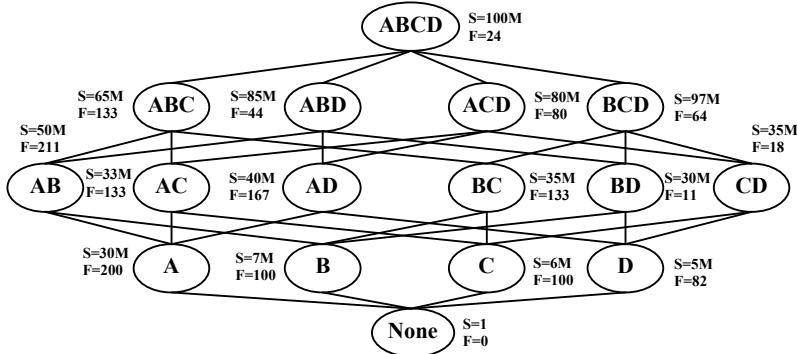
or in-advance approach, is followed in the case of a data warehouse[23], which stores subject-oriented, integrated, time-variant and non-volatile data to support decision making [9]. The queries posed for decision making are usually exploratory and analytical in nature. These queries, being long and complex, consume a lot of time when processed against a large data warehouse. As a result the query response time is high. Although effective solutions for data representation suitable for analytical queries exists, the response time issue still needs to be addressed adequately[16]. To some extent, this problem has been addressed using traditional query optimization[4, 6] and indexing techniques[13]. However, with queries becoming more complex, increasing response times may become unacceptable for decision making[11]. An alternate way to improve the query response time is by using materialized views[14]. Materialized views contain pre-computed and summarized information primarily to improve response time for analytical queries. This necessitates that these materialized views contain relevant and required information for efficient query processing. The selection of such materialized views is referred to as the view selection problem [5, 18] in literature.

View selection is formally defined in [5] as “Given a database schema R, storage space B, and a workload of queries Q, choose a set of views V over R to materialize, whose combined size is at most B”. Since the number of possible views is exponential in the number of dimensions, all possible views cannot be materialized as they would violate the space constraint. There is a need to select an optimal subset of views, from among all possible views, which is shown to be NP-Complete[8]. Alternatively, views can be selected empirically or heuristically[17]. In the former[10], the views are selected based on past query patterns and factors like frequency and data volume, whereas in the latter, the views are selected by pruning the search space based on heuristics like greedy[21], evolutionary[24] etc. This paper focuses on the greedy based selection of materialized views, which is discussed next.

Several greedy based algorithms have been proposed in the literature[1, 2, 3, 7, 8, 12, 15, 16, 19, 20, 22], most of which are focused around the algorithm in [8], which hereafter in this paper will be referred to as HRUA. HRUA selects the top-k beneficial views from a multidimensional lattice. It is based on a linear cost model, where cost, defined in terms of the size of the view, is used to compute the benefit of each view. HRUA, in each iteration, computes the benefit of all the non-selected views and selects the one with highest benefit, with respect to size, for materialization. It may be possible that the selected views though beneficial with respect to size, may not be so beneficial with respect to answering large number of future queries, thereby becoming an unnecessary space overhead. One way to address this problem is by considering the frequency of queries answered by each view, along with its size, to compute the benefit of the view. This would lead to selection of views that are not only beneficial with respect to size but can also provide answers to most future queries.

As an example, consider a four dimensional lattice shown in Fig. 1. The size of the view in million (M) rows, and the query frequency of each view, is given alongside the view. Suppose top-4 views are to be selected.

HRUA considers the size of the views in the multidimensional lattice to select top-4 views for materialization. These selections are shown in Fig. 2.



**Fig. 1.** 4-Dimensional Lattice with size and query frequency of each view

Views	Benefit				Views	QF	QAV	C	TC
	1 <sup>st</sup> Iteration	2 <sup>nd</sup> Iteration	3 <sup>rd</sup> Iteration	4 <sup>th</sup> Iteration					
ABC	280 M				ABCD	24	ABCD	100M	2400M
ABD	120 M	60 M	30 M	30 M	ABC	133	ABC	65M	8645M
ACD	160 M	80 M	60 M	60 M	ABD	44	ABCD	100M	4400M
BCD	24 M	12 M	06 M	06 M	ACD	80	ABCD	100M	8000M
AB	200 M	60 M	30 M	15 M	BCD	64	ABCD	100M	6400M
AC	268 M	128 M	96 M		AB	211	ABC	65M	13715M
AD	240 M	170 M	85 M	60 M	AC	133	AC	33M	4389M
BC	260 M	120 M	60 M	30 M	AD	167	ABCD	100M	16700M
BD	280 M	210 M			BC	133	ABC	65M	8645M
CD	260 M	190 M	95 M	65 M	BD	11	BD	30M	330M
A	140 M	70 M	35 M	03 M	CD	18	CD	35M	630M
B	186 M	116 M	46 M	46 M	A	200	AC	33M	6600M
C	188 M	118 M	83 M	51 M	B	100	BD	30M	3000M
D	190 M	155 M	50 M	50 M	C	100	AC	33M	3300M
				D	82	BD	30M	2460M	
				TQ	1500	TAC			89614M

(a)

(b)

**Fig. 2.** Selection of top-4 views using HRUA

TQ= Total Queries, QF=Query Frequency of view, QAV=Query Answering View, C=Cost of view, TC=Total Cost of query answering by view, TAC=Total Answering Cost of all views.

HRUA selects ABC, BD, AC and CD as the top-4 views. HRUA is able to reduce the total cost of evaluating all these views, referred to as Total View Evaluation Cost (TVEC), from 1600M to 949M, yielding a total benefit of 651M. If we consider the query frequency associated with each view then the number of queries answered, referred to as Total Queries Answered (TQA), by views selected using HRUA is 1121 from among 1500 queries. That is, the remaining 379 queries would need to be answered by the root view ABCD, which is assumed to be materialized as queries on it cannot be answered by any other view in the lattice. Considering the query frequency of each view and the views selected by HRUA, the total cost of answering all the queries, referred to as the Total Answering Cost(TAC), is computed as shown in Fig. 2(b). HRUA requires a TAC of 89614M to process 1500 queries. This value if reduced would result in an improvement in the average query response time. One way to reduce this value is by selecting views that are also capable of answering a greater number of queries so that only few queries would require to be answered by referring to the root view.

Though this may result in an increase in the TVEC value, an acceptable trade-off needs to be achieved between the TVEC and the TQA so that the selected views are not only beneficial with respect to size but are also able to account for large numbers of queries. A Query Answering Greedy Algorithm (QAGA) has been proposed in this paper that attempts to address this problem by selecting top-k views based on both the size of the view and the frequency of previously posed queries answered by the view, referred to as the query frequency of the view. QAGA is focused around HRUA but considers, in each iteration, query frequency and size of each view to compute their profit and then greedily selects the most profitable view for materialization.

The paper is organized as follows: The proposed algorithm QAGA is discussed in section 2 and examples based on it are given in section 3. Section 4 experimentally compares QAGA and HRUA. The conclusion is given in section 5.

## 2 QAGA

View selection aims to select such views that improve the query response time. As discussed above, HRUA considers only the size of the view and may therefore select views that may be unable to provide answers to large numbers of queries resulting in poor query response times. This problem is addressed by the proposed algorithm QAGA, which considers the query frequency of each view, along with its size, to select top-k views for materialization. The algorithm QAGA is given in Fig. 3. QAGA takes the lattice of views, along with the size and query frequency of each view, as input and produces the top-k views as output.

```

Input: lattice of views L along with size and query frequency of each view
Output: top-k views
Method:
    L: set of views
    RootView: root view in the lattice
    Size(V): size of view V in the lattice
    QFreq(V): query frequency of view V in the lattice
    NMA(V): nearest materialized ancestor of view V in the lattice.
    Desc(V): set of all descendant views of the view V in the lattice
    MV: set of materialized views.
    For  $V \in L$ 
        NMA(V) = RootView
    End
    Repeat
        MaxProfit = 0
        For each view  $V \in (L - \{RootView\} \cup MV)$ 
            ProfitableView = V
            Profit(V) = 0
            For each view  $W \in Desc(V)$  and  $(Size(NMA(W)) - Size(V)) > 0$ 
                Profit(V) = Profit(V) + QFreq(NMA(W))*Size(NMA(W)) - QFreq(V)*Size(V)
            End
            If MaxProfit < Profit(V)
                MaxProfit = Profit(V)
                ProfitableView = V
            End
        End
        MV = MV  $\cup$  {ProfitableView}
        For  $W \in Desc(ProfitableView)$ 
            If  $Size(NMA(W)) > Size(ProfitableView)$ 
                NMA(W) = ProfitableView
            End
        End
    Until  $|MV| < k$ 
    Return MV

```

**Fig. 3.** Algorithm QAGA

QAGA, in each iteration, computes the profit of each view as a product of the number of its descendants and the difference in the product of query frequency and size of its smallest nearest materialized ancestor and the view itself. The profit of a view, i.e. Profit(V), is given as:

$$\text{Profit}(V) = \sum \{(Q\text{Freq}(NMA(W)) \times \text{Size}(NMA(W)) - Q\text{Freq}(V) \times \text{Size}(V)) \mid V \in A(W) \wedge (\text{Size}NMA(W) - \text{Size}(V)) > 0\}$$

where

$\text{Size}(V)$ =Size of view V,  $\text{SizeNMA}(V)$ =Size of Nearest Materialized Ancestor of view V,  $Q\text{Freq}(V)$ =Query frequency of view V,  $Q\text{Freq}(NMA(V))$ =Query frequency of Nearest Materialized Ancestor of view V,  $A(W)$ =Ancestor of view W.

The profit takes into consideration the query frequency along with the size of the view, as considering only the size may result in selecting a view that may reduce TVEC without being able to answer an acceptably large number of queries. This may be due to the fact that the query answering ability of the view is not considered for computing its benefit. QAGA considers this fact to compute the profit of each view, in each iteration, and then selects from amongst them the most profitable view for materialization. Examples solved using QAGA are illustrated in the next section.

### 3 Examples

Let us consider the selection of top-4 views, using QAGA from the lattice shown in Fig. 1. The greedy selection of top-4 views using QAGA is given in Fig. 4.

Views	Profit			
	1 <sup>st</sup> Iteration	2 <sup>nd</sup> Iteration	3 <sup>rd</sup> Iteration	4 <sup>th</sup> Iteration
ABC	49960 M			
ABD	10720 M	5360 M	2680 M	2680 M
ACD	32000 M	16000 M	12000 M	12000 M
BCD	30464 M	15232 M	7616 M	7616 M
AB	32600 M	7620 M	3810 M	1905 M
AC	7956 M	17024 M	12768 M	
AD	17120 M	12490 M	6245 M	4280 M
BC	9020 M	15960 M	7980 M	3990 M
BD	8280 M	20770 M		
CD	7080 M	19570 M	9785 M	1770 M
A	7200 M	5290 M	2645 M	1611 M
B	3400 M	15890 M	740 M	740 M
C	3600 M	16090 M	8315 M	4059 M
D	3980 M	10225 M	160 M	160 M

Views	QF	QAV	C	TC
ABCD	24	ABCD	100M	2400M
ABC	133	ABC	65M	8645M
ABD	44	ABCD	100M	4400M
ACD	80	ACD	80M	6400M
BCD	64	ABCD	100M	6400M
AB	211	ABC	65M	13715M
AC	133	AC	33M	4389M
AD	167	ACD	80M	13360M
BC	133	ABC	65M	8645M
BD	11	BD	30M	330M
CD	18	ACD	80M	1440M
A	200	AC	33M	6600M
B	100	BD	30M	3000M
C	100	AC	33M	3300M
D	82	BD	30M	2460M
TQ	1500	TAC		85484M

(a)

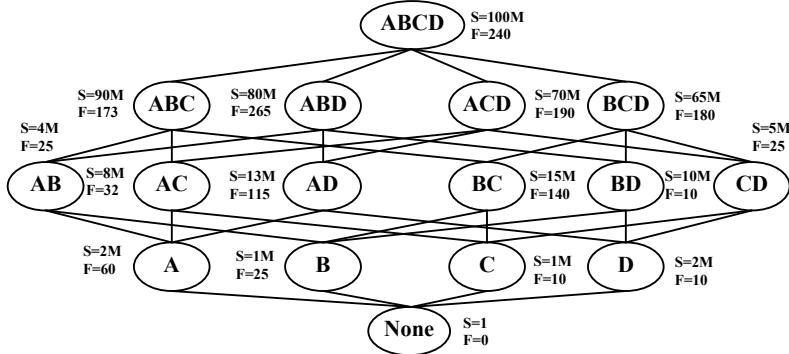
(b)

**Fig. 4.** Selection of top-4 views using QAGA

QAGA selects ABC, BD, AC and ACD as the top-4 views. These views are able to reduce the TVEC from 1600M to 954M, which is slightly more than the TVEC of 949M achieved by views selected using HRUA. Moreover, the views selected using QAGA have a TQA of 1368, which is significantly more than the TQA of 1125 for views selected using HRUA. As a result, views selected using QAGA has a TAC for 1500 queries as 85484M, which is less than 89614M due to views selected using HRUA. This would result in an improvement in the average query response time.

Thus, it can be said that QAGA, in comparison to HRUA, is able to select views that account for a comparatively larger number of queries, resulting in improved average query response time, against a slight increase in the total cost of evaluating all the views. It, accordingly achieves an acceptable trade-off between TVEC and TQA.

QAGA need not always select views with higher TVEC values. As an example, consider the four dimensional lattice shown in Fig. 5.



**Fig. 5.** 4-Dimensional Lattice with size and query frequency of each view

Selection of top-4 views using HRUA and QAGA is given in Fig. 6 and Fig. 7 respectively. The views AB, CD, BCD and AC, selected by QAGA, have a lower TVEC value, i.e. 718M, in comparison to 734M for the views BCD, AC, AB and AD selected using HRUA. Further, the TQA value due to QAGA is 632, which is more than the TQA of 517 attained by HRUA. Also, the TAC value for views selected using QAGA to process all the 1500 queries is 112276M, which is less than 120671M of views selected using HRUA. Thus it can be said that QAGA can select views that not only have a better TQA but also a better TVEC when compared with views selected using HRUA. The higher value of TQA due to views selected using QAGA has resulted in a comparatively lower TAC value.

Views	Benefit			
	1 <sup>st</sup> Iteration	2 <sup>nd</sup> Iteration	3 <sup>rd</sup> Iteration	4 <sup>th</sup> Iteration
ABC	80 M	40 M	30 M	20 M
ABD	160 M	80 M	60 M	40 M
ACD	240 M	180 M	90 M	90 M
BCD	280 M	210 M	105 M	
AB	384 M			
AC	368 M	184 M	92 M	92 M
AD	348 M	174 M	87 M	87 M
BC	340 M	170 M	85 M	50 M
BD	360 M	180 M	90 M	55 M
CD	380 M	285 M		
A	196 M	04 M	04 M	04 M
B	198 M	06 M	06 M	06 M
C	198 M	102 M	07 M	07 M
D	196 M	100 M	05 M	05 M
<b>TQ</b>	<b>1500</b>	<b>TAC</b>	<b>120671M</b>	

Views	QF	QAV	C	TC
ABCD	240	ABCD	100M	24000M
ABC	173	ABCD	100M	17300M
ABD	265	ABCD	100M	26500M
ACD	190	ABCD	100M	19000M
BCD	180	BCD	65M	11700M
AB	25	AB	4M	100M
AC	32	AC	8M	256M
AD	115	ABCD	100M	11500M
BC	140	BCD	65M	9100M
BD	10	BCD	65M	650M
CD	25	CD	5M	125M
A	60	AB	4M	240M
B	25	AB	4M	100M
C	10	CD	5M	50M
D	10	CD	5M	50M
<b>TQ</b>	<b>1500</b>	<b>TAC</b>	<b>120671M</b>	

**Fig. 6.** Selection of top-4 views using HRUA

Views	Profit				Views	QF	QAV	C	TC
	1 <sup>st</sup> Iteration	2 <sup>nd</sup> Iteration	3 <sup>rd</sup> Iteration	4 <sup>th</sup> Iteration					
ABC	67440 M	33720 M	16860 M	8430 M	ABCD	240	ABCD	100M	24000M
ABD	72400 M	11200 M	5600 M	5600 M	ABC	173	ABCD	100M	17300M
ACD	85600 M	42800 M	32100 M	21400 M	ABD	265	ABCD	100M	26500M
BCD	98400 M				ACD	190	ABCD	100M	19000M
AB	95600 M	71000 M			BCD	180	BCD	65M	11700M
AC	94976 M	70376 M	35188 M		AB	25	AB	4M	100M
AD	90020 M	65420 M	32710 M	32710 M	AC	32	AC	8M	256M
BC	87600 M	38400 M	19200 M	9600 M	AD	115	AD	13M	1495M
BD	95600 M	46400 M	23200 M	23200 M	BC	140	BCD	65M	9100M
CD	95500 M	46300 M	34725 M	23281 M	BD	10	BCD	65M	650M
A	47760 M	35460 M	40 M	40 M	CD	25	BCD	65M	1625M
B	47950 M	23350 M	150 M	150 M	A	60	AB	4M	240M
C	47980 M	23380 M	11780 M	336 M	B	25	AB	4M	100M
D	47960 M	23360 M	11760 M	11760 M	C	10	AC	8M	80M
					D	10	AD	13M	130M
					TQ	1500	TAC		112276M

**Fig. 7.** Selection of top-4 views using QAGA

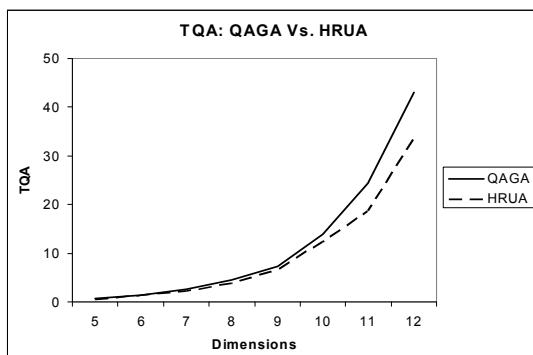
From the above, it can be inferred that the profit computation of a view in QAGA is fairly good as it is able to select views as good as those selected using HRUA.

In order to compare the performance of QAGA with HRUA, both the algorithms were implemented and run on data sets with varying dimensions. The experimental based comparisons of QAGA and HRUA are given next.

## 4 Experimental Results

The algorithm QAGA and HRUA were implemented using JDK 1.6 in a Windows-XP environment. The two algorithms were compared by conducting experiments on an Intel based 2 GHz PC having 1 GB RAM. The comparisons were carried out on parameters like TVEC, TQA and TAC. The tests were conducted by varying the number of dimensions of the data set from 5 to 12. The experiments were performed for selecting top-20 views for materialization.

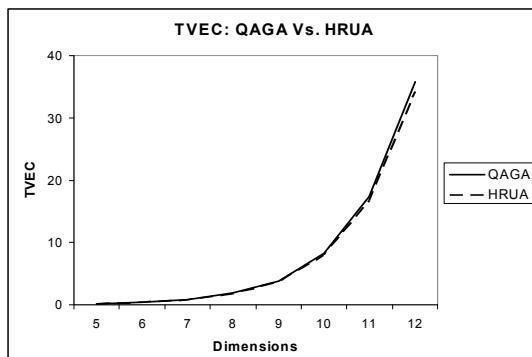
First, graphs were plotted to compare QAGA and HRUA on TQA against the number of dimensions. The graph is shown in Fig. 8.



**Fig. 8.** TQA – QAGA Vs. HRUA

It is observed from the above that the increase in TQA value, with respect to number of dimensions, is more for QAGA vis-à-vis HRUA. For dimensions 11 and 12, the TQA value of QAGA is significantly higher than that of HRUA. This shows that the views selected using QAGA perform better than those selected using HRUA in terms of total queries answered by them.

In order to study the impact of better TQA, due to QAGA, on the TVEC, a graph for TVEC (in million rows) versus Dimensions is plotted and is shown in Fig. 9.

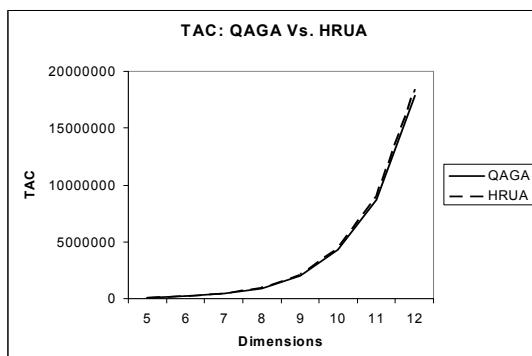


**Fig. 9.** TVEC - QAGA Vs. HRUA

It can be observed from the graph that, with increase in the number of dimensions, the increase in the TVEC value of HRUA is slightly lower than that of QAGA indicating that HRUA performs slightly better than QAGA with respect to total cost of evaluating all the views.

Thus it can be inferred from the above that QAGA, in comparison to HRUA, performs better in TQA with HRUA having a slight edge in TVEC.

Further, in order to compare QAGA and HRUA in terms of the total cost incurred for answering all the queries, a graph for Total Answering Cost (in million rows) Vs. Dimensions is plotted and is shown in Fig. 10.



**Fig. 10.** TAC - QAGA Vs. HRUA

The graph shows that with an increase in the number of dimensions, the TAC due to views selected using QAGA is slightly lower than those selected using HRUA. Though HRUA has a slight edge over QAGA in terms of TVEC, the latter incurs less cost in terms of answering all the queries. As a result views selected using QAGA would result in a better average query response time.

It can be reasonably inferred from the above graphs that QAGA trades significant improvement in TQA with a slight drop in TVEC of views selected for materialization. This has lead to lesser TAC for QAGA, which in turn would result in better average query response time.

## 5 Conclusion

In this paper, an algorithm QAGA, that uses both the size and query frequency of a view to select the top-k profitable views from a multidimensional lattice, is proposed. In each iteration, QAGA computes the profit of each individual view, defined in terms of the size and query frequency of the view. This is followed by selecting, from amongst them, the most profitable view for materialization. Unlike most of the greedy based algorithms for view selection, QAGA selects views which are not only profitable with respect to size but are also capable of answering a large number of queries. Also, the views selected using QAGA not only have a higher TQA but also, generally, have a lower TVEC when compared with HRUA. Accordingly, QAGA has an edge over HRUA in the TAC value.

Furthermore, experimental based comparisons show that QAGA performs significantly better than HRUA vis-à-vis TQA value due to the views selected by the two algorithms. Though HRUA has a slight edge over QAGA in terms of the TVEC value, QAGA seems to perform better on the TAC value. This shows that QAGA is not only able to select views which provide answers to a greater number of queries but it is also able to provide answers to all the queries at comparatively lesser cost leading to an improvement in the average query response time. This in turn would facilitate decision making.

## References

1. Agrawal, S., Chaudhuri, S., Narasayya, V.: Automated Selection of Materialized Views and Indexes in SQL Databases. In: Proceedings of VLDB 2000, pp. 496–505. Morgan Kaufmann Publishers, San Francisco (2000)
2. Aouiche, K., Darmont, J.: Data mining-based materialized view and index selection in data warehouse. Journal of Intelligent Information Systems, 65–93 (2009)
3. Baralis, E., Paraboschi, S., Teniente, E.: Materialized View Selection in a Multidimensional Database. In: Proceedings of VLDB 1997, pp. 156–165. Morgan Kaufmann Publishers, San Francisco (1997)
4. Chaudhuri, S., Shim, K.: Including Groupby in Query Optimization. In: Proceedings of the International Conference on Very Large Database Systems (1994)
5. Chirkova, R., Halevy, A., Suciu, D.: A Formal Perspective on the View Selection Problem. The VLDB Journal 11(3), 216–237 (2002)

6. Gupta, A., Harinarayan, V., Quass, D.: Generalized Projections: A Powerful Approach to Aggregation. In: Proceedings of the International Conference of Very Large Database Systems (1995)
7. Gupta, H., Mumick, I.: Selection of Views to Materialize in a Data Warehouse. *IEEE Transactions on Knowledge and Data Engineering* 17(1), 24–43 (2005)
8. Harinarayan, V., Rajaraman, A., Ullman, J.: Implementing Data Cubes Efficiently. In: Proceedings of SIGMOD 1996, pp. 205–216. ACM Press, New York (1996)
9. Inmon, W.H.: Building the Data Warehouse, 3rd edn. Wiley Dreamtech, Chichester (2003)
10. Lehner, R., Ruf, T., Teschke, M.: Improving Query Response Time in Scientific Databases Using Data Aggregation. In: Proceedings of 7th International Conference and Workshop on Databases and Expert System Applications, pp. 9–13 (September 1996)
11. Mohania, M., Samtani, S., Roddick, J., Kambayashi, Y.: Advances and Research Directions in Data Warehousing Technology. *Australian Journal of Information Systems* (1998)
12. Nadeau, T.P., Teorey, T.J.: Achieving scalability in OLAP materialized view selection. In: Proceedings of DOLAP 2002, pp. 28–34. ACM Press, New York (2002)
13. O’Neil, P., Graefe, G.: Multi-Table joins through Bitmapped Join Indices. *SIGMOD Record* 24(3), 8–11 (1995)
14. Roussopoulos, N.: Materialized Views and Data Warehouse. In: 4th Workshop KRDB 1997, Athens, Greece (August 1997)
15. Serna-Encinas, M.T., Hoya-Montano, J.A.: Algorithm for selection of materialized views: based on a costs model. In: Proceeding of Eighth International Conference on Current Trends in Computer Science, pp. 18–24 (2007)
16. Shah, A., Ramachandran, K., Raghavan, V.: A Hybrid Approach for Data Warehouse View Selection. *International Journal of Data Warehousing and Mining* 2(2), 1–37 (2006)
17. Teschke, M., Ulbrich, A.: Using Materialized Views to Speed Up Data Warehousing. Technical Report, IMMD 6, Universität Erlangen-Nürnberg (1997)
18. Theodoratos, D., Bouzeghoub, M.: A general framework for the view selection problem for data warehouse design and evolution. In: Proceedings of DOLAP, pp. 1–8 (2000)
19. Valluri, S.R., Vadapalli, S., Karlapalem, K.: View Relevance Driven Materialized View Selection in Data Warehousing Environment. In: Proceedings of CRPITS 2002, Darlinghurst, Australia, pp. 187–196. Australian Computer Society (2002)
20. Vijay Kumar, T.V., Ghoshal, A.: A Reduced Lattice Greedy Algorithm for Selecting Materialized Views. In: Communications in Computer and Information Science (CCIS), vol. 31, pp. 6–18. Springer, Heidelberg (2009)
21. Vijay Kumar, T.V., Ghoshal, A.: Greedy Selection of Materialized Views. *International Journal of Computer and Communication Technology (IJCCT)*, 47–58 (2009)
22. Vijay Kumar, T.V., Haider, M., Kumar, S.: Proposing Candidate Views for Materialization. In: Communications in Computer and Information Science (CCIS), vol. 54, pp. 89–98. Springer, Heidelberg (2010)
23. Widom, J.: Research Problems in Data Warehousing. In: Proceedings of ICIKM, pp. 25–30 (1995)
24. Zhang, C., Yao, X., Yang, J.: Evolving Materialized views in Data Warehouse. In: Proceedings of the Congress on Evolutionary Computation, vol. 2, pp. 823–829 (1999)

# Capturing Users' Buying Activity at Akihabara Electric Town from Twitter

The-Minh Nguyen, Takahiro Kawamura,  
Yasuyuki Tahara, and Akihiko Ohsuga

Graduate School of Information Systems, The University of Electro-Communications,  
Tokyo, Japan  
`{minh,kawamura,tahara,akihiko}@ohsuga.is.uec.ac.jp`

**Abstract.** The goal of this paper is to describe a method to automatically capture users' buying activity at Akihabara electric town in each sentence retrieved from twitter. Sentences retrieved from twitter are often diversified, complex, syntactically wrong, have emoticons and new words. There are some works that have tried to extract users' activities in sentences retrieved from weblogs and twitter. However, these works have some limitations, such as inability of extracting infrequent activities, high setup cost, limitation on the types of sentences that can be handled, necessary of preparing a list of object, action and syntax patterns. To resolve these problems, we propose a novel approach that treats the activity extraction as a sequence labeling problem, and automatically makes its own training data. This approach can extract infrequent activities, and has advantages such as scalability, unnecessary any hand-tagged data.

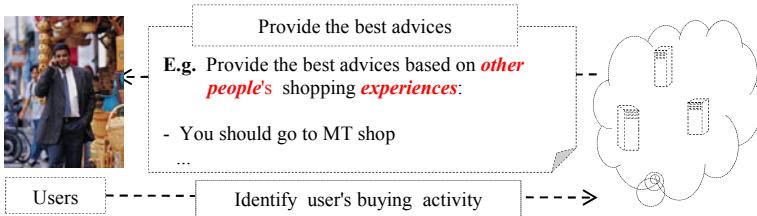
**Keywords:** Real-world Activity, Twitter Mining, Self-Supervised Learning, Context-aware, Users' Behaviors.

## 1 Introduction

The ability of computers to capture users' buying activity is now an important issue in context-aware mobile advertising[4], ubiquitous computing[12], and can be applied to many business models. For example, as shown in Figure 1, the computers can provide the best advices based on other people's shopping experiences.

Today, many uses are posting their buying activities to twitter[1]. Therefore, sentences in twitter are becoming sensors of the real world[5]. Our goal is to automatically extract buying activities in Japanese sentences retrieved from twitter. In this paper, we define a buying activities by two attributes namely *action* and *object*. For example, in the sentence “iPad wo kau (buy an iPad)”, *action* and *object* are “kau (buy)” and “iPad”, respectively. However, sentences retrieved from twitter which are more complex than other text media, are often structurally varying, syntactically incorrect, and have many new words. Thus, there are lots of challenges to extract buying activities in these sentences. Previous

work[27,28] which are based on the co-occurrence of action and object, do not depend on the retrieved sentences syntax. However, this approach can not extract infrequent activities, and have to prepare a list of action and object before extracting. There are some other works[6,11,24,30] have tried to extract human activities from the web and weblogs. These works have some limitations, such as high setup costs because of requiring ontology for each domain [11]. Due to the difficulty of creating suitable patterns, these works are limited on the types of sentences that can be handled [6,24], have to prepare a list of Japanese syntax patterns[30].



**Fig. 1.** Experience sharing service

Based on twitter search API[2], we collect buying activity Japanese sentences which are relating to Akihabara, the most famous electric town in Tokyo, Japan. In this paper, we propose a novel approach that automatically extract buying activities in these sentences. This approach makes its own training data (self-supervised learning), and uses conditional random fields (CRFs)[18] as a learning model. The main contributions of our approach are summarized as follows:

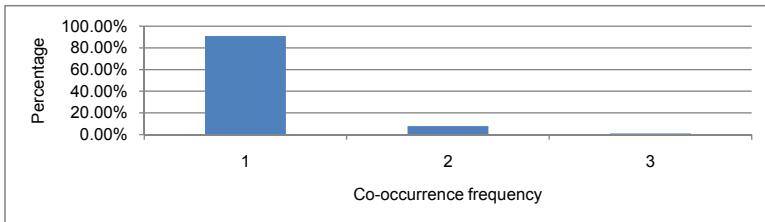
- It can capture users' buying activities at Akihabara from twitter.
- It can extract infrequent activities.
- It does not require *any* hand-tagged data.
- It can extract *all* buying activities by making only a *single pass over its corpus*.

The remainder of this paper is organized as follows. In section 2, we indicate challenges of extracting buying activity in more detail. Section 3 explains how our approach makes its own training data, and extracts buying activity in each sentence retrieved from twitter. Section 4 reports our experimental results. Section 5 considers related work. Section 6 consists of conclusions and some discussions of future work.

## 2 Challenges

Extracting buying activity in each sentence retrieved from twitter has many challenges, especially in Japanese. Below, we explain some of them:

1. By using twitter search API, we collected all buying activity sentences related to Akihabara which are generated in period of June 12-16, 2010. We manually calculated co-occurrence frequency of action (buy) and object in the retrieved sentences. As shown in Figure 2, percentage of single co-occurrence, double co-occurrences, triple co-occurrences are 90.91%, 7.95%, 1.24%, respectively. Therefore, we *can not base on co-occurrence* to extract these buying activities.



**Fig. 2.** Co-occurrence frequency of action and object

2. It is not practical to directly deploy deep linguistic parsers, because sentences retrieved from twitter are often diversified, complex, syntactically wrong, have emoticons and new words.
3. In Japanese, there are not word spaces, and word boundaries are not clear. However, previous works in CRFs assume that observation sequence (word) boundaries were fixed. Therefore, a straightforward application of CRFs is impossible. Additionally, unlike English, Japanese sentences do not follow the Subject-Verb-Object rule; they have many types of structures and are flexible.

### 3 Extracting Buying Activity

#### 3.1 Activity Extraction with CRFs

CRFs[18] is a type of discriminative probabilistic model for segmenting and labeling sequence data. The idea is to define a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. CRFs offers several advantages over hidden Markov models and stochastic grammars, including the ability to relax strong independence assumptions made in those models. Additionally, CRFs also avoids the label bias problem, a weakness exhibited by maximum entropy Markov models (MEMMs) and other conditional Markov models based on directed graphical models. CRF achieves high precision on many tasks including text chunking[19], named entity recognition[20], Japanese morphological analysis[21].

By making a first-order Markov assumption that the dependencies between output variables, and arranging variables sequentially in a linear chain, activity

extraction can be treated as a sequence labeling problem. In this paper, we use CRFs as a learning model for activity extraction. The set of features includes words, verbs, part-of-speech tags and postpositional particles in Japanese. To model long-distance relationships this paper uses a window of size 7. Figure 3 shows an example that activity extraction is treated as a sequence labeling problem. Tokens in the surrounding context are labeled using the IOB2 format. B-X means “begin a phrase of type X”, I-X means “inside a phrase of type X” and O means “not in a phrase”. IOB2 format is widely-used for natural language tasks[26].

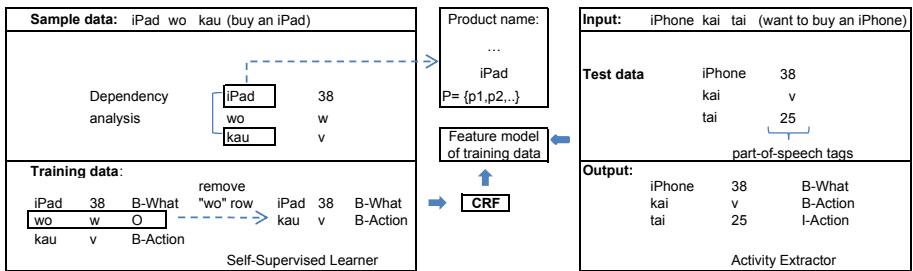
B-What	O	B-Action	O
iPad	wo	kau (buy)	yo

(buy an iPad)

**Fig. 3.** Labeling attributes of a buying activity

### 3.2 Proposed Architecture

As shown in Figure 4, the architecture consists of two modules: *Self-Supervised Learner* and *Activity Extractor*. Based on dependency analytic results and a list of product name (which are automatically extracted from input sentences), the Leaner selects trustworthy attributes to make training data. The Extractor does *not* deploy deep linguistic parser, it bases on the feature model of training data to automatically label object and action of buying activity in each sentence retrieved from twitter. Below, we describe each module in more detail.



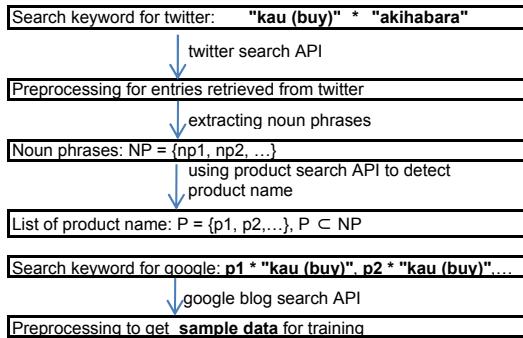
**Fig. 4.** Proposed Architecture

#### 3.2.1 Self-supervised Learner

As shown in Figure 5, the process of making sample data consists of five main steps.

1. We use “kau (buy) \* “akihabara” as a search keyword to collect buying activity sentences from twitter (“\*” means wildcard).
2. After of pre-processing (deleting html tags, stop words), we use Mecab<sup>1</sup> to extract noun phrases in the retrieved sentences.

<sup>1</sup> Available at <http://mecab.sourceforge.net/>



**Fig. 5.** Process of making sample data

3. We use a product search API[29] to detect product name in the retrieved noun phrases.
4. We combine the above detected product name with “kau (buy)” to create search keywords. Based on these keywords, we collect new buying activity sentences by using Google blog search API.
5. We select the sentences which contain “kau (buy)” and at least one product name. Then, we remove html tags, emoticons in these sentences to make sample data.

Given the sample data as input, the Learner automatically make its own training data as follows:

1. It uses deep linguistic parser to analyze dependencies between “kau (buy)” and other noun phrases.
2. In addition to this analytic result, it based on the list of product name to select trustworthy object to make training data.
3. Sentences retrieved from twitter do not often contain “wo” (an important postpositional particles in an activity Japanese sentence). Therefore, we remove “wo” row in the original training data to add a new training data for training.

### 3.2.2 Activity Extractor

As shown in Figure 4, the Extractor consists of two key tasks:

1. In Japanese sentences, there are not word spaces, and word boundaries are not clear. Therefore, the Extractor uses Mecab to get words and their part-of-speech tags.
2. Based on the feature model of training data, the Extractor automatically label all attributes.

## 4 Evaluation

By using twitter search API, we use “kau (buy)” \* “akihabara” as a search keyword to collect sentences related to Akihabara which are generated in period

of June 12-16, 2010. After of pre-processing, we obtained 88 buying activity sentences. We used these sentences for the experiment. Based on product search API, we detected 33 product names in these sentences (33 product names/88 noun phrases = 37.5%).

In this experiment, we say an activity extraction is correct when all attributes (object and action) of this activity are correctly extracted. We use measures of *precision*, *recall* and *F-measure* to compare the results of our proposed approach with that of previous works. The Extractor makes only a single pass over the entire experimental data set, and acquires the results as shown in Table 1.

**Table 1.** Experimental Results

Approach	Recall	Precision	F-measure
Nilanjan[27] (co-occurrence frequency $\geq 3$ )	1.14%	100%	2.25%
Minh[30] (our previous approach)	37.50%	57.89%	45.52%
Using deep linguistic parser (The method of making training data)	23.86%	80.76%	36.84%
Our proposed approach	<b>46.59%</b>	<b>73.21%</b>	<b>56.94%</b>

The experimental results have shown that our approach can automatically extract infrequent buying activities, and outperforms other approaches in terms of *recall* and *F-measure*. Our proposed approach focus on Japanese, but it could also be applied to other languages. However, our approach can only extract activities that are explicitly described in the sentences.

## 5 Related Work

There are two fields related to our research: human activity extraction and relation extraction (RE) from the Web corpus. Below, we discuss the previous researches of each field.

### 5.1 Human Activity Extraction

Previous works on this field are Perkowitz [6], Kawamura [11], Kurashima [24], Nilanjan[27], Fukazawa[28] and Minh [30].

Perkowitz's approach is a simple keyword matching, so it can only be applied for cases of recipe web pages (such as making tea or coffee). Kawamura's approach requires a product ontology and an action ontology for each domain. So, the precision of this approach depends on these ontologies.

Kurashima used JTAG [22] to deploy a deep linguistic parser to extract action and object. It can only handle a few types of sentences, and is not practical for the diversity and the size of the Web corpus. Additionally, because this approach gets date information from date of weblogs, so it is highly possible that extracted time might be not what activity sentences describe about.

Nilanjan[27] and Fukazawa[28] based on co-occurrence of action and object. This approach does not depend on sentences syntax. However, it can not extract infrequent activities, and have to prepare a list of action and object before extracting.

In our previous paper [30], the proposed approaches could not handle complex sentences, and necessary of preparing a list of Japanese syntax patterns.

## 5.2 Relation Extraction

The main researches of RE are DIPRE [15], SnowBall [16], KnowItAll [8], Pasca [7], TextRunner [14], O-CRF [9].

**Table 2.** Comparison with O-CRF

	O-CRF	Our Approach
Language	English	Japanese
Target data	Binary relation	Users' buying activity
Type of sentences can be handled	S-V-O	All types of Japanese sentences
Relation must occur between entities	yes	no
Requirement of determining entities before extracting	yes	no

DIPRE, SnowBall, KnowItAll, and Pasca use bootstrapping techniques applied for unary or binary RE. Bootstrapping techniques often require a small set of hand-tagged seed instances or a few hand-crafted extraction patterns for each domain. In addition, when creating a new instance or pattern, they could possibly extract unwanted patterns around the instance to be extracted, which would lead to extract unwanted instance from the unwanted patterns. Moreover, it is difficult to create suitable instances or patterns for extracting the attributes appeared in sentences retrieved from the Web.

TextRunner is the first Open RE system, it uses self-supervised learning and a Naive Bayes classifier to extract binary relation. Because this classifier predict the label of a single variable, it is difficult to apply TextRunner to extract all of the basic attributes.

O-CRF is the upgraded version of TextRunner. Because of the differences in tasks (activity, binary relation) and languages (Japanese, English), it is difficult to compare our approach with O-CRF. We try to compare them according to the some criteria as shown in Table 2.

## 6 Conclusions

This paper proposed a novel approach that automatically make its own training data. This approach treats activity extraction as a sequence labeling problem,

and uses conditional random fields as a learning model. Without requiring any hand-tagged data, it can extract buying activity in each sentence retrieved from twitter, by making only a single pass over its corpus. This approach outperforms other approaches in terms of recall and F-measure. The experimental results have shown that our approach can capture users' buying activities at Akihabara from twitter.

We are improving the architecture to handle more complex sentences, and to improve *precision*. We will try to extract relationships between activities to build a large human activity semantic network.

## References

1. Twitter, Inc., <http://twitter.com>
2. Twitter, Inc., <http://dev.twitter.com/doc/get/search>
3. Matsuo, Y., Okazaki, N., Izumi, K., Nakamura, Y., Nishimura, T., Hasida, K.: Inferring long-term user properties based on users' location history. In: Proc. IJCAI 2007, pp. 2159–2165 (2007)
4. Jung, Y., Lim, S., Kim, J., Kim, S.: Web Mining based OALF Model for Context-Aware Mobile Advertising System. In: The 4th IEEE/IFIP Int. Workshop on Broadband Convergence Networks (BcN 2009), pp. 13–18 (2009)
5. Zhao, Q., Mitra, P., Chen, B.: Temporal and Information Flow Based Event Detection from Social Text Streams. In: Proc. AAAI 2007, pp. 1501–1506 (2007)
6. Perkowitz, M., Philipose, M., Fishkin, K., Patterson, D.J.: Mining Models of Human Activities from the Web. In: Proc. WWW 2004 (2004)
7. Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge. In: Proc. AAAI 2006, pp. 1400–1405 (2006)
8. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D., Yates, A.: Methods for Domain-Independent Information Extraction from the Web: An Experimental Comparison. In: Proc. AAAI 2004 (2004)
9. Banko, M., Etzioni, O.: The Tradeoffs Between Open and Traditional Relation Extraction. In: Proc. ACL 2008 (2008)
10. Banko, M.: Open Information Extraction for the Web. PhD thesis, University of Washington (2009)
11. Kawamura, T., Tomohiro, Y., Nagano, S., Mizoguchi, Y., Iida, T.: A Proposal for Human Activity Mining from CGM. In: The 22nd Annual Conference of the Japanese Society for Artificial Intelligence (2008)
12. Poslad, S.: Ubiquitous Computing Smart Devices, Environments and Interactions. Wiley, Chichester (2009) ISBN: 978-0-470-03560-3
13. Ozok, A.A., Zaphiris, P.: Online Communities and Social Computing. In: Third International Conference, OCSC 2009, Held as Part of HCI International 2009, San Diego, CA, USA. Springer, Heidelberg (2009) ISBN-10: 3642027733
14. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the Web. In: Proc. IJCAI 2007, pp. 2670–2676 (2007)
15. Brin, S.: Extracting Patterns and Relations from the World Wide Web. In: WebDB Workshop at 6th International Conference on Extending Database Technology, EDBT 1998, Valencia, Spain, pp. 172–183 (1998)
16. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plain-text collections. In: Proc. ACM DL 2000 (2000)

17. Peppers, D., Rogers, M.: *The One to One Future*. Broadway Business (1996) ISBN-10: 0385485662
18. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields Probabilistic models for segmenting and labeling sequence data. In: Proc. ICML, pp. 282–289 (2001)
19. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proc. HLT-NAACL, pp. 213–220 (2003)
20. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons. In: Proc. CoNLL 2003 (2003)
21. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. IPSJ SIG Notes, 89–96 (2004)
22. Fuchi, T., Takagi, S.: Japanese morphological analyzer using word co-occurrence-JTAG. In: Proc. ACL 1998, pp. 409–413 (1998)
23. Kudo, T., Matsumoto, Y.: Japanese Dependency Analysis using Cascaded Chunking. In: Proc. z, pp.63–69 (2002)
24. Kurashima, T., Fujimura, K., Okuda, H.: Discovering Association Rules on Experiences from Large-Scale Weblogs Entries. In: ECIR 2009, pp.546–553 (2009)
25. Hiroyuki, Y., Hideyuki, T., Hiromitsu, S.: An individual behavioral pattern to provide ubiquitous service in intelligent space. WSEAS Transactions on Systems 562–569 (2007)
26. CoNLL: CoNLL 2000 shared task: Chunking (2000),  
<http://www.cnts.ua.ac.be/conll2000/chunking/>
27. Nilanjan, B., Dipanjan, C., Koustuv, D., Anupam, J., Sumit, M., Seema, N., Angshu, R., Sameer, M.: User Interests in Social Media Sites: An Exploration with Micro-blogs. In: Proc. CIKM (2009)
28. Fukazawa, Y., Ota, J.: Learning User's Real World Activity Model from the Web. IEICE SIG Notes, WI2-2009-66 (2009)
29. Kakaku.com, <http://apiblog.kakaku.com/KakakuItemSearchV1.0.html>
30. Nguyen, M., Kawamura, T., Nakagawa, H., Nakayama, K., Tahara, Y., Ohsuga, A.: Human Activity Mining using Conditional Random Fields and Self-Supervised Learning. In: Nguyen, N.T. (ed.) ACIIDS 2010. LNCS (LNAI), vol. 5990, pp. 140–149. Springer, Heidelberg (2010)

# Private Small-Cloud Computing in Connection with WinCE Thin Client

Bao Rong Chang<sup>1,\*</sup>, Hsiu Fen Tsai<sup>2</sup>, Chien-Feng Huang<sup>1</sup>, and His-Chung Huang<sup>1</sup>

<sup>1</sup> Department of Computer Science and Information Engineering

National University of Kaohsiung, Kaohsiung, Taiwan

Tel.: +886-7-5919797; Fax: +886-7-5919514

{brchang, cfhuang15}@nuk.edu.tw, hsc\_1015@yahoo.com.tw

<sup>2</sup> Department of International Business

Shu Te University, Kaohsiung, Taiwan

soenfen@stu.edu.tw

**Abstract.** In this study an open source Ubuntu Enterprise Server with the option Ubuntu Enterprise Cloud is utilized to build a “Private Small-Cloud Computing” (PSCC). In PSCC, a small-cloud is a cloud controller (CLC) via a wired or wireless network connected to several mobile or stationary thin client devices, and afterward a cloud controller (CLC) connected to several small cluster controller (CC), where the cloud can activate SaaS, PaaS, and IaaS services. Instead of Linux-based architecture, we also explored how to use the limited resources of WinCE-based embedded devices to retain the "cloud computing". In addition to the wired Ethernet connection, alternative connection to wireless mobile devices IEEE802.3b/g or 3G is also done here. Here Mysaifu JVM virtual machine is employed to achieve J2ME environment and GNU Classpath is used as the Java Class Libraries. Finally, the experiment on rapid face and fingerprint identifications using intelligent access control security system within 0.5 sec validates the excellent performance of the proposed PSCC.

**Keywords:** Private small-cloud computing; Ubuntu Enterprise Server; WinCE-based embedded devices; Mysaifu JVM virtual machine; intelligent access control security system.

## 1 Introduction

In recent years, both computer software and hardware are booming rapidly, resulting in the blind pursuit of rapid, large-capacity, next-generation computer. However, how to phase out the old computer leads to a big problem. Computer manufacturing as we know has consumed a lot of energy, and the treatment of final recovery must also consume more energy to deal with. Moreover, energy consumption for the new generation of computers are directly proportional to the growth of computer scale where

---

\* Corresponding author.

greater power must be used to stabilize its operation. For example: a new generation of high-end 3D accelerator card, you need to use more than 400-watt power supply, or even higher. "Cloud computing" currently under development is just to solve these problems, because it consists of three areas, namely cloud computing, connectivity, and client equipment, its vision is towards low-cost (saving you money), green energy (energy efficiency), and the ubiquitous (at any time, any place, any device to access any of the services) as the goal. Therefore there is urgent need for the use of cloud computing for low-power PC cluster architecture and low power client devices so that the ratio of energy consumption can significantly be reduced a lot to reduce carbon emissions to be consistent with the trend.

"Cloud computing" is a popular concept in recent years. In fact, these technologies are not entirely new, probably inherited from the nature of "distributed computing" and "grid computing". That is, we divide a large work into small pieces because it is of the incompetence in a single computer, and then these pieces are carried out by a number of computers. After that, compiling their findings to complete the work is done. In addition, we have devoted to connect a variety of different platforms, different architectures, different levels of the computer through the network such that all of computers are cooperated with each other or network makes the computer to do services more far and wide in the cyber space, but the difference is that "cloud computing" has emphasized, even existing the limited resources in a local context, to make use of the Internet to access remote computing resources.

"Cloud computing" is divided into two categories, namely "cloud services" and "cloud technology". "Cloud services" is achieved through the network connection to the remote service. Such services provide users installation and use a variety of operating systems, for example Amazon Web Services (CE2 and S3) services. This type of cloud computing can be viewed as the concepts: "Infrastructure as a Service" (IaaS) "Storage as a service" (StaaS), respectively. Both of them are derived from the concept of "Software as a Service" (SaaS) that is the biggest area for cloud services in demand, while "Platform as a Service" (PaaS) concept is an alternative for cloud computing service. Using these services, users can even simply to rely on a cell phone or thin client to do many of things that can only be done on a personal computer in the past, which means that cloud computing is universal. The "cloud technology" is aimed at the use of virtualization and automation technologies to create and spread computer in a variety of computing resources. This type can be considered as traditional data centers (Data Center) extension; it does not require external resources provided by third parties and can be utilized throughout the company's internal systems, indicating that cloud computing also has the specific expertise. Currently on the market the most popular cloud computing services are divided into public clouds, private clouds, community/open clouds, and hybrid clouds, where Goggle App Eng [1], Amazon Web Services [2], Microsoft Azure [3] - the public cloud; IBM Blue Cloud [4] - the private cloud; Open Nebula [5], Eucalyptus [6], Yahoo Hadoop [7] and the NCDM Sector / Sphere [8] - open cloud; IBM Blue Cloud [4] - hybrid cloud.

## 2 Motivation and Objective

The main purpose of this study is to build a “Private Small-Cloud Computing” (PSCC). The idea of private small-cloud computing is based on three concepts: small clusters, virtualization, and general graphics processor. First, instead of the distributed parallel computing and grid computing, cluster computing has been back to the mainstream, especially significant use in cloud computing. This is because the grid computing has to bear the additional high-cost of data delivery/transportation, and cloud computing has no such burden. Cluster computing with fault-tolerance, robustness, rapid recovery, as well as high-performance computing has become a new industry. In this study, a small-cloud is a cloud controller (CLC) via a wired or wireless network connected to several mobile or stationary client devices, and afterward a cloud controller (CLC) connected to several small cluster controller (CC). This configuration forms a private small-cloud computing. Next, a single machine can not fit for all tasks at the same time, therefore the virtual machine server is a new service designed to run multiple operating systems simultaneously on a single physical x86 computer, and supports multiple and heterogeneous applications running on a variety of physically separate client devices. This definitely has been done to reduce the power consumption, shorter the time and save the cost, and moreover increase availability and improve execution performance. Finally, rather than a general processor (CPU), the use of general-purpose graphics processing unit (GPGPU) [9] can be realized for stream processing. It is of the paradigm of program compile and related to the SIMD computer programming for which general graphics processor allows certain applications easier use of a limited form of parallel processing. This application can use multiple computing units (multi-threaded), for example multiple GPU's floating point units, without the need for a clear allocation of management, synchronization, or communication between these units. In addition, regarding the issue of data storage on a private small-cloud computing, we are designed to store large amounts of data to the controller in the local cluster or cloud controller like Warlus, storage controller, in conjunction to cloud controller. Even we mirror a copy to the real public cloud as a redundant backup, and it can be used to restore a private small-cloud computing when the virtual machine server crash. According to the above-mentioned reasons, based on the thin clients, e.g., several embedded platforms or a number of low-cost computers, if the use of cloud computing controller to connect several small clusters can be implemented, the cloud system can activate SaaS, PaaS, and IaaS services. This cloud system will include the use of Xen [10] virtualization technology, VMGL [11] planning to use general-purpose graphics processors, Open Nebula [5] Management of cluster structure, Eucalyptus [6] implementation of the cloud controller. Despite the recent NCHC Free Software Lab made a similar proposal, based on the deployment of the client desktop PC in the fixed structure in the cluster, but we are committed to deploying wireless mobile devices on the client, rather than fixed equipment. We have shown here to extend the embedded platform client for the cloud computing and other wireless mobile devices being used for cloud computing. From the green energy point of view, therefore, private small-cloud computing also means that the following four characteristics: large amounts of data, low cost, efficiency and reliability.

In many applications, embedded devices often require huge computing power and storage space, just the opposite of the hardware of embedded devices. Thus the only way to achieve this goal is that it must be structured in the "cloud computing" and operated in "cloud services". The program is how to use the limited resources of embedded devices to achieve the "cloud computing", in addition to using the wired Ethernet connection, and further use of wireless mobile devices IEEE802.3b / g or 3G to connect. First, we use the standard J2ME [12] environment for embedded devices, where Mysaifu JVM [13] virtual machine is employed to achieve J2ME environment and GNU Classpath [14] is used as the Java Class Libraries. In order to reduce the amount of data transmission, the acquisition of information processed is done slightly at the front-end embedded devices and then processed data through the network is uploaded to the back-end, private small-cloud computing. After the processing at the back-end is completed, the results sent back to the front-end embedded devices. An open source Ubuntu Enterprise Server [15] with the option Ubuntu Enterprise Cloud is utilized to build the private small-cloud computing. The current Ubuntu Enterprise Cloud has included Xen, VMGL, Open Nebula, and Eucalyptus and other packages, so through Ubuntu Enterprise Cloud we can easily focus on installing the back-end cluster controllers and cloud controller in order to build a private small-cloud computing. Upon completion of the formal operation of a private small-cloud computing, an ISO copy ghosts to a public cloud, e.g., Google App Eng or Yahoo Hadoop, as redundant (backup) purposes. When the virtual machine server at back-end is off, a redundant can be used to restore the private small-cloud computing immediately.

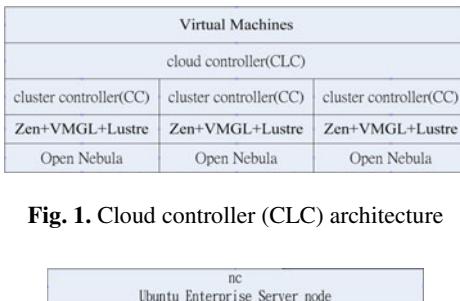
Here we have proposed a one-year program. An embedded platform in a cloud environment is applied to testing the capabilities of fingerprint identification and facial recognition as an intelligent access control security system. The basic structure of PSCC is developed and has been deployed as well. We will then test the performance of the embedded platform operating in cloud computing to check whether or not it can achieve immediate and effective response to required functions. Meanwhile, we continue to monitor the online operation and evaluate system performance in statistics, such as the number of files, file size, the total process of MB, the number of tasks on each node, and throughput. In a cluster implementation of cloud computing, the statistical assessment by the size of each node is listed. According to the analysis of the results, we will adjust the system functions if changes are needed. Once the test to PSCC platform has completed, the platform will be provided to other programs for exploitation, and will open it as one of Open Source in Open Source Software Foundry [16].

### 3 Methods

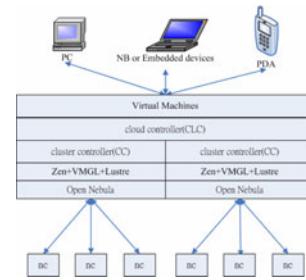
#### 3.1 Deployment of Private Small-C Loud Computing

Deployment of private small-cloud computing (PSCC) that wants some software packages to build the cloud structure generally requires Xen, OpenNebula, Eucalyptus, Euca2ools [17] and so on. Since system installation needs many steps, manipulation often encounter some errors and the configuration is not easy, this study employs an open source, Ubuntu Enterprise Server Edition, because this version of the Ubuntu

has included all of the above packages that are used to deploy cloud structure rapidly and easily. So it has the following advantages: easy, fast, and accurate installation as well as easy to use, good maintenance, easy update, and widely available for free download. I believe it is the mainstream of cloud controller in private small-cloud computing in the future. For private small-cloud computing, installation of cloud/cluster controller using Ubuntu Enterprise Server version can visit the download site [18]. Cloud Controller (CLC) structure as shown in Fig. 1 in which each Cluster Controller (CC) has its own OpenNebula, Zen, VMGL, Lustre [19] and hardware resources, through an unified CLC to manage all of CC. Node Controller (nc) structure as shown in Fig. 2. nc hardware resources will determine the cloud service capabilities; the more powerful nc hardware (more CPU core and more memory), the more virtual machine resources. Furthermore, according to the above Fig.1 and Fig. 2 provides a way to establish CLC and nc, we can completely describe the overall structure of cloud, as shown in Fig. 3. It gives a complete integration CLC+nc that shows the structure of private small-cloud computing.



**Fig. 1.** Cloud controller (CLC) architecture



**Fig. 2.** Node controller (nc) structure

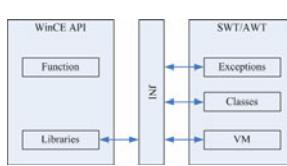
**Fig. 3.** A complete structure of CLC + nc private small-cloud computing

### 3.2 Establishment of Terminal Nodes

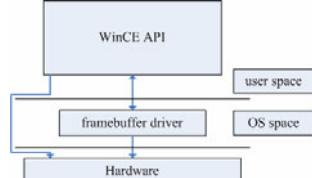
In terms of node device, Mysaifu JVM [13] is treated as the framework of programming development; however the virtual machine Mysaifu JVM has no way to perform the drawing even through their core directly, and thus it must call other graphics library to achieve the drawing performance. The problem we encountered is that Linux-based OS in conjunction with the middleware needs a few packages to work together required many steps for installation, compiling different packages to build system is also difficult, and it is often time-consuming for the integration of a few packages no guarantee to complete the work. Therefore this study has chosen WinCE framework, as shown in Fig. 4, instead of Linux-based architecture, in such a way that achieves GUI interface functions. In Fig. 5, no matter SWT or AWT in Mysaifu JVM they apply Java Native Interface (JNI) to communicate C-written graphics library. Afterward WinCE API through the kernel driver to achieve the drawing performance as shown in Fig. 6. According to the pictures shown in Fig. 5 and Fig. 6, we can string them together to be the structure of a node device as shown in Fig. 4. This part will adopt a low-cost, low-power embedded platform to realize a node device.



**Fig. 4.** Terminal node with WinCE framework



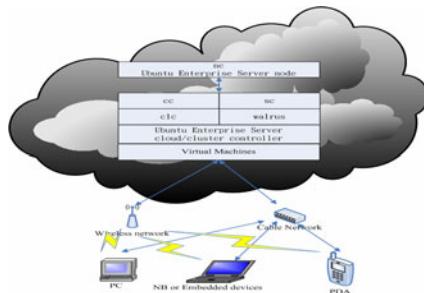
**Fig. 5.** Communication between SWT / AWT and WinCE API



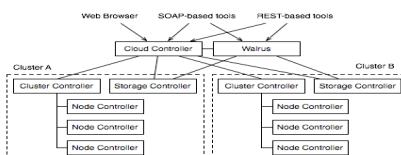
**Fig. 6.** WinCE API communicates with the Framebuffer

### 3.3 Clouds and the Overall Structure of the Node Device

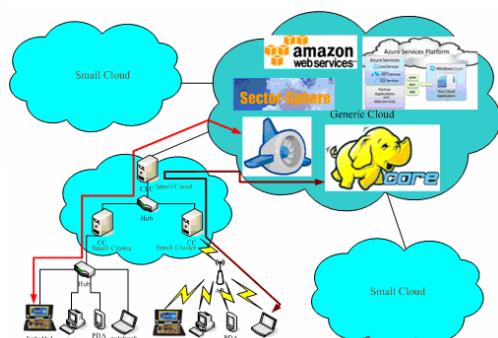
As a private small-cloud together with terminal node in terms of a overall structure is shown in Fig. 7, users (personal computers, laptops, embedded platforms, and smart phones) go over the wireless network IEEE802.3b/g or wired Ethernet network to connect to the private small-cloud computing, PSCC. The storage server in conjunction to Eucalyptus is Walrus [20] that is a compatible storage interface like Amazon S3 storage system and can be managed through the web interface to modify it. In addition, a control unit managing the storage server is called the storage controller (sc) [21] as shown in Fig. 8. As shown in Fig. 9, PSCC can still link to the remote cloud through the Internet such that node device gets remote cloud services via private small-cloud, such as Goggle App Eng, Amazon Web Services, Yahoo Hadoop, Microsoft Azure, IBM Blue Cloud, and NCDM Sector / Sphere. After the formal operation, PSCC may ghost an ISO copy to Google App Eng or Yahoo Hadoop as redundant (backup) purposes. When the virtual machine server crash, the redundant can be used to restore PSCC immediately. Once the test for PSCC has completed, the use of PSCC will be provided to other programs, and PSCC will open itself as an open source and be uploaded to Open Source Software Foundry [16].



**Fig. 7.** PSCC and terminal node devices



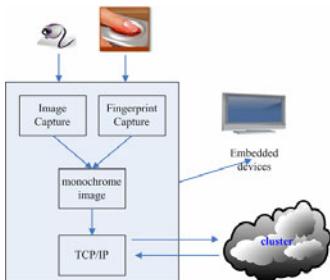
**Fig. 8.** Advanced Eucalyptus setup along with Walrus



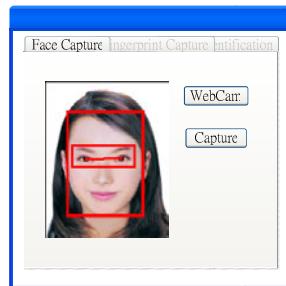
**Fig. 9.** Various cloud services in PSCC and remote cloud

### 3.4 Intelligent Access Control Security System and Cloud Testing

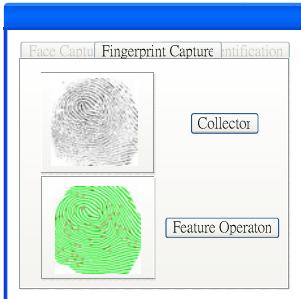
When establishing a private small-cloud computing is done, we will test the cloud employing an embedded platform in a cloud environment to perform fingerprint identification [22] and face recognition [23] to fulfill the intelligent access control security systems [24]. Meanwhile the development of basic structure and deployment for PSCC are valid and even more we test the service performance for an embedded platform collaborated with cloud computing, checking an immediate and effective response to client. The intelligent access control security system is shown in Fig. 10. Steps are as follows: (a) the operation for face recognition is quickly to open the video camera for the first, and then press the capture button, the program will execute binarization automatically as shown in Fig. 11; (b) the rapid fingerprint identification is first to turn on device, then press the deal button for feature extraction that reduces the amount of information as shown in Fig. 12; (c) at first the terminal device test the connection if Internet is working properly, and then we press the identify button and information sent to the cloud, and at last the cloud will return the identification results as shown in Fig. 13.



**Fig. 10.** System architecture



**Fig. 11.** Binarization processing automatically running in program



**Fig. 12.** Processing fingerprint features to reduce the amount of information



**Fig. 13.** Information sent to the cloud and cloud returns the results of recognition

## 4 Experimental Results

In order to deploy a minimum of cloud structure, we need at least two dedicated systems. One will be used as a cloud controller (clc), and contains the entire back-end

cluster controller (cc), storage server Walrus, and the storage controller (sc). This host needs fast disks and a few fast processors to match those disks. Another one is a node controller (nc), used to perform many of the cloud entity. This host takes a lot of capacity with Virtualization Technologies (VT) [25] of the CPU, a large number of CPU computing power, large memory and fast disk. Building the system in the following steps:

#### **1. Installing virtual machine**

In this study we adopt the VMware-Workstation 7 to install virtual machines because VMware currently can only fully support the latest version of Ubuntu Linux and is a paid software.

#### **2. Deploying cloud computing architecture**

To deploy cloud structure will generally need the software with Xen, OpenNebula, Eucalyptus, and Euca2ools. Since system installation needs many steps, manipulation often encounter some errors and the configuration is not easy, this study employs an open source, Ubuntu Enterprise Server Edition, because this version of the Ubuntu has included all of the above packages that are used to deploy cloud structure rapidly and easily. Ubuntu Enterprise Server Edition used to install the cloud/ cluster controller can be visited at the following webpage. ISO file booting system with the English language choice is recommended in order to avoid unusual characters input because command line style is only one input mode for the node controller. There is an illustration for selecting the installation type Cluster.

#### **3. Installing node controller**

Before installing node controller, ISO file booting system with the English language choice is recommended in order to avoid unusual characters input because command line style is only one input mode for the node controller. There is an illustration for selecting the installation type Node.

#### **4. Setting cloud controller**

Back to the cloud controller, and executing commands to find the node controller and examining a link to node controller you created earlier.

#### **5. Setting cloud user through web interface**

Before the user at client side uses the clouds, the client are required to do some of the settings in cloud controller through web interface. Login Management Interface: The default account is admin and password admin. Web managers can do some settings. Applying an account for a new user is found.

#### **6. Setting embedded platform**

Instead of Linux-based architecture for a thin client, we have selected WinCE environment running at the embedded platform as shown in Figs 14, 15, and 16. As the WinCE built-in Java virtual machines are not included, we have to download the virtual machine for WinCE embedded platforms. So this study adopts Mysaifu JVM as shown in Figs 17 and 18, and the related links can be found in [13].

In order to verify the cloud system effectiveness and efficiency, the experiment on fingerprint identification and face recognition by using rapid identification of intelligent access control security system has been done successfully within 0.5 seconds in average to exactly cross-examine the subject identity. As a result the proposed PSCC has been performed very well when it has deployed in local area.

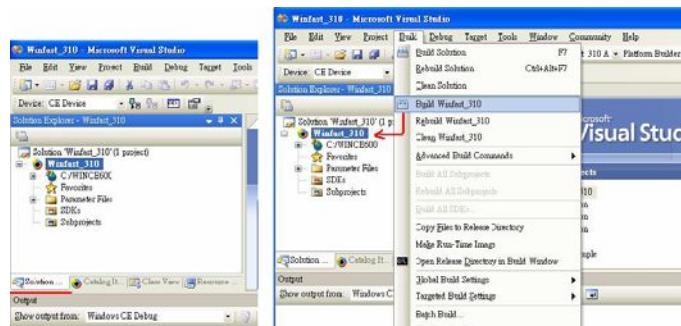


Fig. 14. Buliding WinCE image file NK.bin

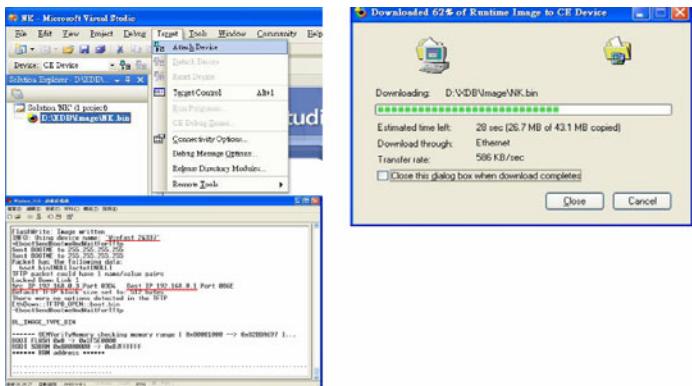


Fig. 15. Porting WinCE into embedded platform



Fig. 16. WinCE in Chinese platform



Fig. 17. Porting Mysaifu

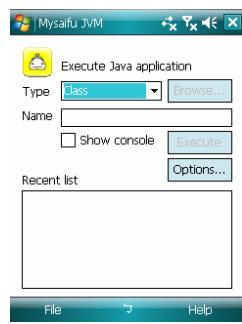


Fig. 18. Virtual machine running in WinCE and installed

## 5 Conclusions

This study is to build a “Private Small-Cloud Computing” (PSCC) in which the cloud can activate SaaS, PaaS, and IaaS services. Instead of Linux-based architecture for

thin client, we also explored how to use the limited resources of WinCE-based embedded devices to achieve the "cloud computing". In addition the wired Ethernet connection, alternative wireless mobile devices IEEE802.3b/g or 3G is also done here. Finally, the experiment on fast fingerprint identification and face recognition by using intelligent access control security system validates the excellent performance of the proposed PSCC.

**Acknowledgments.** This work is supported by the National Science Council, Taiwan, Republic of China, under grant number NSC 99-2218-E-390-002.

## References

1. Google App Engine (2010),  
<http://groups.google.com/group/google-appengine>
2. Amazon Web Services, AWS (2010), <http://aws.amazon.com/>
3. Windows Azure- A Microsoft Solution to Cloud (2010),  
<http://tech.cipper.com/index.php/archives/332>
4. IBM Cloud Compputing (2010), <http://www.ibm.com/ibm/cloud/>
5. OpenNebula (2010), <http://www.opennebula.org/>
6. Eucalyptus (2010), <http://open.eucalyptus.com/>
7. Welcome to Apache Hadoop (2010), <http://hadoop.apache.org/>
8. Sector/Sphere, National Center for Data Mining (2009),  
<http://sector.sourceforge.net/>
9. General-Purpose Computation on Graphics Processing Units (2010),  
<http://gpgpu.org/>
10. Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I., Warfield, A.: Xen and the Art of Virtualization, Technical report (2003),  
<http://www.cl.cam.ac.uk/research/srg/netos/papers/2003-xensosp.pdf>
11. VMGL: VMM-Independent Graphic Acceleration (2007),  
<http://www.cs.toronto.edu/~andreslc/publications/slides/Xen-Summit-2007/vmgl.pdf>
12. Java 2 Platform, Micro Edition (J2ME) (2010),  
[http://www.java.com/zh\\_TW/download/faq/whatis\\_j2me.xml](http://www.java.com/zh_TW/download/faq/whatis_j2me.xml)
13. Mysaifu JVM - A free Java Virtual Machine for Windows Mobile (2010),  
[http://www2s.biglobe.ne.jp/~dat/java/project/jvm/index\\_en.html](http://www2s.biglobe.ne.jp/~dat/java/project/jvm/index_en.html)
14. GNU Classpath, GNU Classpath, Essential Libraries for Java (2010),  
<http://www.gnu.org/software/classpath/>
15. Ubuntu Enterprise Server (2010), [http://docs.sun.com/app/docs/doc/821-1045/ggfrh?l=zh\\_TW&a=view](http://docs.sun.com/app/docs/doc/821-1045/ggfrh?l=zh_TW&a=view)
16. Open Source Software Foundry (2010), <http://www.openfoundry.org/>
17. Euca2ools User Guide (2010),  
[http://open.eucalyptus.com/wiki/Euca2oolsGuide\\_v1.1](http://open.eucalyptus.com/wiki/Euca2oolsGuide_v1.1)
18. ubuntu-9.10-server-i386.iso, Ubuntu 9.10, Karmic Koala (2010),  
<http://releases.ubuntu.com/karmic/ubuntu-9.10-server-i386.iso>
19. Lustre a Network Clustering FS (2009),  
[http://wiki.lustre.org/index.php/Main\\_Page](http://wiki.lustre.org/index.php/Main_Page)
20. Walrus/Eucalyptus (2010),  
[http://open.eucalyptus.com/wiki/EucalyptusStorage\\_v1.4](http://open.eucalyptus.com/wiki/EucalyptusStorage_v1.4)

21. SC/Walrus/Eucalyptus (2010),  
[http://open.eucalyptus.com/wiki/EucalyptusAdvanced\\_v1.6](http://open.eucalyptus.com/wiki/EucalyptusAdvanced_v1.6)
22. VeriFinger SDK, Neuro Technology (2010),  
<http://www.neurotechnology.com/verifinger.html>
23. VeriLook SDK, Neuro Technology (2010),  
<http://www.neurotechnology.com/verilook.html>
24. opencv (open source) (2010),  
<http://www.opencv.org.cn/index.php?title=%E9%A6%96%E9%A1%B5&variant=zh-tw>
25. Virtualization technologies from Intel,  
<http://www.intel.com/technology/virtualization/>

# Honey Bee Mating Optimization Algorithm for Approximation of Digital Curves with Line Segments and Circular Arcs

Shu-Chien Huang

Department of Computer Science, National Pingtung University of Education,  
Pingtung City, Taiwan  
schuang@mail.npue.edu.tw

**Abstract.** Approximation of a digital curve with line segments and analytic curve-pieces is an important technique in image analysis and pattern recognition. A new approximation approach is presented based on honey bee mating optimization. In this method, given the number of breakpoint,  $m$ , find an approximation with  $m$  breakpoints in such a manner that when line segments and circular arcs are appropriately fitted between all pairs of adjacent breakpoints, the approximation error is minimized. Experiments have shown promising results and fast convergence of the proposed method.

**Keywords:** Curve approximation, Integral square error, Honey bee mating optimization.

## 1 Introduction

Curve approximation is an important topic in the area of image processing, pattern recognition, and computer vision. A good representation of an object curve can significantly reduce the amount of data needed to be processed, while at the same time preserving important information about the curve. To fit a digital curve, we first divide it into segments. Then, each digital segment is fitted with a piece of analytic curve, which can be a line segment, a circular arc, or a high order curve.

Polygonal approximation is the simplest approach, as it fits each digital segment with a straight line segment. In recent decades, a number of methods [1-6] have been proposed for polygonal approximation. In general, there are two basic approaches associated with the polygonal approximation of  $n$  ordered points with a uniform error norm: (1) given the number of breakpoints,  $m$ , find a polygonal approximation with  $m$  breakpoints, such that its distance from the curve is minimal among all the approximations with  $m$  breakpoints, and (2) given an error  $\varepsilon$ , find the polygonal approximation with the minimal number of breakpoints such that the polygon is distant from the digital curve by no more than  $\varepsilon$ .

Unfortunately, not all kinds of digital curves are suitable for applying the polygonal approximation. It is hard to fit a smooth curve with a polygon. Using

circular arcs for segment representation produces a better representation at a higher level of computational complexity [7]. The dynamic programming approach proposed by Horng and Li [8] fits a digital curve with line segments and circular arcs based on two perceptual rules. Although their method provides a perceptually better representation of a digital curve, the dynamic programming process is time-consuming. Sarkar et al. [9] proposed a method for approximation of digital planar curves with line segments and circular arcs using genetic algorithms. The desired set of optimal breakpoints is sought such that when line segments and circular arcs are appropriately fitted between all pairs of adjacent breakpoints, the fitting error is minimized.

The remainder of this paper is organized as follows. Section 2 introduces the honey bees mating optimization (HBMO). Section 3 presents the proposed method. The experimental results are presented in Section 4. Finally, conclusions are given in Section 5.

## 2 Honey Bee Mating Optimization

Over the last decades, modeling the behavior of social insects, such as ants and bees, for the purpose of searching and problem solving has been the context of the emerging area of swarm intelligence. The ant colony algorithm [10] and particle swarm optimization [11] are the two most popular approaches in swarm intelligence. The honey bee mating optimization [12-16] may also be considered as a typical swarm-based approach to optimization, in which the search algorithm is inspired by the process of mating in real honey bees.

A honey bee colony typically consists of a single egg-laying long-lived queen, anywhere from zero to several thousand drones, and usually 10,000–60,000 workers. Queens are specialized in egg-laying. A colony may contain one queen or more during its life cycle, and is named a monogynous or polygynous colony, respectively. The drones are practically considered as agents that pass one of their mother's gametes, and function to enable females to act genetically as males. Worker bees specialize in brood care and sometimes lay eggs.

In the marriage process, the queens mate during their mating flights far from the nest. A mating flight starts with a dance performed by the queen who then starts the flight during which the drones follow the queen and mate with her in the air. In each mating, sperm reaches the spermatheca and accumulates to form the genetic pool of the colony. Each time the queen successfully mates with a drone, the genotype of the drone is stored and a variable is increased by one until the required size of the spermatheca is reached.

In order to develop the algorithm, the capability of workers is limited to brood care and thus each worker may be regarded as a heuristic that acts to improve and/or take care of a set of broods. An annealing function is used to describe the probability of a drone ( $D$ ) successfully mating with the queen ( $Q$ ) as shown in Eq. (1).

$$\text{Prob}(Q, D) = \exp \left[ -\frac{\Delta(f)}{S(t)} \right], \quad (1)$$

where  $\Delta(f)$  is the absolute difference in the fitness of  $D$  and the fitness of  $Q$ , and the  $S(t)$  is the speed of the queen at time  $t$ . After each transition of mating, the queen's speed and energy reduce according to the following equation:

$$S(t+1) = \alpha \times S(t), \quad (2)$$

where  $\alpha$  is the decreasing factor. Workers adopt some heuristic mechanisms such as crossover or mutation to improve the brood's genotype. The fitness of the resulting genotype is determined by evaluating the value of the objective function of the brood genotype. The popular five construction stages of the HBMO algorithm were proposed by Fathian et al. [12] and are also used to develop the approximation of line segments and circular arcs in this paper. The five stages are described as follows:

- (1) The algorithm starts with the mating flight, where a queen (best solution) selects drones probabilistically to form the spermatheca (list of drones). A drone is then randomly selected from the list for the creation of broods.
- (2) Creation of new broods by crossover of the drone's genotypes with those of the queen.
- (3) Use of workers to conduct local searches for broods (trial solutions).
- (4) Adaptation of workers' fitness, based on the amount of improvement achieved in the broods.
- (5) Replacement of a weaker queen by fitter broods.

A pseudocode of the proposed Honey Bees Mating Optimization Algorithm for this problem is presented in the following.

```

Generate the initial population randomly
Selection of the best bee as the queen
Select maximum number of mating flights (MF)
for i=1 to MF
    Initialize queen's spermatheca and speed.
    Select α
    Do while speed > 0.01 and spermatheca is not full
        Select a drone
        If the drone passes the probabilistic condition
            Add sperm of the drone in the spermatheca
        end if
         $S(t+1) = \alpha \times S(t)$ 
    enddo
    for j=1 to size of Spermatheca

```

```

Select a sperm from the spermatheca
Generate a brood by using a crossover operator between the
queen's genotype and the selected sperm
Select, randomly, a worker
Use the selected worker to improve the brood's fitness
if the brood's fitness is better than the queen's fitness
Replace the queen with the brood
end if
end for
end for
return the Queen

```

### 3 The Proposed Approach

#### 3.1 Problem Statement

Given a digital closed curve consisting of  $n$  integer-coordinate points enumerated in clockwise order along the curve  $C=\{p_i=(x_i,y_i), i=1,\dots,n\}$ , the digital closed curve can be divided into many segments. Let  $\overline{p_i p_j}$  represent the segment starting at point  $p_i$  and continuing to point  $p_j$  in a clockwise direction. In the present method, each segment is fitted with a line segment or a circular arc. The error between a point  $p_k$  of a digital curve  $C$  and its approximation is denoted by  $e_k$ . If the segment  $\overline{p_i p_j}$  is fitted with a line segment  $\overline{p_i p_j}$ , the error of segment  $\overline{p_i p_j}$ , denoted by  $E(\overline{p_i p_j})$ , is defined as

$$E(\overline{p_i p_j}) = \sum_{p_k \in \overline{p_i p_j}} \{(e_k)^2\}, \quad (3)$$

and 
$$e_k = d(p_k, \overline{p_i p_j}), \quad (4)$$

where  $d(p_k, \overline{p_i p_j})$  is the perpendicular distance from point  $p_k$  to the corresponding line segment. On the other hand, if the segment  $\overline{p_i p_j}$  is fitted with a

circular arc, denoted by  $\overbrace{p_i p_j}$ , of radius R with its center at ( $Cx, Cy$ ), the error of segment  $\overline{\overline{p_i p_j}}$  is defined as

$$E(\overline{\overline{p_i p_j}}) = \sum_{p_k \in p_i p_j} \{ (e_k)^2 \}, \quad (5)$$

and  $e_k = |R - \sqrt{((x_k - Cx)^2 + (y_k - Cy)^2)}|.$  (6)

The circular arc  $\overbrace{p_i p_j}$  has the two breakpoints  $p_i$  and  $p_j$  as its endpoints, and its radius  $R$  can be expressed in terms of the circular arc center ( $Cx, Cy$ ) as

$$R = \sqrt{(x_i - Cx)^2 + (y_i - Cy)^2}. \quad (7)$$

However, a good estimation of the circular arc center proposed by Pei and Horng [7] is available

$$(Cx, Cy) = \left( -\frac{\sum_{l=i}^j K_1 K_2}{\sum_{l=i}^j K_1 K_3}, a * Cx + b \right) \quad (8)$$

where

$$a = -\frac{x_j - x_i}{y_j - y_i}, \quad b = \frac{y_i + y_j}{2} - a \frac{x_i + x_j}{2},$$

$$K_1 = -x_i - ay_i + x_l + ay_l,$$

$$K_2 = x_i^2 + (y_i - b)^2 - x_l^2 - (y_l - b)^2,$$

$$K_3 = -2x_i - 2a(y_i - b) + 2x_l + 2a(y_l - b).$$

If the denominator of Eq. (8) is larger than 0.0001, then the segment is fitted with a circular arc; otherwise the segment is fitted with a line segment in the proposed method. The problem can be stated as follows: Given the number of breakpoints  $m$ , the goal is to

find an approximation with line segments and circular arcs such that the total approximation error is minimized.

### 3.2 Fitting Digital Curves Using the Honey Bee Mating Optimization

Stage 1. Generate the initial drone sets and queen

Generate  $Z$  drones with gene length  $m$ , denoted by the matrix  $D$

$$D=[D_1, D_2, \dots, D_Z] \quad (9)$$

$$D_i = (d_i^1, d_i^2, \dots, d_i^m)$$

where  $d_i^j$  is the  $j$ th gene value of the  $i$ th drone and  $1 \leq d_i^j \leq d_i^{j+1} \leq n$  for all  $j$ , i.e.,  $d_i^j$  indicates that the point  $p_j$  is a breakpoint of the approximation. The value of approximation error of queen, drone or brood is regarded as the fitness value. Among all drones, the one with the minimum approximation error is selected as the queen  $Q$ .

Stage 2. Flight mating

In Stage 2, the simulated annealing method is used to select the best drone set during the flight mating of queen  $Q$ . After the flight mating, the queen's speed will be reduced by Eq. (2). The flight mating continues until the number of sperm in the queen's spermatheca is more than the threshold  $nsperm$ . The value of  $nsperm$  is generally predefined by the user and less than the number of drones. The selected sperms,  $Sperm$ , are described by Eq. (10).

$$Sperm=[SP_1, SP_2, \dots, SP_{nsperm}], \quad (10)$$

$$SP_i=(sp_i^1, sp_i^2, \dots, sp_i^m),$$

where  $SP_i$  is the  $i$ th individual in the queen's spermatheca.

Stage 3. Breeding process

In the breeding process, the  $j$ th individual of the queen's spermatheca is selected to breed if its corresponding random number  $R_j$  is less than a user-defined breeding ratio  $P_c$ . Consider  $SP_j=(sp_j^1, sp_j^2, \dots, sp_j^m)$  and  $Q=(Q^1, Q^2, \dots, Q^m)$ , the brood is generated by selecting one from  $(Q^1, Q^2, \dots, Q^{w-1}, sp_j^w, \dots, sp_j^m)$  and  $(sp_j^1, \dots, sp_j^{w-1}, Q^w, \dots, Q^m)$ , where  $w$  ( $2 \leq w \leq m$ ) is chosen randomly from the set  $CS$ . The set  $CS$  is determined by the following procedure:

$$CS=\emptyset$$

for  $(i=2; i \leq m; i++)$

{ if ( $Q^{i-1} < sp_j^i$ ) and ( $sp_j^{i-1} < Q^i$ )

{  $CS=CS \cup \{i\}$  } }

Stage 4. Brood mutation with the royal jelly by workers

The population of broods is improved by applying the mutation operator as follows:

Step 1. For all broods, the random number  $R_i$  of the  $i$ th brood is generated.

Step 2. If the  $R_i$  is less than the predefined mutation ratio  $P_m$ , the  $i$ th brood needs mutation. The first step is to select  $k$  randomly, where  $1 \leq k \leq m$ . The second step is that  $Brood_i^k$  is changed to a new value, which is between  $Brood_i^{k-1}$  and  $Brood_i^{k+1}$ , such that the corresponding total approximation error is the minimum.

Step 3. The best brood,  $brood_{best}$  with minimum fitness value is selected as the candidate solution.

Step 4. If the fitness of  $brood_{best}$  is superior to the queen, we replace queen by  $brood_{best}$ .

Stage 5. Check the termination criterion

If the termination criterion is satisfied, then finish the algorithm, else discard all previous trial solutions (brood set). Then go to Stage 2 until the assigned iteration is completed.

## 4 Experimental Results

The proposed algorithm, coded in C, was run on a Pentium-4 PC under the Windows XP operating system. In order to assess the performance of this method, we used it to approximate three digital curves namely, a curve with four semicircles, Fig. 1(a); a chromosome-shaped curve, Fig. 1(b); and a figure-of-eight curve, Fig. 1(c). Quantitative comparisons of the proposed method with two other methods, including the number of breakpoints and the integral square error (ISE) is shown in Table 1. Table 2 shows the parameters of the proposed HBMO-based algorithm. The computation time is listed in Table 3. It is observed that the results can be obtained in a reasonable time.

**Table 1.** Results of approximation by different methods; the results quoted for the proposed method are the best obtained in five independent runs

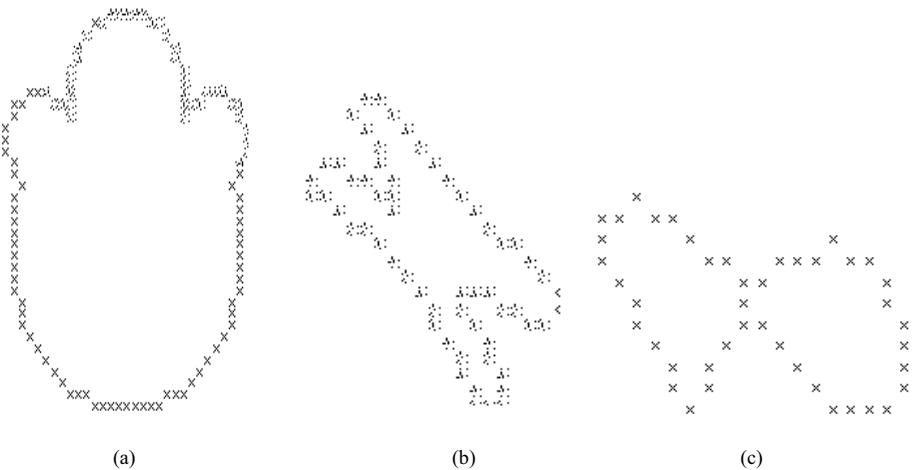
Curve	Yin's method [4]		Sarkar et al.'s method [9]		The Proposed Method	
	# Breakpoints	ISE	# Breakpoints	ISE	# Breakpoints	ISE
Curve with four semicircles(n=102)	12	42.85	12	4.31	12	4.09
			4	6.94	4	6.94
Chromosome-shaped curve(n=60)	15	5.22	15	1.23	15	1.08
			11	2.18	11	1.94
Figure-of-eight curve(n=45)	9	6.84	9	2.03	9	1.87
			8	2.36	8	2.29

**Table 2.** The parameters used in the proposed method

Parameter	Explain	Value
	Number of queue	1
$Z$	Number of drones	800
$MF$	Maximum number of mating flights	200
$nsperm$	Capacity of spermatheca	80
$S(0)$	Speed of queen at beginning of flight	60
$\alpha$	Speed reduction schema	0.99
$P_c$	The breeding ratio	0.90
$P_m$	Mutation ratio	0.15

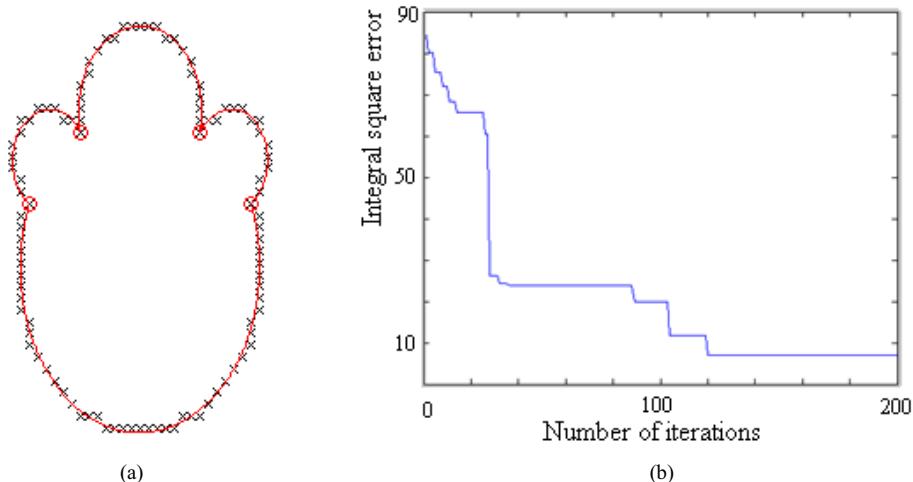
**Table 3.** The computation time

Curve with four semicircles		Chromosome-shaped curve		Figure-of-eight curve	
# Breakpoints	Time (s)	# Breakpoints	Time (s)	# Breakpoints	Time (s)
12	0.578	15	0.937	9	0.766
4	0.235	11	0.531	8	0.750

**Fig. 1.** The three digital test curves: (a) curve with four semicircles with 102 points; (b) chromosome-shaped curve with 60 points and (c) figure-of-eight curve with 45 points

Yin's method [4] is the GA-based method that obtains (near-) optimal polygonal approximations, and only line segments are used for approximation. It is obvious that

the value of ISE is larger compared with the proposed method. Sarkar et al.'s method [9] is also the GA-based method, and both line segments and circular arcs are used for approximation. On average, the proposed method outperforms Sarkar et al.'s method in terms of ISE. For the experiments of the curve with four semicircles, figure 2(a) shows the result obtained by the proposed method when  $m=4$  is employed. The plot of the integral square error in each iteration for the experiment of the curve with four semicircles is shown in Fig. 2(b). The result reveals fast convergence of the proposed method.



**Fig. 2.** (a) The approximation for the curve with four semi-circles and (b) the variation of integral square error versus iteration number

## 5 Conclusion

This paper presents a new curve approximation method based on the honey bee mating optimization algorithm. Two contributions are found from the experimental results. One is that the proposed HBMO-based method can obtain results with fast convergence. Another is that the integral square error using the proposed HBMO-based method is less than that of the two other methods. This result is promising and encourages further research on applying the HBMO-based algorithm to develop algorithms for image processing and pattern recognition.

## References

1. Teh, C.H., Chin, R.T.: On the Detection of Dominant Points on Digital Curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 859–872 (1989)
2. Ray, B.K., Ray, K.S.: Determination of Optimal Polygon from Digital Curve using L1 Norm. *Pattern Recognition* 26, 505–509 (1993)

3. Ho, S.Y., Chen, Y.C.: An Efficient Evolutionary Algorithm for Accurate Polygonal Approximation. *Pattern Recognition* 34, 2305–2317 (2001)
4. Yin, P.Y.: A New Method for Polygonal Approximation Using Genetic Algorithms. *Pattern Recognition Letters* 19, 1017–1026 (1998)
5. Yin, P.Y.: Ant Colony Search Algorithms for Optimal Polygonal Approximation of Plane Curves. *Pattern Recognition* 36, 1783–1797 (2003)
6. Wang, J., Kuang, Z., Xu, X., Zhou, Y.: Discrete Particle Swarm Optimization based on Estimation of Distribution for Polygonal Approximation Problems. *Expert Systems with Applications* 36, 9398–9408 (2009)
7. Pei, S.C., Horng, J.H.: Optimum Approximation of Digital Planar Curves Using Circular Arcs. *Pattern Recognition* 29, 383–388 (1996)
8. Horng, J.H., Li, J.T.: A Dynamic Programming Approach for Fitting Digital Planar Curves with Line Segments and Circular Arcs. *Pattern Recognition Letters* 22, 183–197 (2001)
9. Sarkar, B., Singh, L.K., Sarkar, D.: Approximation of Digital Curves with Line Segments and Circular Arcs Using Genetic Algorithms. *Pattern Recognition Letters* 24, 2585–2595 (2003)
10. Dorigo, M., Gambardella, L.: Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem. *IEEE Transactions on Evolutionary Computation* 1, 53–66 (1997)
11. Claudio, A.P., Aravena, C.M., Vallejos, J.I., Estevez, P.A., Held, C.M.: Face and Iris Localization Using Templates Designed by Particle Swarm Optimization. *Pattern Recognition Letters* 31, 857–868 (2010)
12. Fathian, M., Amiri, B., Maroosi, A.: Application of Honey Bee Mating Optimization Algorithm on Clustering. *Applied Mathematics and Computation* 190, 1502–1513 (2007)
13. Amiri, B., Fathian, M.: Integration of Self Organizing Feature Maps and Honey Bee Mating Optimization Algorithm for Market Segmentation. *Journal of Theoretical and Applied Information Technology* 3, 70–86 (2007)
14. Horng, M.H.: A Multilevel Image Thresholding Using the Honey Bee Mating Optimization. *Applied Mathematics and Computation* 215, 3302–3310 (2010)
15. Afshar, A., Haddad, O.B., Marino, M.A., Adams, B.J.: Honey-Bee Mating Optimization (HBMO) Algorithm for Optimal Reservoir Operation. *Journal of the Franklin Institute* 344, 452–462 (2007)
16. Horng, M.H., Liou, R.J., Wu, J.: Parametric Active Contour Model by Using the Honey Bee Mating Optimization. *Expert Systems with Applications* 37, 7015–7025 (2010)

# Harmful Adult Multimedia Contents Filtering Method in Mobile RFID Service Environment

Namje Park<sup>1</sup> and Youngsoo Kim<sup>2</sup>

<sup>1</sup> Department of Computer Education, Teachers College, Jeju National University,

61 Iljudong-ro, Jeju-si, Jeju Special Self-Governing Province, Korea

namjepark@jejunu.ac.kr, namjepark@gmail.com

<sup>2</sup> Information Security Research Division,

Electronics and Telecommunications Research Institute (ETRI),

161 Gajeong-dong, Yuseong-gu, Daejeon, 305-700, Korea

blitzkrieg@etri.re.kr

**Abstract.** This paper provides a privacy-enhanced method of verifying adults in mobile RFID environment. It can be applied to check whether a user is an adult or not when he or she would like to use some adult contents using mobile terminals playing a role of tag readers in mobile RFID environment. Instead of current adult verification method utilizing the user's own mobile terminals, the proposed method uses any mobile terminal and provides users with anonymity. Additionally, instead of a conventional rating system based on the user's age, we proposed four-level rating system (0 to 3) in detail. The adult content is classified based on each category (for example, swear word, nudity, sex, and foul language) and the rating information gets assigned to the adult content based on each category. The rating information is assigned to the user data region of the RFID tag. Through this method, the rating of multi-media content can be expressed with respect to each category using the detailed rating criteria. The weight factors may be differently applied to the categories according to applications.

## 1 Introduction

Users of mobile RFID networks wishing to access an adult content service at present need to go through an adult verification process prior to using the desired adult content service. The adult verification process is as follows [1]. First, when the user selects the adult content menu through the mobile terminal, a warning message is displayed on the terminal screen of the mobile terminal. The user inputs the last 7 digits of their social identification (ID) number, and then the mobile carrier compares stored user information with the input social ID number and proceeds to identify the user as an adult. When the user is identified, a password input window appears on the terminal screen of the mobile terminal. The user can then access the adult content if the password input is correct.

In this method of adult verification, since the adult verification is achieved through direct use of the mobile terminal, minors can access adult content without control

when they are aware of the social ID number of the official user and have access to the mobile terminal [2]. For example, a teenager may attempt to use his or her parents' mobile terminal to access adult content. In addition, since the user's social ID number is required, threats to maintaining the privacy of the personal information of the user may be serious [3]. Also, the present rating system of multimedia content including adult content is currently divided into three categories based on the user's age: M-rating, R-rating, and X-rating [4]. When these simple ratings are applied to the mobile RFID services, it is expected that a rating value be recorded in a user data region of a RFID tag. However, to protect the minors from harmful materials, more detailed ratings are needed.

In this paper, we provide a system of adult verification in a mobile RFID environment [5] in which the privacy of personal information can be maintained when a user undergoes adult verification needed for accessing adult content. Additionally, we also provide a brand new rating system that rating classifications of adult content are subdivided so that minors can be effectively protected from accessing adult content. The paper is organized as follows. We show a system framework and explain adult verification processes in section 2. Also, we show our rating criteria and expression for adult verification at section 3 and finish it with conclusion in section 4.

## 2 Related Research

### 2.1 Mobile RFID Technology

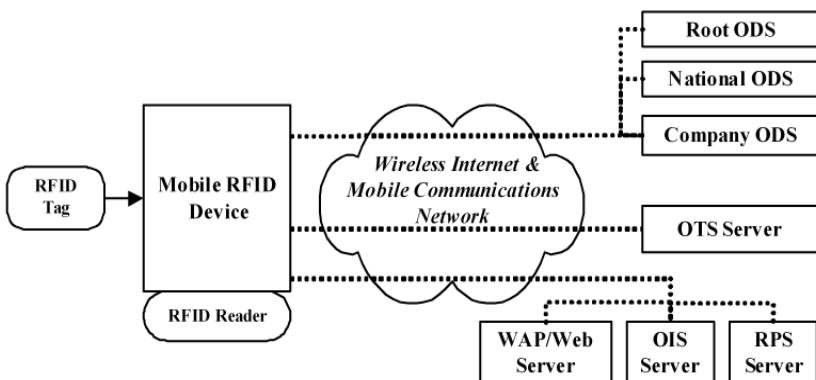
RFID is expected to be the base technology for ubiquitous network or computing, and to be associated with other technology such as telemetric, and sensors. Meanwhile Korea is widely known that it has established one of the most robust mobile telecommunication networks. In particular, about 78% of the population uses mobile phones and more than 95% among their phones have Internet-enabled function. Currently, Korea has recognized the potential of RFID technology and has tried to converge with mobile phone. Mobile phone integrated with RFID can activate new markets and end-user services, and can be considered as an exemplary technology fusion. Furthermore, it may evolve its functions as end-user terminal device, or 'u-device (ubiquitous device)', in the world of ubiquitous information technology [9].

Actually, mobile RFID phone may represent two types of mobile phone devices; one is RFID reader equipped mobile phone, and the other is RFID tag attached mobile phone. Each type of mobile phone has different application domains, for example, the RFID tag attached one can be used as a device for payment, entry control, and identity authentication, and the feature of this application is that RFID readers exist in the fixed positions and they recognize each phone to give user specific services like door opening. In the other hand, the RFID reader equipped mobile phone, which Korea is paying much attention now, can be utilized for providing end-users detailed information about the tagged object through accessing mobile network.

Korea's mobile RFID technology is focusing on the UHF range (860~960MHz), since UHF range may enable longer reading range and moderate data rates as well as relatively small tag size and cost. Then, as a kind of handheld RFID reader, in the selected service domain the UHF RFID phone device can be used for providing object information directly to the end-user using the same RFID tags which have widely spread.

## 2.2 Mobile RFID Services

A mobile RFID service structure aims to support ISO/IEC 18000-6 A/B/C standards for wireless communication between RFID tag and RFID reader. However, there is yet no RFID reader chip supporting all three wireless connection access specifications yet that the communication specification for the cellular phone will be determined by the mobile communication companies [1,9]. It will be also possible to mount the RF communication function to the Reader Chip using SDR technology and develop ISO/IEC 18000-6 A/B/C communication protocol in software to choose from protocols when needed.



**Fig. 1.** Architecture of a Mobile RFID Service

Key issues for mobile RFID terminal are the recognition distance to the RFID reader chip built into the cellular phone, transmission power, frequency, interface, technological standard, PIN specification, UART communication interface, mobile platform extended specification to control reader chip. The middleware functions for the RFID reader chip are provided to the application program in the form of a WIPI API as in figure 1. Here, 'Mobile RFID Device Driver' stands for the device driver software provided by the reader chip manufacturer.

Key issues for the mobile RFID network are the communication protocols such as the ODS (Object Directory Service) communication for code interpretation, the message format for the transmission and reception of contents between the cellular phone terminal and the application server, contents negotiation that supports the mobile RFID service environment and ensures the optimum contents transfer between the cellular phone terminal and the application server, and session management that enables the application to create and manage required status information while transmitting the message and the WIPI extended specification which supports these communication services [2,3,8].

The service model, as shown in figure 1, is a RFID tag, reader, middleware and information server. In the view of information protection, the serious problem for the RFID service is a threat of privacy. Here, the damage of privacy is of exposing the information stored in the tag and the leakage of information includes all data of

the personal possessing the tag, tagged products and location. The privacy protection on RFID system can be considered in two points of view. One is the privacy protection between the tag and the reader, which takes advantage of ID encryption, prevention of location tracking and the countermeasure of tag being forged. The other is of the exposure of what the information server contains along with tagged items. First of all, we will have a look about the exposure of information caused between the tag and the reader, and then discuss about the solution proposing on this paper.

### 3 A System Framework and Adult Verification Processes

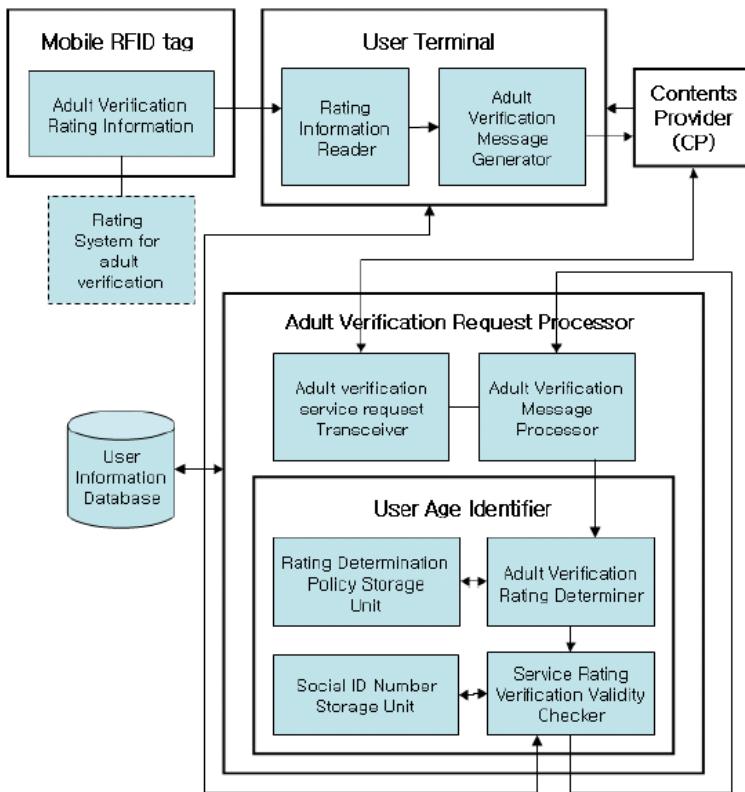
Fig.1 shows our system framework. It consists of 4 components; mobile RFID tag, user terminal, a contents provider(CP) and adult verification request processor. The adult verification processes are summarized as follows.

- The user terminal reads application data (adult verification rating information) indicating the rating recorded in a user region of the RFID tag. It requests the adult verification from the CP.
- The CP redirects the request to the adult verification request processor to process the adult verification.
- The user terminal logs into the adult verification request processor using an ID dedicated to the adult verification, that is an ID for connecting to the adult verification request processor. Then, it is assumed that the user has previously registered in the adult verification request processor.
- The adult verification request processor refers to a user information database and extrapolates the age of the user from the social ID number of the user.
- The adult verification request processor determines whether the user is permitted to have access to the adult content in reference to the extrapolated age of the user and the value of each category of the adult verification rating information and provides the determination result for the CP. The CP provides or does not provide the adult content to the user terminal corresponding to the determination result.

The function modules of the user terminal include a rating information reader and an adult verification message generator. The function modules of the adult verification request processor includes an adult verification service request transceiver, an adult verification message processor(AMP), and a user age identifier.

When the rating information reader included in the user terminal reads the adult verification rating information recorded in the RFID tag [6], the adult verification message generator generates an adult verification request message (inquiry message) including the RFID tag ID and the adult verification rating information and transmits the adult verification request message to the CP.

The CP redirects the adult verification request message to the adult verification request processor and requests the adult verification request processor to process the adult verification. The adult verification service request transceiver receives the message and transmits the message to the adult verification message processor. The adult verification message processor analyzes the adult verification request message, acquires each category value of the adult verification rating information included in the



**Fig. 2.** A system framework for adult verification

user data region, and transmits each acquired category value to the user age identifier. Then, each category value is recorded in the user data region of the RFID tag according to the rating criteria for each category stored in a rating system classifier.

An adult verification rating determiner determines final ratings of the adult verification rating information according to the final rating determination policy of an adult verification rating determination policy storage unit, and transmits the determined final rating and the tag ID to a service rating verification validity checker.

When the service rating verification validity checker requests and receives an input of the ID dedicated to the adult verification from the user terminal, the service rating verification validity checker reads the social ID number of a specific person from a social ID number storage unit and calculates the age of the specific person.

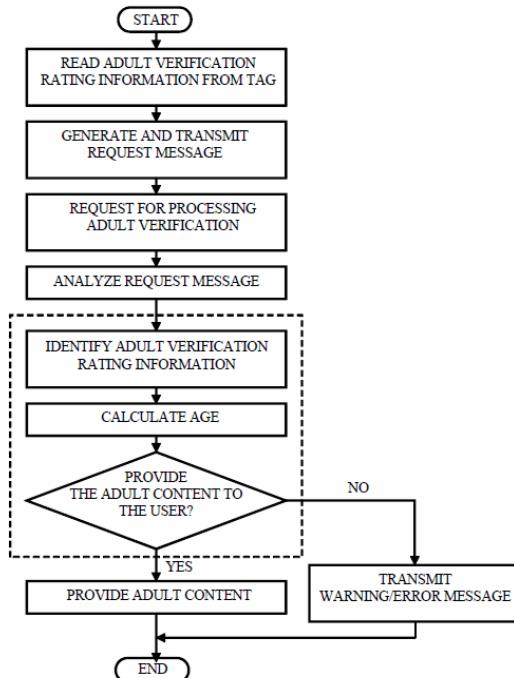
The service rating verification validity checker determines whether the CP provides the corresponding adult content to the user using the determined final rating and the calculated age of the user and transmits the determination result to the adult verification message processor.

The adult verification message processor notifies the CP of the determination result through the adult verification service request transceiver. The CP provides the corresponding adult content to the user terminal or does not provide the adult content by transmitting a warning or error message to the user terminal.

The user undergoes the adult verification through the aforementioned processes and if successful receives the adult content. In the aforementioned processes, any privacy threat to the social ID number is not taken into account as it is assumed that the adult verification request processor is a trusted third party (TTP) [7]. Since the user of the mobile terminal undergoes the adult verification through the dedicated ID stored in the adult verification request processor, the user can receive the adult content without the mobile terminal of the user and undergo the adult verification while maintaining the privacy of the social ID number of the user.

In addition, when a minor undergoes the adult verification using the personal information of a mobile terminal registrant registered in a mobile carrier, the chance that the minor passes the adult verification and gains access to the uncensored adult content when the minor has the mobile terminal of their parents and know the social ID numbers of their parents is eliminated.

Figure 3 shows operation processes. ACM (Adult Certification Management module) reads adult application data included at an RFID tag. Each category values of this data are obtained from CIS (Category Information Storage). AMM(Adult certification Message Management module) generates a query message, contains tag ID and adult application data, and sends it to CSS (Contents Service Server). CSS sends it to SRC of ACRM and request adult verification. AMM analyzes this message and gets each category values. It sends them to ACC. According to grade-policy of GPS, ACC(Adult certification Category Check module) decides and sends final grade to GVV (service Grade Validity Verification module).



**Fig. 3.** Operation Process

GVV requests mobile phone user's ID to input. After getting mobile phone user's ID, GVV refers to user's personal information at PIS and obtains user's age. Through final grade and user's age, GVV decides if CSS can serve contents to user. It sends the decision result to AMM. AMM sends the decision result and tag ID to CSS through SRC. According to the decision result, CSS provides contents or not.

## 4 The Rating Criteria and Expression for Adult Verification

In this section, we can show the rating criteria and expression for adult verification.

### 4.1 The Rating Criteria for Adult Verification

To receive the adult content, as described above, the adult verification rating information indicating the rating of the adult content is assigned to the user data region of the RFID tag. The rating system classifier employs four-level rating system (0 to 3) in detail, which is different from the conventional ratings based on ages. The adult content is classified based on each category (for example, swear word, nudity, sex, and foul language) and the rating information gets assigned to the adult content based on each category. The rating information is assigned to the user data region of the RFID tag. We can show the aforementioned rating criteria as below.

**Table 1.** Detailed Grade Description for Adult Data

Grade	Word(W)	Nudity(N)	Sex(S)	Language(L)
3	-Words for describing abnormal sex acts	-Exposing male/female genitals or pubes	-Explicit sexual acts or crime	-Using extremely loathsome expression -Using exceedingly vulgar expressions
2	-Words for describing normal sex acts	-Exposing female busts	-Sexual touching without taking off clothes	-Using disgusting expression -Using vulgar expressions
1	-Sex-related words for sex education/counsel	-Exposing male busts	-Passionate kissing	-Using light expletives
0	-Not available	-Not available	-Not available	-Not available

Content in which a word describing an abnormal sexual act is used, male/female genitals or pubes are exposed, a direct sexual act or sex crime is described, or outright foul language is used are rated 3. Content in which a normal sexual act description word is used, a female bust exposure picture, a clothed sexual act, or a serious swear word or rude expression is included are rated 2. Content in which a word describing a sexual act used for sex consulting/sex education is included or light foul language is used are rated 1. Content which is not included in the above cases are rated 0. This rating system is based on RSACi Rating System [8].

The rating system classifier records the adult verification rating information assigned based on the four categories (for example, swear words, nudity, sex, and foul

language) in the user data region of the RFID tag. The rating of multimedia content can be expressed with respect to each category using the detailed rating criteria. The weight factors may be differently applied to the categories such as swear words, nudity, sex, and foul language, according to applications.

## 4.2 The Rating Expression for Adult Verification

In figure 4, specific examples of the adult verification rating expression are suggested.

- A TYPE field represents the user data region. Expression format of the adult rating stored in the user data region is defined in the standards of the mobile RFID application data format [9]. In a structure of TYPE-LENGTH-VALUE (TLV), the TYPE value shown in figure 4 is included. Both COMMON and PRIVATE can be assigned to the C/P bit according to the codes included in the tag, products about the code, or kinds of services. Since the adult rating does not include other application data in a VALUE field, an M/P bit is set to MONO. TYPE CLASS may be a product/service and TYPE CODE may be an original code of the product/service.
- A LENGTH field represents length of the user data region, and the value of the adult rating application data allocated to the user data region of the RFID tag can be an integer from 0000(000000002) to 3333(11111112), the length of the LENGTH field is expressed in 8 bits. Accordingly, the value of the LENGTH field is set to a byte value of 110 and expressed in 8 bits as shown in figure 4.
- A VALUE field represents a practical adult rating with respect to the mobile RFID service and is expressed in integer values. The range of the VALUE field is from 0000 to 3333. As shown figure 4, two bits are assigned to swear words W, nudity N, sex S, and foul language L, respectively. A variety of methods of mapping the values of the VALUE field to the adult criteria are available.

In figure 4, an example of the value fields are expressed in the rating expression according to the definition of the mobile RFID application data format standards. This is a case where the value of VALUE field is 1120. For example, in a case where a value of W is 1, a value of N is 1, a value of S is 2, and a value of L is 0, the corresponding final rating may be 2, which is the maximum value, or 1, which is equal to two of the

BIT	C/P	M/P	RESERVED						TYPE CLASS			TYPE CODE				
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
<hr/>																
<hr/>																
C/P		RESERVED						ANY			0 1 0 1 1					
<hr/>																

VALUE = 1120 (W=1, N=1, S=2, L=0)																	
VALUE										LENGTH							
W N S L										BIT 7 6 5 4 3 2 1 0							
BIT 7 6 5 4 3 2 1 0										0 0 0 0 0 0 0 1							
0 1 0 1 1 0 0										110							

Fig. 4. TLV expression and an example

four values. In addition, when the different weight factors are allocated to categories W, N, S, and L, respectively, according to the application, the final rating is determined based on the corresponding weight factor policy. In other words, methods of rating the adult content can be diversified by allocating the weight factors to items W, N, S, and L, respectively.

## 5 Conclusion

In this paper, the adult certificate method in mobile RFID service network is proposed. According to the above sections, the user can receive the adult content without the terminal of the user and undergo the adult verification while maintaining the privacy of the social ID number. Also, when a minor undergoes the adult verification using the personal information of a mobile terminal registrant registered in a mobile carrier, the minor cannot pass the adult verification and gain access to the uncensored adult content even when the minor has the mobile terminal of their parents and know the social ID numbers of their parents. In addition, according to the proposed rating expression method, the content can be expressed with a wider variety of ratings by allocating content ratings to categories including swear words, nudity, sex, and foul language, respectively. Also, the adult rating can be adjusted according to the application and can be applied to various applications.

This paper is for checking a user whether he/she is an adult or not, when a user would like to read some adult-related data using mobile devices playing a role of tag readers on mobile RFID environments. Instead of current adult-certification method using one's own mobile devices, an adult certification in this paper uses any mobile devices and provides user's anonymity. In this paper, the application areas of the proposed platform are discussed briefly in the fields of RFID based LBS (Location Based Service), RFID-based mobile payment, RFID-based mobile CRM (Customer Relationship Management), and mobile ASP (Applications Service Provider).

For further study of this area, the verification and validation of the light-weight security middleware model by design and implementing some pilot-scale service system is necessary. It is also required to develop evaluation and authentication methodologies with the assistance of toolkits for the granularity of the QoS (Quality of Service) of the pilot-scale service system.

## Acknowledgments

This paper is extended from a conference paper presented at the eleventh annual IEEE international symposium on consumer electronics. The authors are deeply grateful to the anonymous reviewers for their valuable suggestions and comments on the first version of this paper.

## References

1. MRF Forum: Adult Certification for Mobile RFID Services. MRFS-4-04 (2005)
2. BBC News: Children can access mobile porn,  
<http://news.bbc.co.uk/1/hi/uk/4671334.stm>

3. Lubinski, A.: Security issues in mobile database access. In: Proc. of the IFIP WG 11.3 Twelfth International Conference on Database Security (1998)
4. Park, N., Kwak, J., Kim, S., Won, D., Kim, H.: WIPI Mobile Platform with Secure Service for Mobile RFID Network Environment. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 741–748. Springer, Heidelberg (2006)
5. Choi, W., et al.: An RFID Tag Using a Planar Inverted-F Antenna Capable of Being Stuck to Metallic Objects. *ETRI Journal* 28(2), 216–218 (2006)
6. Needham, R.M., Schroeder, M.D.: Authentication Revisited. *Operating System Review* 21(1) (1987)
7. MRF Forum: Application Data Format for Mobile RFID Services. MRFS-3-02 (2005)
8. Kim, Y., Park, N., Won, D.: Privacy-Enhanced Adult Certification Method for Multimedia Contents on Mobile RFID Environments. In: Proc. of IEEE International Symposium on Consumer Electronics, pp. 1–4. IEEE, Los Alamitos (2007)
9. Kim, Y., Park, N., Hong, D., Won, D.: Adult Certification System on Mobile RFID Service Environments. *Journal of Korea Contents Association* 9(1), 131–138 (2009)

# Efficient GHA-Based Hardware Architecture for Texture Classification

Shiow-Jyu Lin<sup>1,2</sup>, Yi-Tsan Hung<sup>1</sup>, and Wen-Jyi Hwang<sup>1,\*</sup>

<sup>1</sup> Department of Computer Science and Information Engineering,  
National Taiwan Normal University, Taipei, 116, Taiwan

<sup>2</sup> Department of Electronic Engineering, National Ilan University, I-Lan, 260, Taiwan  
[sjlin.gm@gmail.com](mailto:sjlin.gm@gmail.com), [697470157@ntnu.edu.tw](mailto:697470157@ntnu.edu.tw), [whwang@csie.ntnu.edu.tw](mailto:whwang@csie.ntnu.edu.tw)

**Abstract.** This paper presents a novel hardware architecture based on generalized Hebbian algorithm (GHA) for texture classification. In the architecture, the weight vector updating process is separated into a number of stages for lowering area costs and increasing computational speed. Both the weight vector updating process and principle component computation process can also operate concurrently to further enhance the throughput. The proposed architecture has been embedded in a system-on-programmable-chip (SOPC) platform for physical performance measurement. Experimental results show that the proposed architecture is an efficient design for attaining both high speed performance and low area costs.

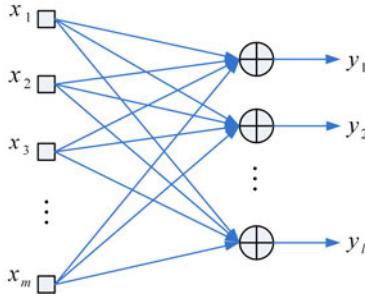
## 1 Introduction

The goal of principal component analysis (PCA) [3] is to reduce the  $m$ -dimensional feature vectors to  $l$ -dimensional feature vectors (where  $l < m$ ) for pattern recognition, classification, or data compression. This transformation is also known as PCA transform or Karhunen-Loeve transform (KLT) [4]. Although the PCA-based algorithms are widely used [5], some of the algorithms may not be well-suited for realtime applications because of their high computational complexity. This is especially true when the dimension of feature vector is large.

A number of algorithms [1] have been proposed for accelerating the computational speed of PCA. However, most of these algorithms are implemented by software. Therefore, only moderate acceleration can be achieved. Although hardware implementation of PCA and its variants are possible, large memory consumption and complicated circuit control management are usually required. The resource consumption may become impractically high as the feature vector dimension grows. An alternative for the PCA implementation is based on the PCA neural network, which is also known as the generalized Hebbian algorithm (GHA) [2,7]. Nevertheless, the GHA may converge slowly, and achieving a good accuracy requires excessive large number of iterations. Long computational time therefore is still required by many GHA-based algorithms.

---

\* To whom all correspondence should be sent.



**Fig. 1.** The neural model for the GHA

The objective of this paper is to present a hardware architecture of GHA for fast PCA. Although large amount of arithmetic computations are required for GHA, the proposed architecture is able to achieve fast training with low area cost. The long datapath for the updating of synaptic weight vectors is separated into a number of stages. The results of precedent stages will be used for the computation of subsequent stages for expediting training speed and lowering the area cost. The input vectors are allowed to be separated into smaller segments for the delivery over data bus with limited width. Both the weight vector updating process and principle component computation process can also operate concurrently to further enhance the throughput.

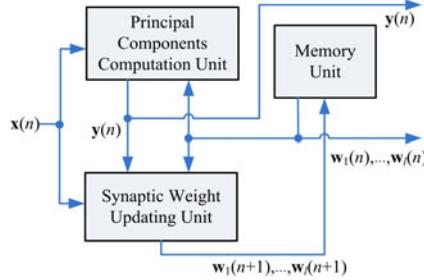
To demonstrate the effectiveness of the proposed architecture, a texture classification system on a system-on-programmable-chip (SOPC) platform is constructed. The system consists of the proposed architecture, a softcore NIOS II processor [8], a DMA controller, and a SDRAM. The proposed architecture is used for finding the PCA transform using the GHA training, where the training vectors are stored in the SDRAM. The DMA controller is used for the DMA delivery of the training vectors. The softcore processor is only used for coordinating the SOPC system. It does not participate the GHA training process. As compared with its software counterpart running on Pentium IV CPU, our system has significantly lower computational time for large training set. All these facts demonstrate the effectiveness of the proposed architecture.

## 2 Preliminaries

Let  $\mathbf{x}(n) = [x_1(n), \dots, x_m(n)]^T$ , and  $\mathbf{y}(n) = [y_1(n), \dots, y_l(n)]^T$ . The neural model of GHA transforms linearly input vector  $\mathbf{x}(n)$  into output vector  $\mathbf{y}(n)$  as depicted in Figure 1. The output vector  $\mathbf{y}(n)$  is related to the input vector  $\mathbf{x}(n)$  by

$$y_j(n) = \sum_{i=1}^m w_{ji}(n)x_i(n), \quad (1)$$

where  $w_{ji}(n)$  stands for the weight from the  $i$ -th synapse to the  $j$ -th neuron at time  $n$ . Each synaptic weight vector  $\mathbf{w}_j(n)$  is adapted by the Hebbian learning rule:



**Fig. 2.** The proposed GHA architecture

$$w_{ji}(n+1) = w_{ji}(n) + \eta[y_j(n)x_i(n) - y_j(n)\sum_{k=1}^j w_{ki}(n)y_k(n)], \quad (2)$$

where  $\eta$  denotes the learning rate. After a large number of iterative computation and adaptation,  $w_j(n)$  will asymptotically approach to the eigenvector associated with the  $j$ -th principal component  $\lambda_j$  of the input vector, where  $\lambda_1 > \lambda_2 > \dots > \lambda_l$ . To reduce the complexity of computing implementation, eq.(2) can be rewritten as

$$w_{ji}(n+1) = w_{ji}(n) + \eta y_j(n)[x_i(n) - \sum_{k=1}^j w_{ki}(n)y_k(n)]. \quad (3)$$

### 3 The Proposed GHA Architecture

The proposed GHA-based architecture, illustrated in Figure 2, is implemented by eq.(1) and (3). It consists of three functional units: the synaptic weight updating (SWU) unit, the principal components computing (PCC) unit, and memory unit. The latest synaptic weight vector  $w_j(n+1)$ ,  $j = 1, \dots, l$ , are stored in the memory unit. The synaptic weight vector  $w_j(n)$ ,  $j = 1, \dots, l$ , and the input vector  $x(n)$  are conveyed concurrently to the PCC unit and SWU unit in order to compute  $y(n)$  and  $w_j(n+1)$ ,  $j = 1, \dots, l$ , respectively.

The design of SWU unit is based on eq.(3). However, the direct implementation of eq.(3) requires large hardware resources. To reduce the resource consumption, we first define a vector  $z_{ji}(n)$  as

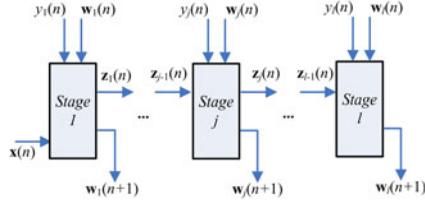
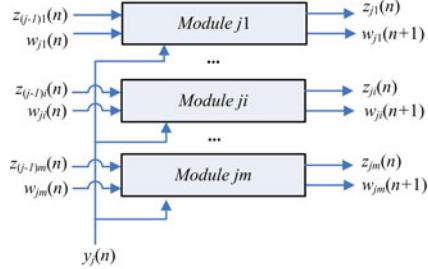
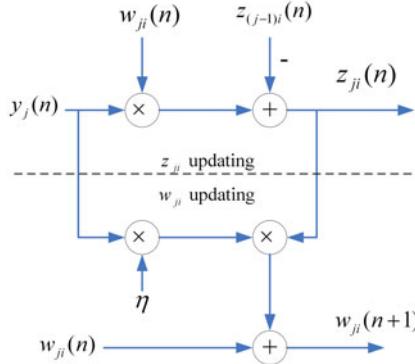
$$z_{ji}(n) = x_i(n) - \sum_{k=1}^j w_{ki}(n)y_k(n), \quad j = 1, \dots, l, \quad (4)$$

and  $\mathbf{z}_j(n) = [z_{j1}(n), \dots, z_{jm}(n)]^T$ . Integrating eq.(3) and (4), we obtain

$$w_{ji}(n+1) = w_{ji}(n) + \eta y_j(n)z_{ji}(n), \quad (5)$$

where  $z_{ji}(n)$  can be obtained from  $z_{(j-1)i}(n)$  by

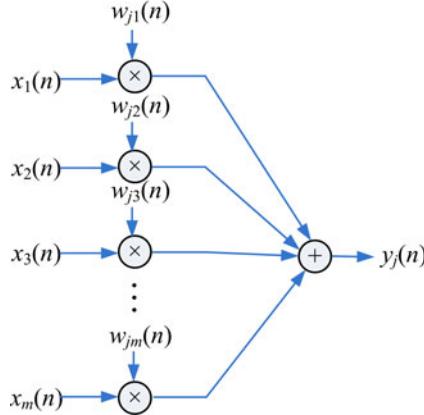
$$z_{ji}(n) = z_{(j-1)i}(n) - w_{ji}(n)y_j(n), \quad j = 2, \dots, l. \quad (6)$$

**Fig. 3.** The architecture of the SWU unit**Fig. 4.** The architecture of stage  $j$  of SWU unit**Fig. 5.** The architecture of the module  $ji$ 

When  $j = 1$ , from eq.(4) and (6), it follows that

$$z_{0i}(n) = x_i(n). \quad (7)$$

The architecture for implementing SWU unit, depicted Figure 3, is derived from eqs.(5) and (6). It consists of  $l$  stages, where each stage  $j$  produces  $\mathbf{w}_j(n+1)$  and  $\mathbf{z}_j(n)$ . There are  $m$  modules at each stage, as shown in Figure 4. Each module  $ji$  produces  $\mathbf{w}_{ji}(n+1)$  and  $\mathbf{z}_{ji}(n)$ . Figure 5 shows the architecture of each module  $ji$ .



**Fig. 6.** The basic architecture of PCC unit

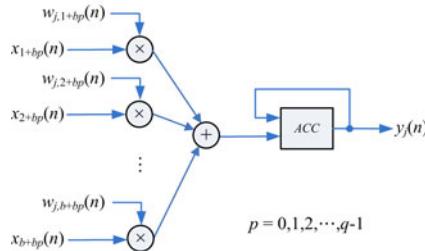
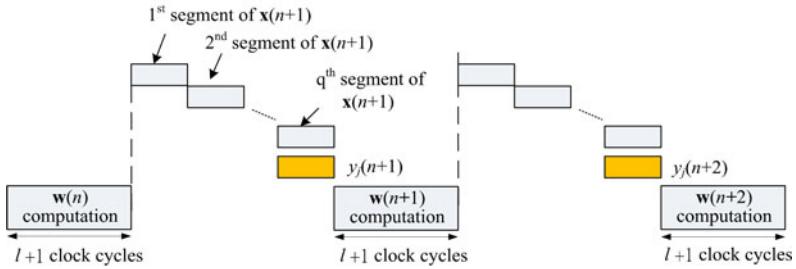
Next we consider the design of PCC unit. A simple implementation of PCC unit is based on eq.(1), as shown in Figure 6. As  $m$  becomes large, it may become difficult to obtain  $x_1(n), \dots, x_m(n)$  at the same time for computing  $y_j(n)$  due to the width limitation on data bus. One way to solve the problem is to separate  $x_1(n), \dots, x_m(n)$  into  $q$  segments, where each segment contains  $b$  elements (pixels), and  $q = m/b$ . The PCC unit then fetches one segment at a time for the computation of  $y_j(n)$ . The operation of the PCC unit can then be described as follows.

$$y_j(n) = \sum_{i=1}^m w_{ji}(n)x_i(n) = \sum_{p=0}^{q-1} \sum_{d=1}^b w_{j,d+bp}(n)x_{d+bp}(n). \quad (8)$$

Based on eq.(8), a modified PCC architecture, termed PCC-I architecture is presented [6]. As shown in Figure 7, the architecture consists of  $b$  multipliers and an accumulator. It will take  $q$  clock cycles to complete the computation of  $y_j(n)$  in eq.(1). Note that the computation of  $y_j(n)$  requires that  $w_{ji}(n)$  and  $x_i(n)$  are fixed. Therefore, the major disadvantage of PCC-I architecture is that the SWU unit should halt for  $q$  clock cycles for calculating  $y_j(n)$ , as shown in Figure 8. The throughput of the GHA will then be degraded.

To solve this problem, a novel PCC architecture, termed PCC-II architecture, is proposed. It removes the accumulator, and employs a  $q$ -stage shift register for the collection of segments of an input vector  $\mathbf{x}(n)$ , as shown in Figure 9. Each stage is able to fetch  $b$  pixels at a time. In this way, it will take only one clock cycle to compute  $y_j(n)$ . The SWU unit then should halt only one clock cycle. The computation of weight vectors can then overlap with the collection of input vectors. The throughput then is significantly improved. The corresponding timing diagram for PCC-II architecture is shown in Figure 10.

The proposed architecture is used as a custom user logic in a SOPC system consisting of softcore NIOS CPU [8], DMA controller and SDRAM, as depicted in [6]. All training vectors are stored in the SDRAM and then transported to the

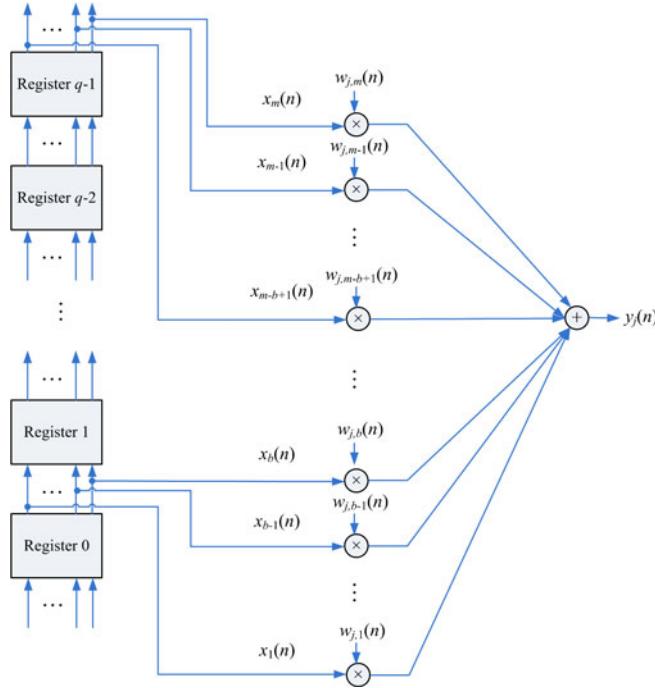
**Fig. 7.** The PCC-I architecture**Fig. 8.** The timing diagram of PCC and SWU units based on PCC-I architecture

proposed circuit via the DMA bus. The softcore NIOS CPU runs on a simple software to coordinate different components , including the proposed custom circuit in the SOPC. The proposed circuit operates as a hardware accelerator for GHA training. The resulting SOPC system is able to perform efficient on-chip training for GHA-based applications.

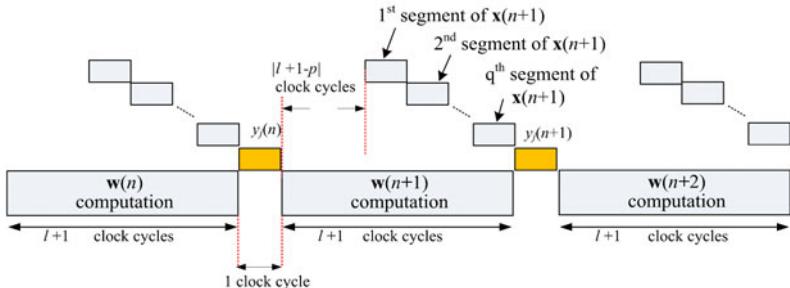
## 4 Experimental Results

This section presents some experimental results of the proposed architecture applied to texture classification. Figure 11 shows the five textures considered in this paper. The dimension of training vectors is  $4 \times 4$ (i.e.,  $m = 16$ ). There are 32000 training vectors in the training set. Due to the limitation of data bus, 4 pixels of each training vector will be conveyed into proposed circuit at a time (i.e.,  $b = 4$ ).The target FPGA device for all the experiments in this paper is Altera Cyclone III [9].

Table 1 shows the hardware resource consumption of the NIOS-based SOPC system with the proposed GHA architectures embedded as the custom user logic. There are two types of GHA architectures: the GHA with PCC-I architecture, and GHA with PCC-II architecture. Three different area costs are considered in the table: logic elements (LEs), embedded memory bits, and embedded multipliers. It can be observed from the table that the area costs grow linearly with  $l$ . In addition, given the same  $l$ , the GHA with PCC-II architecture consumes slightly



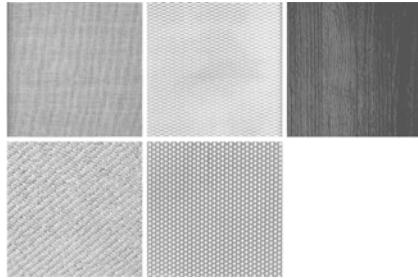
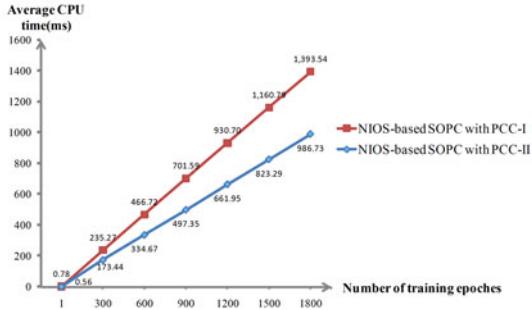
**Fig. 9.** The PCC-II architecture



**Fig. 10.** The timing diagram of PCC and SWU units based on PCC-II architecture

higher hardware resources as compared with the GHA with PCC-I architecture. However, because both GHA architectures share the same SWU and memory units, the differences for the hardware resource consumption is not high.

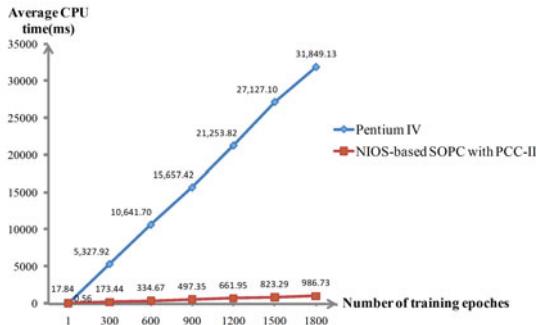
To demonstrate the effectiveness of the proposed architecture, the principal component based  $k$  nearest neighbor (PC- $k$ NN) rule is adopted. Two steps are involved in the PC- $k$ NN rule. In the first step, the GHA is applied to the input vectors to transform  $m$  dimensional data into  $l$  principal components. The synaptic weight vectors after the convergence of GHA training are adopted to

**Fig. 11.** Textures considered in the experiments**Fig. 12.** The CPU time of the proposed hardware architectures for different number of training epoches**Table 1.** Hardware resource consumption of the SOPC system using the proposed GHA architectures as custom user logic

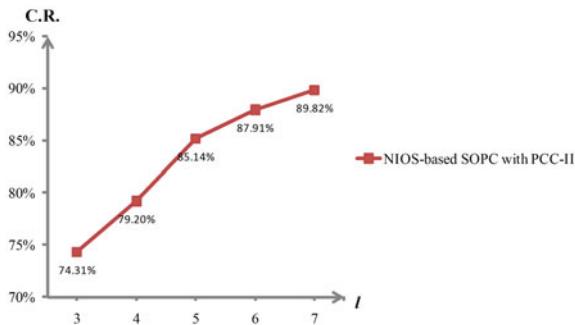
l	GHA with PCC-I			GHA with PCC-II		
	LEs	Memory Bits	Embedded Multipliers	LEs	Memory Bits	Embedded Multipliers
3	33534	2215216	208	33723	1600944	244
4	41204	2420016	276	41510	1805744	324
5	49012	2624816	344	49487	2010544	404
6	57004	2829616	412	57457	2215344	484
7	64805	3034416	480	65384	2420144	564

span the linear transformation matrix. In the second step, the  $k$ NN method is applied to the principal subspace for texture classification.

Figure 12 shows the CPU time of the NIOS-based SOPC system for various numbers of training epoches with  $l = 4$ . It can be observed from the figure that the GHA with PCC-II architecture has lower CPU time as compared with the GHA with PCC-I architecture. In addition, the gap in CPU time enlarges as the number of epoches increases. The GHA with PCC-II architecture has superior



**Fig. 13.** The CPU time of the proposed GHA with PCC-II and its software counterpart for different number of training epoches



**Fig. 14.** The classification success rates of the NIOS-based SOPC system for various numbers of principal components

speed performance because the SWU unit only has to halt for one clock cycle for each computation of  $y_j(n)$ . By contrast, the GHA with PCC-I architecture needs to wait for  $q$  clock cycles.

The CPU time of the software counterpart of our hardware GHA with PCC-II implementation is depicted in the Figure 13. The software training is based on the general purpose 3.0-GHz Pentium IV CPU. It can be clearly observed from Figure 13 that the proposed hardware architecture attains high speedup over its software counterpart. In particular, when the number of training epoches reaches 1800, the CPU time of the proposed SOPC system is 986.73 ms. By contrast, the CPU time of Pentium IV is 31849.13 ms. The speedup of the proposed architecture over its software counterpart therefore is 32.28.

The classification success rates of the NIOS-based SOPC system for various numbers of principal components are shown in Figure 14. We can see from Figure 14 that the classification success rate improves as  $l$  increases. As  $l = 7$ , the classification success rate attains 89.82 %. All these facts demonstrate the effectiveness of the proposed architecture.

## 5 Concluding Remarks

Experimental results reveal that the GHA with PCC-II has superior speed performance over its hardware and software counterparts. In addition, the architecture is able to attain near 90 % classification success rate for texture classification when  $l = 7$ . The architecture also has low area cost for implementing the SWU unit by the employment of an efficient architecture separating weight vector updating process into  $m$  stages. The proposed architecture therefore is an effective alternative for on-chip learning applications requiring both low area cost and high speed computation.

## References

1. Gunter, S., Schraudolph, N.N., Vishwanathan, S.V.N.: Fast Iterative Kernel Principal Component Analysis. *Journal of Machine Learning Research*, 1893–1918 (2007)
2. Haykin, S.: *Neural Networks and Learning Machines*, 3rd edn. Pearson, London (2009)
3. Jolliffe, I.T.: *Principal component Analysis*, 2nd edn. Springer, Heidelberg (2002)
4. Karhunen, J., Joutsensalo, J.: Generalization of Principal Component Analysis, Optimization Problems, and Neural Networks. *Neural Networks*, 549–562 (1995)
5. Kim, K., Franz, M.O., Scholkopf, B.: Iterative kernel principal component analysis for image modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 1351–1366 (2005)
6. Lin, S.J., Hung, Y.T., Hwang, W.J.: Fast Principal Component Analysis Based on Hardware Architecture of Generalized Hebbian Algorithm. In: ISICA 2010. LNCS. Springer, Heidelberg (2010) (to be published)
7. Sanger, T.D.: Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks* 12, 459–473 (1989)
8. NIOS II Processor Reference Handbook, Altera Corporation (2010), <http://www.altera.com/literature/lit-nio2.jsp>
9. Cyclone III Device Handbook, Altera Corporation (2010), <http://www.altera.com/products/devices/cyclone3/cy3-index.jsp>

# Stability by Feedback of the Second Order Singular Distributed Parameter Systems Containing Infinite Many Poles

Feng Liu<sup>1</sup>, Guodong Shi<sup>2</sup>, and Jianchun Wu<sup>3</sup>

<sup>1</sup> Institute of Control Theory and Applications  
Jiangsu Teachers University of Technology,  
Changzhou, 213001, P.R. China  
[liufeng200099@vip.sina.com](mailto:liufeng200099@vip.sina.com)

<sup>2</sup> School of Electrical and Information Engineering  
Changzhou University, Changzhou, 213164, P.R. China  
[sgd@jstu.edu.cn](mailto:sgd@jstu.edu.cn)

<sup>3</sup> Department of Basic Courses  
Jiuquan Vocational and Technical College  
Jiuquan, 735000, P.R.China  
[jqxywujc@163.com](mailto:jqxywujc@163.com)

**Abstract.** Stability by feedback of the second order singular distributed parameter system containing infinite many poles is discussed via the operator theory and subnegative definite matrix, an equivalent proposition for the stability by feedback of this system is given, and the constructive expression of the feedback control law is given by the singular inverse one of bounded linear operator. This research is theoretically important for studying the stability of the singular distributed parameter systems.

**Keywords:** Stability by feedback, second order singular distributed parameter systems, operator theory, Hilbert place.

## 1 Introduction

Singular distributed parameter systems are systems which are broader in range than distributed parameter systems. It appears in the study of the temperature distribution in a composite heat conductor, voltage distribution in electromagnetically coupled superconductive circuits, signal propagation in a system of electrical cables [1]-[3] etc. There is an essential distinction between them and ordinary distributed parameter control systems [1]-[7]. When under disturbance, they not only lose the stability, but also take place great changes in themselves structure, such as presenting impulsive behavior etc.

One of the most important problems is to study the stability by feedback of the singular distributed parameter systems. In [8], we had discussed stability by feedback of the first order singular distributed parameter systems containing infinite non-negative

real part poles; in [9], discussed asymptotical stability for the first order singular systems by the minimized norm theory, give an equipollence conclusion for stability by feedback of the first order singular control system and the first order ordinarily control system, and put forward a practical algorithm for seeking feedback control law; in [10], discussed stability by feedback of the first order singular distributed parameter systems with multi-observers; in [11], discussed uniform exponential stability of the time varying singular distributed parameter systems; in [12], Ge had studied the stabilization by feedback of the second order singular distributed parameter systems, but in which only finite many poles are discuss. In this paper, stability by feedback of the second order singular distributed parameter system containing infinite many poles is discussed via the operator theory and subnegative definite matrix, an equivalent proposition for the stability by feedback of this system is given, and the constructive expression of the feedback control law is given by the singular inverse one of bounded linear operator. This research is theoretically important for studying the stability of the singular distributed parameter systems.

Consider the vibrating problem of the elastic singular beam in which its two terminals are fixed, and mathematics model of this problem can be described by the following equations:

$$M(x) \frac{\partial^2 y(t, x)}{\partial t^2} = \frac{\partial^2}{\partial x^2} (E(x) I(x) \frac{\partial^2 y(t, x)}{\partial x^2}) + b(x) u(t), \quad (1)$$

where  $y(t, x) \in R^n$  is a vector of displacement,  $M(x) \in R^{n \times n}$  is mass distribution constant matrices of the singular beam, where we suppose that it is constant matrix, and usually, this matrix is singular,  $b(x) \in R^n$ ,  $u(t)$  is a scalar-valued function, we have the following initial and boundary conditions:

$$\begin{aligned} y(0, x) &= F_0(x), \left. \frac{\partial y}{\partial t} \right|_{t=0} = F_1(x), \quad y(t, 0) = 0, \quad y(t, l) = 0, \\ \frac{\partial}{\partial x} y(t, 0) &= 0, \quad \frac{\partial}{\partial x} y(t, l) = 0, \end{aligned}$$

Since the system (1) is a singular partial differential equation, i.e. the system (1) is an infinite dimensional singular control system (i.e. singular distributed parameter system), therefore, it can be described by the abstract evolution equation in Hilbert space. In fact, let  $H$  be a separable complex Hilbert space such that  $y(t) \in H$ ,

$$\begin{aligned} E y(t)(x) &= M y(t, x), \quad A y(t)(x) = \frac{\partial^2}{\partial x^2} (E(x) I(x) \frac{\partial^2 y(t, x)}{\partial x^2}), \\ y(t)(x) &= y(t, x), \quad b u(t) = b(x) u(t), \end{aligned}$$

then  $A$  is a linear operator,  $E$  is a bounded linear operator,  $b \in H$ , and we can obtain the following second order singular distributed parameter system in Hilbert space:

$$\begin{cases} E \frac{d^2 y}{dt^2} = Ay + bu(t), \\ y(0) = y_0, \left. \frac{dy}{dt} \right|_{t=0} = y_1. \end{cases} \quad (2)$$

For singular bowstring, singular film, and magnetic resonance spectroscopic imaging in material science and medicine, the vibrating problem also are exist. Therefore, we shall discuss the stability by feedback of the system (2) in general case.

In the following,  $E^*$  denotes the adjoint of  $E$ ,  $\sigma_p(E, A) = \{\lambda : \text{there exists } y \in H \text{ and } y \neq 0 \text{ such that } \lambda E y = A y\}$  denotes the set of all finite generalized eigenvalues of  $E$  and  $A$ ;  $\rho(E, A) = \{\alpha : (\alpha E - A) \text{ is a regular operator}\}$ ;

$$R(\alpha E; A) = (\alpha E - A)^{-1}$$

for  $\alpha \in \rho(E, A)$ ,  $I$  denotes the identical operator in  $H$ .

## 2 Preliminaries

Let the feedback is

$$u(t) = \langle E \frac{dy}{dt}, g \rangle, \quad (3)$$

where  $g \in H$ ,  $g \neq 0$ ,  $\langle \cdot, \cdot \rangle$  denotes the inner product in  $H$ , substituting (3) into (2) gives the following closed loop singular distributed parameter systems:

$$\begin{cases} E \frac{d^2 y}{dt^2} = G \frac{dy}{dt} + Ay, \\ y(0) = y_0, \left. \frac{dy}{dt} \right|_{t=0} = y_1, \end{cases} \quad (4)$$

where  $Gy = \langle Ey, g \rangle$ , the associated generalized eigenequations of closed loop system (4) reads as follows:

$$\lambda^2 E y = \lambda G y + A y. \quad (5)$$

Let  $\sigma_p = \{\lambda : \text{there exists } y \in H \text{ and } y \neq 0 \text{ such that } \lambda^2 E y = \lambda G y + A y\}$  denotes the set of all finite poles of the closed loop system (4).

**Definition 1.** For arbitrary  $y(0) \in H$ , if the solution  $y(t)$  of the system (4) satisfies  $\lim_{t \rightarrow \infty} \|y(t)\| = 0$ , where  $\|y(t)\| = \sqrt{\langle y(t), y(t) \rangle}$ , then the system (2) is called stabilization by feedback.

The problem which will be discussed is whether there exist  $g \in H, g \neq 0$  such that the system (4) is stable, i.e., the system (2) is stable by feedback, and how to seek the constructive expression of  $g \in H, g \neq 0$ .

**Definition 2.** Suppose matrix  $A \in C^{n \times n}$ , if  $\forall 0 \neq y \in C^n$ , always have  $y^T A y < 0$ , then  $A$  is called subnegative definite matrix ( $A$  is not always symmetry matrix).

It is easy to prove that  $A$  is subnegative definite matrix if and only if  $\text{Re}(\lambda(A)) < 0$ .

**Lemma 1.** [6] Let  $x, a \in H, B \in B(H)$ , and there exist  $B^+$ . If  $Bx = a$  is solvable, then the general solutions are  $x = B^+ a + [I - B^+ B]c$ , where  $c$  is any element in  $H$ .

**Hypothesis 1.** Let  $A$  is an invertible and closed densely defined linear operator and there exist  $A^{\frac{1}{2}}$ . let  $E$  is a bounded linear operator,  $E$  and  $A$  only have the finite generalized eigenvalues  $\{\lambda_k\}_1^\infty$ , and any  $\lambda_k$  is single,  $\varphi_k$  is the associated generalized eigenvector, i.e.  $\lambda_k E \varphi_k = A \varphi_k$  ( $k = 1, 2, \dots$ ). Let  $\{\psi_k\}_1^\infty$  denote the set of all generalized eigenvectors of  $E^*$  and  $A^*$  satisfying  $\overline{\lambda_k} E^* \psi_k = A^* \psi_k$  ( $k = 1, 2, \dots$ ) and form a subset of an unconditional bases in  $H$ , and

$$\langle E \varphi_k, \psi_l \rangle = \begin{cases} 1, & k = l \\ 0, & k \neq l \end{cases} \quad (k, l = 1, 2, \dots).$$

Let

$$E_1 = \begin{bmatrix} I & O \\ O & E \end{bmatrix}, \quad A_1 = \begin{bmatrix} O & A^{1/2} \\ A^{1/2} & O \end{bmatrix}, \quad G_1 = \begin{bmatrix} O & O \\ O & G \end{bmatrix},$$

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, \quad z_1 = A^{\frac{1}{2}} y, \quad z_2 = \frac{dy}{dt},$$

from (4), we have

$$\begin{cases} E_1 \frac{dz}{dt} = (A_1 + G_1)z, \\ z_0 = \begin{bmatrix} \frac{1}{A^{\frac{1}{2}}} y_0 \\ y_1 \end{bmatrix}, \end{cases} \quad (6)$$

the associated generalized eigenequations of the system (6) reads as follows:

$$\lambda E_1 z = (A_1 + G_1)z. \quad (7)$$

**Hypothesis 2.** Suppose there exist regular operators  $P, Q$  such that

$$QE_1P = \begin{bmatrix} I_r & O \\ O & O \end{bmatrix}, Q(A_1 + G_1)P = \begin{bmatrix} A_{11} & O \\ A_{21} & A_{22} \end{bmatrix},$$

where the matrix  $A_{22}$  is invertible. From the system (6), we have

$$\begin{bmatrix} I_r & O \\ O & O \end{bmatrix} \begin{bmatrix} \frac{dz_1}{dt} \\ \frac{dz_2}{dt} \end{bmatrix} = \begin{bmatrix} A_{11} & O \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

i.e.

$$\begin{cases} \frac{dz_1}{dt} = A_{11}z_1, z_{10} = A^{\frac{1}{2}}y_0, \\ O = A_{21}z_1 + A_{22}z_2, z_{20} = y_1, \end{cases} \quad (8)$$

the associated eigenequations of the first system in (6) reads as follows:

$$\lambda z_1 = A_{11}z_1. \quad (9)$$

The following lemmas 2 and 3 can be proved directly:

**Lemma 2.** If hypothesis 1 and 2 are satisfied, then  $(\lambda, y)$  is a solution of (5) if and only if  $(\lambda, z)$  is a solution of (7), and if and only if  $(\lambda, z_1)$  is a solution of (9). i.e.  $\sigma_p = \sigma_p(E_1, A_1 + G_1) = \sigma_p(I_r, A_{11})$ .

**Lemma 3.** If hypothesis 1 and 2 are satisfied, then the solution of the system (8) is unique.

**Lemma 4[13].** If  $E_1$  and  $A_1$  only have the finite generalized eigenvalues and  $\alpha \notin \sigma_p(E_1, A_1)$ , then  $\alpha \in \sigma_p(E_1, A_1 + G_1)$  if and only if

$$\langle E_1 R(\alpha E_1, A_1) b'_1, g'_1 \rangle = 1,$$

$$\text{where } b'_1 = \begin{bmatrix} 0 \\ b \end{bmatrix}, g'_1 = \begin{bmatrix} 0 \\ g \end{bmatrix}.$$

### 3 Main Results and Proof

**Theorem 1.** The system (2) is stable by feedback if and only if the matrix  $A_{11}$  is a subnegative definite matrix.

**Proof.** From the above 2, we know that the system (2) can be written as the system (8), and the solution of the system (8) are

$$z_1 = z_{10} e^{A_{11}t}, z_2 = -A_{22}^{-1} A_{21} z_{10} e^{A_{11}t},$$

because  $A_{11}$  is subnegative definite matrix, therefore, the real part of all eigenvalue of the matrix  $A_{11}$  are negative, thereby, we have

$$\lim_{t \rightarrow \infty} \|z_1(t)\| = 0, \lim_{t \rightarrow \infty} \|z_2(t)\| = 0,$$

i.e.  $\lim_{t \rightarrow \infty} \|z(t)\| = 0$ , therefore,  $\lim_{t \rightarrow \infty} \|y\| = 0$ , i.e. the system (2) is stable by feedback.

The reverse is also true.

**Theorem 2.** If hypothesis 1 and 2 are satisfied, and there exist  $E^+$ . Let

$$J = \{\pm 1, \pm 2, \dots, \pm n, \dots\}, \mu_n = \begin{cases} \sqrt{\lambda_n}, & n = 1, 2, \dots \\ -\sqrt{\lambda_n}, & n = -1, -2, \dots \end{cases}, \lambda_n = \lambda_{-n} (n = 1, 2, \dots),$$

and let

$$1) d_k = \frac{1}{\sqrt{2}} \langle b, \psi_k \rangle \neq 0 (k = \pm 1, \pm 2, \dots), \psi_k = \psi_{-k} (k = 1, 2, \dots);$$

$$2) \inf_{i \neq j} |\lambda_i - \lambda_j| \geq \delta > 0;$$

3)  $\{\alpha_i\}_{i \in J}$  is a complex numbers set which satisfy

$$\alpha_i \neq \alpha_j (i \neq j), \alpha_i \notin \sigma_p(E_1, A_1),$$

where  $c$  is a constant real number, and

$$\sum_{i \in J} \left| \frac{\alpha_i - \mu_i}{\frac{1}{\sqrt{2}} \langle b, \psi_i \rangle} \right|^2 \quad (10)$$

be convergence. Let  $c_i = (\mu_i - \alpha_i) \prod_{\substack{j \in J \\ j \neq i}} \frac{\alpha_j - \mu_i}{\mu_j - \mu_i} (i \in J)$  satisfy  $c_i = c_{-i} (i \in J)$ ,

If  $\operatorname{Re} \alpha_i < c < 0$ , then there exists  $g \in H$ , and

$$g = \sum_{k \in J} \frac{-c_k}{\sqrt{2} d_k} \psi_k + [I - (E^*)^+ E^*] a$$

such that the system (2) is stable by feedback, where  $a$  is any element in  $H$ .

**Proof.** Let

$$u_l = \frac{1}{\sqrt{2}} \begin{bmatrix} \operatorname{sgn}(l) E \varphi_l \\ \varphi_l \end{bmatrix}, v_l = \frac{1}{\sqrt{2}} \begin{bmatrix} \operatorname{sgn}(l) E^* \psi_l \\ \psi_l \end{bmatrix},$$

where  $\varphi_l = \varphi_{-l}, \psi_l = \psi_{-l}$ . The following results can be proved directly:

$$\sqrt{\lambda_l} E_1 u_l = A_1 u_l, \sqrt{\lambda_l} E_1^* v_l = A_1^* v_l,$$

$$\langle E_1 u_l, v_k \rangle = \begin{cases} 1, & l=k \\ 0, & l \neq k \end{cases}, \quad \langle b'_1, v_k \rangle = \frac{1}{\sqrt{2}} \langle b, \psi_k \rangle = d_k$$

and  $\{v_k\}_{k \in J}$  forms a subset of an unconditional bases in  $H \times H$ .

Let

$$f(z) = \prod_{i \in J} \left( \frac{z - \alpha_i}{z - \mu_i} \right) = \prod_{i \in J} \left( 1 - \frac{\mu_i - \alpha_i}{z - \mu_i} \right),$$

since series (10) is convergent, we can derive that  $f(z)$  is a semi-scalar function of which  $\mu_i$  is a single pole and  $\lim_{z \rightarrow \infty} \frac{f(z)}{z} = 0$ . Using the results of complex analysis, we have

$$f(z) = f(0) + \sum_{n \in J} c_n \left( \frac{1}{z - \mu_n} + \frac{1}{\mu_n} \right). \quad (11)$$

If  $z \notin \sigma_p(E_1, A_1)$  in (11) and  $z \rightarrow \infty$ , then  $1 = f(0) + \sum_{n \in J} \frac{c_n}{\mu_n}$ .

By (11) we obtain

$$f(z) = 1 + \sum_{n \in J} \frac{c_n}{z - \mu_n}, \quad (12)$$

from (12) we have that

$$c_n = (\mu_n - \alpha_n) \prod_{\substack{j \in J \\ j \neq n}} \frac{\alpha_j - \mu_n}{\mu_j - \mu_n} (n \in J)$$

and  $\alpha_i$  is a zero point of  $f(z)$ . Using (11), we obtain  $1 = \sum_{n \in J} \frac{-c_n}{\alpha_j - \mu_n} (j \in J)$ .

Let  $\Delta_i = \prod_{\substack{j \in J \\ j \neq i}} \left( \frac{\alpha_j - \mu_i}{\mu_j - \mu_i} \right) (i \in J)$ , then

$$|\Delta_i| \leq \prod_{j \in J} \left( 1 + \left| \frac{\alpha_i - \mu_j}{\mu_j - \mu_i} \right| \right) \leq \exp \left( \sum_{\substack{j \in J \\ j \neq i}} \left| \frac{\alpha_i - \mu_j}{\mu_i - \mu_j} \right| \right)$$

$$\leq \exp \left[ \frac{1}{\delta} \sum_{\substack{j \in J \\ j \neq i}} \left| \frac{\alpha_i - \mu_j}{d_j} \right| \right]$$

$$\leq \exp\left[\frac{1}{\delta}\left(\sum_{\substack{j \in J \\ j \neq i}} \left|\frac{\alpha_i - \mu_j}{d_j}\right|^2\right)^{1/2} \left(\sum_{\substack{j \in J \\ j \neq i}} |d_j|^2\right)^{1/2}\right] .$$

$$= M < \infty (i \in J),$$

thus

$$\sum_{i \in J} \left| \frac{c_i}{d_i} \right|^2 = \sum_{i \in J} \left| \frac{\mu_i - \alpha_i}{d_i} \Delta_i \right|^2 \leq M^2 \sum_{i \in J} \left| \frac{\mu_i - \alpha_i}{d_i} \right|^2 . \quad (13)$$

From (13) we have that  $\sum_{i \in J} \left| \frac{c_i}{d_i} \right|^2$  is convergent, therefore

$$g_1 = \sum_{i \in J} \frac{-c_i}{\sqrt{2d_i}} v_i + [I_1 - (E_1^*)^+ E_1^*] a_1$$

is an element in  $H \times H$ , where  $a_1$  is any element in  $H \times H$ , since  $c_i = c_{-i}$  ( $i \in J$ ), we have

$$g'_1 = \begin{bmatrix} 0 \\ \sum_{i \in J} \frac{-c_i}{\sqrt{2d_i}} \psi_i + [I - (E^*)^+ E^*] a \end{bmatrix},$$

therefore  $g = \sum_{i \in J} \frac{-c_i}{\sqrt{2d_i}} \psi_i + [I - (E^*)^+ E^*] a$ , where  $a$  is any element in  $H$ , and

$$\begin{aligned} < E_1 R(\alpha_j E_1, A_1) b'_1, g'_1 > &= \sum_{i \in J} \frac{-c_i}{d_i} < R(\alpha_j E_1, A_1) b'_1, E_1^* v_i > \\ &= \sum_{i \in J} \frac{-c_i}{d_i (\alpha_j - \mu_i)} < R(\alpha_j E_1, A_1) b'_1, (\overline{\alpha_j} E_1^* - A_1^*) v_i > \\ &= \sum_{i \in J} \frac{-c_i}{d_i (\alpha_j - \mu_i)} < b'_1, v_i > \\ &= \sum_{i \in J} \frac{-c_i}{d_i (\alpha_j - \mu_i)} < b, \psi_i > \\ &= \frac{1}{\sqrt{2}} \sum_{i \in J} \frac{-c_i}{d_i (\alpha_j - \mu_i)} < b, \psi_i >. \\ &= \sum_{i \in J} \frac{-c_i}{\alpha_j - \mu_i} = 1 . \end{aligned}$$

From lemma 4 we obtain  $\alpha_i \in \sigma_p(E_1, A_1 + G_1)$ , thus  $\{\alpha_i\}_{i \in J} \subset \sigma_p(E_1, A_1 + G_1)$ .

In order to prove  $\{\alpha_i\}_{i \in J} \supset \sigma_p(E_1, A_1 + G_1)$ , we only need to prove if  $\alpha \notin \{\alpha_i\}_{i \in J}$  then  $\alpha \notin \sigma_p(E_1, A_1 + G_1)$ .

Using disproof, suppose  $\alpha \in \sigma_p(E_1, A_1 + G_1)$ , the following two cases are discussed respectively:

(i) If  $\alpha \notin \sigma_p(E_1, A_1)$ , then by lemma4, we have

$$\langle E_1 R(\alpha E_1, A_1) b'_1, g'_1 \rangle = 1, \quad (14)$$

since

$$R(\alpha E_1, A_1) - R(\alpha_i E_1, A_1) = -(\alpha - \alpha_i) R(\alpha E_1, A_1) E_1 R(\alpha_i E_1, A_1),$$

from (14), we obtain

$$(\alpha - \alpha_i) \langle R(\alpha E_1, A_1) E_1 R(\alpha_i E_1, A_1) b'_1, g'_1 \rangle = 0 (i \in J_N).$$

Using

$$g'_1 = \begin{bmatrix} 0 \\ \sum_{i \in J} \frac{-c_i}{\sqrt{2d_i}} \psi_i + [I - (E^*)^+ E^*] a \end{bmatrix},$$

and  $\alpha \neq \alpha_i$ , we deduce

$$\sum_{j \in J_N} \frac{a_j b_{ij}}{(\alpha - \mu_j)(\alpha_i - \mu_j)} = \langle R(\alpha E_1, A_1) E_1 R(\alpha_i E_1, A_1) b'_1, g'_1 \rangle = 0 (i \in J_N).$$

Let the matrix  $D_{2N} = [d_{jk}]_{2N \times 2N}$ , where  $d_{jk} = \frac{1}{\alpha_k - \mu_j}$ , since

$$\det D_{2N} = (-1)^{\frac{2N(2N-1)}{2}} \frac{\prod_{-N \leq k < j \leq N} (\alpha_k - \alpha_j) \prod_{-N \leq k < j \leq N} (\mu_k - \mu_j)}{\prod_{k=-N}^N \prod_{j=-N}^N (\alpha_k - \mu_j)} \neq 0, \frac{a_j b_{ij}}{\alpha - \mu_j} = 0 (j \in J_N),$$

thus  $a_j = 0 (j \in J_N)$ .

(ii) If  $\alpha \in \sigma_p(E_1, A_1)$ , there exists  $\mu_i (i \in J_N)$  such that  $\alpha = \mu_i$ .

Since  $\alpha \in \sigma_p(E_1, A_1)$ , there exists  $\varphi \neq 0$  such that

$$(A_1 + G_1)\varphi = A_1\varphi + \langle E_1\varphi, g'_1 \rangle b'_1 = \mu_i E_1\varphi, \quad (15)$$

thus

$$\langle (A_1 - \mu_1 E_1)\varphi, \psi_{1,l} \rangle + \langle E_1\varphi, g'_1 \rangle b'_1, \psi_{1,l} \rangle = 0, \quad (16)$$

from  $\langle (A_l - \mu_l E_l)\varphi, \psi_{l,l} \rangle = \langle \varphi, (A_l - \mu_l E_l)^* \psi_{l,l} \rangle = 0$  and (16), we have  $\langle E_l \varphi, g'_l \rangle = 0$ . According to (15), we have  $A_l \varphi = \mu_l E_l \varphi$ , hence  $\varphi = \mu_l$  and  $a_i = \langle E_l \mu_i, g'_l \rangle = 0$ , this is in contradiction with  $a_i \neq 0 (i \in J_N)$ .

From the above prove proof we obtain  $\{\alpha_i\}_{i \in J} \supset \sigma_p(E_l, A_l + G_l)$ , hence

$$\{\alpha_i\}_{i \in J} = \sigma_p(E_l, A_l + G_l).$$

From the lemma 2 and the theorem 1, we know, when

$$g = \sum_{i \in J} \frac{-c_i}{\sqrt{2d_i}} \psi_i + [I - (E^*)^+ E^*]a,$$

the system (2) is stable by feedback.

## 4 Conclusion

In this paper, stability by feedback of the second order singular distributed parameter system containing infinite many poles is discussed via the operator theory and subnegative definite matrix, an equivalent proposition for the stability by feedback of this system is given, and the constructive expression of the feedback control law is given by the singular inverse one of bounded linear operator. This research is theoretically important for studying the stability of the singular distributed parameter systems.

## Acknowledgement

This work was supported by the National Nature Science Foundation of China (.60674018); and supported by the Natural Sciences Research Foundation of Department of Education of Jiangsu Province in China (08KJD510003).

## References

1. Joder, L., Fernandez, M.L.: An Implicit Difference Method for the Numerical Solution of Coupled System of Partial Differential Equations. *Appl. Math. Comput.* 46, 127–134 (1991)
2. Lewis, F.L.: A Review of 2-D Implicit Systems. *Automatic* 28, 345–354 (1992)
3. Trzaska, Z., Marszalek, W.: Singular Distributed Parameter Systems. *IEEE Control Theory and Applications* 40, 305–308 (1993)
4. Ge, Z.Q., Zhu, G.T., Ma, Y.H.: Pole Assignment for the First Order Coupled Singular Control System. *Control Theory and Application* 17, 379–383 (2000) (in Chinese)
5. Wang, K., Lv, T., Zou, Z.: On the Pole Assignment for the Distributed Parameter System. *Science in China Series A* 12, 172–184 (1982) (in Chinese)
6. Ge, Z.Q.: Inverse Problem of Operators and its Applications. Shaanxi Scientific and Technological Press, Shaanxi (1993) (in Chinese)

7. Ge, Z.Q., Ma, Y.H.: Pole Assignment of the First Order Singular Distributed Parameter System. *Chinese Annals of Mathematics* 22, 729–734 (2001) (in Chinese)
8. Shi, G.D., Liu, F.: Stability by Feedback of the first Order Singular Distributed Parameter Control Systems Containing Infinite Non-negative Real Part Poles. In: The Seventh Asian Control Conference, pp. 1416–1419. IEEE Press, New York (2009)
9. Liu, F., Shi, G.D., Ge, Z.Q.: Asymptotical Stability for Singular Control Systems with Minimized Norm State Feedback. In: The 3rd IEEE International Conference on Computer Science and Information Technology, pp. 267–371. IEEE Press, New York (2009)
10. Liu, F., Shi, G.D., Ge, Z.Q.: On the Synthesis of the First Order Singular Distributed Parameter Control Systems with Multi-observers. *Acta Mathematicae Applicatae Sinica* (in press)
11. Liu, F., Shi, G.D.: Uniform Exponential Stability of the Time Varying Singular Distributed Parameter Systems in Hilbert Space. In: The 29th Chinese Control Conference. IEEE Press, New York (2010) (in press)
12. Ge, Z.Q., Zhu, G.T., Feng, D.X.: On the Stabilization by Feedback of the Second Order Singular Distributed Parameter Control Systems. *Dynamics of Continuous, Discrete and Impulsive Systems Series B-Applications & Algorithms* 13, 116–120 (2006)
13. Liu, F., Ge, Z.Q.: Feedback Control and Pole Assignment for the Second Order Singular Control Systems. *Advances in Mathematics* 37, 683–689 (2008) (in Chinese)
14. Liu, F., Ge, Z.Q.: Feedback Control and Pole Assignment for a Class of the Second Order Singular Control Systems in Hilbert Space. *Journal of Systems Science and Mathematical Science* 28, 257–264 (2008) (in Chinese)
15. Liu, F., Li, F.M.: Pole Assignment of Infinite Many Poles for a Class of Coupled Singular Systems in Hilbert Space. *Dynamics of Continuous, Discrete and Impulsive Systems, Series B: Applications and Algorithms* 14, 1356–1358 (2007)

# Mining Generalized Association Rules with Quantitative Data under Multiple Support Constraints

Yeong-Chyi Lee<sup>1</sup>, Tzung-Pei Hong<sup>2,3</sup>, and Chun-Hao Chen<sup>4</sup>

<sup>1</sup> Department of Information Management, Cheng Shiu University, Taiwan

<sup>2</sup> Department of Science and Information Engineering, National University of Kaohsiung, Taiwan

<sup>3</sup> Department of Computer Science and Engineering, National Sun Yat-sen University, Taiwan

<sup>4</sup> Department of Computer Science and Engineering, Tamkang University, Taiwan

yeongchyi@csu.edu.tw, tphong@nuk.edu.tw, chchen@mail.tku.edu.tw

**Abstract.** In this paper, we introduce a fuzzy mining algorithm for discovering generalized association rules with multiple supports of items for extracting implicit knowledge from quantitative transaction data. The proposed algorithm first adopts the fuzzy-set concept to transform quantitative values in transactions into linguistic terms. Besides, each primitive item is given its respective predefined support threshold. The minimum support for an item at a higher taxonomic concept is set as the minimum of the minimum supports of the items belonging to it and the minimum support for an itemset is set as the maximum of the minimum supports of the items contained in the itemset. An example is also given to demonstrate that the proposed mining algorithm can derive the generalized association rules under multiple minimum supports in a simple and effective way.

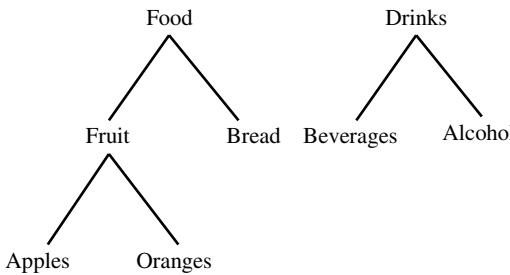
**Keywords:** Fuzzy data mining, generalized association rule, quantitative value, taxonomy, multiple minimum supports.

## 1 Introduction

A variety of mining approaches based on the *Apriori* algorithm [1] were proposed, each for a specific problem domain, a specific data type, or for improving its efficiency [2][5][14][15]. In real applications, different items may have different criteria to judge its importance. Liu *et al.* thus proposed an approach for mining association rules with non-uniform minimum support values [12]. Their approach allowed users to specify different minimum supports to different items. They defined the minimum support value of an itemset as the lowest minimum support among all the items in the itemset. Wang *et al.* proposed a mining approach, which allowed the minimum support value of an itemset to be any function of the minimum support values of items contained in the itemset [17]. Lee *et al.* introduced the mining algorithms based on the *Apriori* approach to generate large itemsets under multiple minimum supports using the maximum constraints [9].

Mining association rules with multiple concept-levels may lead to discovery of more general and important knowledge from data. Relevant data item taxonomies are

usually predefined in real-world applications and can be represented using hierarchy trees. Terminal nodes on the trees represent actual items appearing in transactions; internal nodes represent classes or concepts formed by lower-level nodes. A simple example is given in Fig. 1. For example, “Apples” and “Oranges” are two kinds of “Fruit”. “Fruit” is thus a higher level of concept than “Apples” or “Oranges”. Discovering generalized association rules at different levels may thus provide more information than that only at a single level [4][16].



**Fig. 1.** An example of the given taxonomies

Since transactions in real-world applications usually consist of quantitative values, many approaches have also been proposed for mining fuzzy association rules with taxonomy by using uniform minimum support [6][7][8] or non-uniform minimum support [10][11].

In uniform minimum support mining approaches, firstly, Lee et al. have proposed a mining approach with taxonomy under uniform minimum support in 2001 [6]. The method uses fuzzy concept hierarchies for categorical attributes and generalization hierarchies of fuzzy linguistic terms for quantitative attributes. Then, Hong et al. have proposed an approach, called ML-FRMA, which adopts a top-down progressively deepening approach to finding large itemsets [6]. Each item uses only the linguistic term with the maximum cardinality in later mining processes, thus making the number of fuzzy regions to be processed the same as the number of original items. Based on Hong’s approach, Kaya et al. are proposed an approach, namely wightedFRMA, for mining fuzzy weighted multi-cross-level association rule [7]. Lee et al. proposed a fuzzy multiple-level mining algorithm for extracting knowledge implicit in quantitative transactions with multiple minimum supports of items [11]. Items may have different minimum supports and the maximum-itemset minimum-taxonomy support constraint is adopted to discover the large itemsets.

This paper thus introduces a fuzzy mining algorithm for discovering generalized association rules with multiple supports of items for extracting implicit knowledge from transactions stored as quantitative values. The remaining parts of this paper are organized as follows. The proposed fuzzy mining algorithm for discovering generalized association rules from quantitative data under multiple minimum supports is described in Section 2. An example to illustrate the proposed algorithm is then given in Section 3. Finally, conclusion is given in Section 4.

## 2 The Proposed Algorithm

In the proposed algorithm, the maximum-itemset minimum-taxonomy support constraint is adopted in finding large itemsets. Details of the proposed fuzzy mining algorithm are stated below.

### *The fuzzy mining algorithm for generalized association rules with quantitative data under multiple support constraints*

INPUT: A body of quantitative transaction data  $D$  with  $n$  transactions, a set of items  $I = \{I_1, I_2, \dots, I_m\}$  purchased in  $D$ , a set of predefined taxonomies  $H$  defined on  $I$ , a set of predefined minimum supports  $\tau_j$  on items  $I_j \in I$ , a set of membership functions, a predefined minimum confidence  $\lambda$ .

OUTPUT: A set of generalized fuzzy association rules with multiple minimum supports under the given constraint.

STEP 1: Extend each transaction in the transaction data  $D$  by adding the ancestors of appearing items to transactions and calculate their quantities. Duplicate items in each transaction are then removed. Denote the expanded transaction data as  $D'$ .

STEP 2: Transform the quantitative value  $v_{ij}$  of each transaction datum  $T_i$  ( $i=1$  to  $n$ ), for each expanded item name  $I_j$  appearing into a fuzzy set  $f_{ij}$  represented as  $\left( \frac{f_{ij1}}{R_{j1}} + \frac{f_{ij2}}{R_{j2}} + \dots + \frac{f_{ijh}}{R_{jh}} \right)$  using the given membership functions, where  $h$  is the number of fuzzy regions for  $I_j$ ,  $R_{jl}$  is the  $l$ -th fuzzy region of  $I_j$ ,  $1 \leq l \leq h$ , and  $f_{ijl}$  is  $v_{ij}$ 's fuzzy membership value in region  $R_{jl}$ .

STEP 3: Calculate the scalar cardinality of each fuzzy region  $R_{jl}$  in the transaction data as:

$$c_{jl} = \sum_{j=1}^n f_{ijl}.$$

STEP 4: Find  $count_j = \max_{l=1}^h (c_{jl})$ , for  $j = 1$  to  $m$ , where  $m$  is the number of items in  $T$ . Let  $R_j$  be the region with  $count_j$  for item  $I_j$ , which will be used to represent the fuzzy characteristic of item  $I_j$  in later mining processes.

STEP 5: Check whether the fuzzy count value  $count_j$  of a region  $R_j$ ,  $j = 1$  to  $m$ , is larger than or equal to the respectively predefined minimum support  $\tau_j$ . Notice that the minimum support  $\tau_a$  of an ancestor item  $I_a$  is set to the minimum of minimum supports of the items belonged to it. That is,

$$\tau_a = \min_{I_j \in I_a} (\tau_j).$$

If a region  $R_j$  is equal to or greater than the minimum support value, put it in the large 1-itemsets ( $L_1$ ). That is,

$$L_1 = \{R_j \mid count_j \geq \tau_j, 1 \leq j \leq m\}.$$

STEP 6: If  $L_1$  is null, then exit the algorithm; otherwise, set  $k = 2$ , where  $k$  represents the number of items stored in the current large itemsets, and do the next step.

STEP 7: Both of the fuzzy count values  $count_j$  of regions in  $C_2$  have to larger than or equal to the maximum of the minimum support of each item of  $C_2$  (Lemma 2). All the

possible 2-itemsets are collected as  $C_2$ . Generate the candidate set  $C_2$  from  $L_I$ . Each 2-itemset in  $C_2$  must not include items with ancestor or descendant relation in the taxonomy.

STEP 8: For each newly formed candidate 2-itemset  $s$  with items  $(s_1, s_2)$  in  $C_2$ :

- (a) Calculate the fuzzy value of  $s$  in each transaction datum  $D_i$  as:  $f_{is} = f_{is_1} \wedge f_{is_2}$ , where  $f_{is_j}$  is the membership value of  $D_i$  in region  $s_j$ . If the minimum operator is used for the intersection, then  $f_{is} = \min(f_{is_1}, f_{is_2})$ .
- (b) Calculate the scalar cardinality of  $s$  in the transaction data as:

$$count_s = \sum_{i=1}^n f_{is}.$$

- (c) If  $count_s$  is larger than or equals to the maximum of the minimum supports  $\tau_j$  of the items contained in it, put  $s$  in  $L_2$ .

STEP 9: IF  $L_2$  is null, then exit the algorithm; otherwise, set  $k = k + 1$  and do the next step.

STEP 10: Generate candidate  $k$ -itemset  $C_k$  from  $L_{k-1}$  in the similar way to that in the *Apriori* algorithm [1]. That is, the algorithm first joins  $L_{k-1}$  and  $L_{k-1}$  assuming that  $k-1$  items of the two itemsets are the same and the other one is different. In addition, it is different from the *Apriori* algorithm that the supports of all the large  $(k-1)$ -itemsets comprising a candidate  $k$ -itemset  $I$  must be larger than or equal to the maximum of the minimum supports of these large  $(k-1)$ -itemsets (Lemma 2). Store in  $C_k$  all the itemsets satisfying the above conditions and with all  $(k-1)$ -itemsets in  $L_{k-1}$ .

STEP 11: If  $C_k$  is null, then go to STEP 15; otherwise, go to next step.

STEP 12: Remove the ancestor items that are not contained in any itemset in  $C_k$  from the expanded transaction data  $D'$ .

STEP 13: Do the following substeps for each newly formed candidate  $k$ -itemset  $s$  with items  $(s_1, s_2, \dots, s_k)$  in  $C_k$ .

- (a) Calculate the fuzzy value of  $s$  in each transaction datum  $D_i$  as:

$$f_{is} = f_{is_1} \wedge f_{is_2} \wedge \dots \wedge f_{is_k},$$

where  $f_{is_j}$  is the membership value of  $D_i$  in region  $s_j$ . If the minimum operator is used for the intersection, then  $f_{is} = \min_{j=1}^k f_{is_j}$ .

- (b) Calculate the scalar cardinality of  $s$  in the transaction data as:

$$count_s = \sum_{i=1}^n f_{is}.$$

- (c) If  $count_s$  is larger than or equal to the maximum of the minimum supports of all items contained in it and put  $s$  in  $L_k$ . That is,

$$L_k = \{s \in C_k | count_s \geq \max_{s_k \in s} (\tau_{s_k})\}.$$

STEP 14: If  $L_k$  is null, then go to next step; otherwise, set  $k = k + 1$  and go to step 10.

STEP 15: Construct the association rules for each large  $q$ -itemset  $s$  with items  $(s_1, s_2, \dots, s_q)$ ,  $q \geq 2$ , by the following substeps:

- (a) Form all possible association rules as follows:

$$s_1 \wedge \dots \wedge s_{r-1} \wedge s_r \rightarrow s_r, r = 1 \text{ to } q.$$

- (b) Calculate the confidence values of all association rules by:

$$\frac{\sum_{i=1}^n f_{is}}{\sum_{i=1}^n (f_{is_1} \wedge \dots \wedge f_{is_{r-1}} \wedge f_{is_r} \wedge \dots \wedge f_{is_q})}.$$

STEP 16: Output the rules with confidence values larger than or equal to the predefined confidence value  $\lambda$ .

### 3 An Example

Assume the quantitative transaction dataset includes the six transactions shown in Table 1. Each transaction is presented by a couple of fields, *TID* and *Items*. The field *TID* is used for identifying transactions and the field *Items* lists the items purchased at a transaction. Assume the predefined taxonomy among the items is shown in Fig. 1. Also assume that the predefined minimum support values of items are given in Table 2, and the minimum confidence value is set at 0.7. For convenience, simple symbols are used to represent items and groups the taxonomies in Fig. 1.

**Table 1.** Transaction data set of this example

<i>TID</i>	<i>Items</i>
1	(Apple, 3) (Bread, 4) (Alcohol, 2)
2	(Orange, 3) (Bread, 7) (Beverage, 7)
3	(Orange, 2) (Bread, 10) (Alcohol, 5)
4	(Bread, 9) (Alcohol, 10)
5	(Apple, 7) (Beverage, 8)
6	(Orange, 2) (Bread, 8) (Beverage, 10)

Symbol *A*, *B*, *C*, *E* and *F* represent the primitive items, *Apple*, *Orange*, *Bread*, *Beverage* and *Alcohol*. Symbol *T<sub>1</sub>*, *T<sub>2</sub>* and *T<sub>3</sub>* stands for items in higher concept levels of the taxonomies, *Fruit*, *Food* and *Drink*. Also assume that the fuzzy membership functions are the same for all the items and are as shown in Fig. 2.

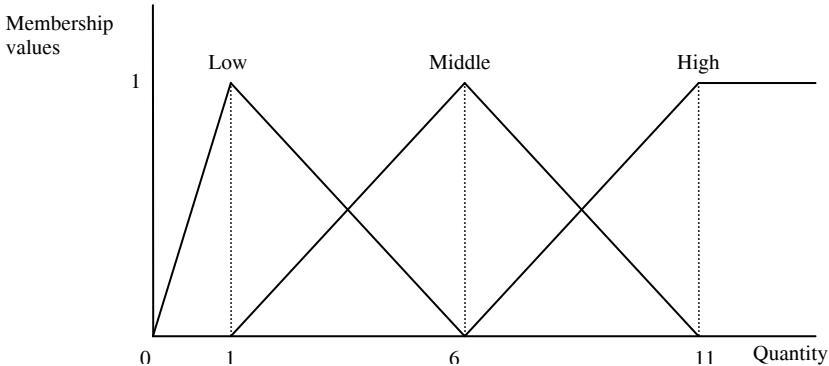
**Table 2.** The predefined minimum supports for items

<i>Item</i>	<i>Apple</i>	<i>Orange</i>	<i>Bread</i>	<i>Beverage</i>	<i>Alcohol</i>
<i>Min-support</i>	1.5	2.1	2.3	1.5	1.8

In this example, amounts are represented by three fuzzy regions: *Low*, *Middle* and *High*. Thus, three fuzzy membership values are produced for each item amount

according to the predefined membership functions. For the transaction data in Table 1, the proposed fuzzy generalized mining algorithm deriving fuzzy generalized association rules is described as following.

The ancestors of appearing items are first added to transactions. Take the items in the first transaction as an example. Since  $T_1$  is the ancestors of  $A$ ,  $(T_1, 3)$  is thus added to the first transaction. Also, since  $T_2$  is the common ancestor of  $A$  and  $C$ ,  $(T_2, 7)$  is added. Similarly,  $(T_3, 2)$  is added.



**Fig. 2.** The membership functions used in this example

The quantitative values of the items are represented using fuzzy sets. Take the first item in transaction 4 as an example. The amount 9 of  $C$  is converted into the fuzzy set  $\left( \frac{0.0}{Low} + \frac{0.4}{Middle} + \frac{0.6}{High} \right)$  using the given membership functions. The scalar cardinality of each fuzzy region in the transactions is then calculated as the *count* value. For instance, the scalar cardinality of fuzzy region  $C.High$  can be calculated as  $(0.0 + 0.2 + 0.8 + 0.6 + 0.0 + 0.4) = 2.0$ . The fuzzy region with the highest count among the three possible regions for each item is then found. Take item  $A$  as an example. Since the count for *Middle* is the highest among the three counts, the region *Middle* is thus used to represent the item  $A$ . This step is repeated for the other items. Thus, "Low" is chosen for  $B$  and  $T_1$ , "Middle" is chosen for  $C, D, E$  and  $T_3$ , and "High" is chosen  $T_2$ .

The *count* values of fuzzy regions then checked against with its predefined minimum support value shown in Table 2. Since the *count* values of fuzzy regions  $B.Low$ ,  $C.Middle$  and  $D.Middle$  are both respectively larger than or equal to their predefined minimum supports, these items are then put in the large 1-itemsets  $L_1$ . In the other hand,  $T_1.Low$  is put in  $L_1$  as well, since its *count* value 2.8 is larger than its minimum support 1.5, which is set at the minimum of the minimum supports of items belonged to it. Therefore, the fuzzy regions are put in large 1-itemset  $L_1 = \{B.Low, C.Middle, D.Middle, T_1.Low, T_2.High, T_3.Middle\}$ .

The candidate set  $C_2$  is then generated from  $L_1$ . An itemset is not put in  $C_2$  when it contains any fuzzy region whose *count* value is less than the maximum of minimum supports among items containing in it. It is not possible to become a large itemset in the following processes. Take the itemset  $(B.Low, C.Middle)$  as an example. This itemset is pruned, since the *count* value of fuzzy region  $B.Low$ , which is smaller than

the maximum of minimum supports of item  $B$  and  $C$ . Additionally, itemsets with an item and its ancestor are pruned. Take itemset  $(B.Low, T_1.Low)$  as an example. Since item  $T_1$  is an ancestor of item  $B$  in the predefined taxonomy, itemset  $(B.Low, T_1.Low)$  is thus pruned.

The fuzzy membership values of each transaction data for the candidate 2-itemsets are calculated. Take  $(B.Low, D.Middle)$  as an example. The derived membership value for the transaction 2 is calculated as:  $\min(0.6, 0.8)=0.6$ . The results for the other transactions are shown in Table 3. The results for the other 2-itemsets can be derived in same fashion. After the *count* value of each candidate 2-itemset in  $C_2$  is calculated. Since the *count* values of  $(B.Low, T_3.Middle)$ ,  $(T_1.Low, T_3.Middle)$  and  $(T_2.High, T_3.Middle)$  are larger than the maximum of both minimum supports of items contain in them, they are thus stored in  $L_2$ .

**Table 3.** The membership values for  $(B.Low, D.Middle)$

<i>TID</i>	<i>B.Low</i>	<i>D.Middle</i>	<i>B.Low</i>	<i>D.Middle</i>
	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0
2	0.6	0.8	0.6	0.6
3	0.8	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0
5	0.0	0.6	0.0	0.0
6	0.8	0.2	0.2	0.2

Since there is no candidate 3-itemset  $C_3$  generated from  $L_2$ , the confidence values of all possible association rules are then calculated. Take “If  $B = Low$ , then  $T_3 = Middle$ .” as an example. The *count* value of  $B.Low \cap T_3.Middle$  is 1.6 and the *count* value of  $B.Low$  is 2.2. The confidence value for the association rule is calculated as 0.73. The confidence values of the possible association rules are then checked against the predefined confidence threshold  $\lambda= 0.7$ . There are two rules are thus kept: “If  $B = Low$ , then  $T_3 = Middle$ , with a confidence value of 0.73” and “If  $T_3 = Middle$ , then  $T_2 = High$ , with a confidence value of 0.86.”

## 4 Conclusion

In this paper, a mining algorithm for discovery of fuzzy generalized association rules from quantitative transactions with multiple minimum supports is proposed. Different items in transactions are given respective minimum supports for user’s support requirements to items. The proposed algorithm meets the mining issues of generating association rules under multiple minimum supports and managing taxonomic relationships among items that usually occur in real mining applications while adopting fuzzy set theory to deal with extracting knowledge from quantitative data. This algorithm handles the minimum support for an item at a higher taxonomic concept (ancestor items) by setting as the minimum of the minimum supports of the items belonging to it. Besides, the minimum support for an itemset is set as the maximum of the minimum supports of the items contained in the itemset. An example is also given to demonstrate that the proposed mining algorithm can derive the generalized association rules under multiple minimum supports in a simple and effective way.

**Acknowledgments.** This research was supported by the National Science Council of the Republic of China under contract NSC 98-2622-E-230 -009 -CC3.

## References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large database. In: The ACM SIGMOD Conference, Washington DC, USA, pp. 207–216 (1993)
2. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market-basket data. In: ACM-SIGMOD International Conference in Management of Data, pp. 207–216 (1997)
3. de Campos, L.M., Moral, S.: Learning rules for a fuzzy inference model. *Fuzzy Sets and Systems* 59, 247–257 (1993)
4. Han, J., Fu, Y.: Discovery of multiple-level association rules from large databases. In: The 21st International Conference on Very Large Data Bases, pp. 420–431 (1995)
5. Hong, T.P., Kuo, C.S., Chi, S.C.: A data mining algorithm for transaction data with quantitative values. *Intelligent Data Analysis* 3(5), 363–376 (1999)
6. Hong, T.P., Lin, K.Y., Chien, B.C.: Mining Fuzzy Multiple-Level Association Rules from Quantitative Data. *Applied Intelligence* 18(1), 79–90 (2003)
7. Kaya, M., Alhajj, R.: Effective Mining of Fuzzy Multi-Cross-Level Weighted Association Rules. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 399–408. Springer, Heidelberg (2006)
8. Lee, K.M.: Mining generalized fuzzy quantitative association rules with fuzzy generalization hierarchies. In: IFSA World Congress and 20th NAFIPS International Conference, vol. 5, pp. 2977–2982 (2001)
9. Lee, Y.C., Hong, T.P., Lin, W.Y.: “ Mining association rules with multiple minimum supports using maximum constraints. *International Journal of Approximate Reasoning* 40(1), 44–54 (2005)
10. Lee, Y.C., Hong, T.P., Wang, T.C.: Mining Multiple-Level Association Rules under the Maximum Constraint of Multiple Minimum Supports. In: IEA/AIE, pp. 1329–1338 (2006)
11. Lee, Y.C., Hong, T.P., Wang, T.C.: Multi-level fuzzy mining with multiple minimum supports. *Expert Systems with Applications* 34(1), 459–468 (2008)
12. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: 1999 International Conference on Knowledge Discovery and Data Mining, pp. 337–341 (1999)
13. Lui, C.L., Chung, F.L.: Discovery of generalized association rules with multiple minimum supports. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 510–515. Springer, Heidelberg (2000)
14. Park, J.S., Chen, M.S., Yu, P.S.: An Effective hash-based algorithm for mining association rules. In: Proceedings of the 1995 ACM-SIGMOD International Conference in Management of Data, pp. 175–186 (1995)
15. Savasere, A., Omiecinski, E., Navathe, S.: An efficient algorithm for mining association rules in large databases. In: The 21st International Conference in Very Large Data Bases (VLDB 1995), pp. 432–443 (1995)
16. Srikant, R., Agrawal, R.: Mining generalized association rules. In: The 21st International Conference on Very Large Data Bases, pp. 407–419 (1995)
17. Wang, K., He, Y., Han, J.: Mining frequent itemsets using support constraints. In: The 26th International Conference on Very Large Data Bases, pp. 43–52 (2000)

# Comparison of Edge Segmentation Methods to Tongue Diagnosis in Traditional Chinese Medicine

Chieh-Hsuan Wang, Ching-Chuan Wei, and Che-Hao Li

Department of Information and Communication Engineering,  
Chaoyang University of Technology, Taichung, Taiwan  
[ccwei@cyut.edu.tw](mailto:ccwei@cyut.edu.tw)

**Abstract.** In Traditional Chinese Medicine, the human tongue is one of the important organs which contain the information of health status. In order to achieve an automatic tongue diagnostic system, an effective segmentation method for detecting the edge of tongue is very important. We mainly compare the Canny, Snake and threshold (Otsu's thresholding algorithm) methods for edge segmentation. The segmentation using Canny algorithm may produce many false edges after cutting; thus, it is not suitable for use. The Snake segmentation in tongue requires a larger convergence number and spends too much time. The threshold method using Otsu's thresholding algorithm and filtering process can achieve an easy, fast and effective segmentation result in tongue diagnosis. Therefore, it may be useful in clinical automated tongue diagnosis system.

**Keywords:** Tongue Diagnosis, Traditional Chinese Medicine, Edge Segmentation Method.

## 1 Introduction

Traditional Chinese Medicine (TCM) has a history of thousands of years and its practitioners have accumulated very rich practical experiences in diagnostic methods. Tongue diagnosis is one of the most important and widely used diagnostic methods in Chinese medicine. This method is also very valuable in clinical applications and self-diagnosis [1]. The human tongue is one of the important organs which contain the health status and physical information. The tongue's information can be divided into tongue color, tongue size, the distribution of tongue coating, etc. To achieve the demand for automation and flexible application of the collection of information in hospitals and pathological analysis, the automatically segmenting the tongue is very important, because the color of the tongue is one of the most important element.

Nowadays, the rapid progress of information technology promotes the automatization of tongue disease diagnosis based on modern image processing and pattern recognition approaches. There has been some work on computerized tongue diagnosis, and many issues of standardization and quantification have been resolved yet there are still some problems [2]. The major problem is: First, the convergence time is too

long, or not have a nice value of convergence to the edge. Second, the tongue coating coverage in tongue body can be divided into white, yellow, black, etc. These different colors can be used for the separation of the tongue coating. However, due to the instability after light correction, it may produce different results.

The paper is organized as follows: Section 2 will introduce a variety of segmentation methods. Section 3 make a comparison between these methods to tongue diagnosis in traditional Chinese medicine. Section 4 gives a conclusion.

## 2 Edge Segmentation Methods to Tongue Diagnosis

Tongue images can be captured using a specific set of image acquisition devices, including advanced camera and other corresponding lighting system. In the classification process, tongue must be extracted from the image region. However, the tongue image includes lips, skin or teeth. Therefore, when using a variety of segmentation methods, most common error is to generate segmentation

Nowadays, there are many ways for segmentation edge. However, the tongue is very difficult to segmentation. The main reason is the similarity of color between tongue and skin. The coupled light source is not stable; hence, making segmentation becomes more difficult.

### 2.1 Canny Algorithm

In 1986, Canny edge detection operator was proposed on optimization algorithms for edge detection. Relatively simple algorithm to make the whole process effectively is executed and has been widely used, but Canny operator has the defect that being vulnerable to various noise disturbances. Thus, there are certain limitations of its concrete application [3].

#### 2.1.1 The Principle of Traditional Canny Algorithm

Canny algorithm for edge detection includes three criterions.

##### 1. The criterion of SNR

The SNR is defined as follow:

$$\text{SNR} = \frac{\left| \int_{-W}^{+W} G(-x)h(x)dx \right|}{\sigma \sqrt{\int_{-W}^{+W} h^2(x)dx}} \quad (1)$$

Where the  $G(x)$  represents the edge function,  $h(x)$  represents the impulse response of the filter of width  $W$ .  $\sigma$  represents the mean square deviation of the Gaussian noise [4].

##### 2. The criterion of positioning accuracy

The positioning accuracy of the edge is defined as follows:

$$\text{Localization} = \frac{\left| \int_{-W}^{+W} G'(-x)h'(x)dx \right|}{\sigma \sqrt{\int_{-W}^{+W} h'^2(x)dx}} \quad (2)$$

Where the  $G'(x)$  and  $h'(x)$  is respectively the derivative of  $G(x)$  and  $h(x)$ . The larger of the positioning accuracy, the result is better.

### 3. The criterion of the singleness edge response

This criterion should meet the following formula:

$$D(f') = \pi \left( \frac{\int_{-\infty}^{+\infty} h^2(x) dx}{\int_{-W}^{+W} h''^2(x) dx} \right)^{\frac{1}{2}} \quad (3)$$

$D(f')$  is the average distance and  $h''(x)$  is the second derivative of  $h(x)$ .

### 2.1.2 The Detecting Process of the Canny Algorithm

The detecting process of the Canny algorithm consists of the following steps[5]:

1. Use the Gaussian filter smoothing image to restrain noise.
2. Calculating the gradient magnitude  $M(x, y)$  and the gradient direction  $H(x, y)$  of the image  $M(x, y)$  is defined as follows:

$$M(x, y) = \sqrt{E_x(x, y)^2 + E_y(x, y)^2} \quad (4)$$

The  $H(x, y)$  is defined as follow:

$$H(x, y) = \arctan \left( \frac{E_x(x, y)}{E_y(x, y)} \right) \quad (5)$$

$E_x$  and  $E_y$  is the result what the image being effected by the filter along the row-column direction.

3. Do non-maximum suppression for the gradient magnitude.
4. Dual-Threshold algorithm is adopted to detect and connect edges.

The main defects of the traditional Canny algorithm are the usage of Gaussian filter. When smooth the noise, some edge is also smoothed. Besides, the detection results have some isolated edges and some false edges.

## 2.2 Active Contour Models (Snakes)

In 1987, Kass et al., introduced the Active Contour Models commonly known as Snakes. Active contour is defined as an energy minimization spline. Its energy depends on its shape and location within the image. Snake can be considered as a number of control points or snakes are linked and free to deform under the constraining forces. The control points  $v(s) = [x(s), y(s)]$  are traditionally placed near the edges of interest because of the poor capture range of a snake [6].

Snake deformation is carried by minimization of an energy function so that the contour will move from the initial position until the energy will stabilize at significant edges. This is to ensure a better performance in edge detection. If the points are too close or too far, the mechanism must delete or add points, respectively, in order to

have a good initial position. Deformation in each step, the process of checking the initial coordinates should be done to ensure better performance of a Snake.

The snake energy is defined as[7]:

$$E_{\text{snake}} = \int_0^1 [E_{\text{int}}(v(s)) + E_{\text{ext}}(v(s))] ds \quad (6)$$

Where  $E_{\text{snake}}$  represents the total energy,  $E_{\text{int}}(v(s))$  represents the internal energy of the spline due to bending,  $E_{\text{image}}(v(s))$  gives rise to the image forces, and  $E_{\text{con}}(v(s))$  gives rise to the external constraint forces.

This is easily done by humans but is not an easy task for a system to intuitively realize. However, if there is enough computer power, we can let it randomly pick up all the edges in an image and then select the required ones. Producing a set of appropriate parameters for snake initialization is another issue. However, after initializing of the snake, it is difficult to have a system to automatically change the parameters. If an object has an edge in the image which shows a concavity, the traditional snake have a tracing problem of concave edge of the part. The reason is that the capture range at these parts of edges is too far from the snake.

### 2.3 Threshold Method Using Otsu's Thresholding Algorithm

Thresholding is the simplest method of image segmentation. From the gray-scale image, the threshold can be created by binary images. It is important in image processing to select an appropriate threshold of gray-level for extracting objects from their background. Histogram has a deep and sharp valley between two peaks representing the objects and background, respectively. So, we can through the histogram select what we are interested in either objects or background image.

The key parameter in the thresholding process is the choice of the threshold value. Users can manually choose a threshold value, or a thresholding algorithm can compute a value automatically. This paper uses threshold method and applies Otsu's method. Otsu's method is used to automatically perform histogram shape-based image thresholding, or the reduction of a gray-level image to a binary image. Otsu's method will exhaustively search for the threshold that minimizes the intra-class variance, defined as a weighted sum of variances of the two classes:

$$\sigma_w^2(t) = \omega_1(t)\sigma_1^2(t) + \omega_2(t)\sigma_2^2(t) \quad (7)$$

Weights  $\omega_i$  are the probabilities of the two classes separated by a threshold  $t$  and  $\sigma_i^2$  are the variances of these classes.

Otsu shows that minimizing the intra-class variance is the same as maximizing inter-class variance[8]:

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = \omega_1(t)\omega_2(t)[\mu_1(t) - \mu_2(t)]^2 \quad (8)$$

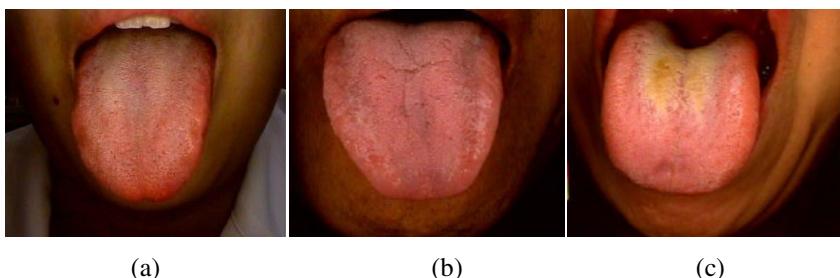
which is expressed in terms of class probabilities  $\omega_i$  and class means  $\mu_i$  which in turn can be updated iteratively.

Otsu's Algorithm step is as follows:

1. Compute histogram and probabilities of each intensity level.
2. Set up initial  $\omega_i(0)$  and  $\mu_i(0)$ .
3. Step through all possible thresholds  $t = 1 \dots$  maximum intensity
  - (a) Update  $\omega_i$  and  $\mu_i$ .
  - (b) Compute  $\sigma_b^2(t)$ .
4. Desired threshold corresponds to the maximum  $\sigma_b^2(t)$ .

### 3 Comparison of These Methods to Tongue Diagnosis

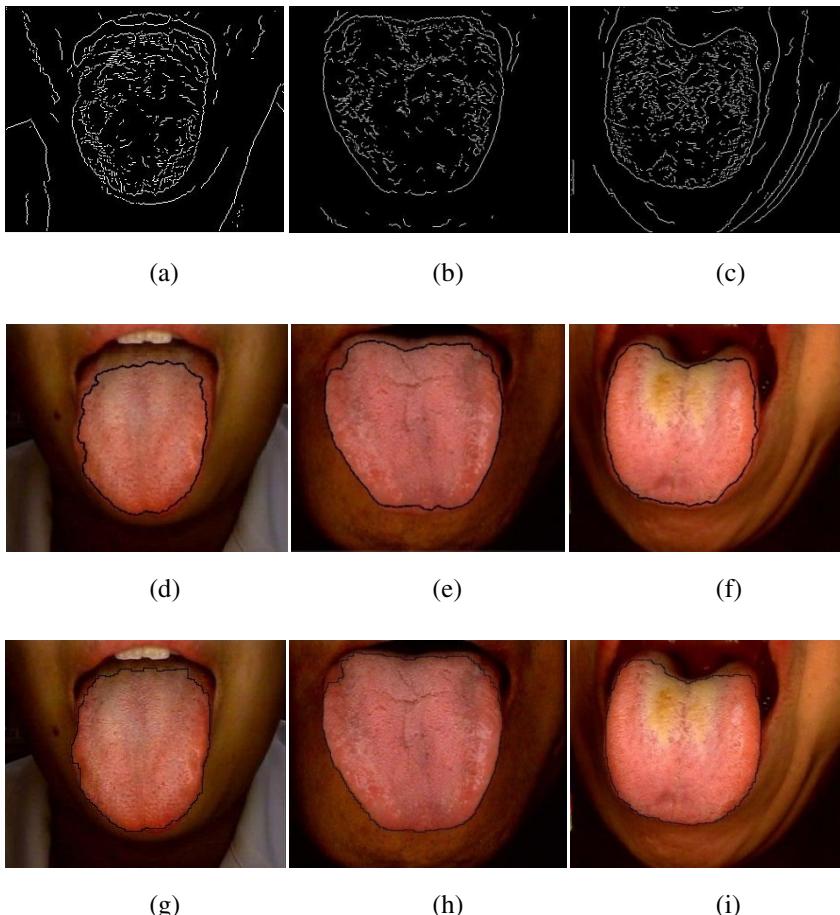
The use of Otsu's method automatically calculates the value and the threshold method, and then Gaussian smoothing filter removes the noise. Comparison of the above methods is listed in this section. We then conduct the following segmentation test of three original images in Fig. 1.



**Fig. 1.** The original tongue images are demonstrated in (a), (b) and (c)

The results of tongue edge segmentation using Canny algorithm are shown in (a)-(c), and the  $\sigma$  is set to be 3. The execution time is respectively for the 0.3, 0.2 and 0.3 seconds in MATLAB 7.9. and LabView 8.5 with dual Intel cores, 2.2G RAM and 1.80 GHz PC. The results using Snake's algorithm are shown in (d)-(f). The execution time is respectively for the 52.0, 105.5 and 91.7 seconds. The results using threshold and Otsu's algorithm are shown in (g)-(i),., and the  $\sigma_b^2(t)$  value is 0.2588, 0.3137 and 0.3333. The execution time is respectively for 0.2, 0.2 and 0.21 seconds.

Finally, the results in Fig. 2(a)-(c) using Canny algorithm can produce more noise, and more false edges. It is not easy to select one edge of the tongue, so is not suitable for use in the tongue image. In Fig. 2(d)-(f) using Snake algorithm is also unsatisfactory in the segmentation. The tip and root of tongue are not the ideal segmentation. The convergence times of this experiment is about 400, and spend too much time. Finally the threshold method with Otsu's algorithm can be achieved automatically and effectively Circle the approximate range. The additional filtering process can achieve good results. In Fig. 2(g)-(i), we can see the tips of tongue segmentation approaching a nice edge, and will not affect the further analysis of the tongue.



**Fig. 2.** The results of tongue edge segmentation using Canny algorithm are shown in (a)-(c). The results using Snake's algorithm are shown in (d)-(f). The results using threshold algorithm are shown in (g)-(i).

## 4 Conclusion

Using Threshold and Otsu's Algorithm, we can effectively segment the tongue without affecting the integrity of the further tongue diagnosis. This is the first step to establish a automated tongue diagnosis system, and will improve the scientific representation of tongue diagnosis in traditional Chinese medicine. In the future, we hope to continue to use this effective method for the separation between tongue and tongue coating, which is useful in the realization of physiological and pathological status within human body from the viewpoint of traditional Chinese medicine.

## References

- [1] Pang, B., Zhang, D., Wang, K.: The bi-elliptical deformable contour and its application to automated tongue segmentation in Chinese medicine. *IEEE Transactions On Medical Imaging* 24(8), 946–956 (2005)
- [2] Pang, B., Zhang, D., Wang, K.: Tongue image analysis for appendicitis diagnosis. *Information Sciences* 175, 160–176 (2005)
- [3] Wang, B., Fan, S.: An improved CANNY edge detection algorithm. In: Second International Workshop on Computer Science and Engineering (2009)
- [4] Canny, J.: A Computational Approach to Edge Detection. *IEEE Transactions On Pattern Analysis And Machine Intelligence PAMI-8(6)* (November 1986)
- [5] Dafu, P., Bo, W.: An Improved Canny Algorithm. In: Proceedings of the 27th Chinese Control Conference, pp. 456–459 (2008)
- [6] Dagher, I., El Tom, K.: WaterBalloons A hybrid watershed Balloon Snake segmentation. *Image and Vision Computing* 26, 905–912 (2008)
- [7] Kass, M., Witkin, A., Terzopoulos, D.: Snakes Active Contour Models. *International Journal of Computer Vision*, 321–331 (1988)
- [8] Otsu, N.: A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions On Systems, Man, And Cyber Netics SMC-9(1)* (January 1979)

# The Grading of Prostatic Cancer in Biopsy Image Based on Two Stages Approach

Shao-Kuo Tai<sup>1,\*</sup>, Cheng-Yi Li<sup>1</sup>, Yee-Jee Jan<sup>2</sup>, and Shu-Chuan Lin<sup>2</sup>

<sup>1</sup> Department of Information Management, Chaoyang University of Technology,  
Wufong Township Taichung County, Taiwan, R.O.C.

{sgdai, s9814606}@cyut.edu.tw

Tel.: +886 4 23323000ext. 4748; Fax: +886 4 23742337

<sup>2</sup> Department of Pathology, Taichung Veterans General Hospital,  
Taichung, Taiwan, R.O.C.

{yeejan, sara}@vgthc.gov.tw

**Abstract.** Prostatic biopsies provide the information for the determined diagnosis of prostatic cancer. Computer-aid investigation of biopsies can reduce the loading of pathologists and also inter- and intra-observer variability as well. In this paper, we proposed a two stages approach for prostatic cancer grading according to Gleason grading system. The first stage classifies biopsy images into clusters based on their Skeleton-set (SK-set), so that images in the same cluster consist of the similar two-tone texture. In the second stage, we analyzed the fractal dimension of sub-bands derived from the images of prostatic biopsies. We adopted the Support Vector Machines as the classifier and using the leaving-one-out approach to estimate error rate. The present experimental results demonstrated that 92.1% of accuracy for a set of 1000 pathological prostatic biopsy images.

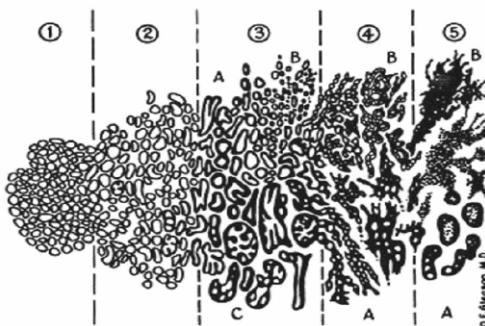
**Keywords:** Gleason grading system, Fractal dimension, Skeleton-set, Prostatic cancer.

## 1 Introduction

Prostate cancers are vary widely detected across the world, with Europe and the United States detecting more frequently than in Asia [1]. There were more than 2 million American men currently living with prostate cancer, and a man dies from prostate cancer every 19 minutes. Although the PSA blood test (prostate-specific antigen) [2] and the digital rectal exam (DRE) can be adopted for screening of prostate cancer. But biopsy is a key step to confirm the diagnosis of malignancy and guiding treatment. By viewing the microscopic image of biopsy, pathologists can determine the histological grade according to the Gleason grading system [3]. Figure 1 is the five basic tissue patterns of the classic Gleason grading diagram. The biopsy Gleason score is a sum of the primary grade that represents the most common tumor and a secondary grade that represents the second most common tumor, and is a number ranging from 2 to 10 [3].

---

\* Corresponding author.



**Fig. 1.** The Gleason grading diagram

Although pathologists will know that how aggressive the cancer is likely to be and how quickly it may spread from the result of Gleason score, human visual grading is time-consuming and very subjective due to variations between inter-observer and intra-observer. Therefore, how to use a computer-aided technique to grade prostatic carcinoma automatically is a topic that one should not ignore. Several methods have been proposed for analyzing pathological images of prostate during the last few years. Stotzka et al. [4] proposed neural network and statistical classification methods to distinguish moderately and poorly differentiated lesions of the prostate. The statistical and structural features are extracted from the spatial distribution of epithelial nuclei over the image area, but the authors have described no algorithm for segmenting the epithelial nuclei [5]. Wetzel et al. [6] proposed methods for content based image retrieval to assist pathology diagnosis. They used Gleason grading of prostate tumor samples as an initial domain for evaluating the effectiveness of the method for specific tasks. Smith et al. [7] proposed a similarity measurement method based on Fourier transform and principle component analysis for Gleason grading of histological slides of prostatic cancer. Jafari-Khouzani et al. proposed a computerized method for grading the pathological images of prostate biopsy samples [8]. In their method, energy and entropy features are calculated from multiwavelet coefficients of an image. These multiwavelet features are tested by using k-nearest-neighbor classifier and leave-one-out approach is used to estimate error rate. Their image set consisted of 100 prostate images of grades 2-5. These methods described above work only when their input images contain only one grade of prostatic cancer, so they cannot detect prostatic cancer from the biopsy images. Huang et al. proposed two feature extraction methods based on fractal dimension to analyze variations of intensity, which called DBC and EBFDE. They achieved 94.1% of CCR evaluated by 5-fold cross-validation for a set of 205 prostate images [9].

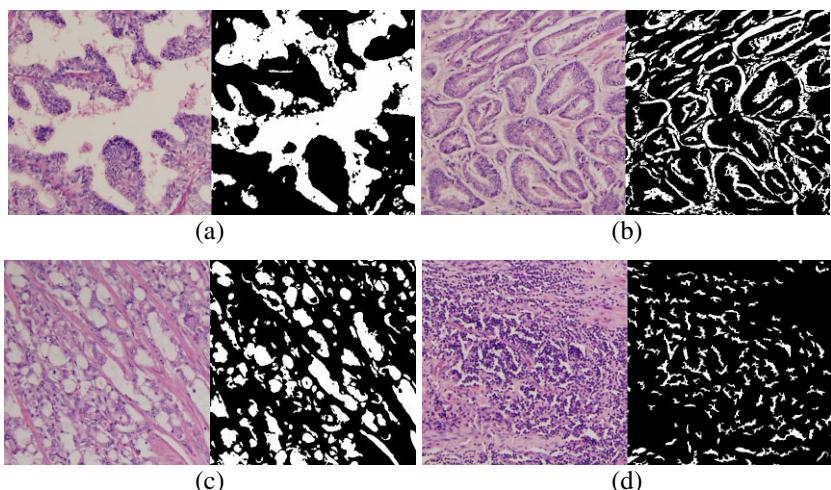
The goal of this paper is to propose an automated system to detect prostatic cancer from biopsy images and provide the grade of them according to Gleason grading system. Due to Gleason grading system mainly based on the texture formed by gland and the lumens are the main part of the gland which exhibit white shapes in the biopsy image. So that texture of the lumen is very important for the grading. Beside lumens, the arrangement and intensity of epithelial cells also have a considerable impact on grading. For example, if these epithelial cells appear in image randomly

then it likely to be classified as grade 5. In this paper we consider these two factors individually. First, we apply the features extracted from Skeleton-set of the lumens to classify these images into clusters. Images belong to the same cluster have similar texture pattern of lumen. Consequently features in respect of the arrangement of epithelial nuclei extracted from images of the same cluster which have rule out the effect of the texture of the lumen. Therefore, in the second stage, we extracted texture feature from gray images to train classifier for each cluster. Since these two factors will not interfere with each other, we can achieve higher precision for the prostatic cancer grade of pathological biopsies. The texture features of gray images used in this study are derived from the fractal analysis of the sub-bands of these images, which denote as SDEBC. Then, the Support Vector Machine is used to classify each image and the leaving-one-out approach is applied to estimate error rate.

This paper is organized as follows. Image clustering is introduced in the next section. Features of Fractal Analysis Extracted from Sub-band are presented in Section 3. The experimental results are presented in Sections 4, respectively. Finally, Section 5 contains conclusions.

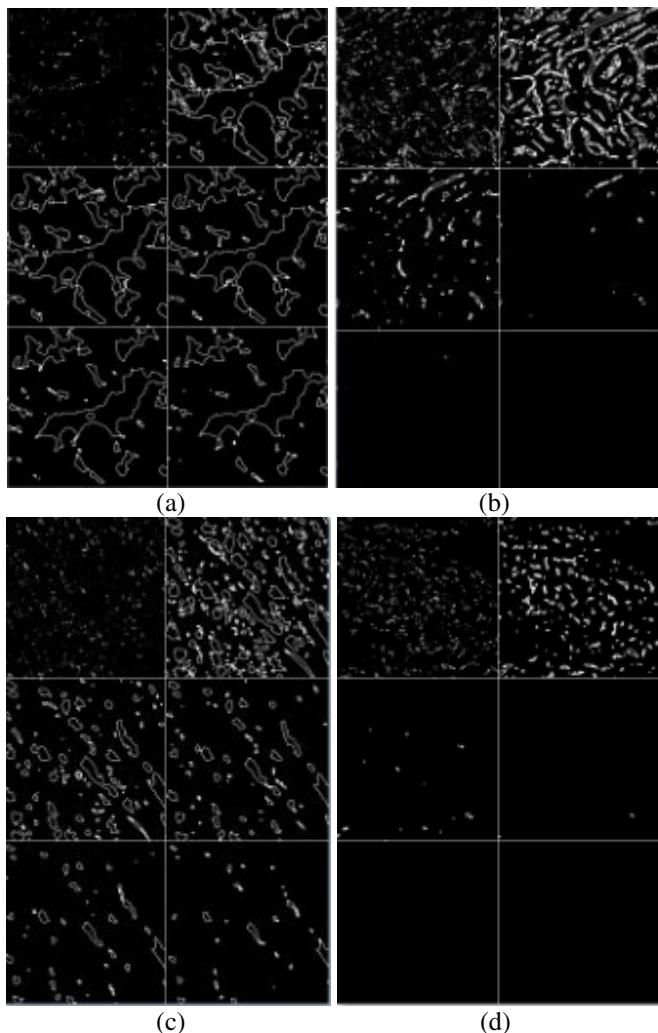
## 2 Image Clustering Based on SK-Set

Prostatic grading basically is according to their lumen and epithelial cells. Proper texture features can distinguish them for grading, but they can be very much influenced by the portion of lumens. Therefore, it is possible to increase the specificity of texture features by controlling the factor of lumens. For this reason, we will cluster the same type of lumen together based on the morphological analysis.



**Fig. 2.** The white shapes in prostatic biopsy image

There are three major colors in prostatic biopsy images, which are white, red and blue. Lumens of prostatic gland in prostatic biopsy images are in white color. Thus we transformed images from RGB color model to two-tone images. As show in figure 2, (a) is the normal prostate; (b), (c), and (d) is the grade 3, 4, and 5 respectively. The white shapes in these images represent the types of gland. Different size, density and shape mean different patterns of gland. Therefore, we adopt three important criteria to distinguish them. In this paper, we extracted Skeleton-set to represent these three criteria. Skeleton-set is described as below.



**Fig. 3.** SK-set of images in figure 2

At first, images transform to two-tone model by keeping all the white shapes in image and others regions convert into background, denoted as  $A_0$ . We perform morphological operation to get the skeleton-set. Let  $B$  is the structure element, and  $A_k = A_{k-1} \Theta B$ ,  $S_k = A_k - (A_k \circ B)$ , where  $\Theta$  is erosion and  $\circ$  denotes opening. Then we can get SK-set,  $\{S_k; k=1 \text{ to } N\}$ . It can reveal these three criteria by counting the pixel of  $S_k$ . Figure 3 are the SK-set of the images in figure 2, where  $N$  is 6 and every picture from top-left to bottom-right is  $S_1$  to  $S_6$  respectively. Figure 3 (a) shows that the pixel number of  $S_1$  is less than others while  $S_6$  is the largest, which represent figure 2 (a) contains large white shapes. The pixel number of  $S_3$  is much larger than  $S_4$  in Figure 3 (b), which means figure 3 (b) contains band-shape components. In the same way, SK-set in figure 3 (c) and (d) represents figure 2 (c) contains some unequal size components and figure 3(d) contains small but high density components. With the SK-set, we apply self-organizing map (SOM) [10] for unsupervised learning to produce clusters of prostatic biopsy images.

### 3 Features of Fractal Analysis Extracted from Sub-band

The term fractal means irregular fragments. Many natural objects exhibit fractal property or self-similarity, which can be described as follows [11, 12]. A bounded set,  $A$ , in Euclidean  $n$ -space is self-similar if  $A$  is the union of  $N_r$  distinct (non-overlapping) copies of itself scaled down by a ratio  $r$ . The fractal dimension  $D$  of  $A$  can be derived from the following basic equation [13]:

$$D = \frac{\log(N_r)}{\log(1/r)} \quad (1)$$

However, natural phenomena or objects do not exhibit deterministic self-similarity in practically due to they can often be classified as random fractals, meaning that each smaller part of it is statistically similar to the whole. Therefore, an object becomes statistically identical to the original one if it is scaled down by a ratio  $r$  in all  $n$  dimensions, so that (1) is satisfied.

There are many ways to estimate the FD of an image. The DBC method is adopted herein because it gives a better approximation for the image intensity surface [14, 15]. In general, compared with the low-grade prostatic carcinoma in pathological image, the high-grade prostatic carcinoma has sharp gray-level variation in neighboring pixels due to sheets of single dark malignant cells invade stroma. Thus, the DBC method can significantly distinguish low-grade and high-grade prostatic carcinoma by measuring the variations of intensity in local regions.

Before using the DBC method, the color pathological images of prostatic tissue are transformed into gray-level images by getting R channel from RGB color space for enhancing the contrast between malignant cells and background tissue. In above pre-processing stage, the malignant cells will be darker than other pathological objects because these are stained blue compared to other pathological objects, such as stroma stained red and lumens which do not stain and belong to white in H&E-stained pathological images. The DBC method is described as follows.

Consider that an image of size  $M \times M$  pixels has been scaled down to a size  $s \times s$ , where  $1 < s \leq M/2$  and  $s$  is an integer. Then, we can get the scale ratio  $r = s/M$ . Consider the image as a three-dimensional (3-D) space that  $(x,y)$  denoting two-dimensional (2-D) position and the third coordinate  $(z)$  denoting gray level of an image. The  $(x,y)$  space is divided into grids of size  $s \times s$ . There is a column of boxes of size  $s \times s \times h$  on each grid, where  $\lfloor G/h \rfloor = \lfloor M/s \rfloor$  and  $G$  is the total number of gray levels in an image. Let the maximum and minimum gray level of an image in the  $(i,j)_{th}$  grid fall in box number  $k$  and  $l$ , respectively. The contribution of  $N_r$  in the  $(i,j)_{th}$  grid is expressed as follows:

$$n_r(i,j) = k - l + 1 \quad (2)$$

After taking contributions from all grids, we can get

$$N_r = \sum_{i,j} n_r(i,j) \quad (3)$$

$N_r$  is counted for different scale ratio  $r$ . Then the fractal dimension  $D$  can be estimated from the least-squares linear fitting of  $\log(N_r)$  versus  $\log(1/r)$  by using (1).

For further analyzing the texture complexity in pathological images for different Gleason grades of prostate carcinoma, we analysis the entropy of image based on the box-counting method. The image is partitioned into several grids of size  $s \times s$ . For each grid in the image, we compute its entropy [16], thus the contribution of  $e_r$  in the  $(i,j)_{th}$  grid is defined as:

$$e_r(i,j) = -\sum_{k=0}^{G-1} p(z_k) \log_2 p(z_k) \quad (4)$$

Where  $Z_k$  is total number of pixels with gray level  $k$  in the  $(i,j)_{th}$  grid of an image, and  $p(Z_k)$  denotes the probability of occurrence of gray level  $k$  in the  $(i,j)_{th}$  grid of an image. After taking the standard deviation of each contribution from all grids, we denoted this entropy feature as  $E_r$ .

$$E_r = STD(e_r(i,j)) \quad (5)$$

Then using (1), the fractal dimension  $D$  can be estimated from the least-squares linear fitting of  $\log(N_r)$  versus  $\log(1/r)$ . The reason for counting  $E_r$  is that we can measure more accurately the variations of texture complexity for the regions with different size, i.e., different values of  $s$ .

Ref. [14] has showed that different textures may have the same FD. This may be due to natural phenomena usually only exhibit the property of random fractals; in other words, natural phenomena are not self-similarity over all scales [17]. Therefore, for finding distinguishing features, multiple FD-based features are calculated from the regions with different size in this paper, such as the grids with different size. The representation of multiple FD-based features is described as follows:

$$MF = (f_D^i, f_E^i), \quad i = 1, 2, 3 \quad (6)$$

where denote the FD of an image calculated from the grids with different size via DBC and Entropy-BC methods, respectively;  $i = 1$ , denotes the grids with different

size  $s$  ( $s = 8, 16$ , and  $32$ ),  $i = 2$ , denotes the grids with different size  $s$  ( $s = 64, 128$ , and  $256$ ), and  $i = 3$ , denotes the grids with different size  $s$  ( $s = 8, 16, 32, 64, 128$  and  $256$ ). For example, denotes the FD of an image calculated from the grids with size  $64, 128$ , and  $256$  by using Entropy-BC method. The texture features of gray images used in this study are fractal dimension (FD).

However, due to Gleason grading system mainly based on the texture formed by gland and the lumens are the main part of the gland which exhibit white shapes in the biopsy image. So that texture of the lumen is very important for the grading. Beside lumens, the arrangement and intensity of epithelial nuclei also have a considerable impact on grading. For example, if these epithelial nuclei appear in image randomly then it likely to be classified as grade 5. However, some undesired textures also exist on the images, such as texture within the stroma. These textures will be ignored by the pathologists, while it affects the grading results. So we have to get rid of it from the feature extraction in order to improve the correctness of automated grading. The patterns of lumen are the lower frequency response. Therefore, we used filter on the original image to get the lower frequency sub-band. Then extract the feature from the sub-bands will be much efficient than from original images. In this paper, we adopted the DB6 [18] as the filter to operate in both vertical and horizontal direction to get the sub-bands, which denoted as  $SB_a, SB_b, SB_c$  and  $SB_d$ , Where  $SB_a$  and  $SB_d$  are the sub-band of low and high frequency in both directions, respectively.  $SB_b$  is the sub-band of low frequency in vertical direction and high frequency in horizontal direction and  $SB_c$  is the sub-band of high frequency in vertical direction and low frequency in horizontal direction. Then, we extracted multiple FD-based features (MF) from each sub-band. We denoted this feature as SDEBC, and described as follows. Although we defined SDEBC as the MF combined four sub-bands here, but it will be adjusted after experiments.

$$SDEBC = (MF_a, MF_b, MF_c, MF_d) \quad (7)$$

Although we defined SDEBC as the MF combined four sub-bands here, but it will be adjusted after experiments.

## 4 Experimental Results

There are 1000 images used in the experiments, including normal image of 118, grade 3 of 251, grade 4 of 524 and grade 5 of 107. Because lower grade 1 and 2 are rare found in the real situation, we did not include them in this experiment. The Support Vector Machine (SVM) [19] is used to classify each image and the leaving-one-out approach is applied to estimate error rate.

At first, we examine the performance of these four sub-bands. As show in table 1,  $SB_a$  is the highest among the four single sub-bands. It reveals that the grading information is mostly contained in the lower frequency. Combining  $SB_d$  on it can only improve the CCR of 0.1, which is meaningless.  $SB_b$  and  $SB_c$  contain the directional information of texture and have lowest CCR, which agree with the fact that the direction of the texture is nothing to the grading of prostate cancer. Pathologists rotate the biopsy and the grading results will remind the same. Even though combined these two feature vector, the CCR still has no any significant improvement. For decreasing the effect of direction, we combined the coefficient of sub-bands,  $SB_b$  and  $SB_c$ , to get the

MF feature. Combining with the feature vector of  $SB_a$  produces the highest CCR of 86.3. Again join the MF feature of  $SB_d$  still without any better results. In other color plane, combining sub-band can even decrease the CCR, so that  $SB_d$  do contain much matterless information and should be ignored.

**Table 1.** The CCR of SDEBC with different combination of sub-bands

Sub-bands	CCR
$SB_a$	85.2%
$SB_b$	57.4%
$SB_c$	56.7%
$SB_d$	66.4%
$SB_a+SB_d$	85.3%
$SB_b+SB_c$	57.4%
$SB_a+SB_bSB_c$	86.3%
$SB_a+SB_bSB_c+SB_d$	86.3%

In this experiment, we adopted self-organizing map (SOM) for unsupervised learning to produce clusters of prostatic biopsy images. And 1000 images are divided into 12 clusters as shown in table 2.

**Table 2.** The number and grading of each cluster

cluster	Normal	Grade 3	Grade 4	Grade 5	Total
1	15	40	142	22	219
2	1	0	44	4	49
3	13	17	0	0	30
4	0	0	69	18	87
5	0	0	41	13	54
6	3	20	7	0	30
7	26	31	7	0	64
8	1	0	28	15	44
9	0	31	142	35	208
10	19	21	2	0	42
11	0	48	34	0	82
12	40	43	8	0	91

Each cluster contains similar lumen texture, so that extracted SDEBC feature will not interfere with the lumens. Besides, some clusters contains less possible grading which also will improve the performance. We compare our method with Huang [9]'s, the result are listed in table 3.

**Table 3.** The comparison of SDEBC and Huang's method

	Huang's method	SDEBC	SDEBC with cluster
Correct Classification Rate (CCR)	81.1%	86.3%	87.4%

Table 3 shows that our method outperforms Huang's method about 6%. But features extracted from sub-band must contain some useless information. We adopted Sequential Forward Floating Search (SFFS) method [20] to reduce the dimension of feature vector. Table 4 shows the results of SFFS, including the dimension of feature and the CCR of both Huang and our methods.

**Table 4.** The comparison of SDEBC and Huang's method with the SFFS

	Huang's method	SDEBC	SDEBC with cluster
Correct Classification Rate (CCR)	81.3%	89.9%	92.1%
Original length of feature vector	6	18	18
Reduced length of feature vector	5	11	11

As showed in table 4, dimension of feature vector of SDEBC has been reduced from 18 to 11, while CCR is increased from 87.4% to 92.1%. This demonstrated that our method can help pathologists to grading the prostatic cancer efficiently.

## 5 Conclusion

This paper developed a two stages approach to classify prostatic cancer according to the Gleason Grading System. The first stage classifies biopsy images into clusters based on their Skeleton-set. Then the second stage, we analyzed the fractal dimension of sub-bands derived from the images of prostatic biopsies, which called SDEBC. This approach can avoid two kinds of texture interfere with each other and produce high performance. The experimental results demonstrated that we can achieve 92.1% of accuracy for a set of 1000 prostatic biopsy images much higher than other method.

## Acknowledgments

This work was supported by National Science Council under Grant NSC 98-2221-E-324-032.

## References

1. American Cancer Society, Cancer Facts & Figures 2007. American Cancer Society, Atlanta, GA (2007)
2. Balk, S.P., Ko, Y.J., Bubley, G.J.: Biology of prostate-specific antigen. *L. Clin. Oncol.* 21(2), 383–391 (2003)
3. O'Dowd, G.J., Veltri, R.W., Miller, M.C., Strum, S.B.: The Gleason Score: A Significant Biologic Manifestation of Prostatic cancer Aggressiveness on Biopsy. *Prostatic Cancer Research Institute: PCRI Insights* 4(1), 1–5 (2001)

4. Stotzka, R., Männer, R., Bartels, P.H., Thompson, D.: A hybrid neural and statistical classifier system for histopathologic grading of prostate lesions. *Analytical Quantitative Cytol. Histol.* 17(3), 204–218 (1995)
5. Tabesh, A., Kumar, V., Pang, H., Verbel, D., Kotsianti, A., Teverovskiy, M., Saidi, O.: Automated prostatic cancer diagnosis and gleason grading of tissue microarrays. In: Proceedings of the SPIE on Medical Imaging, vol. 5747, pp. 58–70 (April 2005)
6. Wetzel, A.W., Crowley, R., Kim, S.J., Dawson, R., Zheng, L., Joo, Y.M., Yagi, Y., Gilbertson, J., Gadd, C., Deerfield, D.W., Becich, M.J.: Evaluation of prostate tumor grades by content-based image retrieval. In: Proceedings of the SPIE on AIPR Workshop Advances in Computer-Assisted Recognition, Washington, DC, vol. 3584, pp. 244–252 (1999)
7. Smith, Y., Zajicek, G., Werman, M., Pizov, G., Sherman, Y.: Similarity measurement method for the classification of architecturally differentiated images. *Computers and Biomedical Research* 32(1), 1–12 (1999)
8. Jafari-Khouzani, K., Soltanian-Zadeh, H.: Multiwavelet grading of pathological images of prostate. *IEEE Transactions on Biomedical Engineering* 50(6), 697–704 (2003)
9. Huang, W., Lee, C.-H.: Automatic Classification for Pathological Prostate Images Based on Fractal Analysis. *IEEE Trans. Medical Imaging* 28(7) (July 2009)
10. Haykin, S.: 9. Self-organizing maps. In: Neural networks - A comprehensive Foundation, 2nd edn. Prentice-Hall, Englewood Cliffs (1999) ISBN 0-13-908385-5
11. Mandelbrot, B.B.: Fractal Geometry of Nature. Freeman, San Francisco (1982)
12. Sarkar, N., Chaudhuri, B.B.: An efficient differential box-counting approach to compute fractal dimension of image. *IEEE Transactions on Systems, Man and Cybernetics* 24(1), 115–120 (1994)
13. Chaudhuri, B.B., Sarkar, N.: Texture segmentation using fractal dimension. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(1) (1995)
14. Voss, R.F.: Random fractals: Characterization and measurement. In: Pynn, R., Skjeltorp, A. (eds.) *Scaling Phenomena in Disordered Systems*. Plenum, New York (1986)
15. Baish, J.W., Jain, P.K.: Fractals and cancer. *Cancer Research* 60, 3683–3688 (2000)
16. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 2nd edn. Prentice-Hall, New Jersey (2002)
17. Gose, E., Johnsonbaugh, R., Jost, S.: *Pattern Recognition and Image Analysis*. Prentice-Hall, Englewood Cliffs (1996)
18. Chui, C.K., Lian, J.A.: A study of orthonormal multiwavelets. *Appl. Numer. Math.* 20, 273–298 (1995)
19. Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. *Neurocomputing* 55(1-2), 169–186 (2003)
20. Pudil, P., Ferri, F.J., Novovičová, J., Kittler, J.: Floating Search Methods for Feature Selection with Nonmonotonic Criterion Functions. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference B: Computer Vision & Image Processing, vol. 2 (1994)

# Automatic Drug Image Identification System Based on Multiple Image Features

Rung-Ching Chen<sup>1</sup>, Cho-Tsan Pao<sup>2</sup>, Ying-Hao Chen<sup>1</sup>, and Jeng-Chih Jian<sup>1</sup>

<sup>1</sup> Department of Information Management Chaoyang University of Technology  
crching@cyut.edu.tw

<sup>2</sup> Graduate Institute of Information Chaoyang University of Technology

**Abstract.** Drugs can be divided into many types, such as different compositions, content and shapes, but users do not always possess or comprehend professional drug facts. Many drug recognition systems offer keyword search but they are difficult for users to understand the medications' names. One possible way would be for users to describe the features of drugs according to their appearance, such as color, shape, etc. In this paper, we propose an automatic drug image identification system (ADIIS) based on multiple image features. ADIIS is able to improve drug identification errors as well as provide drug information. In our primary experiments, by using an image, the system was able to retrieve the top ten similar drugs for the user to identify the specific drug. In addition, out of the ten identified drugs retrieved by ADIIS, the first of the ten drug identifications was 95% of the correct match.

**Keywords:** CBIR, Hamming distance, Gabor filter, Neural network, Weight similarity.

## 1 Introduction

Accidental medication mishaps occur frequently, making medication safety an important issue. Hospitals currently provide a variety of drug counseling in order to rectify these problems.

Searching web sites for medicine information can be performed two ways: (1) database mapping and (2) keyword search [11]. Database mapping allows users to compare drug images with database information [7]. However, this method is time-consuming and not efficient for drug search. Keyword search is done by entering specific words, such as the drug name, color, shape, magnitude, lettering and other features, by which to search [7]. Web search provides lists of drug information for users. However, the search is user-defined, is limited, and can easily result in identification errors.

In this paper, we proposed an Automated Drugs Image Identification System (ADIIS), the system then retrieved the ten of the closest database images similar to the drugs image, allowing the user to correctly identify the drug and obtain the drug information.

This paper is organized into 6 sections. Section 2 describes the relevant drug identification and techniques. Section 3 explains the preprocessing and feature extraction in our proposed system. Section 4 discusses the similarity measurement. Section 5 reports experimental results and discussions. Finally, conclusions and future works are drawn in Section 6.

## 2 Related Research of Drug Identification Systems

### 2.1 Content Based Image Retrieval

In recent years, content-based image retrieval (CBIR) is a popular topic of image recognition. Its concept is calculated by the program, and extracts the useful content from the images, such as human, material, color, etc. Some search methods are use query images by examples or keywords search to find the similar or the same images. Conventional image search method is manual mapping each images from database. So they need to spend much time and manpower. Then CBIR uses computer technology to identify images, in order to reduce the recognition time and manpower. According the concept of CBIR, we use different image processing techniques to extract the features of drug. In this paper, the system will use the concept of CBIR to extract the features of drug.

### 2.2 Related of Drug Identification System

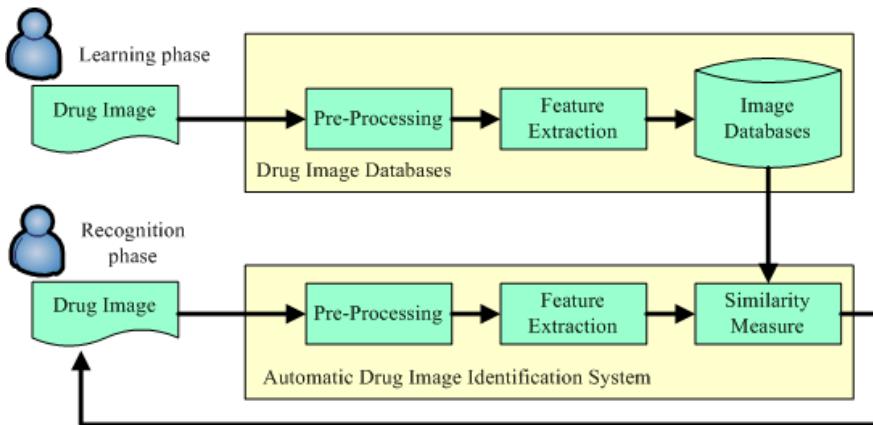
In recent years, some hospitals provide drug information database for people to identify drug functions. The researchers proposed the drug identification system.

Zeno et al. (2001)[3] designed the drug identification system that combines IBM's QBIC (Query By Image Content) and the iMatch system to identify the specific drug. Hsieh et al. (2005)[11] proposed a Real Drug Image Identification System to capture color and texture of drugs and it searches the query images by the examples in the database. Lin et al. (2007)[9] proposed a tablet drug image retrieval system to raise the drug recognition. Lin's system extracts features including the shape, color and size. It used neural networks and combined moment invariants and Zernike moments to identify the drug. Xie et al. (2008)[12] captured drug features that users select for system identification, such as drug size, shape, weight, and color.

The round white drugs are difficult to identify. Because of their popularity, the white and circle drugs and their features are hard to represent. Therefore, we propose a system for extracting color, shape, magnitude, ratio, and texture of a drug image for identification, as far as it is possible to describe the appearance of images of drugs to enhance drug identification.

## 3 The Proposed System

In this paper, we propose an Automatic Drug Image Identification System to identify drugs as shown in Fig. 1. It is divided into two phases. One phase is the learning phase and the other is the recognition phase.



**Fig. 1.** Automatic Drug Image Identification System of Systems Flow Chart

### 3.1 Pre-processing

The pre-processing normalizes the drug image to 800×600 pixels. Then the system automatically defines the drug, and draws a region of interest (ROI). In the ROI, the drug appears as white, on a black background with white graph lines.

In order to begin pre-processing, we calculated the center of gravity. The center of gravity was calculated by rotating the white drug image to align with the horizontal background graph lines. We aligned the longer side of the drug to the horizontal axis and the shorter side to the vertical axis.

As the drug images may have had noise and/or blurred edges, the system used median filters to eliminate noise and effectively preserve the original texture. Through shape enhancement, we used histogram equalization to enhance the drug shape and texture. The image and background images of the drug contrast in order to enhance the edge precision of the drugs used.

### 3.2 Feature Extraction

**Color Feature Extraction.** In this paper, we used HSV (hue, saturation and value of intensity) color space as features for drug identification because it was more stable than the RGB model and the Lab model [10].

If the drug contained two colors on one side, the system calculated the ratio using Equation (1). In our system, when the color feature value was greater than 1, then the drug was composed of compound colors. Because of this equation, the system was able to verify whether the drugs are composed of one or more colors. The system avoided extracting the error of drugs composed of compound colors. An example would be a drug composed of both red and blue colors. The red color value would be multiplied by 0.3, while the blue color value would be multiplied 0.4. The total value would be obtained by using Equation (1). The total value would be 34 when  $A_1$  equals 100 and  $A_2$  equals 10. In our system, when the color feature value was greater than 1, then the drug was composed of compound colors. Because of this equation, the system was able to

verify with feedback to determine the whether the drugs are composed of one or more colors. The system avoided extracting the error of drugs composed of compound colors.

$$Cp = C_1 \times A_1 + C_2 \times A_2 \quad (1)$$

Where  $C_1 < C_2$ ,  $C_1$  and  $C_2$  are different colors in the same drug.  $A_1$  is the weight of  $C_1$ .  $A_2$  is the weight of  $C_2$ . In general, only  $A_2$  is used because of single color of medicine. In order to limit the image of a color feature, we formalized the value of color feature as  $[0, 1]$ .

**Shape Feature Extraction.** Canny edge algorithm [2] is used to define the edge of the drug images and convert them into binary images. The edge was then divided into four equal blocks. MPEG-7 Edge Histogram Descriptor (EHD) was used for the distribution edge of drug images [6].

The values of EHD are sent to a back-propagation neural network to classify the shape of the drug [2]. The neural network has three layers. The input layer has four nodes, while a hidden layer has three neuron nodes, and an output layer has 10 neuron nodes. The output layer indicates ten different types of drugs: triangle, square, pentagonal, hexagonal, octagonal, rectangle, circle, irregular-shaped, oval-shaped and long cylindrical [8]. We then mapped the ten types into numerical values between 0 and 1 with steps 0.1.

**Ratio Feature Extraction.** Ratio is the rate of the maximum width divided by the maximum length. Ratio can effectively strengthen the identification of drugs with the same shape but different ratio. The ratio value is between 0 and 1.

**Magnitude Feature Extraction.** We placed the drugs on a black sheet of paper with white grid lines. Magnitude measurement is determined upon the area of the grid lines covered by the drug. The system uses the grid lines to measure the magnitude of each drug. Each grid is  $1 \times 1$  mm square. After detecting the size of the drugs, they were ordered by size within the database. The system selected the drugs with a magnitude closely matched to the measured drug. Then, Hamming distance was used to calculate the difference, seen in Equation (2).

$$DB_4 = \left( 1 - \left| \frac{Q - D_n}{\max\{Q, D_n\}} \right| \right) \quad (2)$$

In Equation (2),  $Q$  is the magnitude of the query drug, and  $n$  is the index of the candidates in the database.  $D_n$  is the drug magnitude of index  $n$  in the database. So, the value of  $DB_4$  is between 0 and 1. When the value of  $DB_4$  is equal to 1, it means that the two drugs are exactly the same magnitude.

**Texture Feature Extraction.** The Gabor filter was used to calculate drug texture [5]. First, the surface texture of the drug was converted to binary images, and the Canny edge algorithm was used to calculate the value of edge pixels [4]. Secondly, the texture of the drug edge was removed. Finally, the texture input to Gabor filter, and we were processed in a two-dimensional [1]. The output value is between 0 and 1.

## 4 Similarity Measurement

When the system completes the extraction of test drugs features, we defined the vector of features as follows:

$$\text{feature}(Q_n) = (q_1, q_2, q_3, q_4, q_5) \quad (3)$$

Here  $q_1, q_2, q_3, q_4$  and  $q_5$  are the feature of test drugs, in the order of color, shape, ratio, magnitude, texture. In addition, the features in the database are listed as follows:

$$\text{feature}(DB_n^i) = (DB_1^i, DB_2^i, DB_3^i, DB_4^i, DB_5^i) \quad (4)$$

The symbol  $i$  is the index of images in database.  $DB_1^i, DB_2^i, DB_3^i, DB_4^i$  and  $DB_5^i$  were the features of database, in the order of color, shape, ratio, magnitude, texture.

The white circular tablet drugs deem to be difficult to recognize due to the fact that many are similar in size, color and shape. Determining the surface texture of the drugs was useful to correctly verify these drugs. To be able to effectively identify these drugs, the weight of Euclidean distance is used to calculate the difference, shown by Equation (5):

$$\begin{aligned} \text{Sim}_{\text{feature}} &= \\ \sqrt{w_1(q_1 - DB_1^i)^2 + w_2(q_2 - DB_2^i)^2 + w_3(q_3 - DB_3^i)^2 + w_4(q_4 - DB_4^i)^2 + w_5(q_5 - DB_5^i)^2} \end{aligned} \quad (5)$$

Where  $w_1 + w_2 + w_3 + w_4 + w_5 = 1$

$$\begin{aligned} q_1 - DB_1^i &\left\{ \begin{array}{ll} 0 & \text{when } q_1 = DB_1^i \text{ //the same color} \\ 1 & \text{when } q_1 \neq DB_1^i \text{ //different color} \end{array} \right. \\ q_2 - DB_2^i &\left\{ \begin{array}{ll} 0 & \text{when } q_2 = DB_2^i \text{ //the same shape} \\ 1 & \text{when } q_2 \neq DB_2^i \text{ //different shape} \end{array} \right. \end{aligned}$$

$w_1, w_2, w_3, w_4$  and  $w_5$  were the weight of feature, in the order of color, shape, ratio, magnitude, texture.

To obtain the drug images for the database, and for the features of the drug in question, the system used the weight Euclidean distance to measure the similarity between the test image and candidate image. Weights are set based upon fuzzy rules. This decision was the result of the internal parameters set for this study.

Rule: 1

IF color ∈ white and shape ∈ circle

THEN  $w_5$  is  $K$ ,  $w_4$  is  $1-K$ , and  $w_1, w_2, w_3 = 0$ .

$K$  was determined by the magnitude and texture. If the size was larger, the texture was smaller.

In addition, when the shape of the test drug was oval and had two colors, the system classified it as a capsule and increased its weight based upon its color features.

Rule: 2

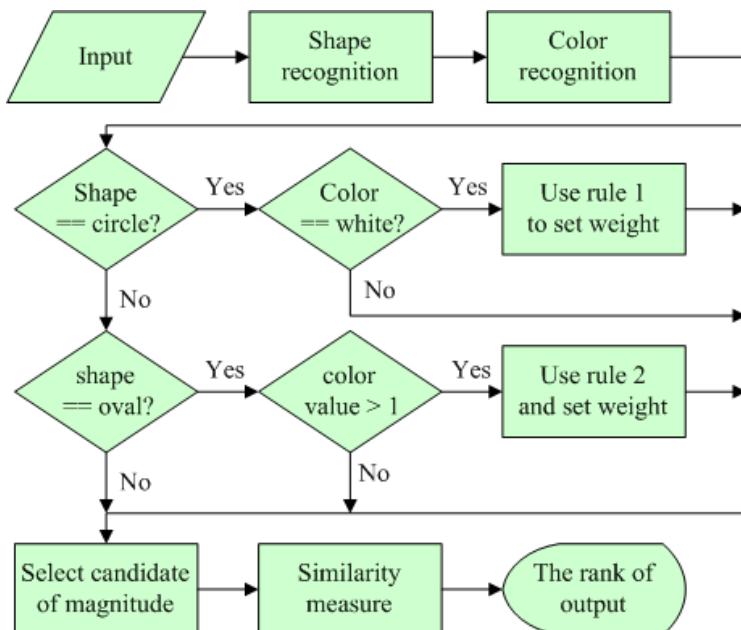
IF shape  $\in$  oval-type and has two colors

THEN  $w_1$  is  $K_1$ ,  $w_4$  is  $1 - K_1$ , and  $w_2, w_3, w_5 = 0$ .

$K_1$  was the weight value, and when the weight of the color was larger, the weight of the magnitude was smaller. Determining the texture of the capsule drugs was difficult, so the system did not consider texture while determining classification.

The Euclidean distance value was smaller when the inputted drug image matched in similarity with drugs in the database. However, when the Euclidean distance value was larger, the similar match with the database is lower. According to the results of the calculation, the system selected the most similar drug images and reported the top ten closest matches to the users. The output information included drug images and drug name. The system was able to effectively identify the specific types of drugs to the users.

Fig. 2 shows our system of similarity measure workflow. The process works by first inputting the feature values of the drug's image. After the image had been inserted, the next step in the process was shape and color recognition. If the color was white and circular, our system used Rule 1 to set the magnitude and texture weights, by following the fuzzy membership value to adjust  $K_1$ . If the drug's color was a compound and shape was a oval, our system used Rule 2 to set the color and magnitude weights, following the fuzzy membership to adjust  $K_1$ . When the shape was not circular or oval, then the weights are equal to 0.2. The next step of the process was to select the candidate of magnitude, and calculate the similarity of the imputed drug image with the database. The drug feature's values and weights were then uploaded into the system, which calculated the similarity measurement. Finally, the system output the lists of matching drugs to the user.



**Fig. 2.** The Workflow of Similarity Measure Based on Rules

## 5 Experimental Result

Drugs were provided by Taichung hospital in Taichung, Taiwan. 90 drugs were used in this research, and the drugs donated treat illnesses such as diabetes, blood pressure, colds, and some drugs were antibiotics. Two photos were taken of each drug to document both sides, resulting in 180 drugs images.

In the similarity measure, we will be divided into the general drugs, the white circular drugs, and the capsules. The feature weights are according classification types to adjust weights. We use cross-validation to prove the classification accuracy of the system. First, divide drugs into two groups of A and B. The group A is the training data, and they provide to systems for drug classification. The group B is the test data. In this step, the recognition average accuracy rate was 97% for general drugs, the white circular drugs were 100%, and the capsules were 96%. Second, let the group B be the training data, and they provide to systems for drug classification. The group A is the test data. In this step, the average recognition accuracy rate is 98% for general drugs, the white circular drugs are 100%, and the capsules are 97%.

Through the validation of the results, our system for the classification of three types has good results. The reason of general drug has lower accuracy than the system identify the cylindrical drugs is the capsule. The Capsules has 97% of accuracy that some capsules are the same color at both ends.

The accuracy rate of the system is according to the follow formula:

$$Accuracy = \frac{\text{The Correct Classification of the Number of Drug Images}}{\text{Total Number of Drug Images of Query}} \quad (6)$$

**Table 1.** The Accuracy of Five Times Test

Rank of results	Test 1	Test 2	Test 3	Test 4	Test 5	Average
Rank 1	95	92.5	97.5	95	95	95
Rank 2	100	97.5	97.5	95	97.5	97.5
Rank 3	100	100	100	97.5	100	99.5
Rank 4	100	100	100	100	100	100

Each test was performed five times, and was conducted by randomly selecting 20 drug images. Table 1 shows five times test results. The recognition average accuracy rate was 95% for the Rank 1. We also designed another experiment, to identify only white circular drugs. We selected 26 white circular drugs out of the 90 testable drugs. Table 2 shows the test results of white circular drugs.

**Table 2.** The Accuracy of White Circular Drugs Test

Rank of results	Test 1	Test 2	Test 3	Test 4	Test 5	Average
Rank 1	80	82.5	75	85	85	81.5
Rank 2	87.5	85	82.5	92.5	90	87.5
Rank 3	92.5	90	87.5	95	97.5	92.5
Rank 4	95	100	87.5	100	97.5	96
Rank 5	100	100	95	100	100	99
Rank 6	100	100	97.5	100	100	99.5
Rank 7	100	100	100	100	100	100

## 6 Conclusions and Future Works

In this paper, we have proposed an automated drug image identification system using multiple image features for proper drug identification. The recognition accuracy rate is 95% on randomly selected tests. Our system used background and drug images. The research confirmed that our method for drug identification was feasible and identification was effective, especially for white circular drugs. However, there were still system identification errors with drugs of similar size, color, and texture.

In future works, we would suggest collecting more images to include in the database and test the system. We would also advocate for using other digital devices such as mobile phones to possibly access and operate the system. We would also recommend including further drug information, such as drug ingredients and side effects to enhance the sense of value of the drug identification system.

**Acknowledgment.** The authors would like to thank the research support from National Science Council, Taiwan, with number: NSC 98-2221-E-324 -031.

## References

1. Arivazhagan, S., Ganesan, L., Priyal, S.P.: Texture Classification Using Gabor Wavelets Based Rotation Invariant Features. *Pattern Recognition Letters* 27(16), 1976–1982 (2006)
2. Gao, X., Xiao, B., Tao, D., Li, X.: Image Categorization: Graph Edit Distance + Edge Direction Histogram. *Pattern Recognition* 41(10), 3179–3191 (2008)
3. Geradts, Z., Hardy, H., Poorman, A., Bijnhold, J.: Evaluation of Contents Based Image Retrieval Methods for A Database of Logos on Drug Tablets. *Image Analysis and Characterization of SPIE* 4232, 553–562 (2001)
4. Lin, C.H., Chen, R.T., Chan, Y.K.: A Smart Content-Based Image Retrieval System Based on Color and Texture Feature. *Image and Vision Computing* 27(6), 658–665 (2009)

5. Pavlou, M., Allinson, N.M.: Automated Encoding of Footwear Patterns for Fast Indexing. *Image and Vision Computing* 27(4), 402–409 (2009)
6. Phan, R., Androutsos, D.: Content-Based Retrieval of Logo and Trademarks in Unconstrained Color Image Databases Using Color Edge Gradient Co-occurrence Histograms. *Computer Vision and Image Understanding* 114(1), 66–84 (2009)
7. Sung, T.Y.L., Hung, F.H., Chiu, H.W.: Implementation of An Integrated Drug Information System for Inpatients to Reduce Medication Errors in Administrating Stage. In: Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 743–746 (2008)
8. Medication Information of Department of Health, Executive Yuan, ROC (2009),  
<http://drug.doh.gov.tw/>
9. Lin, C.H., Huang, E.W., Jiang, W.W.: Pill Image Retrieval using Neural Networks. *The Journal of Taiwan Association for Medical Informatics* 16(7), 29–42 (2007)
10. Chen, W., Chao, P.J., Lin, H.L.: Drug Identification by Network Adaptive Content-Based Image Retrieval. *The Journal of Health Science* 9(2), 133–145 (2007)
11. Hsieh, C.H.: Applying Content Based Image Retrieval to Drug Identification Research, Taipei Medical University Graduate Institute of Medical Information, Thesis of Master (2005)
12. Xie, Y.Z.: Study of Drug Identification System Using The Image Processing Technique, I-Shou University of Department of Electronic Engineering, Thesis of Master (2008)

# On the Development of a Brain Simulator

Wen-Hsien Tseng<sup>1,2</sup>, Song-Yun Lu<sup>1</sup>, and Hsing Mei<sup>1,2</sup>

<sup>1</sup> Web Computing Lab., Department of Computer Science and Information Engineering

<sup>2</sup> Graduate Institute of Applied Science and Engineering,

Fu Jen Catholic University, Taipei, Taiwan, ROC

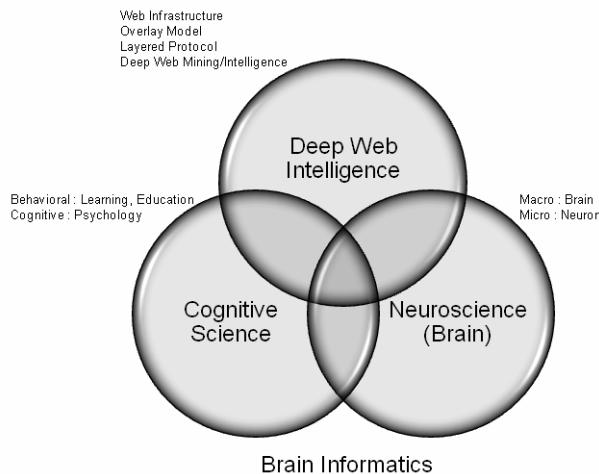
{cstony, jimmy98, mei}@csie.fju.edu.tw

**Abstract.** Developing a whole brain simulator, a computer simulation in modeling brain structure and functionality of human, is the ultimate goal of Brain Informatics. Brain simulator helps researchers cross the bridge between the cognitive behavior/decease, and the neurophysiology. Brain simulators development is still in infant stage. Current simulators mostly consider the neuron as the basic functional component. This paper starts with introducing the background and current status of brain simulator. Then, an extensible brain simulator development framework is proposed. From information technology perspective, we adopt overlay and peer-to-peer network to deal with the complexity of brain network. Moreover, layered design with object-oriented brain class hierarchy forms the flexible development framework of the proposed simulator. The proposed brain simulator is evolved in case-based incremental delivery style. The power of the simulator will grow along with more research cases from cognitive and clinical neuroscience.

**Keywords:** Brain simulator, brain network, object-oriented framework, network analysis.

## 1 Introduction

The fundamental goal of information intelligence is to understand and develop intelligent systems that integrate all the human-level capabilities and achieve the human-level Artificial Intelligence. The next generation of information intelligence research and development needs to understand multiple features of human-level intelligence in advance such as autonomous context-aware interaction, reasoning, planning, and learning. Web intelligence belongs to this category. Therefore, integrated studying the three intelligence related research disciplinary (deep web intelligence, neuroscience, and cognitive science) achieves truly human-level intelligence, namely Brain Informatics [1, 2], as shown in Figure 1. The objective of brain informatics is to understand the integrated human brain in ways of systematic means, as well as the Web Intelligence centric information technologies. In other words, the ultimate goal is to develop a whole brain simulator, a computer simulation in modeling brain structure and functionality of human.



**Fig. 1.** The Brain Informatics Interdisciplinary

Scientists have been collecting data about brain activity with brain activity measuring techniques, such as EEG and functional MRI, in the laboratory for many years. They made many remarkable hypotheses but most of them are not feasible in experiments [3, 4]. Brain simulator, a computer simulation in modeling brain structure and functionality, dramatically enhances the scientific method. It's a tool that scientists can use to not only better understand how cognition works, but rapidly test their ideas on an accurate replica of the brain [5]. Brain simulator is an enormous step forward for Brain Informatics. However, the flexibility and dynamicity of the human brain in time and space are challenges in designing a brain simulator [6]. Therefore, our research combines computer science and neuroscience to build up a flexible and dynamic brain simulator. The difference between two best known state-of-the-art brain simulator projects and our proposed brain simulator will be discussed later.

The brain simulator can be applied to many research topics. For example, scientists have long concerned with brain related disease, such as Alzheimer's disease, epilepsy, and schizophrenia. Topological change, accordingly functional dynamics caused by the diseases and the influence of abnormal release of neurotransmitters on functional regions in the human brain can be simulated by the brain simulator. This gives scientists a new insight to brain related diseases. The ultimate goal of the brain simulator is to achieve the simulation of human-level intelligence and helps to find out resolutions of brain related diseases.

In this section, we introduce the importance of designing a brain simulator. We briefly review some backgrounds of brain informatics in section 2 and then summarize the current status of brain simulator in section 3. The proposed brain simulator is presented in details in section 4. Finally, the conclusion and future work are discussed in Section 5.

## 2 Background

Brain Informatics has been studied as an interdisciplinary research field, including deep web intelligence, neuroscience and cognitive science. Some related research backgrounds are discussed in this section.

### 2.1 Networks Analysis

Recently, complex network analysis has been developed as an emerging research field. It is based on the classical graph theory and mathematical models [7]. Instead of concerning about the purely random or purely ordered networks, the new research field concerns about the real-world networks such as social networks and computer networks that share same non-trivial topological features or patterns which are only shown in the complex networks. Such networks are not only large in their size, but diverse in topological patterns which may change over time. Now, the trend is rapidly translated to the studies of brain networks due to its complexity [8-10]. In order to reduce the difficulties of analyzing brain networks, we divide them into brain connectivity, processing, causal (overlay), and application (behavior/disease) layers. Each layer shows different characteristics which can be further described by applying complex network analysis.

### 2.2 The Human Connectome Project

In 2009, the National Institutes of Health of the United States announced the Human Connectome Project [11, 12]. It's a five-year project which aims to map the connectivity of the human brain. It could be a prominent achievement which can help scientists to better understand the mechanisms of cognition and behaviors of the human brain [13]. However, because of the complexity of the human brain, there are many difficulties waiting to be solved. The challenge tasks are not feasible to be accomplished in just five years.

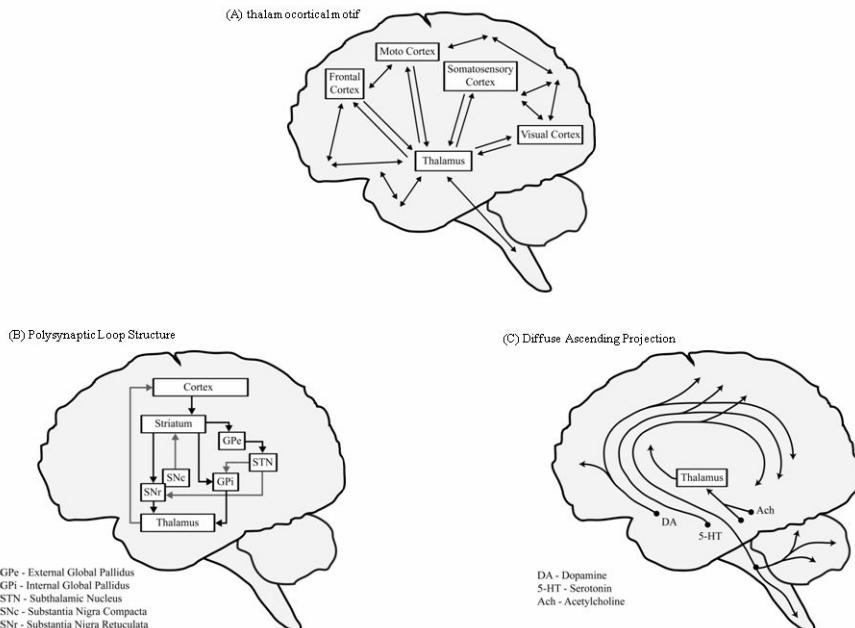
### 2.3 Brain Networks

Brain functions are the emergence of structural connectivity in human brain. There are three basic networks in the human brain [14]:

- **Thalamocortical Motif:** Thalamocortical motif, as shown in Figure 2(A), is fibers between the thalamus and the cerebral cortex [15]. There are loops consists of projections feed forward from the thalamus to the cortex and projections from the cortex back to the thalamus. Neuronal groups within the cerebral cortex not only connect with each other locally but also connect globally to interact for completing a task. These loops are called "reentrant loops" that map different types of stimuli to corresponding brain regions.
- **Polysynaptic Loop Structure:** The major structure of polysynaptic loops, as shown in Figure 2(B), are the basal ganglia which locate in the center of the human brain and consist of the striatum, the STN (subthalamic nucleus), the GP (globus

pallidus), and the SN (substantia nigra) [16]. These loops receive the connections from the motor and somatosensory area of the cerebral cortex and send projections back to the premotor cortex through the thalamus. Direct and indirect pathways are between the striatum and the GP. Inhibitory and excitatory signals are passed through respectively. This mechanism can lead to the result of modulating the activity of the thalamus. Same mechanism works on the SNr (substantia nigra pars reticulata).

- **Diffuse Ascending Projection:** As shown in Figure 2(C), there are several value systems in this projecting structure. Each of them has different kinds of neurotransmitters that relate to rewards and affects many critical body functions [17]. These systems include the locus coeruleus, which release NE (noradrenaline); the raphe nucleus release, which release 5-HT (serotonin); the various cholinergic nuclei, which release Ach (acetylcholine); the dopaminergic nuclei, which release DA (dopamine), and etc.. The path ways of these neurotransmitters construct hair-like networks which are in the diffuse ascending pattern.



**Fig. 2.** Three Brain Networks

## 2.4 Internet vs. Human Brain

Traditionally, we model the human brain by computers. But recent studies subvert traditional thinking and give us a new insight into the human brain, that is, functions are the emergence of the complex networks. Here, we try to make a comparison of the internet and the brain networks. The internet is a complex network which is composed of hundreds of thousands computers that are linked to each other. In the perspective of

higher level, the internet can be viewed as several nets which are inter-connected regionally. The structure is similar to the brain networks. They are not only locally wired but also maintain long-distance connections. Due to their complexity, some common properties such as motif, community, and small structure can be found by applying complex network analysis. And both of them are modeled by layered structure. On the internet, OSI model is a standard which defines us how to implement communication protocols. Likewise, the human brain can be modeled by anatomical structure (nuclei and connectivity), different connection patterns and overlayed networks (three basic brain networks), and output functions and behaviors.

The internet and the human brain are both fault tolerable. They are capable of recovering from the damage to their structure. On the internet, data link and transport layers provide error correction mechanisms that protect the data from unpredictable accidents. On the network layer, disconnection between the routers would trigger a recomputation according to the routing protocol. In the human brain, degeneracy is the ability of structurally different elements of a system to perform the same function or yield the same output.

Although the internet and the human brain share some similarities, they still have some essential differences. Firstly, they are different at scales. The human brain has at least  $10^{11}$  neurons and  $10^{14}$  synaptic connections while the internet is composed of billions of devices. Secondly, the capabilities of the basic elements of the internet and the human brain are quite different. Computers and mobile devices have powerful computational ability which is capable of complex computing tasks such as data analysis or computer graphics. But neurons are only responsible for a single operation which is passing the signals to the others. Thirdly, there are no obvious global functions shown on the internet while diverse functions and behaviors are emerged in the human brain. Finally, the physical structure and the connectivity of the internet are stable, but are dynamic relatively in the human brain due to learning or aging.

**Table 1.** The comparison of the internet and the human brain

	<b>Internet</b>	<b>Human Brain</b>
Scale	Millions of unit elements	$10^{11}$ unit elements
Layered structure	OSI model	Anatomical structure, network overlay, and functional outputs
Mechanisms of fault tolerance	Error correction, recomputation of routing pathway	Degeneracy mechanism, replaceable functional area
Properties of complex networks	Motif, communities, hubs, shortest path way, etc.	Motif, communities, hubs, shortest path way, etc.
Capability of an unit element	Versatile	Specific
Physical structure	Stable	Dynamic

### 3 Current Status of the Brain Simulator

In this section, first we introduce two state-of-the-art brain simulator projects, IBM's C2 and Blue Brain [18-21], and then briefly introduce our proposed brain simulator. Finally, a comparison is presented. C2 and Blue Brain, like other ongoing projects, try to create a synthetic brain in the molecular level. A bottom-up approach is adopted to design their brain simulators. C2 has built up 1 billion neurons connected with by 10 trillion synapses, and 10,000 neurons connected with by 100 million synapses are re-created in Blue Brain. IBM's Blue Gene supercomputer supports the huge demands of computation requirement. They develop a brain simulator in modeling the cortical area, not the whole brain structure and functionality. Therefore, current brain simulators development is still in infant stage.

In contrast, our proposed brain simulator takes whole brain networks, neurotransmitters, brain behaviors and brain diseases into account. Then, an extensible brain simulator development framework is proposed from information technology perspective. We adopt overlay and peer-to-peer network to deal with the complexity of brain network. Moreover, layered design with object-oriented brain class hierarchy form the flexible development framework of the proposed simulator. The proposed brain simulator is evolved in case-based incremental delivery style. While many computer simulations have attempted to work in "brain-like" computation or mimic parts of brain area, our proposed brain simulator is considering the whole brain. The comparison among C2, Blue Brain and our proposed brain simulator is shown in Table 2.

**Table 2.** The comparison among C2, Blue Brain and our propoed brain simulator

	IBM's C2	Blue Brain	Our Proposed Brain Simulator
Perspective	Neuron-Level Microscopic	Neuron-Level Microscopic	Brain-Level Macroscopic
Basic Component	Neuron	Neuron	Nuclei, Region, Tracts
Connection	Synapse	Synapse	Communication Pathway
Communication	Electrical Signal	Electrical Signal	Protocol Data Unit
Architecture	P2P Network	layered Architecture	layered Architecture
Focus Area	Cortex	Neocortical column	Whole Brain
Computation	Supercomputer Blue Gene	Supercomputer Blue Gene	Cloud Computing Environment
Granularity	Fine-grained	Fine-grained	Coarse-grained
Approach	Hardware	Hardware and Software	Software

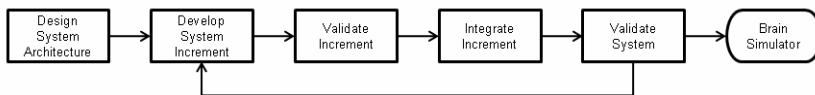
## 4 The Proposed Brain Simulator

Although the study of brain simulator has made significant progress in computation ability, there are still many challenges in building up brain simulator to overcome. The current research did not consider the dynamic structure and complexity of the human brain from the perspective of time and space. The overlay and P2P (peer-to-peer) network architecture have been successfully applied to dynamic and complex network in IT (Information Technology) field. Hence, our proposed brain simulator provides an overlay and P2P network to represent the complexity and dynamicity of brain in time and space. Moreover, it is a flexible architecture that diverse brain networks or models can be easily applied.

### 4.1 Case-Based Incremental Delivery Approach

Most brain simulators are designed merely for one purpose. Due to this reason, they lack of extensibility. Therefore, the scope of application and ability of brain simulator are limited and underestimated. Compared with other brain simulator, the proposed brain simulator is systematically developed in case-based incremental delivery approach, as shown in Figure 3.

First, the development of our brain simulator is based on some experimental data from related research. In this case, the proposed brain simulator can develop part of the required brain functionality as well as others. Rather than stop development, our brain simulator carry on evolving by later research cases. It is the important characteristic of brain simulator in terms of flexibility.

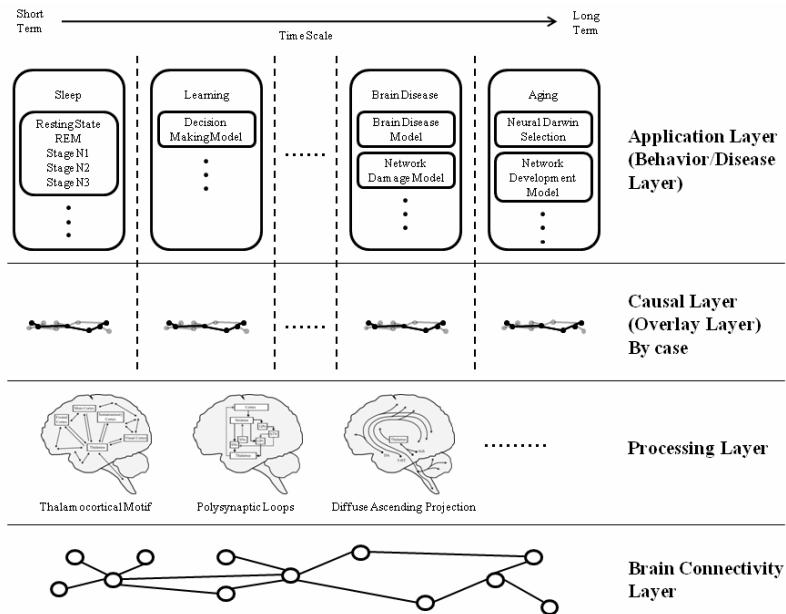


**Fig. 3.** Case-based Incremental Delivery Approach

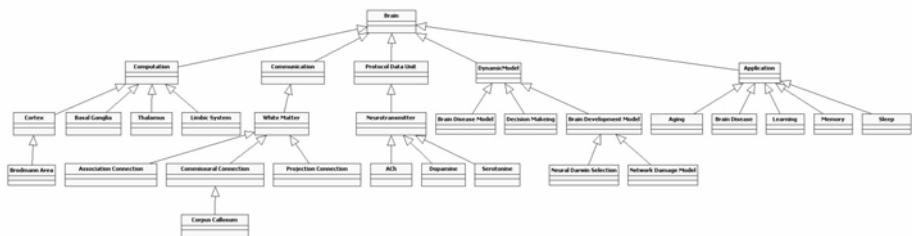
### 4.2 Architecture

A layered architecture is used to design our brain simulator according to our study on human brain, as shown in Figure 4. The layered architecture consists of the following layer:

- **Brain Connectivity Layer:** The lowest layer represents the brain connectivity of physical components. The first step of designing brain simulator is to define the physical computation unit and communication unit. Most researches are taken neurons as computation unit and synapses as communication unit from microscopic point of view. In contrast, we take macroscopic perspective. Considering the common attributes and functions among the components, a hierarchical representation of the brain components is presented by object-oriented design, as shown in Figure 5.



**Fig. 4.** Proposed layered Architecture



**Fig. 5.** Brain Component Class Diagram

The hierarchical structure consists of 5 main categories: computation, communication, protocol data unit, dynamic model and application. The computation units include Brodmann's areas, thalamus, and various nuclei. Tracts, such as corpus callosum, are defined as communication unit. Neurotransmitters are taken as protocol data unit. Dynamic model, such as brain disease model and brain development model, is represented as different theoretical brain models. Finally, the application category includes elements for diverse brain behavior and disease, like sleep, learning, disease and aging.

- **Processing Layer:** Compared the brain network perspective with other brain simulators, the significant improvement of our proposed one is that we consider the various brain networks as an integrated network. The processing layer composed of three brain networks, as we mentioned in section 2.3 Brain Network Fundamentals, represents this integrated network, as shown in Figure 4.

- **Causal Layer:** Causal layer is a layer that consists of diverse brain causal network arising from an evolution of brain applications, as shown in Figure 4. A causal network is an overlay network that builds on top of processing layer and brain connectivity layer. Different brain applications have their own causal networks. Means of brain networks analysis can be applied to this layer to see the causal effects of each brain application.
- **Application Layer:** Application layer, also called Behavior layer, is the top layer in the layered architecture. This layer represents various brain application scopes in different time scale, from long-term to short-term. Because the brain constantly changes but operates in the way to balance, each application, such as aging, brain diseases, learning, or sleep, has its own theoretical models to formulate the dynamicity of brain in time and space. For example, the model of neural Darwin selection focuses on the evolution of brain from the microscopic and macroscopic perspective and the network damage model considers the influence to brain networks from brain diseases. As for decision making model, the characteristics of excitation, inhibition and balance needs to be take into account. The purpose of designing the application layer is to provide a flexible strategy that can study diverse brain behaviors efficiently and effectively. Through the assistance of this layer, we can not only verify the accuracy of our proposed brain simulator comparing with the results from other neural imaging technologies. Moreover, we can better understand the evolution of various brain applications.

## 5 Conclusion and Future Work

In this paper, the importance of brain simulator is pointed out. A brain simulator models brain structure and functionality for better understanding how cognition works and rapidly testing their ideas on an accurate replica of the brain. Then, the current status of brain simulators and the comparison has been presented. The significant shortcomings of ongoing brain simulators are inflexible and maladaptive in time and space. Hence, an extensible brain simulator is proposed. It focuses on the dynamicity of brain in time and space. From the IT point of view, we adopt overlay and peer-to-peer network to deal with the complexity of brain network. Layered design with object-oriented brain class hierarchy forms the flexible development framework of the proposed simulator. Furthermore, the proposed brain simulator is systematically developed in a case-based incremental delivery approach. The proposed brain simulator will be evolved with more research cases from cognitive and clinical neuroscience. These characteristics benefit brain simulator research field significantly.

The development of proposed brain simulator is still ongoing. In order to simulate more accurate human brain structure and function, future research is required. The object-oriented design of brain components has to be refined for better brain structure. The common and specific attributes and functions of different brain components needs further consideration. Then, the theoretical models in different brain applications have to be further studied and implemented into the brain simulator. Decision making is one of the important functions of brain and it is also a complex and difficult design in brain simulator. The achievements in autonomous agent research, such as multi-agent system, game theory, and fuzzy theory provide possibilities for the decision making issue.

## References

1. Zhong, N., Li, K., Lu, S., Chen, L. (eds.): *Brain Informatics*. Springer, Heidelberg (2010)
2. Sendhoff, B., Körner, E., Sporns, O., Ritter, H., Doya, K. (eds.): *Creating Brain-Like Intelligence: From Basic Principles to Complex Intelligent Systems*. Springer, Heidelberg (2009)
3. Catani, M., Ffytche, D.H.: The Rises and Falls of Disconnection Syndromes. *J. Brain* 128, 2224–2239 (2005)
4. Catani, M., Ffytche, D.H.: Graph Theoretical Analysis of Magnetoencephalographic Functional Connectivity in Alzheimer's Disease. *J. Brain* 132, 213–224 (2009)
5. Deuschl, G., et al.: Deep-Brain Stimulation for Parkinson's Disease. *J. Neurology* 249, 36–39 (2002)
6. Tononi, G., Edelman, G.M., Sporns, O.: Complexity and Coherency: Integrating Information in the Brain. *J. Trends in Cognitive Science* 2, 474–484 (1998)
7. Strogatz, S.H.: Exploring complex network. *J. Nature* 410, 268–276 (2001)
8. Bullmore, E., Sporns, O.: Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems. *J. Nat. Rev. Neurosci.* 10, 186–198 (2009)
9. Sporns, O., Chialvo, D.R., Kaiser, M., Hilgetag, C.C.: Organization, Development and Function of Complex Brain Networks. *J. Trends in Cognitive Science* 8, 418–425 (2004)
10. Rubinov, M., Sporns, O.: Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* (2009)
11. NIH Launches the Human Connectome Project to Unravel the Brain Connections,  
<http://www.nih.gov/news/health/jul2009/ninds-15.htm>
12. Human Connectome Project, <http://www.humanconnectomeproject.org/>
13. Sporns, O., Tononi, G., Kotter, R.: The Human Connectome: A Structural Description of the Human Brain. *J. PLoS Comput. Biol.* 1(4), e42 (2005)
14. Edelman, G.M.: *Wider than the Sky: The Phenomenal Gift of Consciousness*. Yale University Press, New Haven (2004)
15. Thalamocortical Radiations,  
[http://en.wikipedia.org/wiki/Thalamocortical\\_radiations](http://en.wikipedia.org/wiki/Thalamocortical_radiations)
16. Basal Ganglia, [http://en.wikipedia.org/wiki/Basal\\_ganglia](http://en.wikipedia.org/wiki/Basal_ganglia)
17. Neurotransmitter, <http://en.wikipedia.org/wiki/Neurotransmitter>
18. IBM Unveils a New Brain Simulator,  
<http://spectrum.ieee.org/computing/hardware/ibm-unveils-a-new-brain-simulator>
19. Blue Brain Project, <http://bluebrain.epfl.ch/>
20. Markram, H.: The Blue Brain Project. *J. Nat. Rev. Neurosci.* 7, 153–160 (2006)
21. King, J.G., et al.: A Component-based Extension Framework for Large-scale Parallel Simulations in NEURON. *J. Frontiers in Neuroinformatics* 3 (2009)

# Obstructive Sleep Apnea Diagnosis from Electroencephalography Frequency Variation by Radial Basis Function Neural Network

Chien-Chang Hsu and Jie Yu

Department of Computer Science and Information Engineering,  
Fu-Jen Catholic University,

510 Chung Cheng Rd., Hsinchuang, Taipei, Taiwan 242  
cch@csie.fju.edu.tw, jhtsay96@csie.fju.edu.tw

**Abstract.** This paper proposes an obstructive sleep apnea diagnosis system based on electroencephalography frequency variations. The system uses a band-pass filter to remove extremely low and high frequency in brainwave. The system then uses baseline correction and the Hilbert-Huang transform to extract the features from the filtered signals. Moreover, the system uses a radial basis function neural network to diagnose the kind of obstructive sleep apnea from electroencephalography. Experimental results show that the system can achieve over 96%, 92%, and 97% accuracy for obstructive sleep apnea, Obstructive sleep apnea with arousal, and arousal. The system provides a feasible way for the technicians of sleep center to interpret the EEG signal easily and completely.

**Keywords:** Obstructive sleep apnea, electroencephalography, frequency variation, Radial basis function neural network.

## 1 Introduction

There are over four hundred thousand sleep disorder patients in Taiwan. The sleep disorder not only influences the sleep quality of these patients but also endanger their lives. They cannot focus or concentration when they are working or driving. As a consequence, brainwave becomes one of the most popular diagnosis tools in sleep disorder [1,2]. When obstructive sleep apnea (OSA) occurs during sleep, it may cause suspensions in the patient's breathing. Moreover, a number of short arousals may also occur [11,12,14,15,17]. Presently polysomnography (PSG) is normally used in OSA diagnosis to detect the patient's physiology sign. Electroencephalogram (EEG), electromyogram (EMG), and electrooculogram (EOG) records are usually examined to identify the different sleep stages. Brainwave analysis can provide an abundance of information in relation to breathing disorders during sleep. Therefore, examination of the brainwaves is an important part in OSA diagnosis. E. Estrada used a band-pass filter to retain 0.5~50Hz brainwave signals and defined the brainwave signals captured every 30 seconds as one algorithm window in his analysis of harmonic

parameters to conduct automatic sleep scoring [3]. He also suggested there were similarities in measurement of EOG signals and EEG signals during various stages of sleep. Using the signals captured every 30 seconds in various sleep stages as one algorithm window, he removed the high frequency signals with a Chebyshev filter, an 8th-order low-pass filter. T. P. Exarchos adopted 355 microseconds as a time interval to detect epileptic surge waves to mine the brainwave signal characteristics of epilepsy [4]. He calculated the standard deviation, average gradient, duration and sharpness of each interval, as well as the power spectral density of 10 frequency domains and used them as the eigenvectors. O. Shmiel came up with a method of research on sleep and brainwaves by using EEG, pulse and blood oxygenation level signals [15]. Initially, 5 seconds was adopted as a window and the series of signals corresponding to the whole night was established. Each window was run through the fast Fourier transform to obtain the spectral power. The pulse signals were also arranged in sequence and expressed in interquartile range to exhibit the degree of dispersion, while the blood oxygenation level data were quantified in equal widths. Finally, Data showing blood oxygenation level below 90%, the Alpha waves transformed from Theta waves and the maximum value of each interquartile were taken as the characteristics.

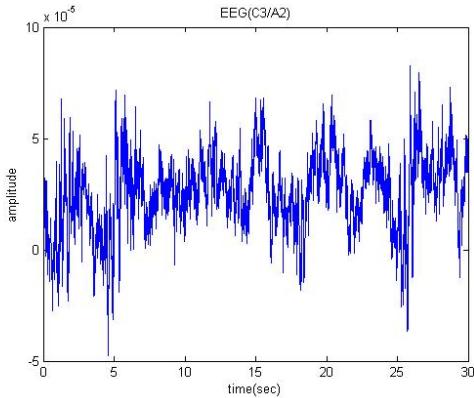
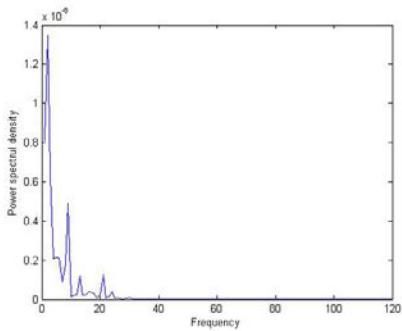
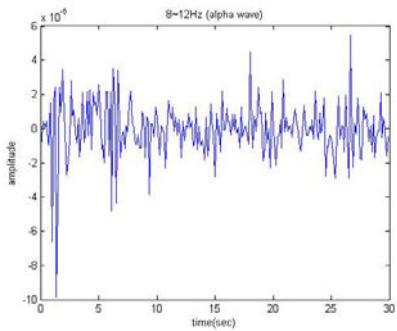
Generally, fast Fourier transformation and wavelet transformation are two major methods for EEG signal processing [6,7,16]. The fast Fourier transformation converts time domain signals into frequency domain signals. Fourier transform is applied to convert time domain signals into frequency domain signals ( $P_{DF}$ ):

$$P_{DF} = \frac{1}{2\pi} \sum_{k=0}^{N_0-1} F^2(k\omega_0) \times \frac{2\pi}{N_0} \quad (1)$$

where  $k$ ,  $N_0$ ,  $F$ , and  $\omega_0$  respectively stand for the initial sequence, signal length, input signal, and angle. It requires division of these values by  $2\pi$ . It is a tendency of loss of spectral energy and lacks the localization function in the time domain. Blind spots therefore exist in determination of OSA time (Fig. 1 and 2). When using the shift function of the wavelet transformation, the signals released at different time points ( $W_{f(a,b)}$ ) can be analyzed.

$$W_{f(a,b)} = \frac{1}{\sqrt{|a|}} \times \int_{-\infty}^{\infty} f(t) \phi\left(\frac{t-b}{a}\right) dt, \quad a, b \in \mathbb{R} \text{ and } a \neq 0 \quad (2)$$

where  $f(t)$ ,  $\phi(t)$ ,  $a$  and  $b$  represent the input signal, wavelet basis function, flexible variant, and shift variant. It may help in analysis of signals released at different time points. But after the wavelet transformation is applied, the numbers of data points of the original signals are compressed and a large amount of data is lost. This apparently brings deterioration in resolution and some data are lost. Therefore, designing a signal processing approach with the localization function as well as high resolution to overcome the aforesaid problems is an important issue in analysis of brainwave signals. Moreover, in brainwave signal processing, lacking an effective method to locate brainwave features will reduce the diagnosis precision.

**Fig. 1.** Raw EEG signal**Fig. 2.** Fast Fourier transform of EEG signal**Fig. 3.** Wavelet transformation of EEG signal

This paper proposes an OSA diagnosis system based on brainwave frequency variation and radial basis function (RBF) neural network. It includes two modules: EEG signal preprocessor and OSA analyzer. The former uses filters and Hilbert-Huang transformation to extract the features from the transformed signal. The latter uses a RBF neural network to diagnose the occurrence of OSA.

The structure of this paper is organized as follows: Section 2 introduces the system architecture. Section 3 verifies the system about 12 cases from Shin-Kong Wu Ho-Su memories hospital in Taiwan. Section 4 concludes the work.

## 2 System Architecture

Figure 4 shows the architecture of the OSA diagnosis system. The system contains two modules: EEG preprocessor and OSA analyzer. The former uses bands pass filter to

eliminate irrelevant EEG signals. Baseline correction and Hilber-Huang transformation are conducted to locate OSA characteristics. The latter then uses RBF neural network to diagnose the OSA of EEG signal.



**Fig. 4.** System architecture

Specifically, the EEG preprocessor performs signal filtering, signal baseline correction, signal conversion, and feature extraction. Throughout the patient's overnight sleep, every 92 seconds of brainwaves signals during OSA is marked as a time window, that is, 31 seconds before and 61 seconds during OSA. The signal filtering removes the extremely low and high frequencies to keep sleep related brainwaves by a 0.5~32Hz band-pass filter. It then filters the signals twice to rectify the signal phase shift phenomenon that occurs after the first filtering [8]. In other words, the output  $y(n)$  from the first filtering of the original signal is filtered again to acquire  $y'(n)$ .

$$y'[n] = y[L_y - n] \quad (3)$$

$$y[n] = h_0x[n] + h_1x[n-1] + \dots + h_kx[n-k] \quad (4)$$

where  $L_y$ ,  $h_k$ , and  $k$  are the length of signal  $y(n)$ , coefficient, and order of the filter. Baseline correction subtracts the fluctuating baseline from the brainwave signals to acquire the EEG signal  $y_0(n)$ .

$$y_o[n] = y'[n] - y_{base}[n] \quad (5)$$

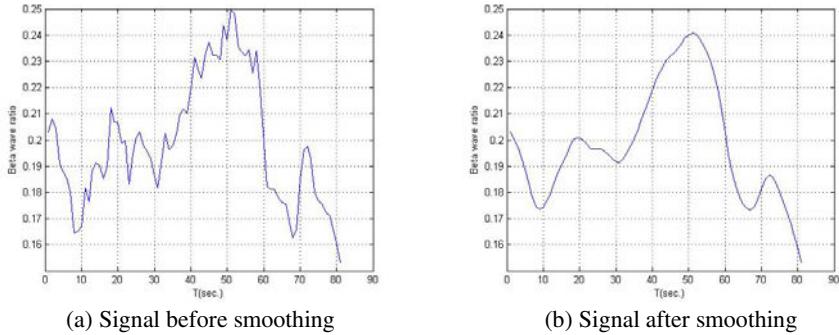
where  $y_{base}[n]$  is the fluctuating baseline [Shih10]. Signal conversion computes the ratio of Delta, Theta, Alpha, Sigma and Beta signal in Hilbert-Huang transformation [Shar08]. Take  $\beta_{ratio}[t]$  as an example:

$$\beta_{ratio}[t] = \frac{P_\beta[t]}{P_\delta[t] + P_\theta[t] + P_\alpha[t] + P_\Sigma[t] + P_\beta[t]} \quad (6)$$

where  $P_\delta[t]$ ,  $P_\theta[t]$ ,  $P_\alpha[t]$ ,  $P_\Sigma[t]$ , and  $P_\beta[t]$  are the total intensity of the moving frequencies of the five bands every ten seconds respectively. The calculation of the total intensity of the moving frequencies every ten seconds, using  $P_\beta[t]$  as an example:

$$P_\beta[t] = \sum_{i=t-9}^t \beta[i] \quad (7)$$

where  $\beta[i]$  is the signal intensity of Beta waves in 10 second. The signal conversion then smoothes the signal by five-point median FIR filters. It can reduce the computation errors resulted from the closeness of the peaks or valleys of the signal.

**Fig. 5.** Beta Wave Smoothing

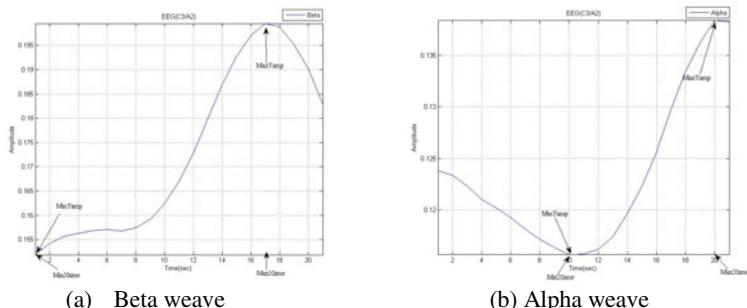
Feature extraction calculates the duration ( $D_i$ ), variation ( $A_i$ ) and slop ( $S_i$ ) of Delta, Theta, Alpha, and Beta weave from Hilbert spectrum [9,10,13].

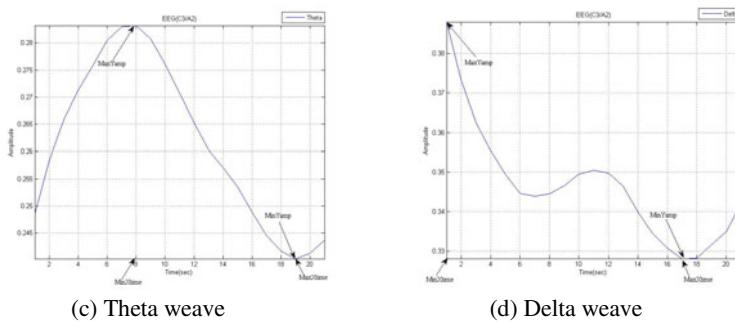
$$D_i = |MaxX_{time} - MinX_{time}| \quad (8)$$

$$S_i = \frac{MaxY_{amp} - MinY_{amp}}{MaxX_{time} - MinX_{time}} * 100 \quad (9)$$

$$A_i = \left| \frac{MinY_{amp} - MaxY_{amp}}{MaxY_{amp}} \right| * 100 \quad (10)$$

where  $MaxY_{amp}$ ,  $MinY_{amp}$ ,  $MaxX_{time}$ , and  $MinX_{time}$  are the maximum and minimum amplitude values and their corresponding time points in the in the Hilbert spectrum frequency (Fig. 6). OSA analyzer is responsible to diagnose OSA from the above features. It uses RBF neural network [5] as classifier to distinguish the input value (Fig. 7). The input layer contains 10 nodes (Table 1). The patients are classified into three classes: OSA, OSA with arousal, and arousal only. The Gaussian function is chosen as basis function.

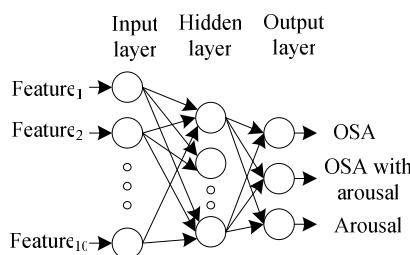
**Fig. 6.** Parameters of duration, variation, and slop computation



**Fig. 6.** (continued)

**Table 1.** Node name of input layer

Feature name	Description
D	Duration time of OSA
S	Sleep stage during OSA
$A_\delta$	Delta weave variation during OSA
$A_\theta$	Theta weave variation during OSA
$A_\alpha$	Alpha weave variation during OSA
$A_\beta$	Beta weave variation during OSA
$S_\delta$	Delta weave slop during OSA
$S_\theta$	Theta weave slop during OSA
$S_\alpha$	Alpha weave slop during OSA
$S_\beta$	Beta weave slop during OSA



**Fig. 7.** RBF neural network

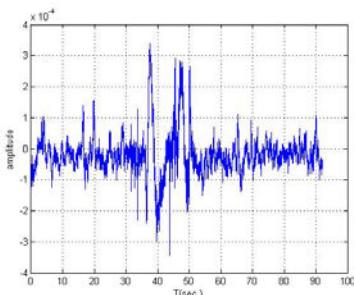
### 3 Experiments

The data presented in this experiment are the overnight sleep C3/A2 EEG signals of OSA patients at the sleep center of Shin Kong Wu Ho-Su Memorial Hospital in

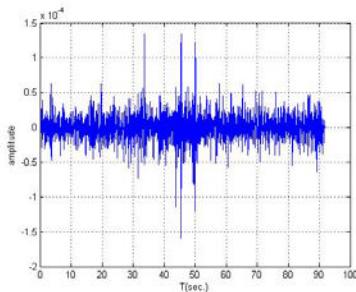
Taiwan. The age of patients are between 40 and 72 years old (Table 2). There are 4,339 data record in total, containing three target attributes and ten continuous-valued attributes. The target attributes are classified into three categories: OSA, OSA with arousal, and arousal. The signal preprocessor filters and transforms the brainwave signals of OSA patients marked out by sleep technologists and extracts the characteristics. Figures 8, 9, and 10 show the original OSA EEG signals, OSA EEG signals after band pass filtering, and weave band of beta, alpha, theta, and delta from Hilbert transformation. Table 3 shows the example data of each feature.

**Table 2.** Experiment data

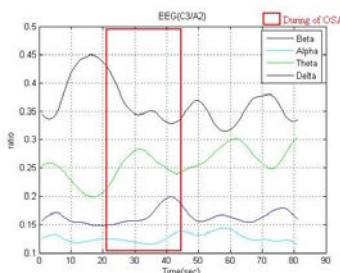
Case No.	Age	RDI	OSA#	OSA with arousal #	Arousal #
1	52	34.9	198	58	166
2	53	33.7	36	12	142
3	58	36.3	134	29	218
4	40	75.8	446	149	308
5	72	47.8	57	8	102
6	48	31.8	176	68	204
7	51	47.2	142	91	180
8	58	30.4	92	73	105
9	60	45.4	78	8	210
10	31	63.2	281	98	164
11	66	61.8	89	2	215



**Fig. 8.** Original OSA EEG Signals



**Fig. 9.** OSA EEG Signals after band pass filtering



**Fig. 10.** Beta, alpha, theta, and delta weave data

**Table 3.** Example data of each input feature

Feature name	D	S	$A_\delta$	$A_\theta$	$A_\alpha$	$A_\beta$	$S_\delta$	$S_\theta$	$S_\alpha$	$S_\beta$
Value	25.1	2	-23.98	33.62	19.68	34.26	-0.52	0.65	0.23	0.25

The OSA analyzer uses RBF neural network to classify the patient data. The hidden and output layer use k-means and least average equal algorithm as the learning algorithm. The total number of patient data is 4339 (Table 2). The RBF neural network is trained by 3471 records from the 13 patients with OSA. It includes 1383, 477, and 1611 records of OSA, OSA with arousal, and arousal. The accuracy rate of each type is 96%, 92%, and 97%. We also use the backpropagation neural network (BPN) to training the same data. The accuracy rate of BPN is 86%, 64%, and 93%. The accuracy of RBF neural network is superior to BPN.

## 4 Conclusions

This paper proposes an OSA diagnosis system by EEG frequency variation and RBF neural network. The system includes two modules: EEG preprocessor and OSA analyzer. The main function of the EEG signal preprocessor is to process the noises in the OSA, OSA with arousal and arousal brainwave signals. A filter is applied to remove the extremely low and high frequencies in the brainwave signals. Also, baseline correction is conducted to eliminate interfering artifacts. The Hilbert-Huang transformation is used to calculate the Hilbert spectrum, find the relative density of each frequency band, and extract the features. The OSA analyzer uses RBF neural network to classify the OSA and arousal.

The contributions of this paper are the use of the Hilbert-Huang transformation to extract features from OSA brainwave signals and RBF neural network to classify different type of OSA data. In this transformation process, the empirical mode decomposition is adopted to extend the original signal data into several intrinsic mode functions to render the signals non-linear or non-stationary. The basis will be able to exhibit the physical characteristics of the original signals. Since sleep brainwave signals can be easily interfered by other signals and short unknown brainwave changes can occur, the total of 10-second mobile frequency intensity of each band during OSA and the relative density of the frequency of each band are calculated to extract the characteristics. At the same time, RBF neural network is also used for data classification to diagnose different type of OSA condition.

**Acknowledgments.** We would like to thank Chia-Mo Lin M. D. and Hou-Chang Chiu M. D., Shin Kong Wu Ho-Su Memorial Hospital, Taiwan, for their professional consultation in medicine. This work is partly supported by National Science Council of ROC under grants NSC 99-2220-E-030-001 and NSC 98-2220-E-030-002.

## References

1. Al-ani, T., Karmakar, C.K., Khandoker, A.H., Palaniswami, M.: Automatic Recognition of Obstructive Sleep Apnoea Syndrome Using Power Spectral Analysis of Electrocardiogram and Hidden Markov Models. In: IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 28–290. IEEE Press, New York (2008)
2. Cho, S.P., Lee, J., Park, H.D., Lee, K.J.: Detection of Arousals in Patients with Respiratory Sleep Disorders Using a Single Channel EEG. In: 27th IEEE Conference on Engineering in Medicine and Biology Annual, pp. 2733–2735. IEEE Press, New York (2005)
3. Estrada, E., Nazeran, H., Nava, P., Behbehani, K., Burk, J., Lucas, E.: Itakura Distance: A Useful Similarity Measure between EEG and EOG Signals in Computer-aided Classification of Sleep Stages. In: 27th IEEE Conference on Engineering in Medicine and Biology Society, pp. 1189–1192. IEEE Press, New York (2005)
4. Exarchos, T.P., Tzallas, A.T., Fotiadis, D.I., Konitsiotis, S., Giannopoulos, S.: EEG Transient Event Detection and Classification Using Association Rules. *IEEE Transactions on Information Technology in Biomedicine* 10(3), 451–457 (2006)
5. Fu, X., Wang, L.: Rule Extraction by Genetic Algorithms based on a Simplified RBF Neural Network. In: Congress on Evolutionary Computation, vol. 2, pp. 75–758 (2001)
6. Garrett, D., Peterson, D.A., Anderson, C.W., Thaut, M.H.: Comparison of Linear, Nonlinear, and Feature Selection Methods for EEG Signal Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 11(2), 141–144 (2003)
7. Hese, V., Philips, P., Koninck, W.D., Walle, J.V.D., Lemahieu, R., Elis, I.: Automatic Detection of Sleep Stages Using the EEG. In: 23rd IEEE International Conference on Engineering in Medicine and Biology Society, vol. 2, pp. 1944–1947. IEEE Press, New York (2001)
8. Hsu, C.C., Shih, P.T.: An Intelligent Sleep Apnea Detection System. In: 9th International Conference on Machine Learning and Cybernetics, pp. 3230–3233. IEEE Press, New York (2010)
9. Hu, M., Li, G., Ding, Q.P., Li, J.J.: Classification of Normal and Hypoxia EEG Based on Hilbert Huang Transform. In: IEEE Conference on Engineering in Neural Networks and Brain, pp. 851–854. IEEE Press, New York (2005)
10. Huang, N.E.: The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis. Royal Society on Mathematical, Physical and Engineering Sciences 454(1971), 903–995 (1998)
11. Karunajeewa, A.S., Abeyratne, U.R., Rathnayake, S.I., Swarnkar, V.: Dynamic Data Analysis in Obstructive Sleep Apnea. In: 28th IEEE Conference on Engineering in Medicine and Biology Society, pp. 4510–4513. IEEE Press, New York (2006)
12. Khandoker, A.H., Karmakar, C.K., Palaniswami, M.: Analysis of Coherence between Sleep EEG and ECG Signals During and After Obstructive Sleep Apnea Events. In: 30th IEEE Conference on Engineering in Medicine and Biology Society, pp. 3876–3879. IEEE Press, New York (2008)
13. Li, Y., Yingle, F., Gu, L., Qinye, T.: Sleep Stage Classification based on EEG Hilbert-Huang Transform. In: IEEE Conference on Industrial Electronics and Applications, pp. 3676–3681. IEEE Press, New York (2009)
14. Penzel, T., Lo, C.C., Ivanov, P.C., Kesper, K., Becker, H.F., Vogelmeier, C.: Analysis of Sleep Fragmentation and Sleep Structure in Patients With Sleep Apnea and Normal Volunteers. In: 27th IEEE Conference on Engineering in Medicine and Biology Society, pp. 2591–2594. IEEE Press, New York (2005)

15. Shmiel, O., Shmiel, T., Dagan, Y., Teicher, M.: Processing of Multichannel Recordings for Data-Mining Algorithms. *IEEE Transactions on Biomedical Engineering* 54(3), 444–453 (2007)
16. Wan, B., Dhakal, B., Qi, H., Shu, X.: Multi-method synthesizing to detect and classify epileptic waves in EEG. In: 4th IEEE Conference on Engineering in Computer and Information Technology, pp. 922–926. IEEE Press, New York (2004)
17. Xavier, P., Behbehani, K., Watenpaugh, D., Burk, J.R.: Detecting Electroencephalography Variations Due to Sleep Disordered Breathing Events. In: 29th IEEE Conference on Engineering in Medicine and Biology Society, pp. 6097–6100. IEEE Press, New York (2007)

# Authentication and Protection for Medical Image

Chih-Hung Lin<sup>1</sup>, Ching-Yu Yang<sup>2</sup>, and Chia-Wei Chang<sup>1</sup>

<sup>1</sup> Dept. of Computer Science and Information Engineering  
Southern Taiwan University  
No.1, Nantai St., Yongkang City, Tainan County, 710 Taiwan  
[{chuck,M98G0216}@mail.stut.edu.tw](mailto:{chuck,M98G0216}@mail.stut.edu.tw)

<sup>2</sup> Dept. of Computer Science and Information Engineering  
National Penghu University of Science and Technology  
No. 300, Liu-Ho Rd., Magong, Penghu, 880 Taiwan  
[chingyu@npu.edu.tw](mailto:chingyu@npu.edu.tw)

**Abstract.** This paper proposes a method to authenticate and protect medical images, especially in the region of interest (ROI). The ROI of a medical image is an important area during diagnosis and must not be distorted during transmission and storage in the hospital information system. The rest of the ROI is used for embedding a watermark that contains the patients data and authentication information, and the authentication information is generated from the ROI by analyzing wavelet coefficients with singular value decomposition (SVD), before embedding the watermark into the discrete wavelet transform (DWT) sub-band. It is important that the ratio of ROI and non-ROI areas is consistent between different systems and doctors, and this ratio is analyzed in this paper. The ROI of watermarked medical image is fragile to any distortion, and patients data and authentication information can be easily extracted from non-ROI. The effectiveness of the new approach is demonstrated empirically.

**Keywords:** Medical image, Watermarking, Singular value decomposition, Image authentication, JPEG compression.

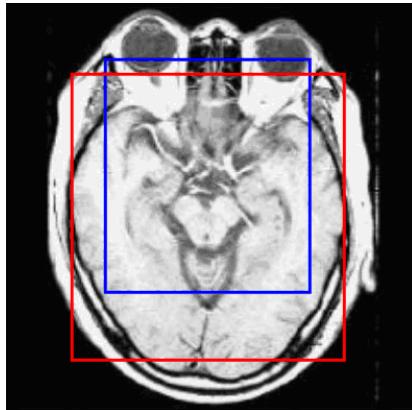
## 1 Introduction

The content of a digital image is easily accessed and modified whilst being transferred via electronic communication and information systems. Therefore, image authentication and protection have become an important issue in recent years. The related schemes can be classified into two categories: digital signature-based [1,2] and watermark-based [3,4] methods. A digital signature can be generated

from a digital media file, and is encrypted as an extra file appended to the media and to be used later for authentication. Watermarking, however, is a copyright protection method, which embeds invisible information into the media. The embedded watermark can then be extracted and used for verification. Both digital signature-based and watermark-based methods are further classified into three broad categories according to the intended purpose: robust, fragile and semi-fragile. Currently, robust watermarking schemes [5,6] are generally used for copyright protection and ownership verification and are tolerated by most image processing operations. Fragile authentication schemes [7] are mainly applied to content authentication and integrity identification because they can be destroyed by common attacks, in other words, they are fragile to any kind of image modification. However, the semi-fragile image authentication scheme [8] is different from the above methods, since it allows non-malicious modification but prevents malicious modification. Generally, images may be unavoidably manipulated by some incidental processes, such as compression. However, some methods are multipurpose, and for example can achieve copyright protection (robust) whilst distinguishing between malicious and non-malicious modification (semi-fragile) simultaneously [9].

Based on the latest technological advances for protecting and authenticating digital images, and the rapid growth of biomedical engineering, the watermarking method has been recently applied to medical images. In the past, many technologies for hiding data have been applied to medical images [10]. However, since advances in data hiding technology have focused on increasing the capacity of embedded information, the security issues have been neglected. This paper adopts the watermarking method for medical image processing, and considers integrity verification and copyright protection. Additionally, the definition of the region of interest (ROI) is considered in this study. In [11], Shih and Wu proposed a robust technique for embedding a watermark containing signature information or textural data around the ROI of a medical image based on genetic algorithms. However, the location of the ROI is decided by the doctor or diagnostician and is therefore subjective. For example, in the magnetic resonance imaging (MRI) brain image shown in Fig. 1 (image source from [11]), doctor A might regard the ROI as the part marked by the blue rectangle, and will then embed the watermark around that ROI. Subsequently, the patient and his medical anamnesis might be transferred to another doctor B who works in another hospital, and who regards the ROI as being the part marked by the red rectangle. If the hospital information systems adopted between doctor A and doctor B are different, then doctor B cannot extract the watermark containing signature information or textural data. Such ambiguous cases are often encountered for these methods, because there is no consistent ratio of location for the ROI and non-ROI areas.

For the reasons mentioned above, we set a consistent ratio of ROI and non-ROI areas for all medical images among different hospital information systems. The image features of the ROI were computed by SVD analysis, and this feature was embedded into the non-ROI along with the patient data. To protect the image content of the ROI, the ROI must not be distorted, and therefore the proposed



**Fig. 1.** MRI brain image

system will locate the modified parts for authentication prior to extraction of the patients data.

The rest of this paper is organized as follows. Section 2 introduces the proposed method, including the method for generating authentication information, the watermark (including authentication information and the patients data) embedding method, and the authentication procedure. Section 3 shows the experimental results and conclusions are drawn in Section 4.

## 2 The Proposed Method

The ROI of a medical image is regarded as the most important and significant area, and is strongly related to the judgment of diagnosis. Therefore, the content of the ROI has to be preserved and protected completely. Moreover, the medical image and patients data are often transmitted to another hospital along with the patient. To preserve the main properties of the ROI, a simple SVD analysis is adopted to generate the features of the ROI, which are then embedded into the non-ROI together with the patients data. In this way a watermarked medical image is generated. The ROI of the watermarked medical image is fragile for slight modification and can be located in the modified parts by extracting the features embedded in the non-ROI. Also, the area ratio between the ROI and non-ROI is constant, avoiding the ambiguous ROI defined by different doctors. The steps introduced above are described in more detail below.

### 2.1 Feature Generation

Let  $I$  be an original medical image with size  $w \times h$  pixels, and  $I^R$  and  $I^{NR}$  denote the ROI and non-ROI regions, respectively. Let  $r$  be an area ratio between  $I^R$

and  $I^{NR}$ , such that  $r = I^R/I^{NR}$ , therefore,

$$I = I^R + I^{NR} = (1 + r)I^{NR} = \left(1 + \frac{1}{r}\right)I^R \quad (1)$$

The calculation of  $r$  is shown in section 2.4. The algorithm for feature generation is as follows.

- Step 1.  $I^R$  is implemented at the 1-level DWT and  $I^{R,LL}$  denotes the LL sub-band of  $I^R$ .
- Step 2.  $I^{R,LL}$  is divided into many non-overlapping blocks of size  $b^R \times b^R$  pixels, and each block is denoted as  $I_m^{R,LL}$ .
- Step 3.  $I_m^{R,LL}$  is decomposed with singular value decomposition according to:

$$SVD(I_m^{R,LL}) = U_m^{R,LL} S_m^{R,LL} V_m^{R,LL} \quad (2)$$

where  $SVD()$  is a singular value decomposition function, and  $U_m^{R,LL}$  and  $V_m^{R,LL}$  are orthogonal matrices. The  $S_m^{R,LL}$  matrix is diagonal and has the singular values  $s_{m,p}^{R,LL}$ , where  $1 \leq p \leq b^R$ .

- Step 4. All the  $s_{m,p}^{R,LL}$  values are cascaded as a feature of  $I^R$ , denoted  $FI^R$ , and then embedded with the patient's data into  $I^{NR}$  together, as described in section 2.2.

## 2.2 Watermark Embedding

The patient's data is a binary image with size  $w^{pa} \times h^{pa}$  pixels, and is cascaded with  $FI^R$  as a watermark and embedded into  $I^{NR}$ . The embedding algorithm referred to in [9] is described as follows.

- Step 1. The result of cascading the patient's data with  $FI^R$  is scrambled with a secret scramble key, which generates the scrambled bit stream regarded as the watermark, denoted as  $W = \{w_1, w_2, \dots, w_k\}$ , where  $w_k$  is a binary bit, 0 or 1.
- Step 2.  $I^{NR}$  is divided into many non-overlapping blocks of size  $b^{NR} \times b^{NR}$  pixels and each block is denoted as  $I_n^{NR}$ .
- Step 3.  $I_n^{NR}$  is implemented at the 1-level DWT, and  $I_n^{NR,LH}$  and  $I_n^{NR,HL}$  denote the LH and HL sub-bands of  $I_n^{NR}$ , respectively.
- Step 4. The watermark is embedded into  $I_n^{NR,LH}$  and  $I_n^{NR,HL}$ . Assume the watermark bit  $w_k$  is embedded into  $I_{n,l}^{NR,LH}$ , where  $I_{n,l}^{NR,LH}$  denotes the  $l$ th wavelet coefficient in  $I_n^{NR,LH}$ , then  $I_n^{NR,LH}$  is quantized by a quantization function  $Q()$ , with a quantization  $q$ . Whether values of interval  $Q(I_{n,l}^{NR,LH}, q)$  are 0 or 1 is decided by the quantization result as follows,

$$Q(I_{n,l}^{NR,LH}, q) = \begin{cases} 0, & \text{if } t \times q \leq I_{n,l}^{NR,LH} < (t+1) \times q, \text{ for } t = 0, \pm 2, \pm 4 \dots \\ 1, & \text{if } t \times q \leq I_{n,l}^{NR,LH} < (t+1) \times q, \text{ for } t = \pm 1, \pm 3 \dots \end{cases} \quad (3)$$

Step 5. The quantization error  $e_{n,l}^{NR,LH}$  that is adopted for calculating the new wavelet coefficient  $I_{n,l}^{'NR,LH}$  is given by:

$$\Delta I_{n,l}^{NR,LH} = \lfloor \frac{I_{n,l}^{NR,LH}}{q} \rfloor \times q \quad (4)$$

$$e_{n,l}^{NR,LH} = I_{n,l}^{NR,LH} - \Delta I_{n,l}^{NR,LH} \quad (5)$$

Then the new wavelet coefficient  $I_{n,l}^{'NR,LH}$  is computed as

$$I_{n,l}^{'NR,LH} = \begin{cases} \Delta I_{n,l}^{NR,LH} + \frac{1}{2}q, & \text{if } Q(I_{n,l}^{NR,LH}, q) = w_k \\ \Delta I_{n,l}^{NR,LH} + \frac{3}{2}q, & \text{if } Q(I_{n,l}^{NR,LH}, q) \neq w_k \text{ and } e_{n,l}^{NR,LH} > \frac{1}{2}q \\ \Delta I_{n,l}^{NR,LH} - \frac{1}{2}q, & \text{if } Q(I_{n,l}^{NR,LH}, q) \neq w_k \text{ and } e_{n,l}^{NR,LH} \leq \frac{1}{2}q \end{cases} \quad (6)$$

The part of the algorithm where the watermark bit  $w_k$  is embedded into  $I_{n,l}^{NR,LH}$  is similar to Steps 4 and Step 5 in Section 2.2. Therefore, the watermark containing features of  $I^R$  and the patients data is embedded in a circle around  $I^{NR}$ , from the inside to the outside as shown in Fig. 2.

1	2	3	4	5	6	7	8
28	29	30	31	32	33	34	9
27	48				35	10	
26	47				36	11	
25	46				37	12	
24	45				38	13	
23	44	43	42	41	40	39	14
22	21	20	19	18	17	16	15

**Fig. 2.** The embedding order in  $I^{NR}$

### 2.3 Authentication

Let  $AI$  denote the medical image authenticated by the system, and let the ROI and non-ROI areas, denoted as  $AI^R$  and  $AI^{NR}$  respectively, be decided by the area ratio  $r$ .

- Step 1. Calculate the feature from  $AI^R$ , denoted  $FAI^R$ . Details of the procedure are similar to those in Section 2.1.
- Step 2. Extract the watermark  $W^{etr}$  from the LH and HL sub-bands of each block of  $b^{NR} \times b^{NR}$  pixels in  $AI^{NR}$ . Assume the watermark bit  $w_k^{etr}$  is extracted from  $AI_{n,l}^{NR,LH}$ , where  $AI_{n,l}^{NR,LH}$  denotes the  $l$ th wavelet coefficient in  $AI_n^{NR,LH}$ , using the algorithm as follows:

$$w_k^{etr} = \begin{cases} 0, & \text{if } t \times q \leq AI_{n,l}^{NR,LH} < (t+1) \times q, \text{ for } t = 0, \pm 2, \pm 4 \dots \\ 1, & \text{if } t \times q \leq AI_{n,l}^{NR,LH} < (t+1) \times q, \text{ for } t = \pm 1, \pm 3 \dots \end{cases} \quad (7)$$

- Step 3.  $W^{etr}$  is unscrambled with the secret scramble key, and then divided into two parts: the extracted feature denoted  $FAI^{R,etr}$  and extracted patient's data denoted  $PA^{etr}$ . Next, these two bit streams,  $FAI^R$  and  $FAI^{R,etr}$ , are formatted as decimal numbers  $s_k$  and  $s_k^{etr}$  by  $DEC(FAI^R) = \{s_{m,k}\}$  and  $DEC(FAI^{R,etr}) = \{s_{m,k}^{etr}\}$ , where  $DEC()$  denotes the decimal function and the meanings of  $m$  and  $k$  are defined in Section 2.1.
- Step 4. Compare  $s_{m,k}$  and  $s_{m,k}^{etr}$ . If  $s_{m,k}$  and  $s_{m,k}^{etr}$  are different, then the corresponding block in  $I^R$  is located and regarded as a non-authentic block; otherwise, it is regarded as an authentic block.
- Step 5.  $PA^{etr}$  is used to output the patient's data.

## 2.4 Discussion and Analysis

The area ratio  $r$  between the ROI and non-ROI parts is constant for the medical images in the proposed method, unlike in the related method where the ROI is set by the doctor. In this section, the area ratio  $r$  is analyzed. Notation follows that of the previous section. The original medical image  $I$  is size of  $w \times h$  pixels and, from (1), we find,

$$I^R = \frac{r}{1+r} I \text{ and } I^{NR} = \frac{1}{1+r} I \quad (8)$$

That is,  $I^R$  has  $(r \times w \times h)/(1+r)$  pixels and  $I^{NR}$  has  $(w \times h)/(1+r)$  pixels. According to Sections 2.1 and 2.2, the size of the watermark (including the feature  $I^R$  and the patient's data) is  $k$  bits and is computed as follows,

$$k = \frac{r \times w \times h}{1+r} \times \frac{1}{4} \times \frac{1}{b^R \times b^R} \times p \times \lceil \log_2 s_{m,p}^{R,LL} \rceil + w^{pa} \times h^{pa} \text{ bits} \quad (9)$$

From Section 2.2, the amount of space used for embedding the watermark in  $I^{NR}$  is:

$$\frac{w \times h}{1+r} \times \frac{1}{b^{NR} \times b^{NR}} \times \frac{1}{2} (b^{NR} \times b^{NR}) = \frac{w \times h}{2 \times (1+r)} \text{ bits} \quad (10)$$

Therefore, we can state that:

$$\frac{w \times h}{2 \times (1+r)} \geq \frac{r \times w \times h}{1+r} \times \frac{1}{4} \times \frac{1}{b^R \times b^R} \times p \times \lceil \log_2 s_{m,p}^{R,LL} \rceil + w^{pa} \times h^{pa} \quad (11)$$

Formula (11) can then be rewritten as:

$$\frac{\frac{w \times h}{2}}{(1+r)} \geq r \times w \times h \times \frac{1}{4} \times \frac{1}{b^R \times b^R} \times p \times \lceil \log_2 s_{m,p}^{R,LL} \rceil + (w^{pa} \times h^{pa}) \quad (12)$$

which gives:

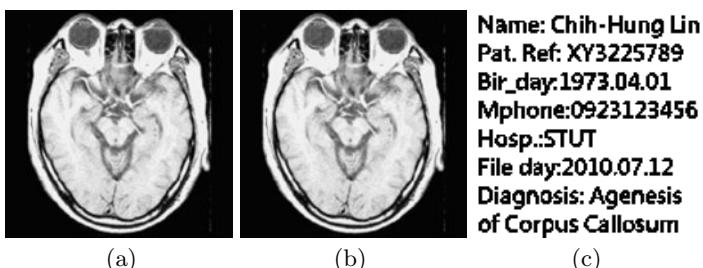
$$r \leq (w \times h \times \frac{1}{4} \times \frac{1}{b^R \times b^R} \times p \times \lceil \log_2 s_{m,p}^{R,LL} \rceil + w^{pa} \times h^{pa})^{-1} \times (\frac{w \times h}{2} - w^{pa} \times h^{pa}) \quad (13)$$

For example, if the size of the original medical image  $w \times h$  is  $512 \times 512$  pixels,  $b^R \times b^R$  is  $2 \times 2$ ,  $p = 1$ ,  $\lceil \log_2 s_{m,p}^{R,LL} \rceil = 9$ ,  $w^{pa} \times h^{pa} = 64 \times 64$ . According to formula (13), the area ratio  $I^R/I^{NR} = r \leq 0.84$ , therefore,  $I^R$  has  $(r \times w \times h)/(1+r) = 119674.4$  pixels and  $I^{NR}$  has  $(w \times h)/(1+r) = 142469.6$  pixels. As a check, we can calculate  $I^R + I^{NR} = 119674.4 + 142469.6 = 262144 = 512 \times 512$  pixels. Also, if  $w^{pa} \times h^{pa} = 128 \times 128$ , the area ratio  $I^R/I^{NR} = r \leq 0.7$ . Therefore, the result of formula (13) can help the system to choose a suitable and consistent area ratio  $r$ .

### 3 Experimental Results

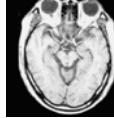
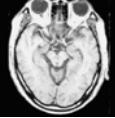
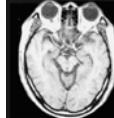
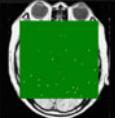
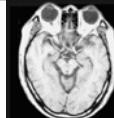
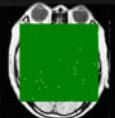
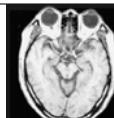
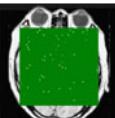
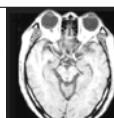
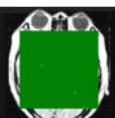
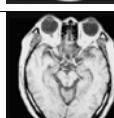
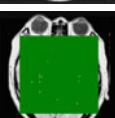
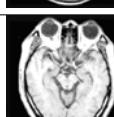
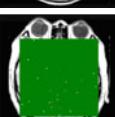
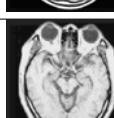
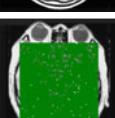
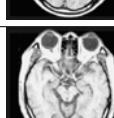
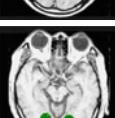
This section describes the experimental results, including the copyright protection and integrity verification. The performance was evaluated according to its imperceptibility, robustness and location. We used a gray-scale medical MRI brain image, of size  $512 \times 512$  pixels, as a test image (Fig. 3(a)). A binary image of size  $128 \times 128$  pixels was used as the patient's data (Fig. 3(c)). The other related factors were defined as:  $b^R \times b^R$  is  $2 \times 2$ ,  $b^{NR} \times b^{NR}$  is  $8 \times 8$ ,  $p = 1$ ,  $\lceil \log_2 s_{m,p}^{R,LL} \rceil = 9$ , and the area ratio  $I^R/I^{NR} = r$  was set as 0.7, computed according to (13). Fig. 3(b) shows the watermarked MRI brain images.

Several common image processing methods, namely JPEG compression with quality factor  $QF$  (denoted as JPEG  $QF$  in Table 1), Gaussian noise and brightness/contrast adjustment, were implemented to watermark the image. The corresponding authentication results are shown in Table 1, demonstrating that in all cases the patient's data can be extracted with good quality. The proposed scheme can robustly resist common image processing problems owing to the embedding of the watermark into the middle frequency sub-band. Additionally, we can clearly locate the tampered region when the received images suffered malicious modification.



**Fig. 3.** (a) Original MRI brain image; (b) Watermarked MRI brain image; (c) Original patient's data

**Table 1.** Authentication results

Processing	Modified image	Verified result image	$PA^{ctr}$
No attack			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
JPEG90			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
JPEG80			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Gaussian noise			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Brightness up			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Brightness down			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Contrast up			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Contrast down			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Cropping			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum
Pasting			Name: Chih-Hung Lin Pat_Ref: XY3225789 Bir_day:1973.04.01 Mphone:0923123456 Hosp.:STUT File day:2010.07.12 Diagnosis: Agenesis of Corpus Callosum

## 4 Conclusions

The protection method for medical images is a little different from general images, because of having to consider the ROI. Generally, the center area is focused in medical images, and also is the most important area used by the doctor for diagnosis. However, different ROIs will be chosen by different doctors in the previous methods, such that the method of protecting and authenticating a medical image is inconsistent and ambiguous between different doctors and hospital systems. In this paper, we proposed an algorithm for dividing the ROI and non-ROI areas at a consistent ratio that is independent of the user, to generate an ROI feature and embed the feature together with the patients data into the non-ROI area. The modified parts of the ROI can be located, and the patients data also can be extracted in the authentication phase. The efficiency of this method was demonstrated in empirically.

**Acknowledgments.** This work is partially supported by the National Science Council, Taiwan, grant NSC99-2221-E-218-039. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

## References

1. Sun, Q., Chang, S.F., Kurato, M., Suto, M.: A New Semi-Fragile Image Authentication Framework Combining ECC and PKI Infrastructures. In: IEEE International Symposium on Circuits and Systems, pp. 440–443. IEEE Press, Phoenix-Scottsdale (2002)
2. El-Din, S.N., Moniri, M.: Fragile and Semi-Fragile Image Authentication Based on Image Self-Similarity. In: IEEE International Conference on Image Processing, pp. 897–900. IEEE Press, Los Alamitos (2002)
3. Bao, P., Ma, X.: Image Adaptive Watermarking Using Wavelet Domain Singular Value Decomposition. *IEEE Trans. on Circuits and Systems for Video Technology* 15, 96–102 (2005)
4. Lu, Z.M., Xu, D.G., Sun, S.H.: Multipurpose Image Watermarking Algorithm Based on Multistage Vector Quantization. *IEEE Trans. on Image Processing* 14, 822–831 (2005)
5. Chang, C.C., Lin, C.C., Hu, Y.S.: An SVD Oriented Watermark Embedding Scheme with High Qualities for the Restored Images. *International Journal of Innovative Computing, Information and Control* 3, 609–620 (2007)
6. Lu, Z.M., Liao, X.W.: Counterfeiting Attacks on two Robust Watermarking Schemes. *International Journal of Innovative Computing, Information and Control* 2, 841–848 (2006)
7. Lin, C.H., Hsieh, W.S.: Applying Projection and B-spline to Image Authentication and Remedy. *IEEE Trans. on Consumer Electronics* 49, 1234–1239 (2003)
8. Lin, C.H., Hsieh, W.S.: Image Authentication Scheme for Resisting JPEG, JPEG 2000 Compression and Scaling. *IEICE Trans. on Information and Systems* E90-D, 126–136 (2007)

9. Lin, C.H.: Multi-Purpose Digital Watermarking Method – Integrating Robust, Fragile and Semi-Fragile Watermarking. *International Journal of Innovative Computing, Information and Control* 6, 3023–3036 (2010)
10. Fallahpour, M., Megias, D., Ghanbari, M.: High capacity, reversible data hiding in medical images. In: 16th IEEE International Conference on Image Processing, pp. 4241–4244. IEEE Press, Cairo (2009)
11. Shih, F.Y., Wu, Y.-T.: Robust Watermarking and Compression for Medical Images based on Genetic Algorithms. *Information Sciences* 175, 200–216 (2005)

# FLC-Regulated Speaker Adaptation Mechanisms for Speech Recognition

Ing-Jr Ding

Department of Electrical Engineering, National Formosa University  
No.64, Wunhua Rd., Huwei Township, Yunlin County 632, Taiwan, R.O.C.  
ingjr@nfu.edu.tw

**Abstract.** The exploitation of fuzzy logic control (FLC) mechanism in the fields of speaker adaptation (SA) is thoroughly investigated in this study, specifically in the reliable determination of HMM acoustic parameters. For enhancing the performance of speaker adaptation, the FLC mechanism is engineered into the MAP estimate of HMM parameters for Bayesian-based adaptation; also into the MLLR estimate for transformation-based adaptation. The speech recognition system using an adaptation scheme with the support of FLC will still be able to keep a satisfactory recognition performance even in an ordinary case.

**Keywords:** Speech recognition; speaker adaptation; fuzzy logic control.

## 1 Introduction

Computing techniques for automatic speech recognition have existed for years and, with the ever growing maturity, have found more and more applications in current daily life [1]. Nevertheless, the recognition performance of all speech recognition systems ever built is undeniably inferior to a human listener as already pointed out in [2].

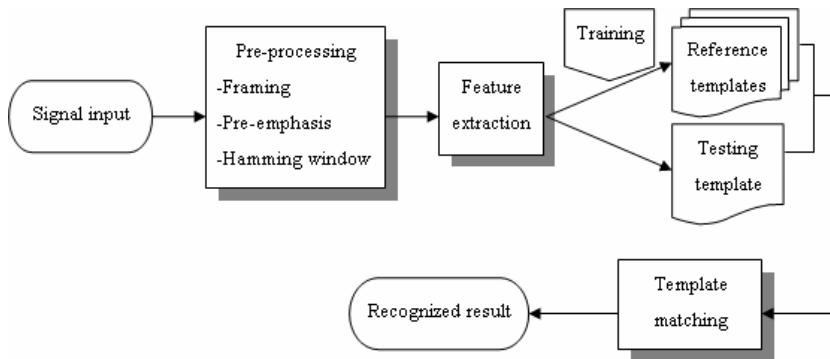
Fig. 1 depicts the operating structure of a typical speech recognition system for capturing specific short phrases or primitive statements only. Note that during the operation any disturbances causing a mismatch between the pre-established reference templates and the testing template would compromise the recognition performance and the sources of disturbances may include

- speech from speakers strange to the system
- speech from speaker known to the system, only in poor “vocal shape”
- various interferences in the background
- channel distortion induced in the acquisition process

and so forth.

Countermeasures can be taken in two aspects:

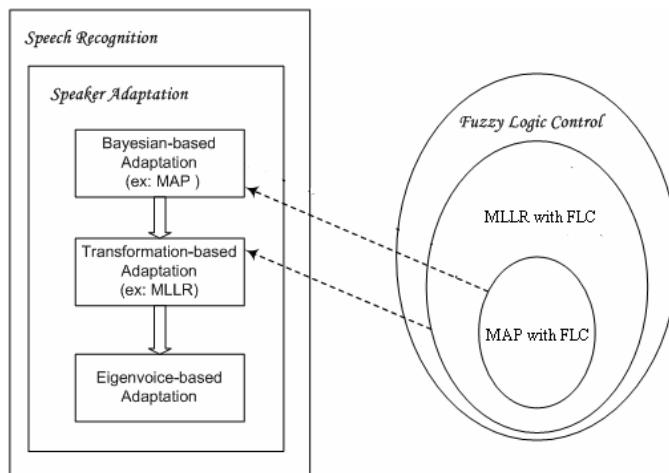
- (1) Signal filtering and normalization are deployed so that the operating condition is in as much alignment with the referential condition as could be done.
- (2) Internal tuning of the referential settings is undertaken so that the system adapts toward the actual operating environment when new speakers appear.



**Fig. 1.** The operating structure of a typical speech recognition system

Approaches in the second category use sample utterances collected from the new speaker (the end-user of the system) for adapting the system internal parameter settings of the pre-established speech model. Consequently, they are referred to as model-based adaptation or speaker adaptation.

Fig. 2 reveals the chronological development of the three major speaker adaptation schemes:



**Fig. 2.** Three categories of speaker adaptation techniques in speech recognition

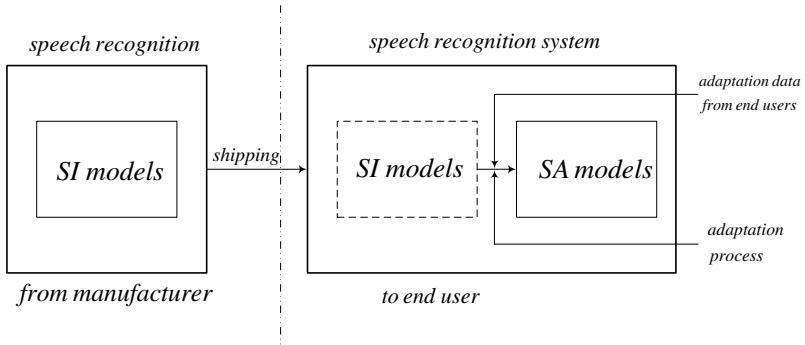
MAP adaptation, appearing around 1991 and the representative of Bayesian-based adaptation, works better than the ML (maximum likelihood) estimate of the adaptation by taking into account the information of prior means of the model. By the nature of MAP computation, in the speech model only the portions associated with the adaptation samples get updated, for which case VFS scheme came into the play as a supplement to MAP by extending the coverage of adaptation in the model space. The

MAP-VFS adaptation in general offers more satisfaction in recognition performance than MAP alone given the same adaptation data. MLLR adaptation first appeared in 1995 and became the representative of transformation-based adaptation, where linear regression was employed to derive the transformation matrix using ML-estimate. Note that through the transformation by matrix multiplication, the entire model space is adapted at one time despite the fact that the sample utterances might convey very limited information for adaptation. In a sense, MLLR adaptation provides with an overall but somewhat coarser speech model adaptation, in contrast to MAP adaptation which brings about a local and yet specific effects of adaptation, given the same adaptation samples.

One thing that is common to both MAP and MLLR is that the quality of adaptation depends on the amount and adequacy of the adaptation samples: the more the samples, the better the adaptation quality which in turn determines the recognition performance. When the adaptation utterances from a new speaker are insufficient, the effects of either MAP or MLLR adaptation would be questionable: the recognition rate of which would fall below the baseline, i.e., worse than no adaptation at all as shown by the author's experiments.

Eigenvoice-based adaptation [3, 4] is a relatively young member in the speaker adaptation family, first appearing around 2000, and is also known as speaker-clustering-based adaptation where a speaker dependent (SD) speech model is established for every member in a group of speakers, from which feature vectors called as eigenvoices are extracted through PCA for building the eigenvoice speech model. The adaptation to the speech model (an eigenvoice vector space) then can be undertaken when adaptation data is available.

To summarize, speaker adaptation is a process that turns speaker-independent (SI) speech models into speaker-adapted (SA) ones, as is clearly seen in Fig. 3.



**Fig. 3.** Speaker adaptation scheme

To ensure the quality of the adaptation at the scarcity of adaptation samples, the author proposed a framework of fuzzy mechanism that is applicable to some major speaker adaptation schemes for resolving the unreliable adaptation due to insufficient training samples; the implementation of which, a series of fuzzy logic controllers

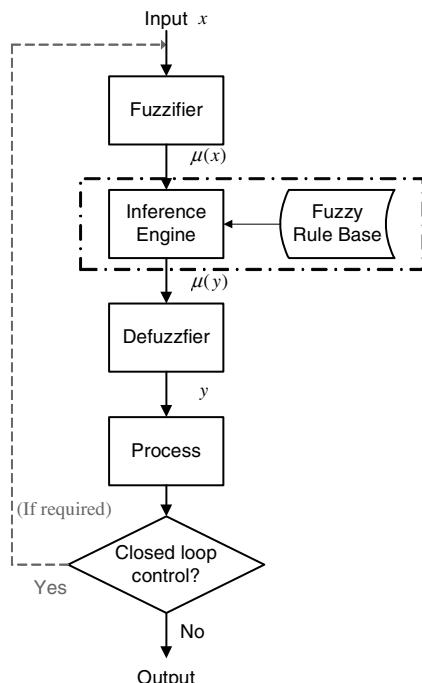
(FLC) embedded in MAP and MLLR adaptations, has proven themselves by achieving far superior recognition rate at extreme adverse conditions.

The same framework will also be applied to the eigenvoice-based adaptation scheme in the future research work to design the adaptation algorithm based on fuzzy-logic thoughts.

## 2 Fuzzy Schemes and Speech Recognition

Fuzzy approaches have been widely applied to the field of speech recognition for many years, playing a variety of roles from data clustering, logic reasoning, to neural network configuration for speech recognition. Although many fuzzy approaches have been widely used in various sub-areas of speech recognition as mentioned hereinbefore, it has not been seen for the use of FLC in HMM speaker adaptation, or even in speech recognition. Based on the methodology of FLC, a series of speaker adaptation computations under FLC regulation are designed.

A fuzzy logic control concerns the automation of a control process for which the operator's expertise/experience regarding the process control imparted in oral or written form has to be translated so as to fit in the framework of fuzzy logic control, together with other accommodation or extension in the fuzzy set theory specific to this particular application. Fig. 4 shows the architecture of a typical FLC:



**Fig. 4.** Architecture of a typical FLC

Various types of FLCs have been proposed with variations in the module design considerations. The renowned Mamdani and Sugeno FLC are, for instance, different in consequence design in every individual rules. Since T-S fuzzy model has been seen in the control of the system as complicated as an electric power plant with success [5], and thus this type of FLC is employed in the author's research in speaker adaptation schemes, as will be detailed in the next sections.

### 3 MAP Speaker Adaptation with FLC

#### 3.1 MAP (Maximum a Posteriori)

MAP adaptation is a kind of direct model adaptation, which attempts to directly re-estimate the model parameters [6, 7]. However, it is noted that MAP adaptation re-estimates only the portion of model parameter units associated the adaptation data, and therefore, MAP adaptation usually needs a large amount of data for adaptation and the performance will be improved as adaptation data increases and gets covering the model space. When the amount of data is sufficiently large, the MAP estimation yields as good recognition performances as that obtained using maximum-likelihood estimation [7]. As shown in (1),

$$\hat{\mu}_k = \frac{N_k}{\tau + N_k} \bar{y}_k + \frac{\tau}{\tau + N_k} \mu_k, \quad (1)$$

the MAP estimate of the mean is essentially a weighted average of the prior mean and the sample mean, and the weights are functions of the number of adaptation samples, given that  $\tau$  being fixed. When  $N_k$  is equal to zero (i.e., no additional training data are available for adapting the k-th Gaussian), the estimate is simply the prior mean of the k-th Gaussian alone. Conversely, when a large number of training samples are used for the k-th Gaussian ( $N_k \rightarrow \infty$ , to be exaggerative), the MAP estimate in (1) then converges asymptotically to the maximum likelihood estimate, i.e., the sample mean parameter with the k-th Gaussian,  $\bar{y}_k$ .

Now consider the other way round with  $N_k$  being fixed, the parameter  $\tau$  controls the balance in the interpolation between the  $\bar{y}_k$ -term and the  $\mu_k$ -term, (as  $N_k$  does). It is referred to as the "adaptation speed parameter" in [8] in that the speed of adaptation can be increased or held-back by choosing a small or a large value of  $\tau$ . The parameter  $\tau$  is also known as a "prior density parameter" since it determines, to which side of, and for how close to  $\bar{y}_k$  or  $\mu_k$ , the MAP-estimate of  $\hat{\mu}_k$  would be.

As a general remark, that the recognition performance of adaptation, regardless of whatever adaptation schemes in consideration, would not be as good as desired given insufficient training samples  $N$  is a consensus among all. The robustness of MAP adaptation against relatively small  $N$  should not be overlooked either, and as yet in conventional schemes for MAP adaptation ([9], e.g.), a common value of  $\tau$  was used for all the Gaussians of a given state, or for all states of an HMM, or even for all HMMs.

With all the aforementioned thoughts in mind and looking at (1), it would be quite natural for one to come out with the idea that  $\hat{\mu}_k$  should stay in the vicinity of  $\mu_k$  when  $N$  is somewhat small (by the choice of a large  $\tau$ ) to avoid the performance

degrading caused by the potentially poor estimate of  $\bar{y}_k$ , and on the other hand when  $N$  is large enough, the adaptation should move toward  $\bar{y}_k$  speedily. Putting such notion in terms of simple rules in plain words leads to statements as follows

- (1) When  $N$  is small,  $\tau$  should be large such that  $\hat{\mu}_k$  sticks more to  $\mu_k$ .
- (2) When  $N$  is medium,  $\tau$  should be medium such that  $\hat{\mu}_k$  locates between  $\bar{y}_k$  and  $\mu_k$  accordingly.
- (3) When  $N$  is large,  $\tau$  should be small such that  $\hat{\mu}_k$  adapts toward  $\bar{y}_k$ .

This is where fuzzy methodology comes into play, and how the statements of linguistic terms with uncertainty to some degree can be formulated in quantized forms for subsequent computations will be explained in the next section.

### 3.2 MAP with FLC

Within the framework of fuzzy process, the formulation of the problem at hand is given as a set of three fuzzy IF-THEN rules and the system output  $\tau(\cdot)$ .

Rule 1: If  $N$  is  $M_1(N)$ , then  $\tau_L = f_1(N)$ ,

Rule 2: If  $N$  is  $M_2(N)$ , then  $\tau_M = f_2(N)$ ,

Rule 3: If  $N$  is  $M_3(N)$ , then  $\tau_S = f_3(N)$ ,

where  $M_1(N)$ ,  $M_2(N)$  and  $M_3(N)$  are the membership functions representing the degree of how much  $N$  is involved in the classes of linguistically “small”, “medium” and “large” respectively, and are defined as

$$M_1(N) = \begin{cases} 1 & N \leq N_1, \\ \frac{N_2 - N}{N_2 - N_1} & N_1 \leq N \leq N_2, \\ 0 & N \geq N_2, \end{cases}$$

$$M_2(N) = \begin{cases} 0 & N \leq N_1 \text{ or } N \geq N_3, \\ \frac{N - N_1}{N_2 - N_1} & N_1 < N \leq N_2, \\ \frac{N_3 - N}{N_3 - N_2} & N_2 \leq N < N_3, \end{cases} \quad (2)$$

$$M_3(N) = \begin{cases} 0 & N \leq N_2, \\ \frac{N - N_2}{N_3 - N_2} & N_2 < N < N_3, \\ 1 & N \geq N_3. \end{cases}$$

$f_i(N), i=1, 2, 3$  are output functions in each rule for regulating the  $\tau$  value and are defined as

$$\begin{aligned} f_1(N) &= \frac{b}{\log(N) + a}, \\ f_2(N) &= \frac{N}{c}, \\ f_3(N) &= \frac{\log(N)}{N}. \end{aligned} \tag{3}$$

Note that the definitions in (3) is an empirical choice among many possibilities.

For the system output,  $\tau(\cdot)$  is defined as [10]

$$\tau = \frac{\sum_{i=1}^3 M_i(N) f_i(N)}{\sum_{i=1}^3 M_i(N)}. \tag{4}$$

## 4 MLLR Speaker Adaptation with FLC

### 4.1 MLLR (Maximum Likelihood Linear Regression)

In the transformation-based model adaptation, certain appropriate transformations have to be derived from a set of adaptation utterances acquired from a new speaker and then applied to clusters of HMM parameters. An affine transformation (also called linear transformation) over HMM parameters in general offers a more appropriate model than a bias transformation does and there have been numerous adaptation schemes using affine transformations. In the work by Leggetter et al. [11], MLLR adaptation was firstly proposed under the framework of affine transformation, which has become quite popular and successful for its rapid adaptation. However, it is necessary to have sufficient adaptation data to ensure the estimate of the MLLR transformation, and various solutions have been suggested for further reinforcement. For instance, instead of using the maximum likelihood (ML) estimate in the MLLR scheme, the maximum a *posteriori* estimate is used to estimate the transformation parameters by maximizing the posterior density [12, 13]. In addition, it is suggested in [14, 15] that a prior distribution for calculating the mean transformation matrix parameters is used, which is generally dubbed as the MAPLR technique.

MLLR makes use of the simplicity of ML criterion, which states that the transformed model  $\hat{\eta}_{ML}$  should maximize the likelihood of the adaptation data  $p(Y | \Lambda, \eta)$ , i.e.

$$\hat{\eta}_{ML} = \arg \max_{\eta} p(Y | \Lambda, \eta). \tag{5}$$

Consider the Gaussian mean vector of the model at state  $s$ ,  $\mu_s$ , and the associated affine transformation action as follows

$$\hat{\mu}_s = A_s \cdot \mu_s + b_s, \tag{6}$$

which sometimes is written as

$$\hat{\mu}_s = W_s \cdot \xi_s, \quad (7)$$

and  $\xi_s$  is the extended mean vector in the form

$$\xi_s = [\omega, \mu_{s_1}, \dots, \mu_{s_n}]^T, \quad (8)$$

where  $\omega$  is the offset term of the regression, usually being set as 1.

The transformation matrix  $W_s$  is to be estimated such that the likelihood of the adaptation data is maximized, for which a closed form solution is available in [11] by solving the following equation,

$$\sum_{t=1}^T \sum_{r=1}^R \gamma_{s_r}(t) \Sigma_{s_r}^{-1} o_t \xi_{s_r}' = \sum_{t=1}^T \sum_{r=1}^R \gamma_{s_r}(t) \Sigma_{s_r}^{-1} W_s \xi_{s_r} \xi_{s_r}', \quad (9)$$

where  $\gamma_{s_r}(t)$  is the total occupation probability for the state  $s_r$  at time  $t$  given the observation vectors of adaptation data  $o_t$  at time  $t$ ,  $\Sigma_{s_r}^{-1}$  is the covariance matrix of the output probability distribution, and  $R$  is the number of states.

## 4.2 MLLR with FLC

When only a limited and insufficient amount of adaptation data is available, the quality of the acquired transformation matrix  $W_s$ , especially derived by the MLLR approach, would be in doubt; poor estimation of  $W_s$  could lead to the corruption of underlying structure of the acoustic space. The problem due to the scarcity of adaptation data can be alleviated by utilizing the MAPLR scheme instead, if one disregards the heavy computation involved.

With insufficient training data, one would naturally tend to be more “conservative” while using the transformation matrix thus derived, i.e., the effect of the adaptation should be restricted in this case so that the adapted mean vector would not vary too much from the state prior to adaptation. Accordingly, an incremental approach to MLLR model transformation is proposed as follows

$$\tilde{\mu}_s = \alpha \cdot \mu_s + (1 - \alpha) \cdot W_s \cdot \xi_s, \quad 0 \leq \alpha \leq 1, \quad (10)$$

where  $\mu_s$  is the initial mean vector,  $\xi_s$  is the extended mean vector as defined in (8) and  $W_s$  is the transformation matrix derived from (9). The form of incremental MLLR adaptation in (10) is very similar to the one in (1), essentially an MAP-like adaptation. A weight parameter  $\alpha$  is devised to govern the balance of the maximum likelihood estimate of the mean from the adaptation data and the prior mean, as is the role of  $\tau$  in MAP estimate. By using a weighted sum of the initial mean vector and the MLLR adapted mean vector, it is expected that a satisfactory performance will also be achieved even when only a little amount of training data is available for adaptation. Note that the weight  $\alpha$  is to vary in a way depending on how much confidence one has in  $W_s$ . A possibly not so well estimated  $W_s$  due to insufficient adaptation data would preferably goes with  $\alpha$  approaching 1 so that  $\tilde{\mu}_s$  stays closer to  $\mu_s$ , instead of drifting away drastically. On the opposite, 0-approaching  $\alpha$  should be taken for full advantage of fast adaptation by  $W_s$ .

A rule base with three fuzzy implications is given to govern  $\alpha$  regulation under the circumstance of  $N$  training samples (in terms of acoustic frames) observed for all Gaussian mixture components as follows.

- Rule 1: If  $N$  is small,  
Then  $\alpha$  is large,
- Rule 2: If  $N$  is medium,  
Then  $\alpha$  is medium,
- Rule 3: If  $N$  is large,  
Then  $\alpha$  is small.

With Takagi-Sugeno FLC (T-S FLC) in consideration, let  $M_1(N)$ ,  $M_2(N)$  and  $M_3(N)$  be membership functions associated respectively with small, medium and large amounts of training data available for adaptation as has been shown in (2), and  $\alpha_L$ ,  $\alpha_M$  and  $\alpha_S$  be the  $\alpha$  values determined respectively by functions  $f_1(N)$ ,  $f_2(N)$  and  $f_3(N)$  in each of the three cases. Then the previous set of rules can be further clarified as

- Rule 1: If  $N$  is  $M_1(N)$ ,  
Then  $\alpha_L = f_1(N)$ ,
- Rule 2: If  $N$  is  $M_2(N)$ ,  
Then  $\alpha_M = f_2(N)$ ,
- Rule 3: If  $N$  is  $M_3(N)$ ,  
Then  $\alpha_S = f_3(N)$ ,

where the implication functions

$$\begin{aligned}f_1(N) &= a_1 \cdot N + b_1, \\f_2(N) &= a_2 \cdot N + b_2, \\f_3(N) &= a_3 \cdot N + b_3,\end{aligned}\tag{11}$$

and the final system output [10]

$$\alpha = \frac{\sum_{i=1}^3 M_i(N) f_i(N)}{\sum_{i=1}^3 M_i(N)}.\tag{12}$$

## 5 Conclusions and Future Work

This study presents a work that an FLC mechanism is considered for governing the speaker adaptation procedure in an automatic speech recognition system. Speaker adaptation with the FLC support will properly adjust HMM acoustic parameters and alleviate the mismatch phenomenon between training and testing environments, so that

a high recognition performance will be maintained. In this work, FLC has been successfully embedded into the Bayesian-based and the transformation-based adaptation.

In the realm of speaker adaptation, there are many other techniques available for parameter tuning and not covered in the scope of this paper. Eigenvoice-adaptation, a younger cousin of eigenface methodology for face detection or face recognition in image process, has been a new focus in recent years for instance. How FLC mechanism could be incorporated in the framework of eigenvoice process is an open issue and will be a promising subject for the next research work.

## References

1. Rabiner, L.R.: The Power of Speech. *Science* 301, 1494–1495 (2003)
2. Lippmann, R.P.: Speech Recognition by Machines and Humans. *Speech Communication* 22, 1–15 (1997)
3. Kuhn, R., Junqua, J.-C., Nguyen, P., Niedzielski, N.: Rapid Speaker Adaptation in Eigenvoice Space. *IEEE Transactions on Speech and Audio Processing* 8, 695–707 (2000)
4. Mak, B., Hsiao, R.: Kernel Eigenspace-based MLLR Adaptation. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 784–795 (2007)
5. Kermiche, S., Saidi, M.L., Abbassi, H.A., Ghodbane, H.: Takagi-Sugeno Based Controller for Mobile Robot Navigation. *Journal of Applied Science* 6, 1838–1844 (2006)
6. Gauvain, J.L., Lee, C.H.: Maximum a *Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing* 2, 291–298 (1994)
7. Lee, C.H., Lin, C.H., Juang, B.H.: A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing* 39, 806–814 (1991)
8. Takahashi, J.-I., Sagayama, S.: Vector-field-smoothed Bayesian Learning for Fast and Incremental Speaker/Telephone-channel Adaptation. *Computer Speech and Language* 11, 127–146 (1997)
9. Woodland, P.C.: Speaker Adaptation: Techniques and Challenges. In: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 85–90 (1999)
10. Takagi, T., Sugeno, M.: Fuzzy Identification of Systems and Its Application to Modeling and Control. *IEEE Transactions on Systems, Man and Cybernetics* 15, 116–132 (1985)
11. Leggetter, C.J., Woodland, P.C.: Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language* 9, 171–185 (1995)
12. Chien, J.T., Lee, L.M., Wang, H.C.: Estimation of Channel Bias for Telephone Speech Recognition. In: *Proceedings of International Conference on Spoken Language Processing*, pp. 1840–1843 (1996)
13. Chien, J.T., Wang, H.C.: Telephone Speech Recognition Based on Bayesian Adaptation of Hidden Markov Models. *Speech Communication* 22, 369–384 (1997)
14. Chesta, C., Siohan, O., Lee, C.H.: Maximum a *Posteriori* Linear Regression for Hidden Markov Model Adaptation. In: *Proceedings of the European Conference on Speech Communication and Technology*, pp. 211–214 (1999)
15. Chou, W.: Maximum a *Posteriori* Linear Regression with Elliptically Symmetric Matrix Priors. In: *Proceedings of the European Conference on Speech Communication and Technology*, pp. 1–4 (1999)

# Monitoring of the Multichannel Audio Signal\*

Eugeniusz Kornatowski

West Pomeranian University of Technology, Szczecin, Faculty of Electrical Engineering,  
Department of Signal Processing and Multimedia Engineering, 26 Kwietnia 10 St.,  
71-126 Szczecin, Poland  
korn@zut.edu.pl

**Abstract.** The paper describes operation algorithm as well as hardware and software realization of a detector, enabling a visual evaluation of acoustic perspective and dominating sound source direction in multi-channel registration and transmission of spatial sound. Proposed system may be successfully used in modern sound engineering studios, including “live” transmissions.

**Keywords:** Sound engineering, spatial sound, acoustic perspective, virtual sound source.

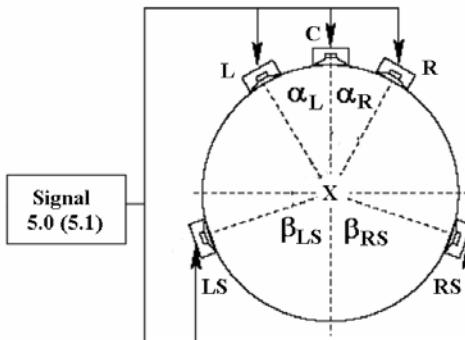
## 1 Introduction

The purpose of this paper is to present a monitoring method for multi-channel audio signal quality. The system for recording, transmission and archiving of stereo sound productions, known for many years, may be considered to be fully refined, also in the aspect of solutions developed to enable a full objective evaluation of final result quality features. The term “signal quality” used in this paper should be construed as a spatial effect quality; therefore, for a listener, it is a parameter pertaining to a sensation of being present in the location of acoustic event (i.e. concert hall, recording studio, etc.). Among other devices, an instrument to enable visual control of spatial effect in a real time, known as goniometer, is used for this purpose for stereophonic productions. However, in case of multi-channel production events (i.e. 5.1), which is five main channels and one LFE (Low Frequency Effects) channel, there are no efficient tools to evaluate signal quality within the context mentioned above. The problem is quite significant since multi-channel productions are utilized more and more often, frequently in “live” events (with the use of i.e. Dolby Digital coding).

Evaluation of the above described quality criterion is done by means of electric system analysis being associated with particular audio channels (for 5.0 or 5.1 systems), i.e. left front (L), central front (C) and right front (R), as well as left rear (LS) and right rear (RS)). Schematic diagram of this sound playing system is shown in Figure 1.

---

\* Scientific work financed by the Ministry of Higher Education and Science (Poland) from funds for the science in years 2009 - 2010, as a research project No. N N505 364336.



**Fig. 1.** Concept of sound reproduction 5.1 (5.0), according to ITU-R-BS.775-1 standard:  $\alpha_L = \alpha_R = 30^\circ$ ,  $\beta_{LS} = \beta_{RS} = 75^\circ$ ; (x - listener)

## 2 Multi-channel Sound Monitoring

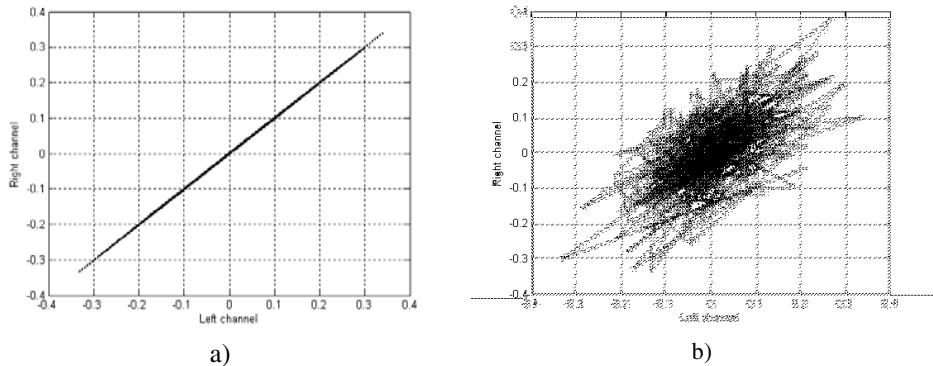
A comprehensible method used in sound engineering for monitoring of multi-channel sound spatiality is the use of immediate field sound monitors. One can conclude without difficulty that this method is totally subjective - since it does not always provide the assurance of getting a highest quality final product. It is therefore obvious that there is a need for a method that would bring explicit result of obtained effect observations. Sound "spatiality" may be construed as an ability to receive clearly defined planar acoustic perspective pertaining to virtual sound sources surrounding the listener. Similar interpretation, known as a lateral efficiency indicator:

$$LE(\theta) = \frac{\int_{25ms}^{80ms} p_8^2(t) \cos^2(\theta) dt}{\int_{0ms}^{80ms} p^2(t) dt} \cdot 100\% \quad (1)$$

was proposed in 1980 [1] to describe a measure of spatiality illusion, i.e. a concert hall effect. LE is related to a ratio of reflections coming from a side to total direct and deflected reflections reaching the listener.  $q$  is an angle between direction of coming sound reflection and „ear-to-ear” axis of the listener,  $p$  - acoustic pressure measured with omnidirectional microphone,  $p_8$  – acoustic pressure measured with a bidirectional figure-eight response microphone (microphone axis set at  $q$  angle).

In the period of dynamic stereo sound development, namely in sixties and seventies of the twentieth century a goniometer was used for visual evaluation of the acoustic space within so-called stereo base. These simple devices have been used by the sound engineering until today.

The goniometer is displaying Lissajous figures or patterns on its oscilloscope screen. Oscilloscope X and Y channels receive left and right signals from a stereophonic system. In case of a “strong” stereo effect, the displayed figure has considerably large area, while with “mono” signals, it is represented as a straight line. The related examples are shown in Figure 2.



**Fig. 2.** “Lissajous” figures for: (a) mono signal, (b) stereo signal

It is obvious that this method can be also applied in case of multi-channel recordings. In that case, it would be prudent to use separate goniometers for various channel pairs, such as: L-C, R-C, L-LS, R-RS, LS-RS. However, a large number of charts makes it impossible to provide a quick analysis.

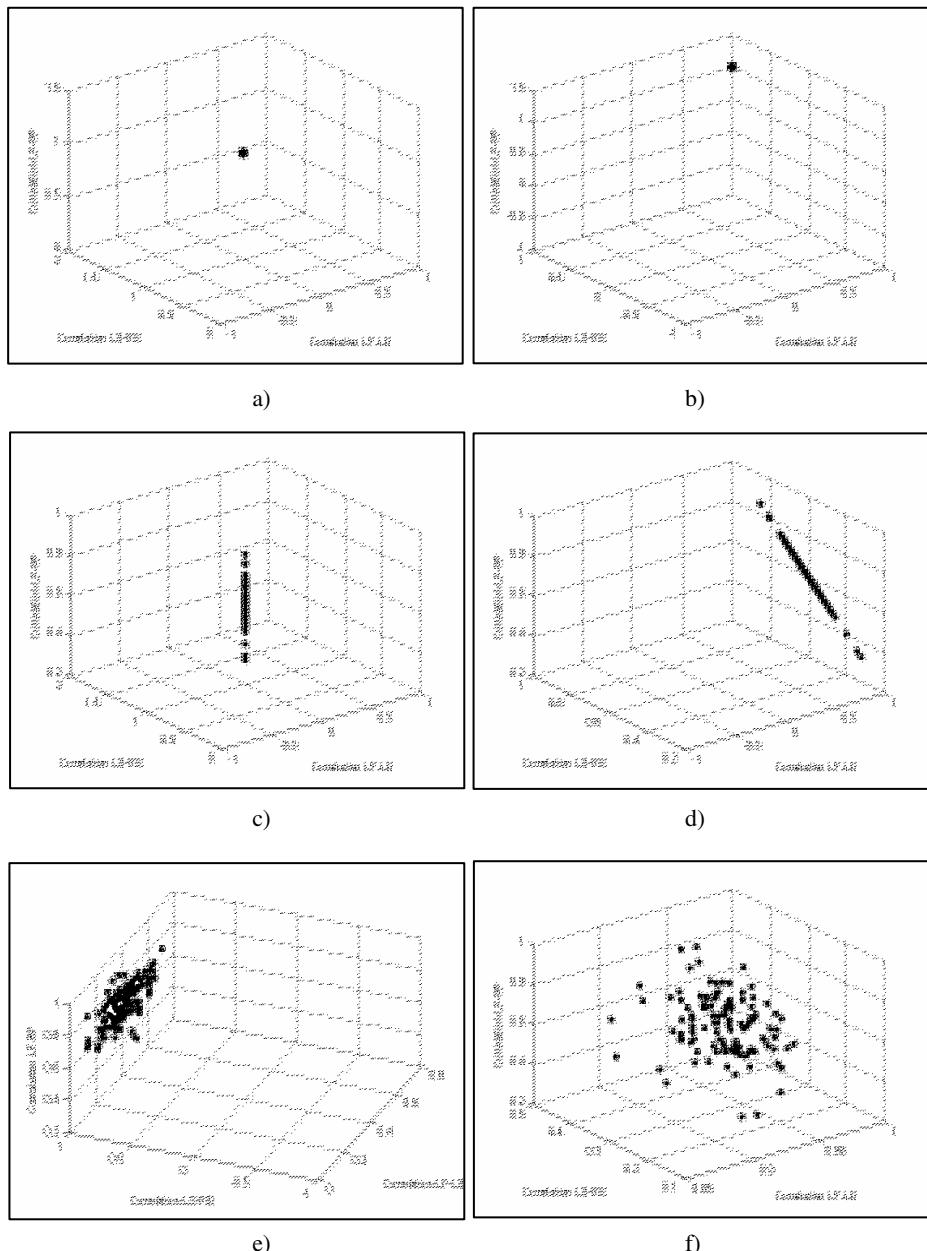
In [3], author this work propose to use a known technique to measure the amount of antiphase components, namely, the correlation coefficient, which is given in any text book (e.g. [5]) on statistics as:

$$\text{Corr} = \frac{\sum_i [(ch1_i - \bar{ch1}) \cdot (ch2_i - \bar{ch2})]}{\sqrt{\sum_i [(ch1_i - \bar{ch1})^2 \cdot (ch2_i - \bar{ch2})^2]}} \quad (2)$$

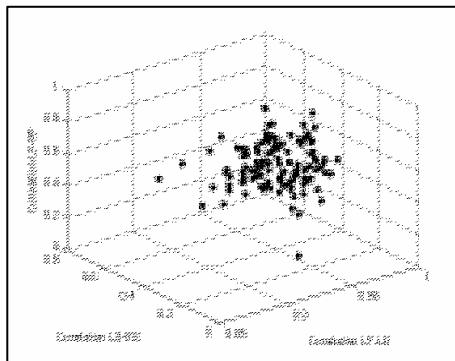
where  $ch1$  or  $ch2$  are one from set of multi-channel signal;  $\bar{ch1}$  and  $\bar{ch2}$  are the mean values of  $ch1$  and  $ch2$ , respectively.

On Fig. 3 and 4 few exemplary diagrams made by the method described is presented.

Analysis of diagrams shown leads to following conclusion: algorithm RSS (Real Surround Sound proposed in [3], Fig. 3f) makes possible to receive spatial effects similar to multi-channel recordings registration (Dolby Digital 5.0, Fig. 4) This conclusion can be also confirm by the listeners.



**Fig. 3.** Changes of correlation coefficient for following cases: a) mono signal:  $L=R$ ,  $LS=RS=0$ , b) mono:  $L=R=LS=RS$ , c) stereo L and R:  $LS=RS=0$ , d) stereo:  $L=LS$ ,  $R=RS$ , e) effect of matrix Dolby Surround decoder, f) effect of algorithm in [3] described (RSS)



**Fig. 4.** Changes of correlation coefficient for multi-channel signal: Dolby Digital 5.0

There are other solutions known to resolve multi-channel sound spatiality, e.g. [2]. A special software and PC computers are used for visualization in all known solutions. Apart from that, a displayed image requires that the observer has an advanced knowledge within the area of sound theory [3, 6, 7], can interpret meaning of the displayed and is able to relate the image to acoustic impressions. The optimum solution would be if the principle of the visualizing device operation was similar to the one used at present in sound engineering, i.e. the goniometer.

### 3 Visual Sound Spatiality Detector

A concept of the proposed solution is based on the following assumptions:

1. Visualization pertains to 5.0 system (Fig. 1),
2. Acoustic wave (wave front) is propagated along straight line,
3. Sound sources (loudspeaker sets) are considered as points and the generated sound wave is ball-shaped,
4. There are no reflections, interference and diffractions within audio room.

Those assumptions provide for creation of a detection system, which enables observation of sound spatiality in “sterile” conditions, i.e. in a room without its own acoustics. Detector input signals are: L, C, R, LS and RS (Fig. 1). Sound monitors and listener are located within a coordinate system XY. Assuming that the acoustic pressure is inversely proportional to distance from a sound source and taking into consideration simple trigonometric relationships, one could calculate the components of value proportional to the acoustic pressure of the virtual sound source in XY coordinate system:

$$P_x = -\frac{\sin \alpha_L}{r_L} |L| + \frac{\sin \alpha_R}{r_R} |R| - \frac{\sin \beta_{LS}}{r_{SL}} |LS| + \frac{\sin \beta_{RS}}{r_{RS}} |RS| \quad (3)$$

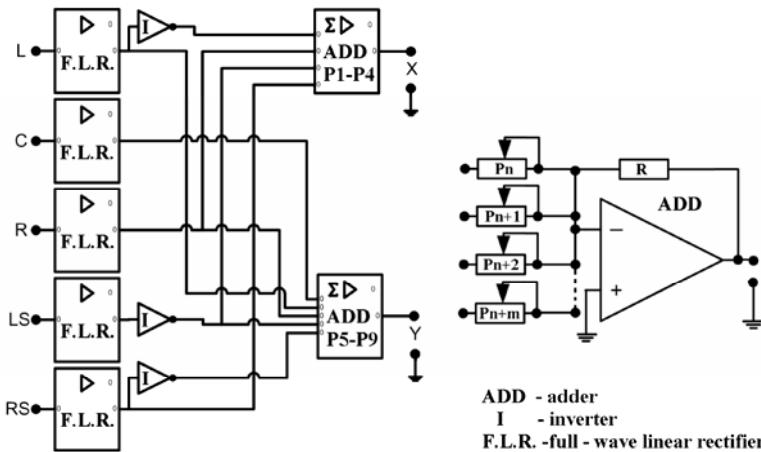
$$P_y = \frac{1}{r_C} |C| + \frac{\cos \alpha_L}{r_L} |L| + \frac{\cos \alpha_R}{r_R} |R| - \frac{\cos \beta_{LS}}{r_{LS}} |LS| - \frac{\cos \beta_{RS}}{r_{RS}} |RS| \quad (4)$$

where:  $|l|$  is a rectified full-wave signal from channels L, R, C, LS, or RS;  $\alpha_x$  and  $\beta_x$  – are related angles shown in Figure 1, and

$$r_n = \frac{r'_n}{r'_C} \quad (5)$$

is a standardized distance from related loudspeaker monitors to the listener<sup>1</sup> ( $n$  denotes: L, C, R, LS or RS), while  $r'_n$  and  $r'_C$  – are „physical” distances in meters.

For the formulated equations one can design an electronic system with operation similar to that of a goniometer.

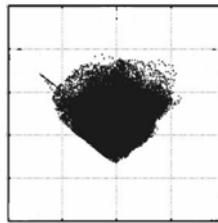


**Fig. 5.** Electronic „goniometer” system for 5.0 signals

The main components of this diagram are full-wave linear rectifiers. Potentiometers located at summing inputs are used to set the coefficients of factors  $|l|$  in equations 3 and 4. Settings of potentiometers depend on speaker set distribution geometry ( $\alpha$ ,  $\beta$  angles and  $r$  distances). For ITU-R-BS.775-1 standard:  $\alpha_L = \alpha_R = 30^\circ$ ,  $\beta_{LS} = \beta_{RS} = 75^\circ$ . Therefore, resistances P1 to P9 should be:  $P1 = P2 = P3 = P4 = 2.00R$ ,  $P5 = R$ ,  $P6 = P7 = P8 = P9 = 3.86R$  for loudspeaker sets located on a circle.

For those formulated conditions and equations 3, 4 and 5, an image displayed on the oscilloscope screen will be similar to that shown in Figure 6:

<sup>1</sup> The listener is located in the middle of coordinate system from Figure 1.

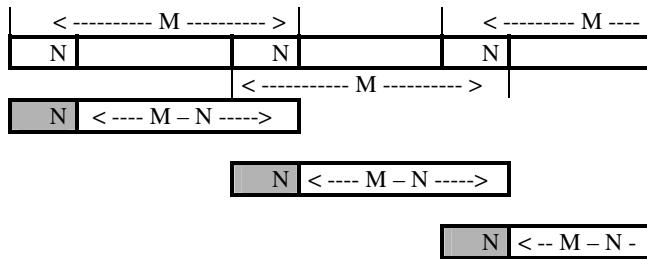


**Fig. 6.** Example of oscilloscope screen shot for 5.0 signals

Therefore, an image quality, or in fact, legibility of the obtained results, may be significantly improved using PC visualization with appropriate software. Naturally, the algorithm base is still constituted by the equations 3 – 5, but in this case signals from all channels are given in a digital form. Displaying the data points on a chart is done with so called overlapping. The overlapping process is carried out according to the following pattern:

1. Signals from all channels are divided to segments (windows) having equal numbers of M samples,
2. Currently displayed image consists of M window components for each channel,
3. Subsequent windows are overlapped on each other along the length of N components. It means that subsequent images contain M-N of "new" samples and N samples from the previous image.

The process is shown in Figure 7.



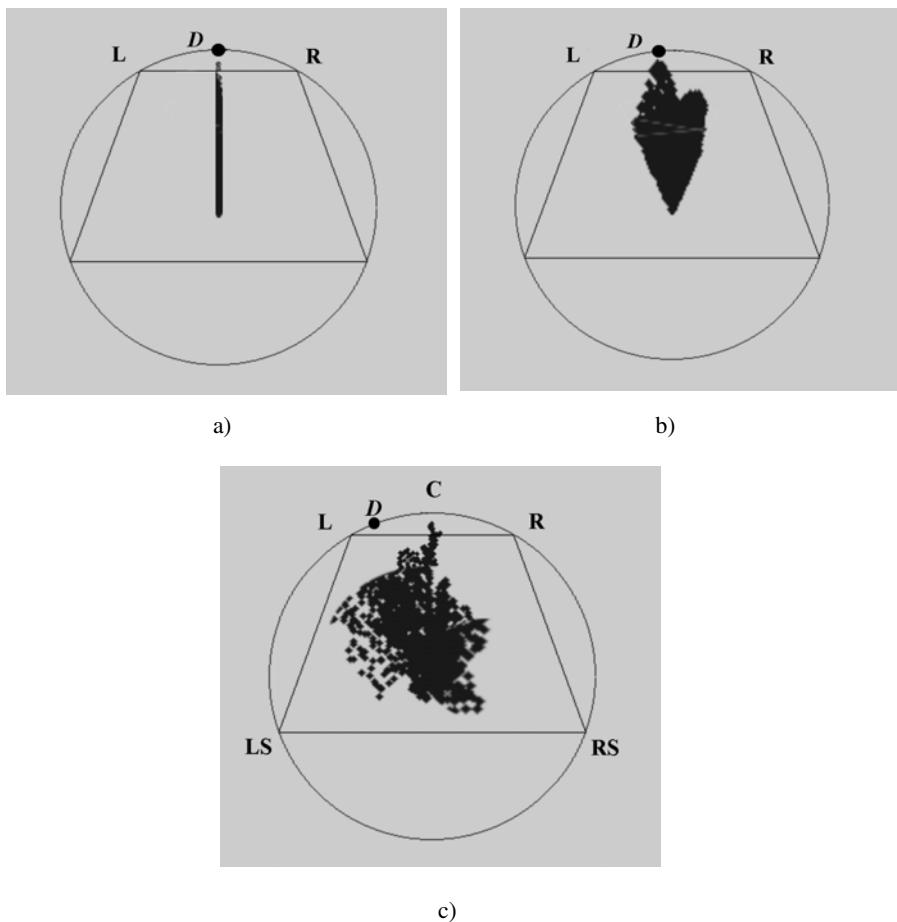
**Fig. 7.** Overlapping process for one of the channels

In addition, what is quite significant, it is also possible to show other information, e.g. direction of a dominating sound source. The location point  $D(\phi, r)$  on the circle placed on a straight line between the listener and the dominating sound source is clearly marked in the graph. Coordinates of that point are defined by the circle radius (equal to 1 for ITU-R-BS.775-1 standard) as well as:

$$\varphi = \arctan \frac{\max(P_y)}{\frac{M}{\max(P_x)}} \quad (6)$$

for a current window showing M number of samples.

The described algorithm provides a “visualization” of acoustic perspective, with examples shown in Figure 8.



**Fig. 8.** Example of acoustic perspective and dominating sound direction for: a) mono:  $L = R$ ,  $LS = RS = C = 0$ , b) stereo:  $L \neq R$ ,  $LS = RS = C = 0$ , c) 5.0 signals

#### 4 Conclusions

Quality of the proposed solution was tested by the way of experiment. The test pertained to confrontation of displayed images with audio impressions of a group of “experts” (28 individuals). The listeners were presented test recordings, and their task was to indicate (by means of a laser indicator) where a dominating sound is coming from. Tests conducted in the professional Laboratory of Sound Engineering and Ambiphonics at the Faculty of Electrical Engineering (West Pomeranian University of Technology in Szczecin). Agreement of detector readings and listener observations

for the sound front stage was 4.8%, while 9.7% was registered for the rear stage. The resulting differences could be explained by the fact that a human can localize the front sound source position with accuracy of about  $3^0$ , and the rear sound source position – with accuracy of  $6^0$  [4].

A concept of the described solution was presented at the Szczecin Studio of the Polish Television and was met with an appreciation by the producers and engineers.

## References

1. Jordan, V.: Acoustical Design of Concert Halls and Theaters. ASP Ltd., London (1980)
2. Aarts, R.M., Irwan, R.: Two – to - Five Channel Sound Processing. Journal of the Audio Engineering Society (11), 914–926 (2002)
3. Kornatowski, E.: Spatial information filtration algorithm and surround sound effect indication. Electrical Engineering. Poznan University of Technology Academic Journals (58), 7–15 (2008)
4. Sztekmiler, K.: Podstawy nagłośnienia i realizacji nagrań, NCK Warszawa (2003)
5. Wojnar, A.: Teoria sygnałów, Wydawnictwo Naukowo – Techniczne, Warszawa (1980)
6. Yamamoto, K., Iwakiri, M.: Real-Time Audio Watermarking Based on Characteristics of PCM in Digital Instrument. Journal of Information Hiding and Multimedia Signal Processing 1(2), 59–71 (2010)
7. Wey, H., Ito, A., Okamoto, T., Suzuki, Y.: Multiple Description Coding Using Time Domain Division for MP3 coded Sound Signal. Journal of Information Hiding and Multimedia Signal Processing 1(4), 269–285 (2010)

# Watermark Synchronization Based on Locally Most Stable Feature Points

Jiansheng Qian, Leida Li, and Zhaolin Lu

School of Information and Electrical Engineering  
China University of Mining and Technology  
Xuzhou 221116, P.R. China  
[reader1104@hotmail.com](mailto:reader1104@hotmail.com)

**Abstract.** A novel feature based watermark synchronization scheme is presented in this paper. The feature points are first extracted from the image and the idea of locally most stable feature points (LMSP) is proposed to generate some non-overlapped circular areas. The local regions are geometrically invariant so that they can be directly used for efficient watermark embedding and extraction. Simulation results have demonstrated the effectiveness of the proposed scheme.

**Keywords:** Watermarking, geometric attack, synchronization, locally most stable feature point.

## 1 Introduction

Extensive algorithms on watermarking have been proposed since its first appearance in the 1990s. However, geometric transformation is still one of the most challenging issues. Unlike traditional signal processing attacks, which affect watermark detection by reducing watermark energy, geometric attacks act on a watermarking scheme by changing the spatial distribution of the original pixels. In other words, the watermark signal still exists in the carrier image, but the position has changed. As a result, the synchronization should be achieved first before watermark detection.

Three kinds of methods are commonly involved in the current countermeasures, namely invariant domain embedding [1–5], template based embedding [6], and image feature based embedding [7–14]. Due to its good invisibility and generalized robustness, feature based embedding has been an active research field. The idea is to determine the positions for both embedding and extraction by referring to intrinsic image features. This kind of scheme first extracts feature points from the image and then decomposes it into a set of disjointed local regions. Then the watermark is embedded into the regions repeatedly. Tang *et al.* adopt the Mexican Hat wavelet scale interaction to extract feature points [10]. Then local circular regions are generated and the watermark is embedded. In this method, a feature point has a higher priority for watermark embedding if it has more neighboring feature points inside its circular disk. This can produce feature points (regions) in textured areas. However, the features adopted do not

necessarily have the best robustness. Similar method can be found in [13]. In [14], the scale-invariant feature transform (SIFT) is employed to generate some scale-adapted characteristic regions. A circular neighborhood constraint is applied to control the distribution of extracted features, and the value from the difference of Gaussian (DoG) function is used to measure the strength of each feature point. Other feature point based synchronization schemes can be found in [11, 12].

In feature point based watermark synchronization schemes, the selection of features is crucial for the overall performance. In order to achieve better invariance, the feature points with better robustness should be adopted. We claim that there should be a proper measure of robustness during the selection of the features. In this paper, we present a novel method to achieve this goal. The Harris features are first detected from the image and the detector response is adopted to measure the robustness, based on which the idea of locally most stable feature points is then proposed to achieve reliable watermark synchronization. The performance of the proposed scheme is then demonstrated by simulation results.

## 2 Proposed Scheme

The proposed watermark synchronization scheme is based on the Harris detector [15] and image normalization [16]. Therefore, we will first introduce these two technologies and then we present the proposed method.

### 2.1 Harris Corner Detector

Given a digital image  $f(x, y)$ , the Harris detector extracts feature points from the second-moment matrix  $M$  [15], which is defined as follows.

$$M = \begin{bmatrix} A & C \\ C & B \end{bmatrix} = \begin{bmatrix} (\frac{\partial f(x,y)}{\partial x})^2 & \frac{\partial f(x,y)}{\partial x} \cdot \frac{\partial f(x,y)}{\partial y} \\ \frac{\partial f(x,y)}{\partial x} \cdot \frac{\partial f(x,y)}{\partial y} & (\frac{\partial f(x,y)}{\partial y})^2 \end{bmatrix} \quad (1)$$

where  $\frac{\partial f(x,y)}{\partial x} = f(x, y) * [-1 \ 0 \ 1]$ ,  $\frac{\partial f(x,y)}{\partial y} = f(x, y) * [-1 \ 0 \ 1]^T$ , and  $*$  is the convolution. The detector response  $R_H$  is calculated using the following formula.

$$R_H = \text{Det}(M) - k \cdot \text{Trace}^2(M) \quad (2)$$

where  $\text{Det}(M)$  and  $\text{Trace}(M)$  are the determinant and trace of the matrix  $M$ , respectively. The parameter  $k$  is a constant (typically 0.04-0.06). The feature points can be extracted by comparing the responses with a threshold. Large threshold will produce few feature points, while smaller threshold produces more features.

### 2.2 Image Normalization

In the proposed scheme, scale normalization is employed in generating the local invariant regions. The reason is that the Harris features are not scale invariant.

Therefore, we propose to normalize the original image with regard to the scale, and then the feature points are detected.

Given an image  $f(x, y)$ , its  $(p+q)$ th order geometric moment, denoted by  $m_{pq}$ , is defined as:

$$m_{p,q} = \sum_x \sum_y x^p y^q \cdot f(x, y) \quad (3)$$

The corresponding  $(p+q)$ th order central moment, denoted by  $\mu_{pq}$ , is defined by

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q \cdot f(x, y) \quad (4)$$

where  $(\bar{x}, \bar{y})$  is the centroid of the image, with  $\bar{x} = \frac{m_{1,0}}{m_{0,0}}$  and  $\bar{y} = \frac{m_{0,1}}{m_{0,0}}$ . Scale normalization can be achieved by transforming  $f(x, y)$  into a new image  $f(x/a, y/a)$ . The scale normalization factor  $a$  can be obtained using  $a = \sqrt{\beta/m_{0,0}}$ , where  $\beta$  is a predetermined value.

For rotation normalization, define two tensors  $t^1$  and  $t^2$ .

$$t^1 = \mu_{12} + \mu_{30}, \quad t^2 = \mu_{21} + \mu_{03}, \quad (5)$$

Then the normalizing angle  $\theta$  is defined as:

$$\theta = \arctan(-t^1/t^2) \quad (6)$$

It is obvious that there are two solutions to the above equation, say  $\theta$  and  $\theta + \pi$ . In order to obtain a unique angle, another tensor  $t_3$  can be defined.

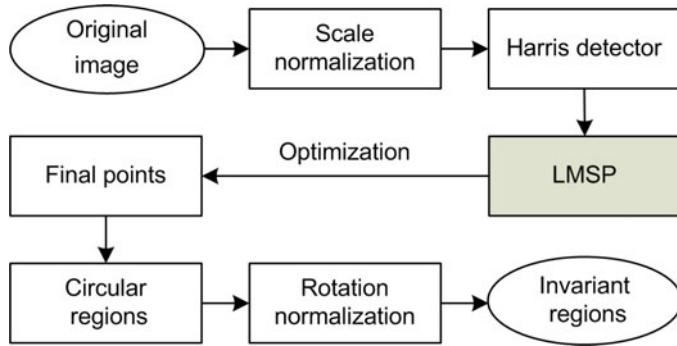
$$t_3 = -t_1 \sin \theta + t_2 \cos \theta \quad (7)$$

For  $\theta$  and  $\theta + \pi$ , only one can satisfy  $t_3 > 0$ . As a result, we obtain a unique angle  $\theta$  by making  $t_3 > 0$ . If  $t_3 < 0$ , then  $\theta = \theta + \pi$ . Given the normalization angle  $\theta$ , the image can be normalized by rotating it clockwise by angle  $\theta$ .

### 2.3 Watermark Synchronization

The diagram of the proposed watermark synchronization scheme is illustrated in Fig.1. It consists of four main phases, namely scale normalization, LMSP extraction, feature optimization and rotation normalization. The Harris detector is employed to extract feature points. However, Harris features are sensitive to image scaling. As a result, we propose to detect the feature points from the scale normalized image.

Watermark embedding requires that the local regions are independent, i.e. the local regions should not overlapped with each other. As many feature points can be extracted, we propose to use the locally most stable ones to generate the local regions. These feature points are generated as follows. For each detected feature point, a circular region is first determined. Then the detector response at the feature point is compared with the responses at other positions inside the



**Fig. 1.** Diagram of watermark synchronization

circular region. If the response of the central point achieves local maximum, it is stored as a LMSP. Otherwise, it is dropped. These LMSPs are denoted by the set  $\Omega_1$ .

$$\Omega_1 = \{(x, y) | R_H(x, y) > R_H(s, t), \forall (s, t) \in U_{x,y}\} \quad (8)$$

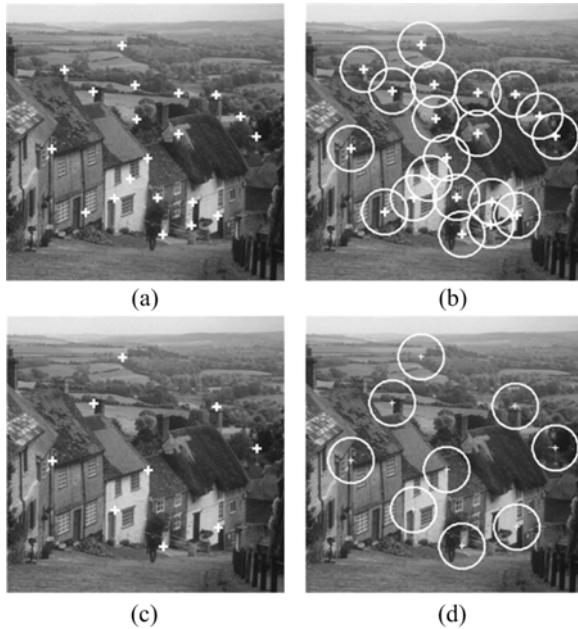
where  $U_{x,y}$  is the circular region centered at  $(x, y)$ . Fig.2(a) shows an example of the LMSPs detected from image House and Fig.2(b) shows the circular regions determined by these LMSPs. The feature point centered at each circular region is the LMSP.

In Fig.2(b), some of the circular regions overlap with others. In order to obtain non-overlapped circular regions, some of the regions must be dismissed. As a result, these LMSPs are then optimized by the following operations to generate the finally feature point set and the corresponding non-overlapped local circular regions.

- Step 1:** Choose, from  $\Omega_1$ , the feature point with the biggest response, say  $P_0$ ;
- Step 2:** Dismiss the points whose regions overlap with that of  $P_0$ ;
- Step 3:** Update  $\Omega_1$  by dismissing  $P_0$ ;
- Step 4:** If the circular regions generated using the updated points in  $\Omega_1$  still overlap with others, repeat step 1-3, otherwise go to step 5;
- Step 5:** Generate non-overlapped regions using the reserved feature points.

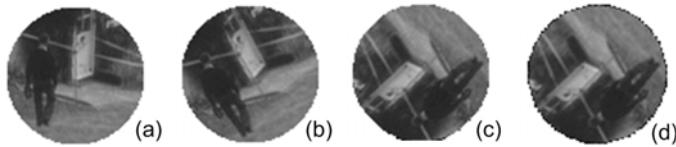
Fig.2(c) shows the finally selected points and Fig.2(d) shows the non-overlapped local circular regions. These regions have some desirable properties, such as rotation and scale invariance, which will be shown in the experiments. In our scheme, watermark embedding and extraction are implemented in these local circular regions.

It should be further noted that due to image rotations, the region from the original image and the region from the rotated image have different orientations. Fig.3(a) and Fig.3(b) show two matched regions from the original image and the image that has been rotated by 30 degree. For watermarking purpose, the distortion should be reduced before information hiding. This can be done



**Fig. 2.** An example of watermark synchronization. (a) Locally most stable points, (b) Circular regions generated using (a), (c) Finally selected points, (d) Non-overlapped circular regions generated using (c).

through rotation normalization. The local regions are normalized with respect to rotation, so that they have standard orientation. Fig.3(c) and Fig.3(d) show the normalized regions. These regions are scale and rotation invariant, so that they can be directly used for watermark embedding.

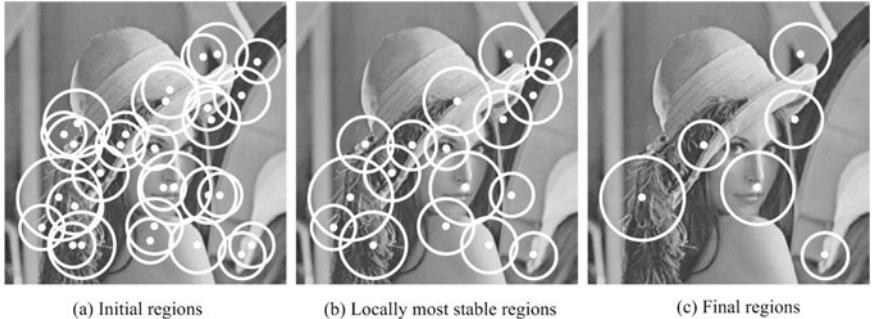


**Fig. 3.** Rotation normalization: (a) original region, (b) region from 30 rotated image, (c) normalized region of (a), (d) normalized region of (b)

## 2.4 Extension to Scale-Space Features

Although the proposed idea of locally most stable feature point is based on the Harris feature extraction method, it can be easily extended to other feature extraction methods, such as the Harris-Laplace detector [17] and the scale-invariant

feature transform (SIFT) [18]. Fig.4 illustrates the steps on how to generate the non-overlapped local circular regions using multi-scale feature points. In order to measure the robustness of the scale-space feature points quantitatively, the detection function of the Harris-Laplace detector and the Difference-of-gaussian (DoG) can be used respectively for the criteria of robustness.



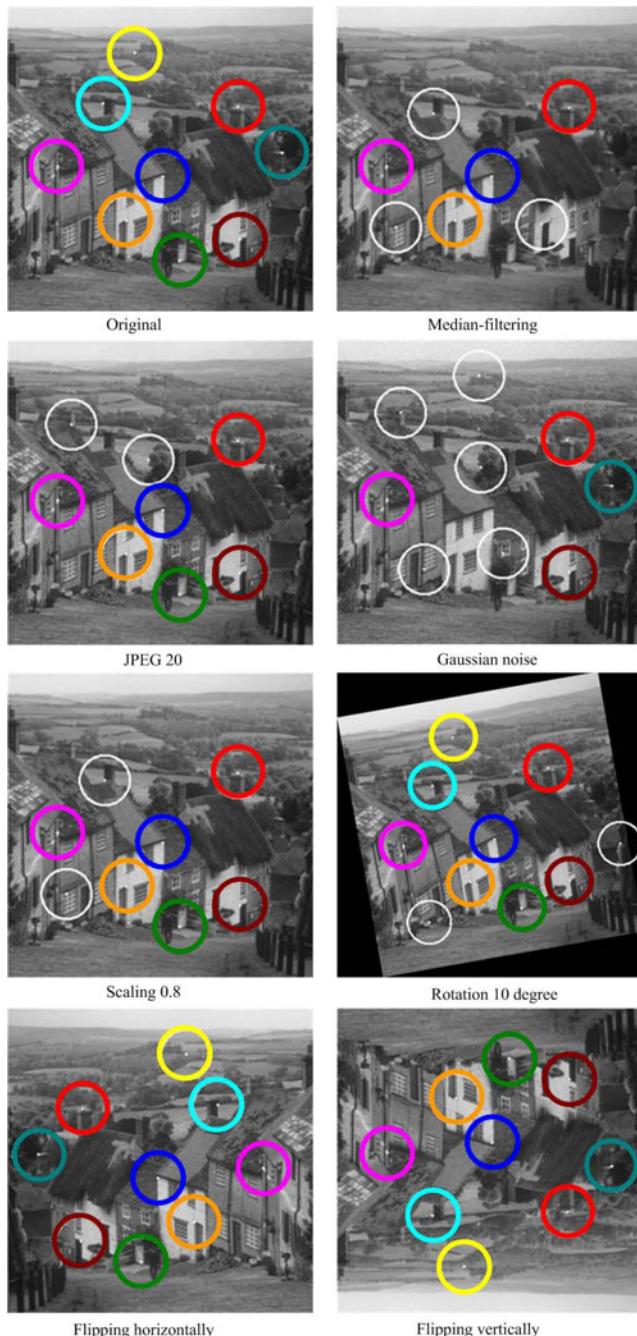
**Fig. 4.** Local regions extracted from the original image and some distorted images

### 3 Simulation Results

In order to evaluate the performance of the proposed scheme, we have done extensive experiments. The local regions are first extracted from the original images. When the images are distorted, we detect the local regions again. Then the regions from both the original and the distorted images are compared in regard of their position to determine whether they are matched.

Fig.5 shows an example of watermark synchronization on standard image House, where the matched regions are marked with the same color. In the figures, the regions with white borders denote the regions that can not be matched to the original ones. Note that the size of the test image is  $512 \times 512$ , and the radius of the local region is 40 in pixel. It can be seen Fig.5 that due to the distortions, the regions extracted from the distorted images are different from those extracted from the original images. Some new regions may be detected while some other ones may be missing. In all cases, at least four local regions can be re-detected. Especially for the flipping attacks, all the regions can be synchronized. As a result, if the watermark is embedded into the regions in the original image, they can always be extracted from the synchronized regions.

Table 1 lists the simulation results on other test images, where the denominator denotes the number of original regions and the numerator denotes the number of synchronized regions. It can be seen from both Fig.5 and Table 1 that the performance of the proposed watermark synchronization scheme is satisfactory, regardless of traditional signal processing attacks or geometric attacks. This also provides a strong basis for robust watermark embedding and extraction.



**Fig. 5.** Local regions extracted from the original image and some distorted images

**Table 1.** Performance evaluation on watermark synchronization

Attacks	Image					
	Peppers	Boat	Girl	Barbara	Baboon	House
No attack	7	9	8	7	8	9
Median filter ( $3 \times 3$ )	5/7	6/9	5/8	6/7	2/8	5/9
Added Gaussian noise	7/7	7/9	6/8	7/7	7/8	8/9
JPEG compression 50	6/7	7/9	8/8	7/7	5/8	7/9
Scaling 0.8	5/7	6/9	6/8	6/7	6/8	7/9
Rotation 10 deg	6/7	9/9	8/8	5/7	8/8	7/9
Flip horizontally	6/7	8/9	8/8	7/7	8/8	9/9
Flip vertically	7/7	9/9	8/8	6/7	8/8	8/9
Rotation 10 deg + Scaling 0.8	6/7	7/9	5/8	5/7	8/8	6/9

## 4 Conclusion

This paper presents a novel method to achieve watermark synchronization based on feature points. The proposed method is based on the quantitative measure of the feature points. Instead of searching the feature space as a whole, our method search for the feature points with the best robustness within a neighboring area, i.e. locally most stable feature points. Both signal processing attacks and geometric attacks have been involved in the experiments, and the simulation results show that the proposed method can achieve watermark synchronization purposes effectively.

## Acknowledgements

This work is supported by National Natural Science Foundation of China (60802077) and Youth Science Research Foundation of China University of Mining and Technology.

## References

1. O’Ruanaidh, J., Pun, T.: Rotation, scale and translation invariant digital image watermarking. *Signal Process.* 66(3), 303–317 (1998)
2. Lin, C., Wu, M., Bloom, J., Cox, I., Miller, M., Lui, Y.: Rotation, scale, and translation resilient watermarking of images. *IEEE Trans. Image Process.* 10(5), 767–782 (2001)
3. Zheng, D., Zhao, J., El-Saddik, A.: RST-invariant digital image watermarking based on log-polar mapping and phase correlation. *IEEE Trans. Circuits Syst. Video Technol.* 13(8), 753–765 (2003)
4. Kim, H., Lee, H.: Invariant image watermark using Zernike moments. *IEEE Trans. Circuits Syst. Video Technol.* 13(8), 766–775 (2003)

5. Xin, Y., Liao, S., Pawlak, M.: Circularly orthogonal moments for geometrically robust image watermarking. *Pattern Recognit.* 40(12), 3740–3752 (2007)
6. Pereira, S., Pun, T.: Robust template matching for affine resistant image watermarks. *IEEE Trans. Image Process.* 9(6), 1123–1129 (2000)
7. Kutter, M., Bhattacharjee, S., Ebrahimi, T.: Towards second generation watermarking schemes. In: Proc. IEEE Int. Conf. Image Process., Kobe, Japan., vol. 1, pp. 320–323 (1999)
8. Bas, P., Chassery, J., Macq, B.: Geometrically invariant watermarking using feature points. *IEEE Trans. Image Process.* 11(9), 1014–1028 (2002)
9. Qi, X., Qi, J.: A robust content-based digital image watermarking scheme. *Signal Process.* 87(6), 1264–1280 (2007)
10. Tang, C., Hang, H.: A feature-based robust digital image watermarking scheme. *IEEE Trans. Signal Process.* 51(4), 950–959 (2003)
11. Seo, J., Yoo, C.: Localized image watermarking based on feature points of scale-space representation. *Pattern Recognit.* 37(7), 1365–1375 (2004)
12. Seo, J., Yoo, C.: Image watermarking based on invariant regions of scale-space representation. *IEEE Trans. Signal Process.* 54(4), 1537–1549 (2004)
13. Wang, X., Wu, J., Niu, P.: A new digital image watermarking algorithm resilient to desynchronization attacks. *IEEE Trans. Inf. Forensics Security* 2(4), 655–663 (2007)
14. Lee, H., Kim, H., Lee, H.: Robust image watermarking using local invariant features. *Opt. Eng.* 45(3), 037002(1–11) (2006)
15. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proceedings of 4th Alvey Vision Conference, pp. 147–151 (1988)
16. Alghoniemy, M., Tewfik, A.: Geometric invariance in image watermarking. *IEEE Trans. Image Process.* 13(2), 145–153 (2004)
17. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. *Int. J. Comput. Vis.* 60(1), 63–86 (2004)
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60(2), 91–110 (2004)

# Audio Watermarking with HOS-Based Cepstrum Feature

Bo-Lin Kuo, Chih-Cheng Lo, Chi-Hua Liu, Bin-Yih Liao, and Jeng-Shyang Pan

Department of Electronic Engineering

Kaohsiung University of Applied Science, Taiwan

plkuo@hotmail.com, hua5216@yahoo.com.tw, louccc@hotmail.com,  
byliao@cc.kuas.edu.tw, jspan@cc.kuas.edu.tw

**Abstract.** In this paper, we propose a more robust scheme of audio watermarking which is based on cepstrum (or cepstral coefficients) and HOS (higher-order statistics) schemes. This scheme is a zero-watermarking one for the reason to maintain the audio quality. The audio signal is firstly kurtosis-estimated and feature-recognized, and then analyzed via CC and HOS, respectively, to extract the essential parameters and characteristics, which are then used for information embedding and extracting. The achievement of the proposed scheme could outperform the previous innovative one [1].

## 1 Introduction

Schemes for embedding watermark/message into multimedia contents are widely available and mainly focus on image and video but less on the area of audio. The main reason is the human auditory system (HAS) is extremely more sensitive than human visual system (HAS). Traditional audio-watermarking mechanism is to hide the watermark directly into the original information media for copyright protection. No matter what schemes, in spatial or frequency domain, are used for watermarking, two main inevitable problems will be met. One is the watermarking process will alert the original audio message, which inevitably reduce the audio quality. The other one is the necessary tradeoff (conflict) between robustness and imperceptibility of the watermarking schemes.

In order to solve the above two problems, zero-watermarking schemes [1-3] are proposed, in which, instead of directly embedding the watermark message into the host audio media, some essential and important features are extracted from it and they are then used for watermark extraction or detection. Zero-watermarking schemes certainly maintain both the quality of audio signal and achieve the embedding ability to some degree.

The cepstrum is a common transform used to gain information about the peak and fine variations of spectrum from a speech signal.

## 2 Previous Works

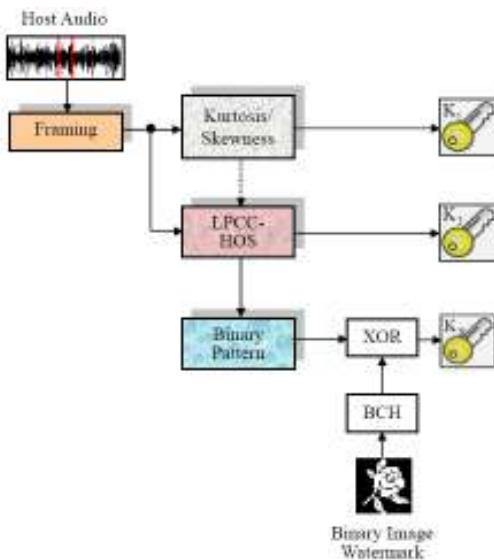
Zero-watermarking schemes use essential and robust features of audio signal as reference to recover watermark. During embedding process, different features (more robust)

are constructed from the host audio media to be watermarked and then managed into some sets of binary patterns, and finally, exclusive-or (XOR) operation is performed between these sets of binary form of pattern and watermark to generate secret key(s). Those robust features form some keys which may be independent ones or used together to generate a main key which is related to the binary image watermark to be embedded. During watermark recovering, we also firstly frame the test audio to extract kernel features and binary them as in embedding phase. Then, those secret keys are used to identify and test these features for further process. At last, we perform the XOR operation to the final binary pattern and the main key to recover the binary watermark.

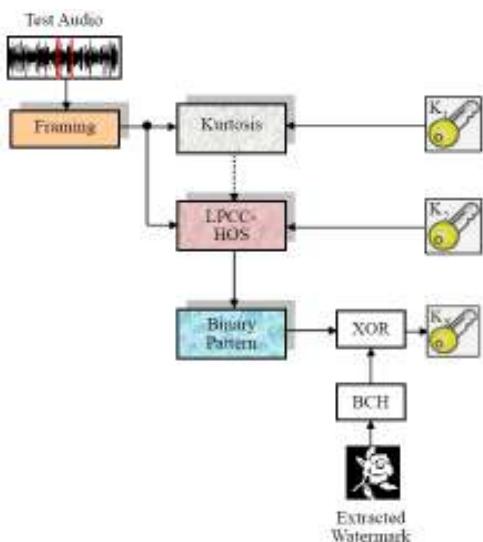
In [1], the authors take advantage of three important features, the multi-resolution characteristic of DWT, the energy compression effect of DCT, and the Gaussian noise suppression property of HOS, to extract from the audio signal and use them for watermark recovery. With the scheme in [1], the audio signal is firstly segmented into frames and the energy value of each frame is calculated. Some of these energy values are then sorted and selected the indices as the first secret key. Next, each selected frame is performed with DWT to get its coarse signal and then with DCT to take the low-frequency coefficient. Then, these coefficients are operated with 4<sup>th</sup>-order cumulant (HOS) and the yielded values are further absolutely manipulated. These absolute values are sorted and selected the indices as the second secrete key. Finally, the above cumulant values are signed with only 0 and 1. This binary pattern is XOR operated with the binary watermark to get the third (final) key. This novel zero-watermarking scheme certainly demonstrates its effectiveness in terms of inaudibility, detection reliability, and robustness with various experiments.

### **3 Audio Zero Watermarking with HOS-Based Linear Prediction Cepstral Feature**

In our study, we propose a more robust and computation-efficient zero-watermarking scheme to overcome the problems of magnitude variation and complex procedure for secret-key generation mentioned above. The audio signal is kurtosis feature estimated first for its value tends to remain constant whenever the audio content is not altered significantly or in Gaussian noise environment, and then further recognized and analyzed via linear prediction cepstral coefficients (LPCC) and HOS (4<sup>th</sup>-order cumulant), respectively, to extract the essential invariant and noise-immunity parameters and characteristics, which are then used for information embedding and recovering. At the same time, we use BCH coding algorithm (invented in 1959 by Hocquenghem, and independently in 1960 by Bose and Ray-Chaudhuri, it is a multilevel cyclic variable-length digital error-correcting code used to correct multiple random error patterns) to encode the binary image watermark to reduce the length of secret key. The embedding/recovering is shown in Figs. 1 and 2, respectively. The merit of our proposed HOS-LPCC scheme achieves the variation compression with kurtosis, invariant feature keeping with LPCC and Gausian noise compression with HOS. 3. LPCC analysis is mainly to calculate the linear predicted coefficients and which is



**Fig. 1.** Proposed structure of audio zero-watermarking for message embedding



**Fig. 2.** Proposed structure of audio zero-watermarking for message recovering

one of the main parameters in speech signal recognition. LPCC states that the predicted speech signal can be estimated with the combination of previous  $p$  samples. The main purpose of LPCC is to represent the features of peak and fine variation in speech spectrum domain.

Also, assume the host audio signal as

$$\mathbf{A} = \{a(i) | i = 0, \dots, L_A - 1\} \quad (1)$$

Then, it is segmented into  $L$  frames,

$$\mathbf{A} = \{\mathbf{F}_m | m = 0, \dots, L - 1, L > 2MN\} \quad (2)$$

Then, the absolute kurtosis value of each frame is calculated. Some of these kurtosis values are then sorted and selected the indices of larger ones as the first secret key

$$\mathbf{K}_1 = \{i(k) | i(k) \in \{0, \dots, L - 1\}, k = 0, \dots, T - 1\} \quad (3)$$

Let's define the predicted audio one for each frame as

$$\tilde{f}_m(i) = a_2 f_m(i-1) + a_3 f_m(i-2) + \dots + a_{p+1} f_m(i-p) \quad (4)$$

The audio signal is firstly transformed to its frequency components, then natural logarithm operation is employed to these components, and finally, inversely frequency transformation is used to obtain the cepstral coefficients as

$$\begin{aligned} c_{f_m,1} &= a_2, \\ c_{f_m,n} &= -a_n + \frac{1}{n} \sum_{k=1}^{n-1} [-(n-k)a_k c_{n-k}], 1 \leq n < p \\ c_{f_m,n} &= -a_n + \sum_{k=n}^p \left[ -\frac{n-k}{n} a_k c_{n-k} \right], n \geq p \end{aligned} \quad (5)$$

These coefficients of each frame are the kernel features we need for the following watermarking as it is less invariable to many attacks. These coefficients of each frame are the kernel features we need for the following watermarking as it is less invariable to many attacks.

LPCC coefficients in Eq. (5) are operated with 4<sup>th</sup>-order cumulant (HOS) and the yielded values

$$\mathbf{C}_{i(k)}^{(LH)} = \{c_{i(k)}^{(LH)}(m), k = 0, \dots, T - 1, m = 0, \dots, \frac{L_f}{2^H} - 1\} \quad (6)$$

which are further manipulated with absolute operation. These absolute values are sorted and selected the indices as the second secret key.

$$\mathbf{K}_2 = \{i_{i(k)}^{(LH)}(p) | i_{i(k)}^{(LH)}(p) \in \{0, \dots, \frac{L_f}{2^H} - 1\}, k = 0, \dots, T - 1\} \quad (7)$$

Here, if Gaussian noise  $\sum_{k=2}^p G_p$  is added to corresponding items, respectively, and

then employing higher-order cumulant operator  $cum_{\geq 3}\{\bullet\}$  to them, it will yield

$$\begin{aligned}
cum_{\geq 3}(c_{f_m,1} + G_2) &\approx cum_{\geq 3}(a_2) = 0 \\
cum_{\geq 3}(c_{f_m,n} + \sum_{k=1}^{n-1} G_k) &\approx \frac{1}{n} \prod_{k=1}^{n-1} - (n-k)a_k cum_{\geq 3}(c_{n-k}), 1 \leq n < p \\
cum_{\geq 3}(c_{f_m,n} + \sum_{k=n}^p G_k) &\approx \prod_{k=n}^p - \frac{(n-k)}{n} a_k cum_{\geq 3}(c_{n-k}), n \geq p \\
&\text{if } c_{n-k} \text{ is non-Gaussian distribution}
\end{aligned} \tag{8}$$

Finally, the absolute values of the above cumulant ones are sorted and arranged in decreasing order. Then, the corresponding cumulant values

$$\mathbf{D}_{i(k)}^{(LH)} = \{d_{i(k)}^{(LH)}(p), k = 0, \dots, T-1, p = 0, \dots, P-1\} \tag{9}$$

with respect to these indices are signed with binary ones, 0 or 1. This binary pattern,

$$\mathbf{B}_{i(k)}^{(LH)} = \{b_{i(k)}^{(LH)}(p), k = 0, \dots, T-1, p = 0, \dots, P-1\} \tag{10}$$

where

$$b_{i(k)}^{(LH)}(p) = \begin{cases} 1, & \text{for } d_{i(k)}^{(LH)}(p) \geq 0 \\ 0, & \text{for } d_{i(k)}^{(LH)}(p) < 0 \end{cases} \tag{11}$$

is XOR operated with the binary image watermark  $\mathbf{W}$  to get the third (final) key  $\mathbf{K}_3$ .

Also, during watermark recovering phase, the procedure can be carried out without the host audio. The test audio signal  $\tilde{\mathbf{A}} = \{\tilde{a}(i) | i = 0, \dots, L_A - 1\}$  is also firstly divided into frames  $\tilde{\mathbf{A}} = \{\tilde{\mathbf{F}}_m | m = 0, \dots, L-1, L > 2MN\}$ , and then, the first key  $\mathbf{K}_1$  is used to select frames,  $\tilde{\mathbf{F}}_{i(k)} = \{\tilde{f}_{i(k)} | k = 0, \dots, T-1\}$ , with larger absolute kurtosis values. Next, LPCC and then HOS operations are performed on the selected frames with larger absolute kurtosis values, yielding

$$\tilde{\mathbf{C}}_{i(k)}^{(LH)} = \{\tilde{c}_{i(k)}^{(LH)}(m), k = 0, \dots, T-1, m = 0, \dots, \frac{L_F}{2^H} - 1\} \tag{12}$$

With the second secret key  $\mathbf{K}_2$ , we select  $P$  elements from  $\tilde{\mathbf{C}}_{i(k)}^{(LH)}$  and obtains

$$\tilde{\mathbf{D}}_{i(k)}^{(LH)} = \{\tilde{d}_{i(k)}^{(LH)}(p), k = 0, \dots, T-1, p = 0, \dots, P-1\} \tag{13}$$

Finally, the estimated binary pattern

$$\tilde{\mathbf{B}}_{i(k)}^{(LH)} = \{\tilde{b}_{i(k)}^{(LH)}(p), k = 0, \dots, T-1, p = 0, \dots, P-1\} \tag{14}$$

is generated according to  $\tilde{b}_{i(k)}^{(LH)}(p) = 1$  for  $\tilde{d}_{i(k)}^{(LH)}(p) \geq 0$  or  $\tilde{b}_{i(k)}^{(LH)}(p) = 0$  for  $\tilde{d}_{i(k)}^{(LH)}(p) < 0$ . The final step is to perform XOR operation between the estimated binary pattern

$\tilde{\mathbf{B}}_{i(k)}$  and the third key  $\mathbf{K}_3$  (the watermark detection key) to recover back the binary image watermark  $\tilde{\mathbf{W}}$ .

The watermark recovering procedure is reversed as embedding one and the performance evaluation, without loss of generality, is also with SNR (signal-to-noise ration), NC (normalized cross-correlation) and BER (bit error rate). They are defined as follows, respectively.

$$SNR(\mathbf{A}, \tilde{\mathbf{A}}) \equiv 10 \log \left\{ \frac{\sum_{i=0}^{L_A-1} a^2(i)}{\sum_{i=0}^{L_A-1} [a(i) - \tilde{a}(i)]^2} \right\} \quad (15)$$

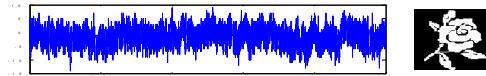
$$NC(\mathbf{W}, \tilde{\mathbf{W}}) \equiv \frac{\sum_{i=0}^{M-1N-1} \sum_{j=0}^{N-1} w(i, j) \tilde{w}(i, j)}{\sqrt{\sum_{i=0}^{M-1N-1} \sum_{j=0}^{N-1} w^2(i, j)} \sqrt{\sum_{i=0}^{M-1N-1} \sum_{j=0}^{N-1} \tilde{w}^2(i, j)}} \quad (16)$$

$$BER \equiv \frac{B}{M \times N} \times 100\% \quad (17)$$

where  $\tilde{\mathbf{A}}$ ,  $\tilde{\mathbf{W}}$  and  $B$  are the attacked audio, extracted watermark, and the number of erroneously extracted watermark bits, respectively.

## 4 Experimental Results

The feasibility of the proposed scheme is also demonstrated according to the performance, detection reliability, and robustness. The experimental results are also compared to those of scheme [1]. The experiment is arranged with the host audio and binary image watermark with  $64 \times 64$  pixels are shown as in Fig. 3. The attacks follow the “Stirmark for Audio v0.2” and the manipulation with “CoolEditor”.

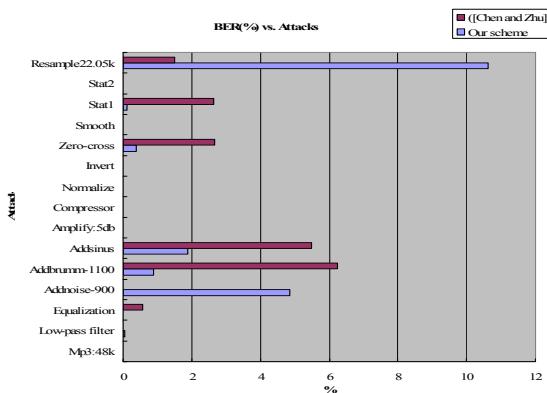


**Fig. 3.** Host audio and watermarking halftone image for test

Various attacks are summarized in Table 1, which shows that the proposed scheme outperforms the scheme in [1]. Also, as shown in Fig. 4, the BER comparison between these two schemes demonstrates the robustness of ours.

**Table 1.** Experimental results and comparison with Chen and Zhu's scheme [1]

No.	Attacks	[Chen and Zhu]		Our scheme		
		SNR	NC	SNR	NC	Extracted watermark
1	Mp3:48k	45.54	1	48.12	1	
2	Low-Pass filter	76.54	1	73.25	0.99	
3	Equalization	11.71	0.97	73.25	0.99	
4	Add noise-900	16.22	0.99	12.34	1	
5	Add brumm-1100	12.30	0.98	24.01	0.92	
6	Add sinus	10.77	0.92	20.62	0.99	
7	Amplify:5db	3.56	1	19.10	0.97	
8	Compressor	18.75	0.99	6.02	1	
9	Normalize	59.88	1	13.49	1	
10	Invert	-6.01	0.997	57.46	1	
11	Zero-cross	20.27	0.97	33.13	0.99	
12	Smooth	25.08	0.999	33.13	0.99	
13	Stat1	20.24	0.94	22.32	1	
14	Stat2	32.04	1	19.79	0.99	
15	Resample 22.05k	8.67	0.6	10.22	0.84	

**Fig. 4.** Bit error rate (BER) vs. attacks

## 5 Conclusion

Experimental results demonstrate the proposed audio watermarking scheme achieves five essential and general properties of transparency, robustness, security, reliability,

and blindness. Transparency means maintaining the quality of audio signal. Robustness means the watermark will survive well under various attacks. The kernel features of the audio signal itself and keys guarantee the security of the proposed scheme. The proposed scheme is a blind watermarking one since the watermark can be recovered without need of the original audio.

## References

- [1] Chen, N., Zhu, J.: A robust zero-watermarking algorithm for audio. EURASIP Journal on Advances in Signal Processing 2008, 1–7, Article ID 453580
- [2] Sun, T., Quan, W., Wang, S.-X.: Zero-watermark watermarking for image authentication. In: Proceedings of the Signal and Image Processing, Kauai, Hawaii, USA, pp. 503–508 (August 2002)
- [3] Wen, Q., Sun, T.-F., Wang, S.-X.: Concept and application of zero-watermark. Tien Tzu Hsueh Pao/Acta Electronica Sinica 31(2), 214–216 (2003)

# Processing Certificate of Authorization with Watermark Based on Grid Environment

Heng-Sheng Chen, Tsang-Yean Lee, and Huey-Ming Lee

Department of Information Management, Chinese Culture University  
55, Hwa-Kung Road, Yang-Ming-San, Taipei (11114), Taiwan  
{chenhs, tylee, hmlee}@faculty.pccu.edu.tw

**Abstract.** Based on the grid computing architecture, we divided grid nodes into supervisor grid node and execute grid nod. In this study, we propose the safe treatment of the certificate of authorization in this article. Both of the owner and grantee keep the certificate of authorization. We insert the watermark in the certificate of authorization to be one file. We set encryption data table and use it to encrypt the file with different key to produce the files of cipher text. Both of the owner and grantee keep their files. We create certificate information database in supervisor grid node. When both sides want to confirm, we use their keys stored in supervisor grid node to decrypt both files. If the certificate of authorization and watermark are the same, then they are correct. The steps that the certificate of authorization is encrypted and decrypted will be safer.

**Keywords:** Authorization, Certificate, Decryption, Encryption, Grid.

## 1 Introduction

“Grid” was used to denote a proposed distributed computing infrastructure in the mid 1990[6]. In grid environment, users may access the computational resources at many sites [5]. Lee et al. [7] proposed a dynamic analyzing resources model which can receive the information about CPU usage, number of running jobs of each grid node resource to achieve load-balancing. The functions of security system are security, authenticity, integrity, non-repudiation, data confidentiality and access control [1-3]. McEliece [9] used algebraic coding theory to propose public key. Miyaguchi [10] developed the fast data encipherment algorithm (FEAL-8). Lee and Lee [8] used the basic computer operations, such as insertion, rotation, transposition, shift, complement and pack to design encryption and decryption algorithms. Chen et al. [4] insert the encryption data table into cipher text and uses this encryption data table to encrypt and decrypt.

In this study, we have the certificate of authorization in both of owner and grantee. We select the watermark and combine the certificate of authorization to one file. We send this combined file and information of owner and grantee to supervisor. We create certificate information data base in supervisor grid node. We set encryption data table and use it to encrypt the combine file to produce the different encrypted files. Each file is kept to both owner and grantee sides. When we want to confirm, we send these

encrypted files to supervisor and use their keys stored in the certificate information data base of the supervisor to decrypt both files and get two kinds of certificate of authorization and watermark. If they are the same, they are correct.

## 2 The Proposed Method Description

The proposed method is to process the certificate of authorization (COA) in security based on grid environment.. We insert the selected watermark (WM) and print two copies to be kept in the owner and grantee sides. We combine the COA and WM. We send this file and information of owner and grantee to supervisor and create certificate information data base (CIDS). We set the encryption data table (EDT). We get different keys of owner and grantee and store in the CIDS. We use the EDT and these different keys to encrypt this combined file to produce two encrypted files. We also insert the EDT to encrypted files. We send these encrypted files back and each file is kept by owner and grantee sides. When we want to confirm, we receive these encrypted files from owner and grantee sides. We send these encrypted files to supervisor to verify. We extract the EDT from these encrypted files. We use the EDT and different key stored in the data base to decrypt these two encrypted files to produce two kinds of COA and WM. If these two copies of files are the same, we return message that the COA is true. We explain the process.

### 2.1 Supervisor Grid Node

In the supervisor grid node, we explain to process the certificate of authorization. We build files and tables of supervisor grid node as follows:

1. Build files and tables of supervisor grid node
  - (1) Source information. We combine COA and WM to one file.
  - (2) Update information
    - (a) Receive information  
Get the code (N), title, book number, owner-id, password, grantee-id, password and length of COA and WM as Table 1.

**Table 1.** Information of COA

Code N	Title	Book No.	Owner ID Password	Grantee ID Password	Length of COA and WM
-----------	-------	-------------	----------------------	------------------------	-------------------------

- (b) Create COA  
Get the code (F), title, book number; owner-id, password, grantee-id, password and file as Table 2.
- (c) Delete COA  
Get code (D), title, book number, owner-id, password, grantee-id, password as Table 3.

**Table 2.** File of COA

Code F	Title	Book No.	Owner ID Password	Grantee ID Password	File
-----------	-------	-------------	----------------------	------------------------	------

**Table 3.** Information of certificate of authorization

Code D	Title	Book No.	Owner ID Password	Grantee ID Password
-----------	-------	-------------	----------------------	------------------------

## (d) Conform COA

Get code (C), title, book number, owner-id, password, file of owner, grantee-id, password and file of grantee as Table 4.

**Table 4.** Cipher text of COA

Code C	Title	Book No.	Owner ID Password	File of Owner	Grantee ID Password	File of Grantee
-----------	-------	-------------	----------------------	------------------	------------------------	--------------------

## (e) Inquire information

Get code (I), title and book number as Table 5.

**Table 5.** Inquire user information

Code I	Title	Book No..
-----------	-------	--------------

## (f) Return error message

Return code (E) of error message as Table 6.

**Table 6.** Return error message

Code E	Message
-----------	---------

## (3) Create certificate information data base (CIDB).

- (a) Get data of Table 1.
- (b) Get key code of owner and grantee;
- (c) Create CIDB as Table 7

**Table 7.** Certificate information data base

Title	Book NO.	Owner ID Password	Key Code	Grantee ID Password	Key Code	Length of COA WM
-------	-------------	----------------------	-------------	------------------------	-------------	---------------------

(4) Encryption Data Table (EDT)

The encryption data table contains format code (FC), length of left shift table (LLFT), number blocks (NB), rotated byte (RB), left shift table (LST), displacement offset (DO) and direction flag (DF) as Table 8.

**Table 8.** Encryption data table (EDT)

FC	LLST	NB	RB	LST	DO	DF
----	------	----	----	-----	----	----

2. Processes of supervisor grid node

To process certificate of authorization, we have the following operations.

(1) Create certificate information data base

- (a) Receive code (N) as Table 1, use title and book number as key to find the entry in the CIDB. If it exists then it is error and exit;
- (b) Get key code of owner and grantee;
- (c) Use title and book number as key to insert to CIDB as Table 7.

(2) Create encrypted certificate of authorization

- (a) Receive code (F) as Table 2;
- (b) Use title and book number as key to find the entry in the CIDB;
- (c) If non-exist, returns error and exit;
- (d) Get key codes of owner and grantee. (e) Set encryption data table.
- (f) Use these key codes and encryption data table to encrypt the file to produce two encrypted files and return.

(3) Delete certificate of authorization

- (a) Receive code (D) as Table 3, use title and book number as key to find the entry in the CIDB;
- (b) If it exists then delete the entry, else returns error.

(4) Confirm certificate of authorization

- (a) Receive code (C) as Table 4;
- (b) Use title and book number as key to find the entry in the CIDB; no
- (c) If it non-exists then returns error and exits;
- (d) Use the key code of owner and grantee to decrypt these files to produce two copies of certificate of authorization and watermark;
- (e) If two copies are same, returns correct, else returns error.

(5) Inquire information.

- (a) Receive code I as Table 5;
- (b) Use title and book number as key to find the entry in the CIDB;
- (c) We list the information of owner and grantee. .

(6) Return error message

- (a) Set code E as Table 6; (b) Set error message.

## 2.2 Execute Grid Node

In the execute grid node, we explain to process the certificate of authorization as follows.

## 1. Build files and tables of execute grid node

## (1) Source information

(a) COA file. File of the primitive COA is as Table 9.

**Table 9.** Certificate of authorizationCertificate of authorization

(b) WM. File of watermark is as Table 10. It can be either selected from watermark database or a new one and is created to watermark database,

**Table 10.** WatermarkWatermark

(c) COA file with WM. Combine COA file with WM as Table 11.

**Table 11.** COA with WM

Certificate of authorization	Watermark
------------------------------	-----------

## (2) Update information

- (a) Receive information as Table 1.
- (b) Create encrypted certificate of authorization as Table 2..
- (c) Delete certificate information of authorization as Table 3.
- (d) Confirm certificate of authorization as Table 4.
- (e) Inquire information as Table 5.

## (3) Return message

Receive return message as Table 12.

**Table 12.** Return messageReturn Message

## 2. Processes of execute grid node

To process certificate of authorization, we have the following operations.

## (1) Create certificate of authorization

- (a) Get the primitive COA as Table 9;
- (b) Select the WM as Table 10 and the position of watermark first, type two copies of the COA with WM and are kept in owner and grantee sides;
- (c) Combine the COA and WM to one file as Table 11.

## (2) Send information of owner and grantee.

Send code N as Table 1 to supervisor to create data base.

## (3) Create encrypted certificate of authorization.

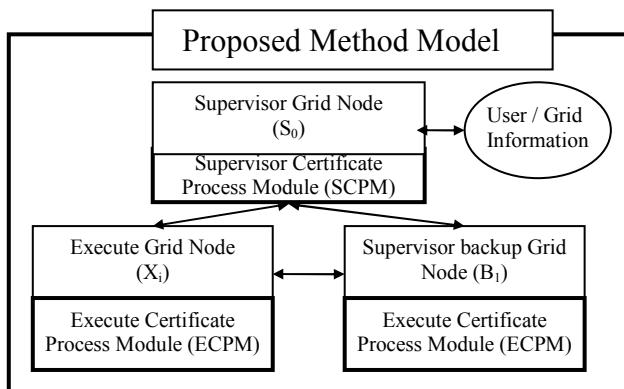
Send code F as Table 2 to supervisor grid node;

Receive encrypted certificate of authorization from supervisor grid node.

- (4) Delete certificate of authorization  
Send code D as Table 3 to supervisor grid node;
- (5) Confirm certificate of authorization
  - (a) Get the files of certificate of authorization to be kept in both sides;
  - (b) Send code C as Table 4 to supervisor grid node;
  - (c) Receive return message from supervisor grid node as Table 6..
- (6) Inquire information
  - (a) Send code I as Table 5; (b) Get information of owner and grantee.

### 3 The Proposed Model

In this section, we present the framework of COA with WM process model based on grid environment. We divide grid nodes into supervisor grid node ( $S_0$ ) and execute grid nodes ( $X_i$ ). We also present the supervisor certificate process module (SCPM) on the supervisor grid node, execute certificate process module (ECPM) on the execute grid nodes, as shown in Fig. 1.



**Fig. 1.** Framework of the proposed model

#### 3.1 Supervisor Grid Node

We present the supervisor certificate process module (SCPM) on the supervisor grid node. It has 4 components and their functions are as the follows:

1. Supervisor receive information component (SRIC):  
SRIC receives information from the execute grid node. It calls subprogram depending on code.
  - (a) If code is C and F, It calls SPEDC.;
  - (b) If code is N, D and I, it calls SPIC ; (c) others is error
2. Supervisor process encryption/decryption component (SPEDC):  
SPEDC processes encryption and decryption of file depending on code.

- (a) If code is (F), uses title and book number as key to find the entry in the CIDB. If non-exist, returns error and exit. Get key code of owner and grantee in CIDB. Use these key codes to call CAEC (Certificate of Authorization Encryption Component) to encrypt the file to produce two files of cipher text and return.
  - (b) If code is (C), it uses title and book number as key to find the entry in the CIDB. If it is non-exist then returns error and exits. Uses these key codes of owner and grantee in CIDB and calls CADC (Certificate of Authorization Decryption Component) to decrypt these files to produce two copies of certificate of authorization and watermark. If two copies are same, returns correct, else returns error.
3. Supervisor process information component (SPIC)  
 SPIC processes information of certificate of authorization.
- (a) If code is (N), it receives information from the execute grid node. It randomly gets key codes of owner and grantee and store to CIDB.
  - (b) If code is (D), it uses title and book number as key to find the entry in the CIDB. If it exists then delete the entry, else returns error;
  - (c) If code is (I), it uses title and bookumber as key to find the entry in the CIDB. If it is non-exist then returns error and exits. Lists information of owner and grantee
4. Supervisor send information component (SSIC):  
 SSIC sends information to grid node.

## 3.2 Execute Grid Node

We present the execute certificate process module (ECPM) on the execute grid node. It has 6 components and their functions are as the follows:

1. Execute receive certificate component (ERCC):  
 ERCC receives information. If the executed grid node receives information from supervisor, it calls EPSIC (Execute Process Supervisor Information Component). If it wants to create COA, it calls ECCC (Execute Create Certificate Component), otherwise it calls EPIC (Execute Process Information Component).
2. ECCC combines COA as Table 9 and WM as Table 10 to one file as Table 11.
3. Execute process information component (EPIC):  
 EPIC processes to send information to supervisor. It has the following formats.
  - (1) Send information of owner and grantee. Set code as N and input title, book number, owner-id, password, grantee-id and password as Table 1
  - (2) Receive encrypted COA. Set code as F and title; book number; owner-id, password, grantee-id, password and file as Table 2.
  - (3) Delete COA. Set code as D and title, book number, owner-id, password, grantee-id, password as Table 3.
  - (4) Confirm COA. Set code as C and title, book number, owner-id, password, file of owner, grantee-id, password and file of grantee as Table 4.
  - (5) Inquire information. Set code I and input title and book number as Table 5.

4. Execute process supervisor information component (EPSIC)

EPSIC gets information from supervisor. From the received code, it has following process depending on the codes.

- (1) Code N. Store information in supervisor grid node..
- (2) Code D. Delete title and book number.
- (3) Code C. Receives return code from supervisor to verify COA.
- (4) Code I. Receives information from supervisor.
- (5) Code F. Gets encrypted COA.
- (6) Code E. Gets error message.

5. Execute send information component (ESIC):

ESIC sends or returns information to the supervisor.

## 4 Encryption and Decryption Algorithms

### 4.1 Encryption Algorithm (CAEC Certificate of Authorization Encryption)

Based on Chen et al.[4], we propose the encryption algorithm as following steps.

1. Get files.

Get COA with WM. Store it to ST (symbol table).

2. Set key code and encryption data table (EDT)

Set key code of this file. Set fields of EDT in Table 8.

3. Set left shift table.

Set value of left shift table (LST) and its length is LLST. Store LLST to EDT.

4. Rotate symbol table

(1) Get NB from EDT. Divide ST (symbol table) into NB blocks;

(2) Get RB from EDT. From the beginning block, we repeat to rotate each block left RB bytes and right RB bytes.

(3) We get the symbol table after rotation (STAR);

5. Left shift each byte

(1) Get left shift table from LST

(2) Following each half byte of LST, left shift each byte of STAR;

(3) We get symbol table after shift (STAS).

6. Position exchange:

(1) Get displace offset (DO) of EDT;

(2) Extract each byte of STAS offset DO bytes and store to symbol table after extract (STAE) to end of data;

(3) Decrease DO by 1, repeat above process;

(4) Process above until DO equal to 0 and get symbol table after extract (STAE).

7. Output direction change

Get DF from EDT. If DF is set, reverses STAE to get symbol table after direction (STAD).

8. Create encrypted certificate of authorization file

From key code and length of file, we compute LP and insert EDT to LP point of STAD. STAD is the encrypted COA file.

## 4.2 Decryption Algorithm (CADC Certificate of Authorization Decryption)

Decryption algorithm is the reverse of encryption algorithm. The steps are as following.

1. From key code and length of encrypted COA, we find the entry point and extract EDT from encrypted COA
2. We use EDT to decrypt the remaining COA. The process is the reverse of encryption steps.
3. We get the original COA and WM.

## 4.3 Key Code (KC) and Location Point (LP)

From the key code (KC), the length COA (LCOA) and watermark (LWM) in CIDS and length of EDT (LEDT), we set the value LP as following rules:

$$LF = LCOA + LWM + LEDT \text{ (Length of encrypted file)}$$

If  $LF \geq KC$  then  $LP = KC$ ;

If  $LF < KC$  then  $LP = \text{mod}(KC/LF)$ .

LP is the point to insert EDT in the encrypted COA file..

## 4.4 Format Code

The fields in EDT have FC, LLST, NB, RB DO and DF. The different format of EDT is depending on format code (FC) as Table 13. The EDT is stored in encrypted COA. The FC and LLST are in the fixed area.

**Table 13.** Encryption data table format (EDTF)

FC	Fixed	Fields
1	LLST	NB,RB, LST, DO, DF
2	LLST	NB,RB, LST, DF, DO,
3	LLST	NB,RB, DO, DF ,LST
...	...	...

## 5 Comparison

The difference between the proposed method with others is as following:

- (1). The same plaintext has different cipher text.
  - (a) The different length and content of left shift table and encryption data table;
  - (b) The different selected watermark.
- (2). Owner and grantee have different EDT and are stored in the cipher text and used to encrypt and decrypt.
- (3). We use key code to store EDT to cipher text and each file has different key code.

## 6 Conclusion

In this study, we use the basic computing operations to design these encryption and decryption algorithms. It doesn't need any special hardware. Finally, we make some comments about this study.

- (1). The certificate of authorization may be any combination of letters, graphic and any other figures.
- (2). Watermark may be more than one.
- (3). It is more safer, because we must know the following to do the decryption;
  - (a) The EDT in cipher text; (b) The different format of EDT depends on format code; (c) The value of each field in EDT.
- (4). In supervisor grid node, it may have the function of execute grid node to process COA;

## References

1. Biham, E., Shamir, A.: Differential Cryptanalysis of DES-like Cryptosystem. In: Menezes, A., Vanstone, S.A. (eds.) CRYPTO 1990. LNCS, vol. 537, pp. 2–21. Springer, Heidelberg (1991)
2. Biham, E., Shamir, A.: A Differential Cryptanalysis of the Data Encryption Standard. Springer, Heidelberg (1993)
3. Biham, E., Shamir, A.: Differential Cryptanalysis of Data Encryption Standard. Springer, Berlin (1993)
4. Chen, H.-S., Lee, T.-Y., Lee, H.-M.: Collect and broadcast news in security. In: Proceedings of 2nd International Conference on Interaction Science: Information Technology, Culture and Human. ACM international Conference Processing Series, vol. 403, pp. 912–917 (2009)
5. Foster, I., Kesselman, C.: Globus: A Metacomputing Infrastructure Toolkit. International Journal of Supercomputer Application 11(2), 115–128 (1997)
6. Foster, I., Kesselman, C., Tuecke, S.: GRAM: Key concept (July 31, 1998), <http://www-unix.globus.org/toolkit/docs/3.2/gram/key/index.html>
7. Lee, H.-M., Lee, T.-Y., Hsu, M.-H.: A Process Schedule Analyzing Model Based on Grid Environment. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) KES 2006. LNCS (LNAI), vol. 4253, pp. 938–947. Springer, Heidelberg (2006)
8. Lee, T.-Y., Lee, H.-M.: Encryption and Decryption Algorithm of Data Transmission in Network Security. WSEAS Transactions on Information Science and Applications 12(3), 2557–2562 (2006)
9. McEliece, R.J.: A Public-Key System Based on Algebraic Coding Theory, 114-116. Deep Space Network Progress Report, 44, Jet Propulsion Laboratory, California Institute of Technology (1978)
10. Miyaguchi, S.: The FEAL-8 Cryptosystem and Call for Attack. In: Brassard, G. (ed.) CRYPTO 1989. LNCS, vol. 435, pp. 624–627. Springer, Heidelberg (1990)

# Probability Collectives Multi-Agent Systems: A Study of Robustness in Search

Chien-Feng Huang\* and Bao Rong Chang

Department of Computer Science and Information Engineering,

National University of Kaohsiung, Taiwan, R.O.C.

{cfhuang15, brchang}@nuk.edu.tw

**Abstract.** We present a robustness study of the search of Probability Collectives Multi-agent Systems (PCMAS) for optimization problems. This framework for distributed optimization is deeply connected with both game theory and statistical physics. In contrast to traditional biologically-inspired algorithms, Probability-Collectives (PC) based methods do not update populations of solutions; instead, they update an explicitly parameterized probability distribution  $p$  over the space of solutions by a collective of agents. That updating of  $p$  arises as the optimization of a functional of  $p$ . The functional is chosen so that any  $p$  that optimizes it should be  $p$  peaked about good solutions. By comparing with genetic algorithms, we show that the PCMAS method appeared superior to the GA method in initial rate of decent, long term performance as well as the robustness of the search on complex optimization problems.

**Keywords:** Probability collectives; multi-agent systems; optimization; robustness.

## 1 Introduction

Biologically-inspired algorithms, such as Genetic algorithms(GA) [1], Ant Colony Optimization (ACO) [2], Particle Swarm Optimization (PSO) [3], have been used as computational models to mimic evolutionary and social learning systems and as adaptive algorithms to solve complex problems. The core component of this class of optimization algorithms is a population of solutions that are employed to search for optimal solutions to the problem at hand. In the past decade, the research on Multi-agent Systems (MAS) has advanced the population-based algorithms with theoretical principles that shed light on the mechanism of system's decomposition and interacting agents' coordinating behavior in complex systems. As the complexity of a system grow, a generally more effective way to handle the system is through decomposition of the system into distributed and decentralized sub-systems and a given optimization task can be accomplished collectively by the sub-systems. In this scenario, the smaller subsystems can be regarded as a group of learning agents. These agents are self-interested and are

---

\* Corresponding author.

dedicated to optimizing their individual rewards or payoffs that in turn collectively optimize the global goal on the systems level.

Probability Collectives (PC) theory [4] refers to the systems-level objective as the world utility, which measures the performance of the whole system. PC is a broad framework for modeling and controlling distributed systems, and has deep connections to game theory, statistical physics, distributed control and optimization. Typically the search of adaptive, distributed agent-based algorithms is conducted by having each agent run its own reinforcement learning algorithm [5]. In this methodology the global utility function  $G(x)$  in the system maps a joint move of the agents,  $x \in X$ , to the performance of the overall system. Moreover, in practice the agents in a MAS are bounded rational. The equilibrium they reach typically involves mixed strategies rather than pure strategies; i.e., they don't settle on a single point  $x$  that optimizes  $G(x)$ . This suggests formulating a framework to explicitly account for the bounded rational, mixed strategy characteristics of the agents. PC adopts this perspective to show that the equilibrium of a MAS is the minimizer of a Lagrangian  $L(P)$  (derived using information theory) that quantifies the expected value of  $G$  for the joint distribution  $P(x_1, x_2, \dots, x_n)$  [4][6].

PC considers a bounded rational game in which each agent  $i$  chooses its move  $x_i$  independently at any instant by sampling its probability distribution (mixed strategy),  $q_i(x_i)$ . Accordingly, the probability distribution of the joint-moves is a product distribution; i.e.

$$P(x) = P(x_1, x_2, \dots, x_n) = \prod_{i=1}^N q_i(x_i), \quad (1)$$

if there are  $N$  agents participate in the game. In this representation, all coupling between the agents occurs indirectly; it is the separate distributions of the agents  $q_i$  that are statistically coupled, yet the actual moves of the agents are independent. The core of the PCMAS<sup>1</sup> methodology is thus to approximate the joint distribution by the product distribution, and to concentrate on how the agents update the probability distributions across their possible actions instead of specifically on the joint action generated by sampling those distributions. This embodies a notion of a region where the optimum is likely to be located as well as an uncertainty due to both imperfect sampling, and the stochastic independence of the agents' actions.

The PC approach differs from traditional optimization methods such as gradient descent, GA and PSO that concentrate on a specific choice for the design variables (i.e., pure strategies) and on how to update that choice. Since the PC approach operates directly on probability distributions that optimize an associated Lagrangian, it offers a direct treatment for incorporating uncertainty, which is also represented through probabilities [7]. This is the most salient feature that this class of algorithms possesses – the search course is guided by a probability

---

<sup>1</sup> PCMAS is used here to reflect the nature of PC in the context of multi-agent systems. In this paper, PC and PCMAS are two exchangeable terms.

distribution over  $x$ , rather than a single value of  $x$ . By building such a probabilistic model of promising solutions and sampling the built model to generate new candidate solutions, PC allows the agents to significantly expand the range of exploration of the search space, and simultaneously focus on promising areas of solutions. As a result, this class of algorithms may provide a more robust and scalable solution to many important classes of optimization problems.

Recently, Kulkarni and Tai [8] compared PC with other optimization algorithms, including Chaos Genetic Algorithm (CGA) [9], Punctuated Anytime Learning (PAL) system [10] and Loosely Coupled GA (LCGA) [11]. They stated several advantages of PC that make it a competitive choice over other algorithms.

In this paper we report a further comparison on the search performance and robustness on function optimization between the PCMAS and GA. The goal is to investigate how agents in the PCMAS search for solutions in complex optimization problems. The organization of this paper is as follows. In Section 2, a review of PC is provided. Section 3 presents the experimental results. We then conclude this paper in Section 4.

## 2 Review of Probability Collectives

### 2.1 Non-cooperative Game Theory

Assume that a set of  $N$  agents participate in a noncooperative game in which a mixed strategy that agent  $i$  adopts is a distribution  $q_i(x_i)$  over its allowed pure strategies [12]. Each agent  $i$  also has a private utility function  $g_i(x)$  that maps a joint move of all the  $N$  agents,  $x$  (i.e., pure strategies adopted by all the agents) into the real numbers. Given mixed strategies of all the agents, the expected utility of agent  $i$  is

$$\begin{aligned} E(g_i) &= \int dx P(x) g_i(x) \\ &= \int dx \prod_j q_j(x_j) g_i(x). \end{aligned} \quad (2)$$

Given the mixed strategies of the other agents, a Nash equilibrium indicates that every agent adopts a mixed strategy to maximize its own expected utility. In theory, Nash equilibria require the assumption of full rationality, that is, every agent  $i$  can calculate the strategies of the other agents and its own associated optimal distribution.

However, in the absence of full rationality the equilibrium is determined using the knowledge available to the agents. As a means to quantify this knowledge, the Shannon entropy,

$$S(P) = - \int dy P(y) \ln(P(y)),$$

describes a unique real-valued amount of syntactic information in a distribution  $P(x)$ . Given incomplete prior knowledge about  $P(x)$ , the maximum entropy (maxent) principle states that the best estimate of  $P(x)$  is the distribution with the largest entropy, constrained by the prior knowledge. This approach has proven useful in domains ranging from signal processing to supervised learning [13].

In the case of maximal prior knowledge available to agent  $i$ , the actual joint-strategy of the agents and thus all of their expected utilities are known. For this situation, trivially, the maxent principle states that the estimated  $q$  is that joint-strategy (it being the  $q$  with maximal entropy that is consistent with the prior knowledge).

Removing agent  $i$ 's strategy from this maximal prior knowledge leaves the mixed strategies of all agents other than  $i$ , together with agent  $i$ 's expected utility. For prior knowledge concerning agent  $i$ 's expected utility  $\varepsilon_i$ , the maxent estimate of the associated  $q_i$  is obtained by the minimizer of the Lagrangian:

$$\begin{aligned} L_i(q_i) &\equiv \beta_i[\varepsilon_i - E(g_i)] - S_i(q_i) \\ &= \beta_i[\varepsilon_i - \int dx \prod_j q_j(x_j) g_i(x)] - S_i(q_i), \end{aligned} \tag{3}$$

where  $\beta_i$  denotes the inverse temperatures (i.e.,  $\beta_i = 1/T_i$ ) implicitly set by the constraint on the expected utility.

The solution is a set of coupled Boltzmann distributions:

$$q_i(x_i) \propto e^{-\beta_i E_{q(i)}[g_i|x_i]}, \tag{4}$$

where the subscript  $q(i)$  on the expectation value indicates that it is evaluated according to the distribution  $\prod_{j \neq i} q_j$ ; and the expectation is conditioned on agent  $i$  making move  $x_i$ .

The first term in  $L_i$  of Eq. (3) is minimized by a perfectly rational agent, whereas the second term is minimized by a perfectly irrational agent, i.e., by a perfectly uniform mixed strategy  $q_i$ . Thus  $\beta_i$  in the maxent Lagrangian explicitly specifies the balance between the rational and irrational behavior of the agent. When  $\beta \rightarrow \infty$ , the set of  $q$  that simultaneously minimize the Lagrangians will recover the Nash equilibria of the game, which is the set of delta functions about the Nash equilibria.

In team games, the individual private utilities  $g_i$ 's are the same and can be replaced with a single world utility  $G$ . In this case, the mixed strategies minimizing the Lagrangian are related to each other via

$$q_i(x_i) \propto e^{-E_{q(i)}[G|x_i]}, \tag{5}$$

where the overall proportionality constant for each  $i$  is set by normalization, and

$$G(x) \equiv \sum_i \beta_i g_i(x). \tag{6}$$

Eq. (5) and (6) show that the probability of agent  $i$  choosing pure strategy  $x_i$  depends on the effect of that choice on the utilities of the other agents. Therefore, in the PC framework, even though the actual moves of the agents are independent, the probability distributions are coupled through expectation. Once all  $q_i(x_i)$ 's are determined, the probability distribution of the joint-moves by all the agents is a production distribution and can be obtained through Eq. (1).

## 2.2 Formulation of Optimization Problems

Given that the agents in PC are bounded rational, if they play a team game with world utility  $G$ , their equilibrium will be the optimizer of  $G$ . Furthermore, if constraints are included, the equilibrium will be the optimizer of  $G$  subject to the constraints. The equilibrium can be found by minimizing the Lagrangian:

$$L_q \equiv \sum_i \beta_i [E_q(g_i) - \varepsilon_i] - S(q), \quad (7)$$

with the prior information set being empty, i.e. for all  $i$ ,  $\varepsilon_i = 0$  [4], [6].

Specifically for the unconstrained optimization problem,

$$\min_{\mathbf{x}} G(\mathbf{x})$$

assumes each agent sets one component of  $\mathbf{x}$  as that agent's action. The Lagrangian  $L_i(q_i)$  for each agent as a function of the probability distribution across its actions (replacing  $\beta_i$  with  $1 / T_i$ ) is,

$$\begin{aligned} L_i(q_i) &= E[G(x_i, x_{(i)})] - T_i S(q_i) \\ &= \sum_{x_i} q_i(x_i) E[G(x_i, x_{(i)}) | X_i] + T_i \sum_{x_j} q_i(x_j) \ln(q_i(x_j)), \end{aligned}$$

where  $G$  is the world utility (system objective) which depends upon the action of agent  $i$ ,  $x_i$ , and the actions of the other agents,  $x_{(i)}$ , simultaneously. The expectation  $E[G(x_i, x_{(i)})]$  is evaluated according to the distributions of the agents other than  $i$ :

$$p(x_{(i)}) = \prod_{j \neq i} q_j(x_j).$$

Each agent then addresses the following local optimization problem:

$$\min_{q_i} L_i(q_i)$$

$$s.t. \sum_{x_i} q_i(x_i) = 1, q_i(x_i) \geq 0, \forall x_i.$$

During the minimization of the Lagrangian, the temperature  $T$  offers a means to adjust the degree of the exploitation of existing promising solutions (low temperature) and that of the exploration of the search space (high temperature). One can employ gradient descent or Newton updating to minimize the Lagrangian since both the gradient and the Hessian are obtained in closed forms. Using Newton updating and enforcing the constraint on total probability, the following update rule at each iteration is obtained [14]:

$$\begin{aligned} q_i(x_i) &\rightarrow q_i(x_i) - \alpha q_i(x_i) \times \\ &\{(E[G|x_i] - E[G])/T_i + S(q_i) + \ln q_i(x_i)\} \end{aligned} \quad (8)$$

where  $\alpha$  plays the role of a step size. The step size is required since the expectations result from the current probability distributions of all the agents. The update rule ensures that the total probability sums to unity but does not prevent negative probabilities. To ensure this, all negative components are set to a small positive value, typically  $1 \times 10^{-6}$ , and then the probability distribution is re-normalized.

To perform the gradient descent in probability space each agent must estimate the expected value of any of its actions,  $E[G|x_i]$ , from Monte-Carlo samples. Briefly, optimization proceeds in alternating rounds of Monte-Carlo sampling blocks, and updates to the agent's probability distribution over the parameter value. To draw a Monte-Carlo sample each agent chooses the value for its parameter  $x_i$  from its current probability distribution, and the world cost function  $G(x)$  is evaluated.

The number of samples in each Monte-Carlo block determines accuracy of the expected cost estimate. If sampling the objective function is costly, one may wish to gain the most information from the least number of samples. The kernel density estimation implies and exploits weak prior knowledge about smooth interpolation between the sample points. Additionally, as long as each iteration update does not dramatically change the PD, samples from the previous iterations may be re-used by geometrically weighting them according to their “age” in iterations. The imperfections that these augmentations introduce can be considered as another contribution to the bounded rationality term that broadens the probability distribution. The primary free parameters in the optimization are the Gaussian kernel width ( $\tau$ ), the rate of cooling ( $\delta T/T$ ), the number of Monte-Carlo samples per iteration, the proportional step size in the gradient descent ( $\alpha$ ), and data-aging rate ( $\gamma$ ).

The procedure for updating temperature  $T$  in PCMAS, referred to as the annealing schedule, plays an important role in the efficiency and reliability of the approach [15]. If the temperature is reduced too rapidly, the PCMAS is more likely to find a local minimum; however, if too slowly, then a large number of iterations and Monte-Carlo samples are required. Typically, a geometric schedule is applied, which involves multiplying the temperature by some fixed factor every several iterations. The detailed steps of the PCMAS may be found in [15].

### 3 Experimental Results

In this section, we report a comparison of the PCMAS with a genetic algorithm in searching for the global minimum of a complex function. The study of search efficiency usually involves defining a performance measure that embodies the idea of rate of improvement, so that its change over time can be monitored for investigation. In many practical problems, a traditional performance metric is the “best-so-far” curve that plots the fitness of the best individual that has been seen thus far by generation  $n$  for the GA, i.e., a point in the search space that optimizes the objective function thus far. The best-so-far curves presented for each testbed are the mean over 50 runs and error bars on the graph show the 95% confidence intervals about the mean.

For the traditional GA adopted in [16], we examined a range of population sizes (50,100, 200 and 500) to bracket a range of initial descent rates and long term performance. The 200 member GAs generally had the same long-term performance but converged faster than the GAs with 500 members. The 50 member populations generally descended quickly but converged to sub-optimal solutions. Parameter values were finely discretized to approximate a continuous range, and encoded as bit strings. (Various bit lengths were tried before settling upon 20 or 50 bits.) The GA experiments employ a binary tournament selection [17], one-point crossover and mutation rates of 0.7/pair and 0.005/bit, respectively.

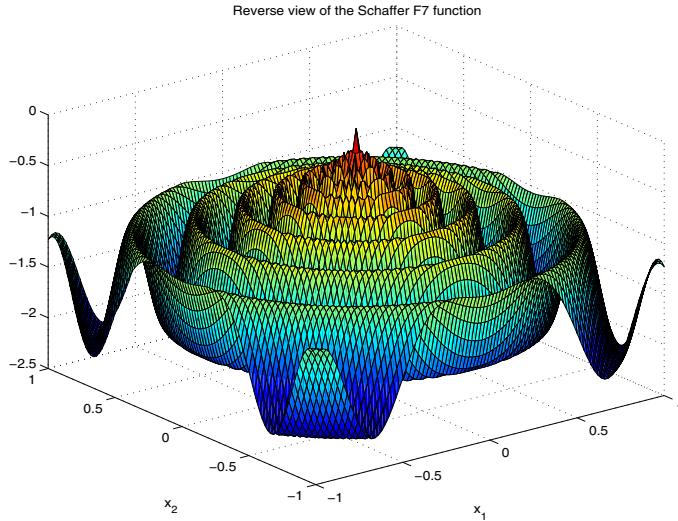
In the following examples the optimization free parameters for the PC were set as follows: step size  $\alpha = 0.2$ , data-ageing rate  $\gamma = 0.5$ , cooling rate  $\delta_T/T = 0.01$ , Gaussian kernel width  $\tau$  is set to 1% of the range of the search parameter, and  $T = 0.1$  was a sufficiently high starting temperature. Monte-calro block sizes of 50 and 25 were examined. Interestingly, using more samples per iteration did not significantly improve the best-so-far value for any given iteration. Thus only the results for the 25 monte-carlo blocks are reported since they use fewer samples per iteration.

Here we compare the PCMAS with the GA using Schaffer’s test function  $F_7$  [18], which is defined as:

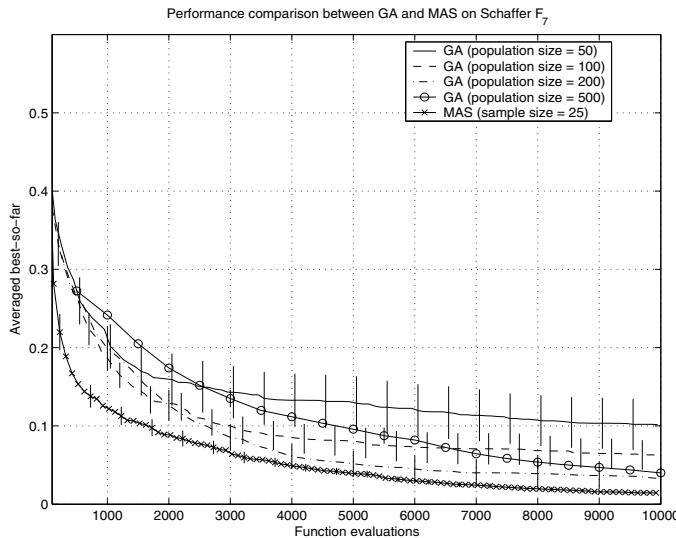
$$f(\bar{x}) = (x_1^2 + x_2^2)^{0.25} [\sin^2(50(x_1^2 + x_2^2)^{0.1}) + 1],$$

where  $-1 \leq x_i \leq 1$  for  $1 \leq i \leq 2$ . Figure 1 displays the surface which is plotted upside down for easier viewing of the inverted minimum as a peak. Since there are many local optima in the search space, the population in the GA can easily converge on any of them. The barriers would also present considerable difficulty to search approaches that evolve a single point  $x$  using local gradient information.

For this simple 2-dimensional case one could feasibly model the probability distribution in the full joint PC space rather than approximating it as a product distribution. However since we are, in fact, exploring multi-agent systems, instead two agents will carry out the search in the two parameters independently. For the GA, each variable is encoded by 50 bits; thus each agent in the GA consists of a bit string of length 100 (two blocks of 50 bits are concatenated to form a string).

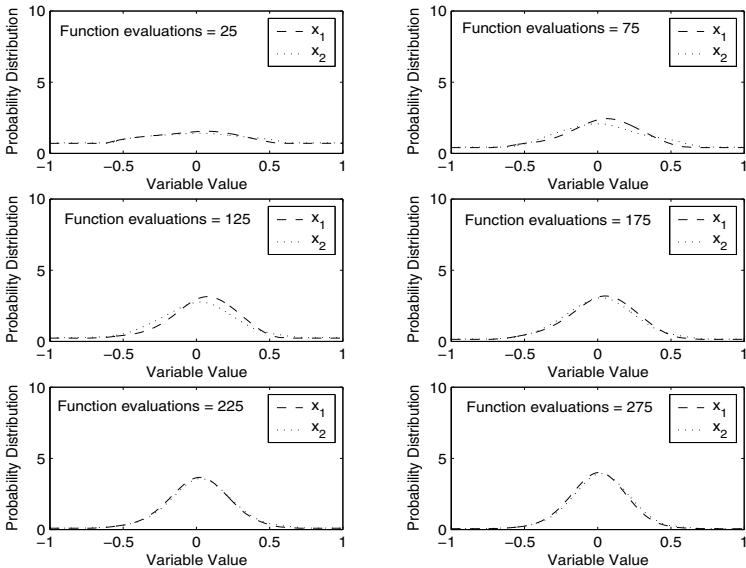


**Fig. 1.** Reverse view of the Schaffer  $F_7$  function



**Fig. 2.** Best-so-far performance on Schaffer  $F_7$

Figure 2 displays the best-so-far values attained by the PCMAS and the GA as a function of the number of sample evaluations of the objective function. The curves are the mean values over 50 repetitions and the vertical bars are the 95% confidence intervals on the means. Curves for different population sizes of the GA are shown. The methods distinguish themselves with different rates of initial descent of the objective function (on left) and the long-term performance



**Fig. 3.** Evolution of probability distribution

(on right). Notably, the run-to-run variation of the performance trajectory is much lower on the PCMAS than for the GA (see vertical bars). As a result, the PCMAS best-so-far trajectory was far more reproducible than that of the GAs, thereby implying the robustness of the search process in PCMAS.

Figure 3 displays the evolution of the probability distribution of the two variables for a typical MAS run. As can be seen, the probability density quickly centers about the optimum. This explains why the PC-based MAS is able to locate the optimum in a rather short period of time.

## 4 Conclusions

We presented a comparative study of two agent-based adaptive algorithms – the PCMAS and the GA approaches. The PCMAS approach introduces a methodology by which the search course in this system is guided by probability distribution over variables, rather than using single values derived from those variables. We have shown that the PCMAS method appeared superior to the GA method both in initial rate of decent and more significantly on long term performance. We have also shown the evolution of probability distributions of the strategies that the agents take in solving the optimization problem, thus providing a clear picture on how the PCMAS works.

In addition, since the PC best-so-far trajectory was far more reproducible than that of the GA, the search process of the PCMAS is thus shown to be more robust than that of the GA on complex optimization problems.

## References

- [1] Holland, J.H.: *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor (1975)
- [2] Dorigo, M.: Optimization, learning and natural algorithms. Ph.D. thesis, Politecnico di Milano (1992)
- [3] Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proc. of IEEE International Conference on Neural Networks, pp. 1942–1948 (1995)
- [4] Wolpert, D.H.: Theory of collectives. In: *The Design and Analysis of Collectives*, Springer, New York (2003), <http://ic.arc.nasa.gov/dhw>
- [5] Wolpert, D.H.: Bounded rational games, information theory, and statistical physics. In: Braha, D., Bar-Yam, Y. (eds.) *Complex Engineering Systems*
- [6] Wolpert, D.H., Tumer, T.: Optimal payoff functions for members of collectives. *Advances in Complex Systems* 4(2/3), 265–279 (2001)
- [7] Bieniawski, S., Wolpert, D.H., Kroo, I.: Discrete, continuous, and constrained optimization using collectives. In: Proc. of the 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, AIAA Paper 2004-4580 (2004)
- [8] Kulkarni, A., Tai, K.: Probability collectives: a multi-agent approach for solving combinatorial optimization problems. *Applied Soft Computing* 10(3), 759–771 (2010)
- [9] Cheng, C., Wang, W., Xu, D., Chau, K.: Optimizing hydropower reservoir operation using hybrid genetic algorithm and chaos. *Water Resource Management* 22, 895–909 (2008)
- [10] Blumenthal, H., Parker, G.: Benchmarking punctuated anytime learning for evolving a multi-agent team's binary controllers. In: *World Automation Congress* (2006)
- [11] Bouvry, P., Arbab, F., Seredyński, F.: Distributed evolutionary optimization. Manifold: Rosenbrocks Function Case Study, *Information Sciences* 122, 141–159 (2000)
- [12] Fudenberg, D., Tirole, J.: *Game Theory*. MIT Press, Cambridge (1991)
- [13] Mackay, D.: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge (2003)
- [14] Wolpert, D., Bieniawski, S.: Distributed control by lagrangian steepest descent. In: Proc. of the 43rd IEEE Conference on Decision and Control, pp. 1562–1567 (2004)
- [15] Bieniawski, S.: *Distributed optimization and flight control using collectives*. Ph.D. thesis, Stanford University, CA (2005)
- [16] Huang, C.F., Bieniawski, S., Wolpert, D.H., Strauss, C.: Comparative study of probability collectives based multi-agent systems and genetic algorithms. In: Proc. of the 2005 Genetic and Evolutionary Computation Conference (GECCO 2005), pp. 751–752 (2005)
- [17] Goldberg D. E., Deb K.: A comparative analysis of selection schemes used in genetic algorithms. *Foundation of Genetic Algorithms*, 69–93 (1991)
- [18] Schaffer, J.D., Caruana, R.A., Eshelman, L.J., Das, R.: A study of control parameters affecting online performance of genetic algorithms for function optimization. In: Proc. 3rd International Conference on Genetic Algorithms, pp. 51–60. Morgan Kaufmann, San Francisco (1989)

# An Improved Ant Algorithm for Fuzzy Data Mining

Min-Thai Wu<sup>1</sup>, Tzung-Pei Hong<sup>1,2</sup>, and Chung-Nan Lee<sup>1</sup>

<sup>1</sup> Department of Computer Science and Engineering  
National Sun Yat-sen University

No. 70, Lienhai Rd., 80424 Kaohsiung, Taiwan

<sup>2</sup> Department of Computer Science and Information Engineering  
National University of Kaohsiung

No. 700, Kaohsiung University Rd., 81148 Kaohsiung, Taiwan

d953040015@student.nsysu.edu.tw, tphong@nuk.edu.tw,  
cnlee@cse.nsysu.edu.tw

**Abstract.** In the past, two mining algorithms were proposed to find suitable membership functions for fuzzy association rules based on the ant colony systems. In the two approaches, the coding of the possible solutions is by binary strings, which form a discrete solution space. The paper extends the original approaches to continuous search space, and a fuzzy mining algorithm based on the improved ant approach is proposed. The improved ant approach doesn't have fixed paths and nodes and produces some paths in a dynamic way according to the distribution functions of pheromones. The experimental results show that the mining process based on the improved ant approach gets better results than that based on the previous two algorithms.

**Keywords:** Ant colony system; continuous space; data mining; fuzzy set; membership function.

## 1 Introduction

Knowledge Discovery and Data Mining (KDD) means the application of nontrivial procedures for identifying effective, coherent, potentially useful, and previously unknown patterns in large databases [8]. Since 1993, the practice of inducing association rules from transaction data has been commonly used in KDD [1]. An association rule is an expression  $X \rightarrow Y$ , where  $X$  is a set of items and  $Y$  is usually a single item. It means in the set of transactions, if all the items in  $X$  exist in a transaction, then  $Y$  is also in the transaction with a high probability.

Recently, the fuzzy set theory has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [8][15]. Fuzzy set theory has been applied in many fields such as manufacturing, engineering, medical diagnostics, economics, and others. It can also be easily combined with other techniques. For example, Delgado et al. proposed a fuzzy rule based back-propagation method for training binary multilayer perceptrons [5]. Hu proposed a genetic algorithm to find useful fuzzy concepts for pattern classification [13]. Many others can be found in the literature.

The role of fuzzy sets in data mining helps transform quantitative values into linguistic terms, thus reducing possible item sets in the mining process. Hong and Kuo thus proposed a mining approach that integrated fuzzy-set concepts with the Apriori mining algorithm [1]. In fuzzy data mining, the given memberships functions may have a critical influence on the final mining results. In [9], a GA-based fuzzy data-mining method is proposed to automatically get membership functions and association rules. Besides, Kaya et al. also proposed several GA-based algorithms for fuzzy data mining [16][17].

The ant system is another popular nature-inspired tool for solving optimization problems. It was first introduced by Colorni et al. in 1991 [2][3] and then extended to the ant colony system [6]. The idea of the ant system was from the observation on the real colonies of ants searching for food. In the past, Hong et al. proposed two ACS algorithms [11][12] to optimize membership functions in discrete solution space. In this paper, we propose a continuous coding ant colony algorithm to find membership functions in continuous solution space. Experimental results show the proposed approach can get better membership functions than the previous ones.

## 2 Review of the Ant Colony System

As mentioned above, the ant system was first introduced by Colorni et al. in 1991 [2][3]. Ants are capable of cooperating to solve complex problems such as searching for foods and carrying food. They can find a good path between their nest and food without vision. When ants move, they will deposit pheromone on the paths for their companions. The next ants can thus select the path with high density of pheromone to follow. They can determine the next node on the route according to the pheromone density. Once all the ants have terminated their tours, the amount of pheromone on the tours will have been modified.

The ant algorithms have thus been designed to simulate the above ant behavior for solving optimization problems. Especially, the process of modifying the amounts of pheromone on the tours is called the updating rule, which is designed to give more pheromone to a better path. Currently, the ant algorithms have widely been applied to solve difficult NP-hard problems. It has also been applied to the researches of data mining, such as classification, cluster, fuzzy control and so on [11][12].

The Ant Colony System (ACS) [6] is based on the above ant system. It applies the following rules in the execution process.

1. State transition rule: The rule is used for an ant to decide its next state in a probabilistic way.
2. Local updating rule: When an ant chooses an edge between two nodes, it will immediately update the pheromone of the edge for avoiding local optimization.
3. Global updating rule: After all ants have completed their tours, the pheromone of the best tour passed will be updated. There are two kinds of global updates. The first one takes the best tour among the ones passed by the ants in all the executed iterations to update. The other one takes the best tour among the ones passed by the ants in each iteration to update.

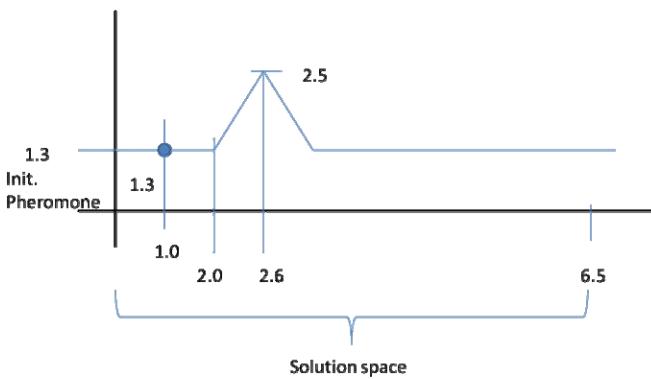
There are many variants of the original ant system algorithm. The difference among these variant ant algorithms lie in the ways of updating pheromone, the rules of state transition, and the heuristic functions used, among others.

### 3 The Improved Ant Colony System for Continuous Search Space

The traditional ACS algorithms use a fixed route map on which the nodes and the paths are discrete. An ant makes a choice among the fixed number of nodes according to the pheromones accumulated in the paths. However, in actual situations, ants may search for their food in a continuous search space. If the ant colony system can be extended to continuous nodes, it will be able to solve more problems. In this paper, an improved ant colony system for continuous search space is proposed to achieve the above purpose.

As mentioned above, in the traditional ACS algorithms, fixed numbers of paths and nodes are predefined in the search graphs and the values of pheromones are stored in the paths. In the improved ACS, because the numbers of paths and nodes are infinite, the chosen paths and nodes are produced and stored whenever ants generate and choose them. It is thus not easy to store the information of pheromone in the paths between pairs of nodes. Instead, the pheromone is stored at the nodes. A node with a larger amount of pheromone will be chosen with a higher probability.

However, when the node number becomes infinite, the pheromone values will be expressed by a distribution function in the proposed approach. For example, an simple problem is shown below:  $x + y = 10$ ,  $0 \leq x, y \leq 6.5$  and  $x, y$  are not necessarily integers. There will be a distribution of pheromone in either  $x$  or  $y$  dimension in the solution space. For example, a distribution of the  $x$  dimension in the execution process is shown in Figure 1.



**Fig. 1.** An example of the distribution of pheromone for the  $x$  dimension

In Figure 1, the distribution of pheromone is triangular because the update rules introduced later adopt this kind of update function for pheromone. Note that other kinds of functions may also be used. In this example, the node with its  $x$  value being 1.0 has

a pheromone density of 1.3, which is the initial amount of pheromones. After an ant selects the value of  $x$  using  $x$ 's pheromone distribution function, it will link to  $y$ 's pheromone distribution function according to the selected value of  $x$ . The updating rules adopted in the proposed improved ACS are described below.

### 3.1 Global Updating Rule

In the traditional ACS, the global updating rule will increase the pheromone value of the best tour and decrease those of the others. This concept will be extended to the improved ACS approach except that the update is not performed only on a node (a single  $x$  or  $y$  value), but on a set of neighboring nodes (a range of  $x$  or  $y$  values). An influence (update) function is then defined for achieving the purpose. In real implementation, we keep the axis values at which peaks appear and their peak values as well to reduce the storage space. The superimposed distribution function can be easily constructed from these kept values.

### 3.2 Local Updating Rule

When an ant constructs a path, the improved approach will immediately reduce (volatilize) the pheromone in the range of the influence function for each node on the path. The action can be easily achieved by reducing the corresponding peak values. If the reduced peak value is less than the initial amount of pheromones, the approach will remove the peak from the distribution function. Because of the process, the proposed approach can remove some unimportant information and reduce the cost of storage.

### 3.3 Generating a Path

In the improved approach, a partial path for an ant is produced whenever the ant needs to make a choice of its next node. A resulting path is finally determined by the requirement of the problem to be solved and by the density of the current distribution function. The algorithm first calculates the total area  $A$  of each current pheromone distribution function in the solution space. It represents the current total amount of pheromone in solving the problem. The algorithm then generates a random number  $n$  between 0 to  $A$ . It then finds the axis value to which the integral of the distribution function from the minimum value of the dimension equals  $n$ . According to the above process of node selection, an ant will then form a feasible solution when all the nodes in a path from the start to the end are selected. Each node that an ant selects is thus a value of an attribute in this feasible solution.

## 4 Fuzzy Data Mining Based on the Improved ACS

The above improved ACS is then applied to fuzzy data mining for getting appropriate membership functions. In the past, Hong *et al.* proposed two approaches to find appropriate membership functions in fuzzy data mining by the traditional ACS [11][12]. In their approach, the sets of membership functions were encoded to a 01 string. An ant needs to select several nodes to form a route, which represents a possible pair of center and span of a membership function. When the range of the solution space is large, the encoding length of a path will become long, affecting the quality of

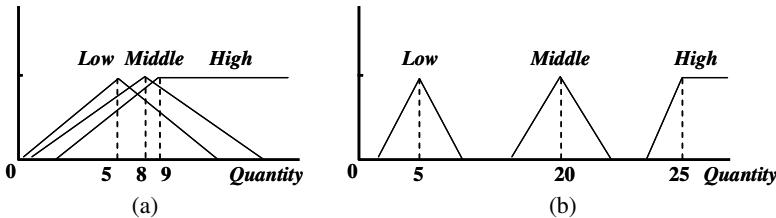
solutions. In addition, the binary strings can only map the possible solutions to a discrete space, thus reducing the accuracy of the solution.

Below, we will use the improved ACS to effectively determine appropriate membership functions. We assume each membership function is an isosceles triangle. The proposed approach based on the improved ACS approach thus encodes a membership function by a pair of numbers (*Center*, *Span*). If there are  $l$  linguistic terms for an item, the solution is easily represented by the concatenation of the codes for the membership functions. The resulting code is thus ( $Center_1$ ,  $Span_1$ ), ( $Center_2$ ,  $Span_2$ ), ..., ( $Center_l$ ,  $Span_l$ ). Each ant will sequentially decide the center and the span of each membership function. Thus if an ant needs to build  $l$  membership functions for a solution, it needs to select  $2l$  edges to finish a route. The center and the span of each membership function will be encoded in the continuous space of real numbers.

In this work, we use the fitness function proposed by Hong et al. [9] to obtain a good set of membership functions. The fitness value of a possible solution is defined as:

$$f = \frac{|L_1|}{\text{suitability}},$$

where  $|L_1|$  is the number of large 1-itemsets obtained by using the set of membership functions obtained. The suitability factor used in the fitness function is designed to reduce the occurrence of the two bad kinds of membership functions shown in Figure 2, where the first one is too redundant, and the second one is too separate.



**Fig. 2.** Two bad kinds of membership functions

The suitability of the membership functions includes two items, the overlap factor designed for avoiding the first bad case and the coverage factor designed for avoiding the second bad case. The details can be found in [9]. The proposed algorithm for mining membership functions based on the improved ACS is given as follows.

#### *The proposed algorithm for mining membership functions:*

##### **INPUT:**

1.  $n$  quantitative transaction data,
2. a set of  $m$  items, each of which is with a predefined number of linguistic terms,
3. a support threshold  $\alpha$ ,
4. a generated path number  $d$  for each ant,
5. a confidence threshold  $\lambda$ ,
6. a maximum number of iterations  $G$ , and
7. a number  $q$  of ants.

**OUTPUT:** An appropriate set of membership functions for all the items in fuzzy data mining.

**STEP 1:** Let  $p = 1$ , where  $p$  is used to keep the identity number of the item to be processed.

**STEP 2:** Initially set the pheromone distribute function of each variable (center and span of each membership function) as a constant, say 1.0.

**STEP 3:** Set the initial generation number  $g = 1$ .

**STEP 4:** Select the edges from start to end according to the state transition rule to build a complete route for each artificial ant by the following substeps.

**STEP 4.1:** Set  $s = 1$ , where  $s$  is used to keep the identity number of the  $s$ -th linguistic term of the  $p$ -th item.

**STEP 4.2:** Link to the pheromone distribution function for the center of the  $s$ -th linguistic term of the  $p$ -th item.

**STEP 4.3:** Produce  $d$  paths according to the pheromone distribution function.

**STEP 4.4:** Select a path from the above  $d$  paths according to the pseudo random proportional rule, which is used in the traditional ant colony system [7].

**STEP 4.5:** Update the peak value of the selected path in the distribute function according to the local updating rule.

**STEP 4.6:** Link to the pheromone distribution function for the span of the  $s$ -th linguistic term of the  $p$ -th item.

**STEP 4.7:** Produce  $d$  paths according to the pheromone distribution function.

**STEP 4.8:** Select a path from the above  $d$  paths according to the pseudo random proportional rule.

**STEP 4.9:** Update the peak value of the selected path in the distribute function according to the local updating rule.

**STEP 4.10:**  $s = s + 1$ .

**STEP 4.11:** If every ant has constructed its own solution, then go to STEP 5; else if  $s = l_p$  (the number of linguistic terms for the  $p$ -th item), then go to STEP 4.2; else go to STEP 4.1 for the next ant.

**STEP 5:** Evaluate the fitness value of the solution (membership functions) obtained by each artificial ant according to the fitness formula mentioned above.

**STEP 6:** Use the ant with the highest fitness value to produce its corresponding isosceles triangle influence function and then to update the distribute functions.

**STEP 7:** If the generation  $g$  is equal to  $G$ , output the current best set of membership functions of the  $p$ -th item for fuzzy data mining; otherwise,  $g = g + 1$  and go to STEP 4.

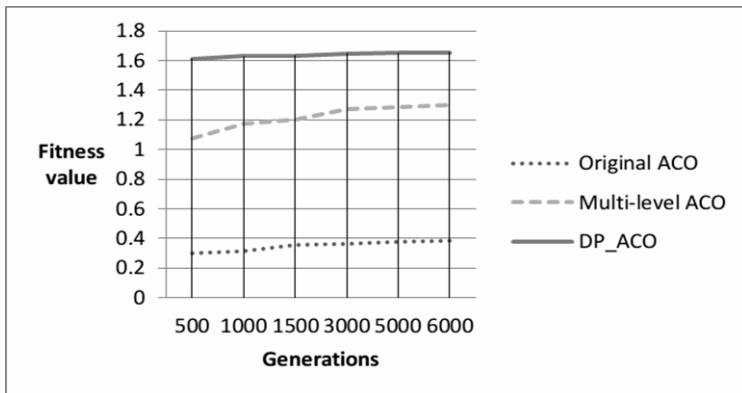
**STEP 8:** If  $p \neq m$ , set  $p = p + 1$  and go to STEP 2 for another item; otherwise, stop the algorithm.

The final set of membership functions output in STEP 7 and the 1-itemsets obtained can then be used to mine fuzzy association rules from the given database.

## 5 Experiments

Experiments were made to show the performance of the improved ACS approach for mining membership functions. The original ACS [12] and the multi-level ACS [11]

approaches for deriving membership functions were also run for a comparison. The experiments were implemented in C/C++ on a personal computer with Intel Core 2 Quad 6600 CPU and 4 GB RAM. The parameters used in the three approaches were set as follows. The minimum support for association rules was set at 0.04, the size of ants was 10, the initial pheromone was 1.0, the base pheromone was 0.7, the evaporation ratio was 0.2, and the local updating ratio was 0.2. The results were averaged for ten runs. The average fitness values along with different numbers of generations are shown in Figure 3. It can be easily observed that the proposed method had a better performance on the average fitness values than the original two methods.



**Fig. 3.** The average fitness values along with different numbers of generations for minimum support = 0.04

## 6 Conclusion

In this work, we have proposed a fuzzy mining approach based on the improved ACS algorithm to get appropriate membership functions of item quantities. The improved ACS supports the search in continuous solution space and is different from the past ant algorithms in that it doesn't have fixed paths and nodes. It will produce some paths in a dynamic way according to the distribution functions of pheromones. The experimental results show that the mining process based on the improved ACS gets better results than that based on the previous two ACS algorithms. It is because the proposed approach has a larger search space than the previous ones. In the future, we may attempt to study the effects of different encoding methods and different variant ant algorithms on the proposed algorithm. We may also apply the algorithm to solve some other real-world optimization problems.

## References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of International Conference on the Management of Data ACM SIGMOD, pp. 207–216 (1993)
2. Colomi, A., Dorigo, M., Maniezzo, V.: Distributed optimization by ant colonies. In: The First European Conference on Artificial Life, pp. 134–142 (1991)

3. Colomi, A., Dorigo, M., Maniezzo, V., Trubian, M.: Ant system for job-shop scheduling. *Belgian Journal of Operations Research, Statistics and Computer Science* 34, 39–53 (1994)
4. Cordon, J.C., Herrera, F.: Learning fuzzy rules using ant colony optimization. In: The Second International Workshop on Ant Algorithms, pp. 13–21 (2002)
5. Delgado, M., Mantasa, C.J., Moraga, C.A.: Fuzzy rule based backpropagation method for training binary multilayer perceptrons. *Information Sciences* 113, 1–17 (1999)
6. Dorigo, M., Maniezzo, V., Colomi, A.: Ant system: optimization by a colony of cooperating agents. *Transactions on Systems, Man, and Cybernetics-Part B* 26, 29–41 (1996)
7. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. *Transactions on Evolutionary Computation* 1, 53–66 (1997)
8. Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.: Knowledge discovery in database: an overview. In: AAAI Workshop on Knowledge Discovery in Databases, vol. 13, pp. 1–30 (1997)
9. Hong, T.P., Chen, C.H., Wu, Y.L., Lee, Y.C.: A GA-based fuzzy mining approach to achieve a trade-off between number of rules and suitability of membership functions. *Soft Computing: A Fusion of Foundations, Methodologies and Applications* 10, 1091–1101 (2006)
10. Hong, T.P., Kuo, C.S., Chi, S.C.: Trade-off between time complexity and number of rules for fuzzy mining from quantitative data. *Uncertainty, Fuzziness, and Knowledge-Based Systems* 9, 587–604 (2001)
11. Hong, T.P., Tung, Y.F., Wang, S.L., Wu, Y.L.: A Multi-level Ant-based Algorithm for Fuzzy Data Mining. In: North American Fuzzy Information Processing Society Annual Conference, pp. 1–5 (2009)
12. Hong, T.P., Tung, Y.F., Wu, M.T., Wang, S.L., Wu, Y.L.: An ACS-based framework for fuzzy data mining. *Expert Systems with Applications* 36, 11844–11852 (2009)
13. Hu, Y.C.: Finding useful fuzzy concepts for pattern classification using genetic algorithm. *Information Sciences* 175, 1–19 (2005)
14. Jiang, W.J., Xu, Y.H., Xu, Y.S.: A novel data mining algorithm based on ant colony system. In: Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, vol. 3, pp. 1919–1923 (2005)
15. Kandel, A.: Fuzzy expert systems, pp. 8–19. CRC Press, Boca Raton (1992)
16. Kaya, M., Alhajj, R.: A clustering algorithm with genetically optimized membership functions for fuzzy association rules mining. In: The IEEE International Conference on Fuzzy Systems, pp. 881–886 (2003)
17. Kaya, M., Alhajj, R.: Genetic algorithm based framework for mining fuzzy association rules. *Fuzzy Sets and Systems* 152, 587–601 (2005)
18. Martens, D., Backer, M.D., Haesen, R., Vanthienen, J., Snoeck, M., Baesens, B.: Classification with ant colony optimization. *Transaction on Evolutionary Computation* 11, 651–665 (2007)
19. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: An ant colony based system for data mining: application to medical data. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 791–798 (2001)
20. Stützle, T., Hoos, H.H.: MAX-MIN ant system. *Future Generation Computer System* 16, 889–914 (2000)

# Information-Driven Collective Intelligences

Francesca Arcelli Fontana, Ferrante Formato, and Remo Pareschi

University of Milano Bicocca, University of Sannio, University of Molise  
arcelli@disco.unimib.it, formato@unisannio.it,  
remo.pareschi@unimol.it

**Abstract.** Information-driven collective intelligences derive from the connection and the interaction of multiple, distributed, independent agents that produce and process information, and eventually turn it into knowledge, meant in the broad sense of capability of conceptual representation. For instance, the Worldwide Web can be viewed as an information-driven collective intelligence emerging from the digital network known as the Internet. The point here is how to extend such a capability for knowledge generation from the participating agents to the collective intelligence itself.

In this paper we show how this can be obtained with graph-based algorithms for the detection of communities of agents so as to support a dynamic, self-organized form of concept-discovery and concept-incarnation. In particular, we show how to strengthen community ties around concepts in order to increase their level of socialization and, consequently, of “fertility” in the generation of new concepts. Since there exists a direct relationship between concept discovery and innovation in human intelligences, we point out how analogous innovation capabilities can now be supported within information-driven collective intelligences, with direct applications to product innovation.

**Keywords:** Collective intelligences, social networks, web communities, evolution, innovation, knowledge generation.

## 1 Introduction

Information-driven collective intelligence [17], can be seen as the coordinated activities of independent intelligent units capable of information-processing and generating and managing ideas into the activity of larger intelligences. The coming of age of information-driven collective intelligences is a very important evolutionary milestone, since with the growing interest in complex adaptive systems, artificial life, swarms and simulated societies, the concept of “collective intelligence” is coming more and more relevant [13].

Our starting point here is that ideas are generated not just by individuals but also by communities, which are themselves a fundamental ingredient of collective intelligences. The ideation capabilities of communities within organizations have been widely studied, among others, by Nonaka and Takeuchi in their book in [21], where they introduce an interesting and useful construction called the “Knowledge Spiral” that relates the emergence of new ideas (“Knowledge Creation” in their terminology) with innovation.

In a collective intelligence new ideas are produced all the time, and are immediately internalized by being freely exchanged. At the basis of our approach there is one simple consideration that links conceptual discovery with community mapping, namely: once we have identified strong internalization and socialization points that do not correspond to any so far known idea, then we have found new ideas that need to be accounted for. Hence, we can extend our collective intelligence with computational tools that will recognize ideas once they are likely to have been already produced in different parts of the community that underlies the collective intelligence itself, as shown in [3].

On the other hand, once a new idea has been combined with others, this will be for the purpose of applying the new conceptual structure to a larger domain than the sum of the ones from which the originating ideas came from; this is the purposeful activity of planning and extending knowledge domains which is just as important in individual intelligences, where it is the result of consciousness, as in collective intelligences, where it is the output of organized teams of planners and knowledge experts that complement communities and social networks as components of the intelligence. However, for this to happen successfully, new conceptual structures must be seeded into communities that will feed them back into the creative fertilization necessary for the externalization of new ideas, thus fulfilling the process of “concept incarnation”.

Concept incarnation is indeed the focus of this paper, where it is accomplished by providing ways to set up suitable community structures whenever there is need of matching corresponding conceptual structures. Intuitively, this corresponds to automating the process of “marketing of new ideas”. Indeed, the aim of the process of concept discovery, which we have described in [3], was of making explicit concepts in order to pursue bottom-up extensions of ontologies. By contrast the aim of the process of concept incarnation, described in this paper, is of populating with communities concepts that are still “unpopular” but nevertheless they may become very popular.

**Sections:** 1. Introduction; 2. Networks and Communities; 3. A Cognitive Prosthesis for Concept Incarnation; 4. Application: Product Innovation; 5. Conclusion and Future Developments; 6. References.

## 2 Networks and Communities

One crucial point at the basis of our approach is that the notion of community that we adopt here not only assumes networks as a form of representation, but is itself a specialization of the notion of network. In fact, we adhere to the view, coming from the tradition of network theory, that a community can be defined in topological terms as a region of a complex network (which is itself a special kind of graph, as we shall characterize below), where links are denser than in the surrounding regions. In other words, communities are directly identified with highly interconnected regions of dynamic networks. This allows us to model communities hierarchically. For instance, at a social level communities can derive from social networks whose nodes map directly into human individuals, such as family clans, but can also correspond to digital communities where the role of humans is crucial but indirect, in that the primary

community members are Web sites pointing one to the other. As will be shown later on, the most immediate applications for our cognitive prostheses supporting abstraction capabilities in collective intelligences are indeed in the domain of this kind of Web communities.

We now provide a formal characterization of networks and of communities-as-networks that sets the premises for the implementation of the cognitive prostheses for concept discovery described in [3] and concept incarnation described in this paper. We refer to such networks as “complex networks”. Complex networks have been widely applied to describe a vast variety of real-life systems ranging from protein structures to airline routes, therefore it is not surprising that they would fit quite well the case of collective intelligences, which is itself a network-based phenomenon. As a matter of fact modeling social networks as “complex networks” is a largely investigated area (as shown, among others, in [10,19,20]), fully complementing traditional social network analysis [22]. The fact that the same model can be applied to the analysis of a variety of biological networks (for instance, in [6] complex networks are applied to the functional analysis of eukaryotic cells) provides a common denominator at the modeling level that fits naturally with the evolutionary inspiration behind our approach.

## 2.1 Complex Networks

Although there is a general agreement on considering as “complex” several kind of networks as for example the networks of power distribution and the networks of social relations, at present there is not a general accepted definition of “complex network”, probably because such definition is “complex” itself.

According to Erdős and Rényi [9], in a random graph with  $N$  nodes and connection probability  $p$ , the probability  $P(k_i=k)$  that node  $i$  has degree  $k$  follows a binomial distribution. In fact we have:

$$P(k_i = k) = C_{N-1}^k p^k (1-p)^{N-1-k}. \quad (1)$$

In this work we adopt the following definition: a network is *complex* when the distribution of nodes and edges does not follow the model of Erdős and Rényi. In fact, this model entails a uniform (random) distribution of edges between the nodes of a network; by contrast, non-uniform, “unfair” or “biased” distribution of edges between nodes is precisely what makes complex networks adequate in the modelling of real-life phenomena, social ones *in primis*.

A (complex) network is *scale free* when the degree of distribution of its nodes does not depend from the size of the network. Examples of scale-free networks are air routes, power distribution and Internet. In particular, Barabasi and Réka [6, 8] showed that in scale-free networks, the following holds:

$$P(k_i = k) = \frac{1}{k^\gamma}. \quad (2)$$

Where  $i$  is the node and  $P(k_i=k)$  is the probability that a new arc is added to a node with degree  $k_i$ . This means that a complex network can never be a random graph. Rather, the percentage of nodes with, for example, 8 arcs is:

$$p(8) = \frac{1}{8^{2.1}} \approx 0.15. \quad (3)$$

Where  $\gamma$  is a real number and  $1 \leq \gamma \leq 2$ .

Thanks to the works of Watts and Strogatz [23] and Barabasi and Reka [7], a taxonomy of complex networks is in the course of being defined.

In particular, two categories of networks have emerged:

- *scale free networks*, as characterized above, and as exemplified by airline routes, power grids and the Internet ([7]);

- *small world networks*, where any pair of nodes are not far. In [23] a small-world network is described as a regular lattice combined with a random graph. Therefore, in small world networks nodes are highly clustered. Some of the edges—called *weak ties*—are randomly distributed throughout the network, thus letting nodes “jump out” from their immediate neighborhoods into wider social circles. The importance of weak ties has been studied in social networks in [12]. Weak ties are useful for example to find a job, or getting in new businesses, for people who have a lot of weak ties are generally “well-connected” persons. Beside social networks in general, the World-wide Web itself can be modelled as a graph [8] with characteristics both of small-worlds [1] and scale-free networks.

## 2.2 Parametrization of Complex Networks

We now introduce a set of parameters defined in [4] that we use in order to classify and characterize the structure of a complex network.

**Degree.** The *degree* of a node of a network is the number of its connections. The in-degree (resp. out-degree) is the number of incoming (resp. out-coming) edges. By indicating with  $k$ ,  $k_i$  and  $k_o$  the degree, in-degree and out-degree of a vertex, we have:  $k = k_i + k_o$

We now indicate with  $P(k)$ ,  $P(k_i)$  and  $P(k_o)$  the distribution of the degrees, in-degrees and out-degrees, respectively.

**Shortest path.** Let  $G$  be a graph. Given a pair of nodes  $\mu$  and  $v$ , we call *geodesic distance*  $d_g(\mu, v)$  the length of the shortest-path in  $G$  between  $\mu$  and  $v$ . Observe that if  $G$  is a directed graph then  $d_g$  is not a distance.

We now formalize the concept of small-world network.

Let  $\Omega_n$  be the set of all the networks with  $n$  vertices and  $N$  as set of vertices. Let  $\omega \in \Omega_n$ . We call  $\pi(\mu, v, \omega)$  the distribution of the shortest paths over  $\Omega_n$ . Then we define the *average shortest path in  $\omega$*  as:

$$\langle l, \omega \rangle = \sum_{\mu, v \in N_\omega} \pi(\mu, v, \omega) d_g(\mu, v) \quad (4)$$

Finally,

$$\langle l \rangle = \sum_{\omega \in \Omega_n} \langle l, \omega \rangle \quad (5)$$

$\langle l \rangle$  is also called the diameter of the network. It can be shown that:

$$\langle l \rangle \approx \frac{\ln n}{\ln z} \quad (6)$$

Where  $n$  is the cardinality of the nodes of the network and  $z$  is the average number of nearest neighbours of a vertex. By setting:

$$\sigma = \frac{\ln n}{\ln z} \quad (7)$$

we capture the property of *small world*, as defined by Watts and Strogatz and in [23].

**Clustering.** Let  $G = (V, E)$  be a graph. It is obvious that, in a network with  $n$  nodes, the number of possible connections is  $\frac{n(n-1)}{2}$ . Let  $\sigma$  be the fraction of actual connections in the networks. The ratio between the number  $\sigma$  and the clique:

$$C_\lambda = \frac{2\sigma}{n(n-1)} \quad (8)$$

is the clustering coefficient with respect to node  $i$ . Then the clustering coefficient of a network is:

$$\sum_{v \in N} \frac{C_v}{n}. \quad (9)$$

In a random graph, Watts and Strogatz [23] define a statistic parameter.

$$\langle C \rangle = \sum_{i \in N} P(i)C_i \quad (10)$$

Clustering is an important parameter when we want to search for Web communities. In fact, it measures the average number of clusters in the graph. In particular, when  $\langle C \rangle$  is not high, it is unlikely that communities can be found in the graph.

**Edge degree distribution.** Another important parameter of a network is the edge distribution. In a random graph this distribution is poissonian, i.e

$$P(k_i = k) = C_{N-1}^k p^k (1-p)^{N-1-k} \quad (11)$$

Barabasi and Rèka [7] have shown that in many networks such distribution is not uniform, but is a function with a tail exponentially decreasing.

$$P(k_i = k) = \frac{1}{k^i} \quad (12)$$

**Spectrum.** The spectrum of a network  $G$  is the set of eigenvalues of the matrix associated with  $G$ . In particular, as shown by Kleinberg in [16], the principal eigenvalue individuates a pair of vectors  $\mathbf{h}$  and  $\mathbf{a}$ , while the remaining eigenvalues characterize the clustering of the graphs. Furthermore, given a secondary eigenvalue  $\lambda$ , for any

eigenvector  $\mathbf{v}$  of  $\lambda$  there exist a pair of integers  $i$  and  $j$  such that  $v_k \geq i$  and  $v_k \leq j$  are a cluster in  $G$ .

### 2.3 Extracting Communities

The automatic extraction of community out of a complex network requires a formal graph-based definition of community. At present there is not a general agreement and several graph-based definitions of community have been proposed. We report them here and recall that an *N-clique* is a subgraph  $G'$  such that any pair of nodes in  $G'$  is linked by a path of length not greater than  $n$ .

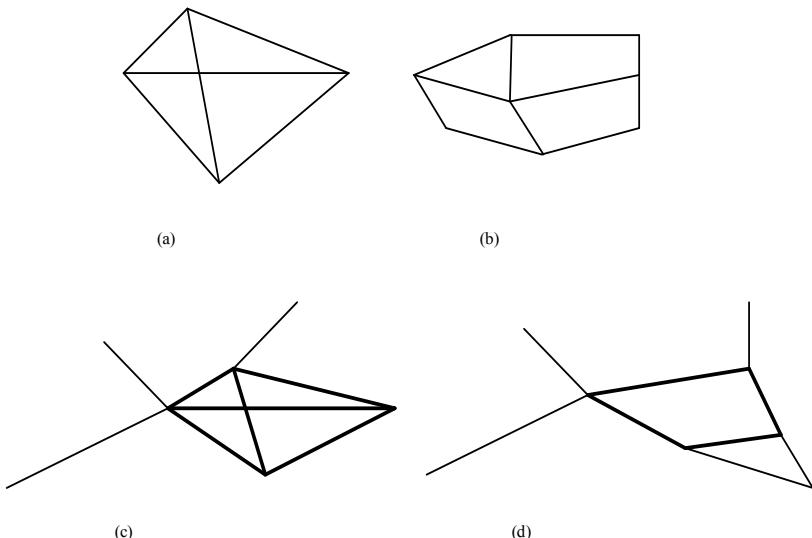
**Definition 2.1.** Let  $n$  be an integer. An  $n$ -community is a  $n$ -completely connected subgraph, or  $n$ -clique.

In Figure 1 several examples of communities-as-graphs are depicted. Figure 1 (a) is a 1-clique community. Figure 1-(b) is a 2-clique community, Figure 1-(c) is a strong community and Figure 1-(d) is a weak community.

**Definition 2.2.** Let  $G$  be a graph. A strong community is a subgraph  $G'=(N',E')$  of  $G$  such that, for any  $i \in N'$ , the number of internal links of  $G'$  are greater than the number of external links of  $G'$ .

Finally, the following is an extension of definition 2.2.

**Definition 2.3.** Let  $G$  be a graph. A weak community is a subgraph  $G'=(N',E')$  of  $G$  such that the sum of internal links in  $G'$  is greater than the sum of external links in  $G'$ .



**Fig. 1.** A taxonomy of communities

### 3 A Cognitive Prosthesis for Concept Incarnation

In [3] we shown how to merge ontologies and communities, and how to build up collective intelligence from a community layer. In this way we described a cognitive prosthesis for concept discovery.

If we look at things upside-down, namely from the standpoint of concepts versus communities, then we notice that ontologies as built and updated by experts in the phase of Knowledge Combination need then to be internalised properly in order to provide effective feedback into the Knowledge Spiral [21] of an information-driven collective intelligence. As a matter of fact, the experience with our prototype Gelsomino [11] revealed that, in the course of incrementally building and refining an automotive ontology, some relevant concepts in the field of automotive - such as “SCR-based ecologically sustainable engines” - were not tokenized enough, namely were not associated with communities of Web sites. Thus, we face an opposite problem with respect to the one addressed within the process of concept discovery, where the objective is making explicit concepts in order to pursue bottom-up extensions of ontologies. By contrast here we need to populate with communities concepts that are still “unpopular” but nevertheless may become at a certain point very relevant, as may indeed the case of a newly discovered R&D concepts such as “SCR-based ecologically sustainable engines”. Our answer to this is the definition of a cognitive prosthesis for concept incarnation with the goal to equalize the structure of a community with respect to the ontology it refers to in the context of our approach to information-driven collective intelligences.

For this purpose, we take the set of parameters (Clustering, Diameter, Average degree, Degree distribution and Spectrum) introduced in the previous section in order to characterize the structure of networks as the basis for defining a graph-transformation operator that equalizes the set of tokens  $S(c)$  associated with a concept  $c$  (such as, in our specific example, a set of Web sites) by extending it into a properly tokenised network, thus implementing the cognitive prosthesis for concept incarnation. The validity of the output of the operator can be checked against the value of the relevant network parameters.

Thus we now define an operator that, given a set of nodes  $S$  and the induced graph  $I(S)$ , takes  $I(S)$  as input and yields a set of smallest communities containing  $I(S)$ .

**Definition 3.1.** Given a graph  $G = (V, E)$  we define the operator  $Com$  and we set  $Com(G)$  as follows:

**Until**  $G$  is a strong community

**Do**

**Begin**

**For** any node  $v$  , let  $Out(n)$  be the number of outlinks of  $n$ .

**IF**  $Out(n) > In(n)$  **then**

Add  $Out(n) - In(n)$  edges to the  $n+1 \dots n+|Out(n)-In(n)|$  outside  $G$

Close (N,E)

**End**

The operator  $Com$  transforms a graph  $G$  into a weak community by adding to each node the number of internal edges in  $G$  necessary to outnumber the external edges in  $G$ .

### **Proposition 3.1**

- $Com(G)$  is a community
- If  $G'$  is a community and  $G' \supseteq G$  then  $G' \supseteq Com(G)$

*Proof:* Immediate

We immediately observe that the distribution degree is not an invariant of operator  $Com$ . In other words, the degree distribution of the graph  $G$ , that we indicate with  $P_G$ , is different from the degree distribution of  $Com(G)$ , that we indicate with  $P_{Com(G)}$ . To prevent this, we define a new –non-deterministic–operator  $Com'$ , as follows.

**Definition 3.2.** Given a graph  $G = (V, E)$ , we set  $Com'(G)$  as follows:

**Until**  $G$  is a web Community

**Do**

**Begin**

For any node  $v$ , let  $Out(n)$  be the number of outlinks of  $n$ .

**IF**  $Out(n) > In(n)$  **then**

Choose  $Out(n) - In(n)$   $n+1 \dots n+|Out(n)-In(n)|$  according to  $P_G$

Add  $Out(n) - In(n)$  edges to the  $n+1 \dots n+|Out(n)-In(n)|$  outside  $G$

Close  $(N, E)$

**End**

The random choosing can be made as follows:

Given  $P_G$ , generate a pseudo-random bit  $a$

**If**  $a = 1$  **then** add the node to  $Com_G$ .

**If**  $a = 0$  **then** do not add the node.

The random choosing is made in order to preserve the distribution degree of the graph. In this way, for instance, scale-free networks are transformed into scale-free communities.

### **3.1 Network Parametrization of the Community Operator**

We now analyse the operator  $Com$  with respect to the parameters introduced in section 2.2. By doing so, we prove that the graph of the community is denser than the induced graph of the seed  $G$ .

**Proposition 3.2.** Let  $G$  be a graph. Then

$$\langle C(Com(G)) \rangle \geq \langle C(G) \rangle$$

*Proof.*

By definition, we have that

$$\langle C \rangle = \sum_{i \in N} P(i)C_i .$$

By construction,

$$\langle Com'(G) \rangle = \sum_{i \in Com'(N)} P'(i)C_i$$

Since  $Com'(N) \supseteq N$  and  $P(i) = P'(i)$  we have that

$$\sum_{i \in Com'(N)} P(i)C_i \geq \sum_{i \in N} P(i)C_i$$

The thesis follows.

Proposition 3.2 shows that the graph yield by Com is denser then the input graph. Clustering preservation is a sufficient condition for the following property:

**Proposition 3.3.** Let  $W$  be a Web community and  $W \subseteq I(S)$  Then

$$\langle C(W) \rangle \leq \langle C(Com'(I(S))) \rangle$$

*Proof.* Immediate.

Proposition 3.3 says that when operator  $Com'$  is applied to a set  $S$  that is mostly rarefied but contains some Web community  $W$ , it yields a community  $I(S)$  denser than  $W$ .

## 4 Application: Product Innovation

The most direct application of our approach we are in the course of experimenting with is in a class of information-driven collective intelligences of enormous social and economic relevance: corporate manufacturing organizations. That such organizations must merge, through digital data networks, with their stakeholder communities (end-users, business partners, employees etc.) into information-driven collective intelligences is an evident step if they want to preserve themselves from the risk of death from irrelevance — a threat, as recently argued and corroborated with a wealth of empirical evidence in [18], which is becoming all the more real in a context where people use online social technologies (blogs, social networking sites, YouTube, podcasts) to discuss products and companies, write their own news, and find their own deals. More than that, manufacturers might even get a chance to fully reverse the turn of the game by leveraging collective intelligences to find new ways to innovate. Indeed, so far two main approaches have emerged for product innovation. In the so-called “linear model” the traditionally recognized source is manufacturer innovation. This is where an agent (person or business) innovates in order to sell the innovation. Another source of innovation, only now becoming widely recognized, is end-user innovation. This is where an agent (person or company) develops an innovation for their own (personal or in-house) use because existing products do not meet their needs. Eric von Hippel has identified end-user innovation as, by far, the most important and critical in his classic book on the subject, *The Sources of Innovation* [14] as well as in the more recent *Democratizing Innovation* [15]. One outstanding example of end-user innovation is open-source and free software.

In reality, we believe that the approaches will merge into one comprehensive approach of circular innovation, with users feeding fresh concepts into organizations and organizations combining new concepts with established ones in order to launch new products. In [3] we have addressed the issue of feeding innovation from users into organizations, while here we focus into the complementary effort of organizations shaking hands with user communities in the most effective way.

## 5 Conclusion and Future Developments

We have introduced a formal and computational framework for information-driven collective intelligences that has as motivation the networking of individual intelligences through digital data networks and fits within the background of a complex view of evolution composed of symbiotic interaction and integration among multiple species as well as of natural selection within a single species. Symbiosis accounts for big evolutionary jumps and explains the coming of age of information-driven collective intelligences as analogous to the integration of bacteria and prokaryotic cells into eukaryotic cells as happened 2.5 billions years ago through biochemical networks, thus paving the way for the variety of forms of life on earth that have developed since then. We take the integration of multiple intelligent information-processing units into a digital data network as a similar starting point for collective intelligences and from there we embark into the task of speeding up the process of natural selection within the species by adding new features to their genotype through the introduction of software layers acting as cognitive prostheses that implement higher-level thought processes. Our ingredients are a class of widely studied and applied graphs known as “complex networks” for describing the structure of collective intelligences coupled with computational operators for, respectively, graph analysis and transformation and for content analysis and classification. Thus our contribution fits into a fully computational perspective consistent with the tradition of artificial intelligence and cognitive science. We specifically address the process of innovation, which is the most characterizing of the species *homo sapiens* and as such is suitable to be transferred to information-driven collective intelligences where humans play a crucial role. Our cognitive prostheses support innovation both in the phase of discovery of new concepts and in the way they are then set within the general conceptual organization of collective intelligences.

## References

1. Adamic, L.A.: The Small World Web. In: Abiteboul, S., Vercoustre, A.-M. (eds.) ECDL 1999. LNCS, vol. 1696, pp. 443–452. Springer, Heidelberg (1999)
2. Arcelli, F., Formato, F., Pareschi, R.: Ontology Engineering: Co-evolution of Complex Networks with Ontologies. In: Proceedings of the Workshop on Ontologies for e-Technology (OET 2009), Italy (May 2009)
3. Arcelli, F., Formato, F., Pareschi, R.: Boosting Concept Discovery in Collective Intelligences. In: Zhong, N., Li, K., Lu, S., Chen, L. (eds.) BI 2009. LNCS, vol. 5819, pp. 214–224. Springer, Heidelberg (2009)

4. Arcelli, F., Formato, F., Pareschi, R.: Equalizing the structures of web communities in ontology development tools. In: Proceedings of the International Conference on Intelligent Systems Design and Applications ISDA 2009, Pisa (November 2009)
5. Barabasi, A.L., Réka, A., Hawoong, J.: The diameter of the World Wide Web Nature, vol. 401, pp. 130–131 (September 9, 1999)
6. Barabasi, A.L., Oltvar, Z.N.: Netowrl biology: understanding the cells functional organization. *Nature Review* 5 (2004)
7. Barabasi, A.L., Réka, A.: Emergence of Scaling in Random Networks. *Science* 286, 509–512 (1986)
8. Broder, A., Kumar, R., Maghoul, F., Prabhakar, R., Rajagopalan, P., Stata, R., Tomkins, A., Wiener, J.: The Web as a Graph. In: Proc. of the 9th International Web Conference, Amsterdam, May 5 (1999)
9. Erdős, P., Rényi, A.: On Random Graphs. I, *Publicationes Mathematicae* 6, 290–297 (1959)
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99, 7821–7826 (2002)
11. Gelsomino: Gelsomino, <http://www.essere.disco.unimib.it/> on request
12. Granovetter, M.: The Strength of weak ties. *American Journal of Sociology* 78 ( May 1973)
13. Heylighen, F.: Collective intelligence and its implementation on the web: algorithms to develop a collective mental map. In: Computational & Mathematical Organization Theory, vol. 5.3, pp. 253–280. Kluwer Academic Publishers, Dordrecht (1999)
14. von Hippel, E.: The Sources of Innovation. Oxford University Press, Oxford (1988)
15. von Hippel, E.: Democratizing Innovation. MIT Press, Cambridge (2005)
16. Kleinberg, J.: Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM) Archive* 46(5) (September 1999)
17. Levy, P.: Collective Intelligence: Mankind's Emerging World in Cyberspace. Helix Books (1998)
18. Li, C., Bernoff, J.: Groundswell: Winning in a World Transformed by Social Technologies Harvard Business Press (2008)
19. Newman, M.E.J.: Detecting community structures in networks. *Eur.Phis. J.* 38 (2004)
20. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* 74, 036104 (2006)
21. Nonaka, I., Takeuchi, F.: The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, Oxford (1995)
22. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press, Cambridge (1994)
23. Watts, D., Strogatz, S.: Collective dynamics of 'small world' networks. *Nature* 393 (1998)

# Instructional Design for Remedial English e-Learning

Chia-ling Hsu<sup>1</sup>, Ai-ling Wang<sup>2</sup>, and Yuh-chang Lin<sup>3</sup>

<sup>1</sup> Tamkang University, Graduate Institute of C&I, Associate Professor  
151, Ying-chuan Road, Tamsui, Taipei County, 25137, Taiwan  
chlhsu@mail.tku.edu.tw

<sup>2</sup> Tamkang University, Dept. of English, Associate Professor  
151, Ying-chuan Road, Tamsui, Taipei County, 25137, Taiwan  
wanga@mail.tku.edu.tw

<sup>3</sup> Aletheia University, Center for General Education, Instructor  
32, Chen-Li St, Tamsui, Taipei County, 25103, Taiwan  
au1258@mail.au.edu.tw

**Abstract.** Society and learning change with the development of technology. In order to keep up the latest development of technology in education, this study focused on a remedial English e-learning course in a university. Hopefully, by using a chance building model, few but important elements would be acquired. The chance building was based on the text mining theory and KeyGraph technology to present a visualized scenario. The participants were graduate school students who took the e-learning course. A questionnaire is composed of the ARCS model and preference selection with material topics and types. The results indicated that from the interaction between students' characters and teaching characters some attributes tended to create innovated scenarios. With the nodes and links, different innovated scenarios would be interpreted. As a result, the chance points also appeared in the graph. More studies in chance building model and in empirical experiment are needed.

**Keywords:** Instructional Design, Chance Building, e-learning, Remediation Instruction, Course Evaluation.

## 1 Introduction

The purposes of using technologies in education are not only for gaining students' attention but also for providing remedial instruction. Studies indicated that using multimedia in courses would increase students' learning motivation especially for using computers as educational media [1][2]. The computer assisted instruction (CAI) is a good example. Learning with CAI, students would learn by themselves with their own learning speed and level. Many studies, therefore, indicated that CAI could be used for remedial education and for individual learning [3][4].

Since Web 2.0 made the technology change radically, the instructional design in using educational technology appeared different. E-learning is a new trend in education. For example, blogs, wiki systems, and game-based environments have been

developing rapidly. Hence, how to adapt these technologies into instructional design will be a new and important issue.

This study is not to develop new platform but to promote a way of evaluation and analyzing. In other words, this study proposed a model of reanalyzing the data collected from an e-learning course and of representing a chance building map. Therefore, the purposes of this study were to evaluate the remedial English e-learning course by a motivation questionnaire, to analyze the data to present a chance building map, and to provide instructional design strategies in remedial e-learning course.

What are chances? According to chance discovery theory, the definition is that chances are the rare but import events or factors [5][6]. However, Watts emphasizes the importance of linking [7]. Hsu's studies has been using chance discovery model in education especially in class activities. The findings were novel and would provide teachers with chances to design the activities in different point of view [8] [9]. The model we proposed would help instructional designers evaluate the course development as well as find chances to improve the quality of e-learning course. The data were analyzed in text format and represented in a chance building map. In this study we took students' motivation as an indicator to evaluate the course. The KeyGraph was another indicator to improve the course. With evaluation and reflection, the quality of the remedial e-learning English course would be better. Hopefully, few but important elements would be found.

## 2 Relative Literature

The instructional design and motivation models were applied in this study. In addition, the approach of text mining and keyGraph technologies were used in this study, too. A further illustration of the models and technologies mentioned above is offered as follows.

### 2.1 Instructional Design

Instructional design is to provide teaching blueprints, and to examine teaching and offer solutions. Accordingly, the practice of instructional design is to target specific learners, select specific approaches, contents, and strategies, and make an effective teaching policy [10]. Instructional design is often presented and explained through models [11].

### 2.2 ARCS Model

Keller proposed the attention (A), relevance (R), confidence (C) and satisfaction (S) motivation model in order to solve motivational problems for instructional designer [12][13]. There are two main issues facing with the ARCS model: one is attitude, and the other is evaluation. The evaluation for the motivation is not to evaluate the learning efficiency but to evaluate the motivation character of the learning motivation in

the instructional design model. The ARCS Model identifies four essential components for motivation instruction:

- Attention: strategies for arousing and sustaining curiosity and interest
- Relevance: strategies that link to learners' needs, interests and motives
- Confidence: strategies that help students develop a positive expectation for successful achievement
- Satisfaction: strategies that provide extrinsic and intrinsic reinforcement for efforts.

The ARCS was also applied in adult learning [14]. Hence this study used the ARCS model to evaluate the graduate students' motivation in a remedial e-learning English course in university.

### **2.3 Text Mining and KeyGraph Technology**

Weiss, Induskhya, Zhang, & Damerau [15] indicated that data mining technology would find out the pattern in structure data base but not in none or semi structure data base. Hence, Hearst [16] pointed out that data mining would not be satisfied the human needs of pursuing information and knowledge. However, text mining applied the language and statistics to analyze text data in order to attain new information [17]. Therefore, text mining technology became one of the important issues.

Ohsawa in 1998 proposed the KeyGrph technology as a kind of data visualization tool in order to discover the chance [18]. The KeyGraph technology brought the text mining research into a new age. Montero and Araki, in 2005, showed that a text could be divided into some different subgroup and there was a connection link to each subgroup [19]. Some phases would be connected to each other but some were not. At the same time, 2005, Sakakithara and Ohsawa sorted different subgroups and presented these subgroups into KeyGraph format [20]. They also defined the high frequency element as a “black node”, and the number of baskets which contain the two elements and the high frequency co –occurrence as a “black link”. In this model, the KeyGraph technology becomes a very powerful tool in many areas. Oshawa in 2002 pointed out that the value of KeyGraph technology as an extractor of causalities from an event - sequence, and as a words abstractor from a document [21]. Moreover, the main point of the KeyGraph technology provided some chances which would reverse the bad situation into a better one, especially in a feeble industry.

Wang, Hong, Sung, and Hsu applied this method to get the validity of KeyGraph [22]. The results indicated that although the statistics data showed no significant difference, the KeyGraph technology provided more information. Hsu and other educators also applied the KeyGraph technology in education setting. The results pointed out that the learners' scenario map would tell more information than the traditional statistics results. Huang, Tsai, & Hsu also applied the KeyGraph technology to

exploring the learners' thinking [23]. Tsai, Huang, Hong, Wang, Sung, and Hsu [24] used Keygraph technology and tried to find the chances in instructional activity.

### 3 Research Method

This study was to analyze the students' motivation and preference in a remedial English e-learning course. Although the traditional ARCS questionnaire was applied, this study employed a new technology, KeyGraph approach, with a view to acquiring more information in learners' motivation. In addition, this study intends to combine the data of ARCS motivation with interesting study materials to form a possible teaching plan. The participants, instrument, and scoring system were expressed as follows.

#### 3.1 Participants

A remedial English e-learning in university for graduate students who did not achieve the graduation required level of English tests such as TOEIC, TOEFL and GEPT. 186 students enrolled in this class. At the end of the course, students filled out online the questionnaire voluntarily. Hence, 179 questionnaires were collected. Table 1 showed population of the students.

**Table 1.** Population of the students

School	Number of students
Engineer	104
Science	32
Management	21
Business	9
Education	5
Liberal Arts	3
others	5
Total	179

#### 3.2 Instrument

The ARCS questionnaire contains 34 items. The Attention factor contains item 1, 4\*, 10, 15, 21, 24, 26\*, and 29. The items marked with \* mean the inverse items. The Relevance factor contains item 2, 5, 8, 13, 20, 22, 23, 25, and 28. The Confidence factor contains item 3, 6\*, 9, 11\*, 17\*, 27, 30, and 34. The Satisfaction factor contains item 7, 12, 14, 16, 18, 19, 31\*, 32, and 33. The score is calculated for each item by 5 scale points, from non-agree to very agree.

Besides, two more questions were asked. One was a multiple-choice question about the material types. The other was an open question about the material topics.

### 3.3 Scoring System

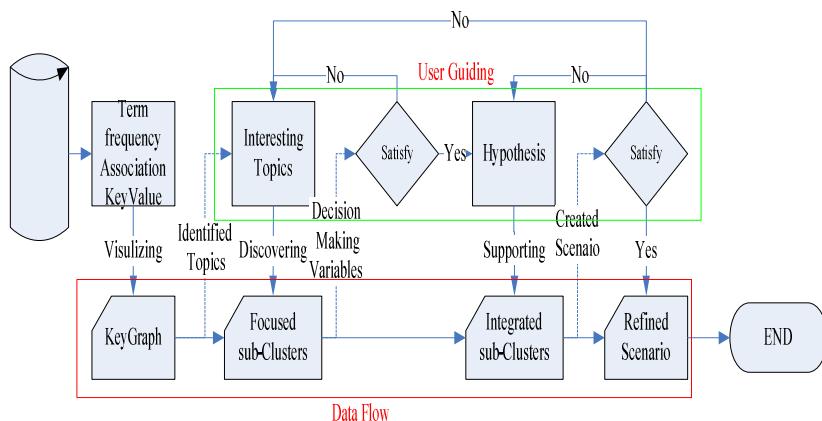
The scoring system in this study was used to translate the participants, items, and scores into KeyGraph model. Each subject was represented by one sentence in KeyGraph scoring system. 179 sentences in total were collected. The ARCS questionnaire contained 34 items as 34 words in KeyGraph model, that is, each item was a word. The score of each item was scaled from 1 to 5, which was considered as the number of times that each item appeared. Namely, in this study the score was the frequency. For example, if item x was 5 points, this meant that this item x would appear 5 times. In KeyGraph wording, the frequency of Word x was 5. The results of transformation were shown in table 2.

**Table 2.** The word frequency from ARCS transformation

Word	Frequency	Word	Frequency
C3confidence_finish	750	R1expectation_goal	652
C3effort	735	A1media	649
R3relevance_knowledge	726	S2feedback	649
R1value	722	S3fairness	646
A3arousal_concept	713	S2satisfaction	644
A3content_arousal	696	R1goal_orientation	639
S1assignment	693	surprise	633
C3confidence_learning	684	S3fairness_grade	627
R3pre-knowledge	679	R2initiative	620
S2quality_knowledge	674	A1curiosity	606
R1key_point	673	S1disappointment	547
C3challenge	667	daydream	536
R1achievement_goal	666	R2benefit	526
A2inquiry	663	A3attention	468
C3reward	662	C2difficulty	434
S2study_finish	659	C3luckiness	346
S1study	655	C2grade	290

## 4 Chance Building Model

Using KeyGraph technology would provide scenarios for possible chance. The KeyGraph was a kind of data-driven technology which meant that the scenarios were not driven by objectives. However, simply based on the KeyGraph technology, the chance building model tended not to emerge. With the objectives and the interesting topics we added to the original data-driven information, a more meaningful scenario would be formed. Figure 1 presented the procedure of chance building model. Each step was explained as follows.



**Fig. 1.** The chance building model

The procedure of the research model was described below:

Step 1: The data of questionnaires were collected from the WebCT platform. Then, the reversed items in ARCS questionnaire were taken care of. After that, the numeral data were transferred by the scoring system into text data form in order to find the relationship between students' character and teaching' character.

Step 2: Using the KeyGraph technology, the thresholds of key frequency and jaccard relation were determined.

Step 3: Based on the research interest, the sub-clusters were considered as part of the value focus system. In this study the relationship between the learners' characters and the teaching' characters was taken into consideration.

Step 4: The final KeyGraph was generated and would propose an innovated scenario.

In this case study, the revised items in ARCS questionnaire were encoded into statistics consistent numbers. After the data were keyed in, the numeral data were transferred by a scoring system into a text data format in order to find the relationship between the students' character and the teaching' character. The results showed that the frequency of ARCS data compared with the frequency of students' and teaching character data was significant difference. The numeral value of material types was return to the original question items, short-essay, conversation, and both. The material topics collected were the answers from students, including travel, health, technology, business, international, entertainment, education, and campus. KeyGraph technology was used to decide the frequency threshold of word-term 54 and the jaccard threshold value 0.3. Figure 2 presented the threshold of using KeyGraph technology and the sensitive analysis of the word term frequency. Figure 3 showed the threshold of using KeyGraph technology and the sensitive analysis of the jaccard relation valued.

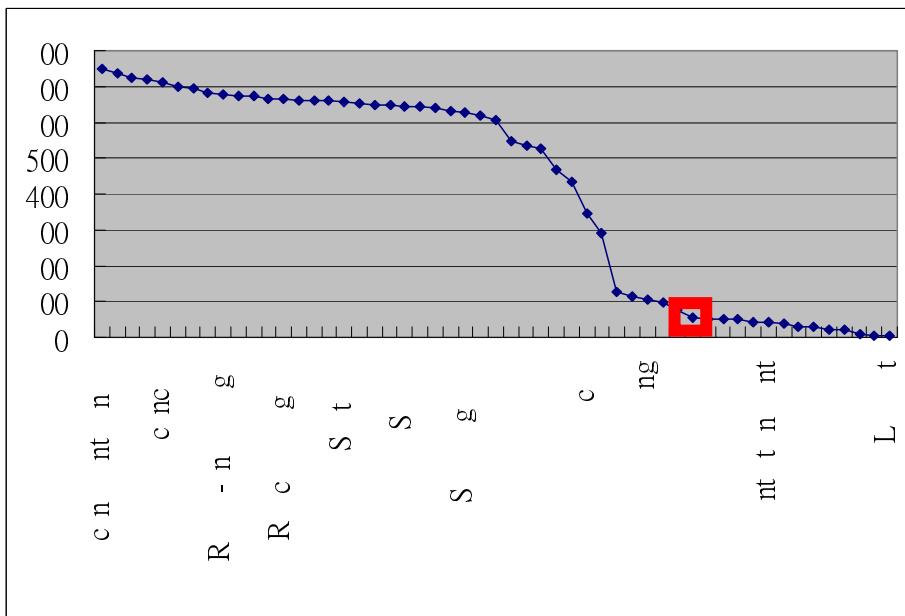


Fig. 2. The frequency of word term

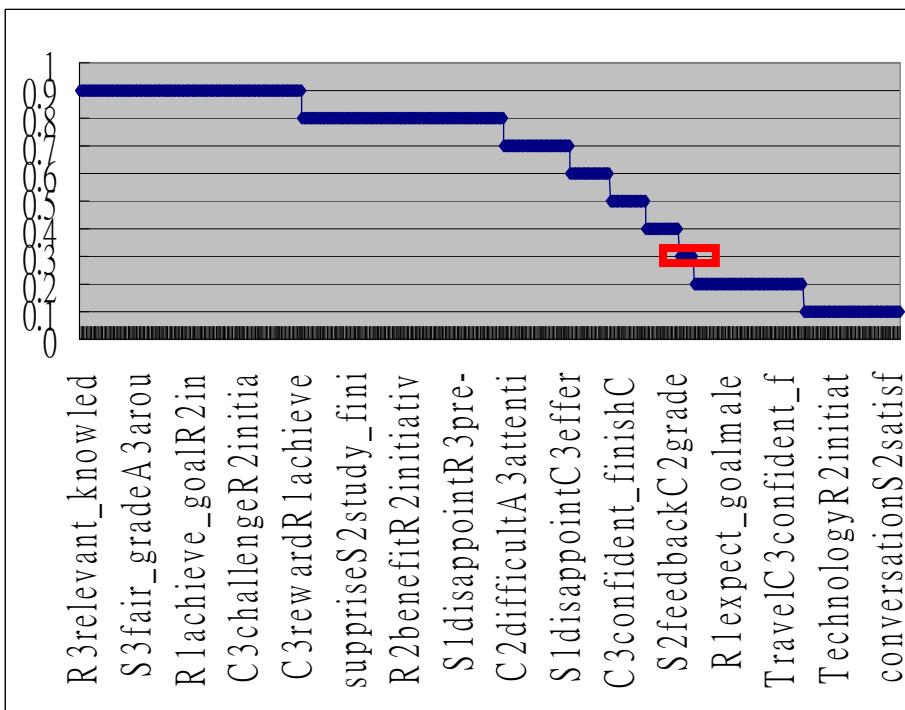
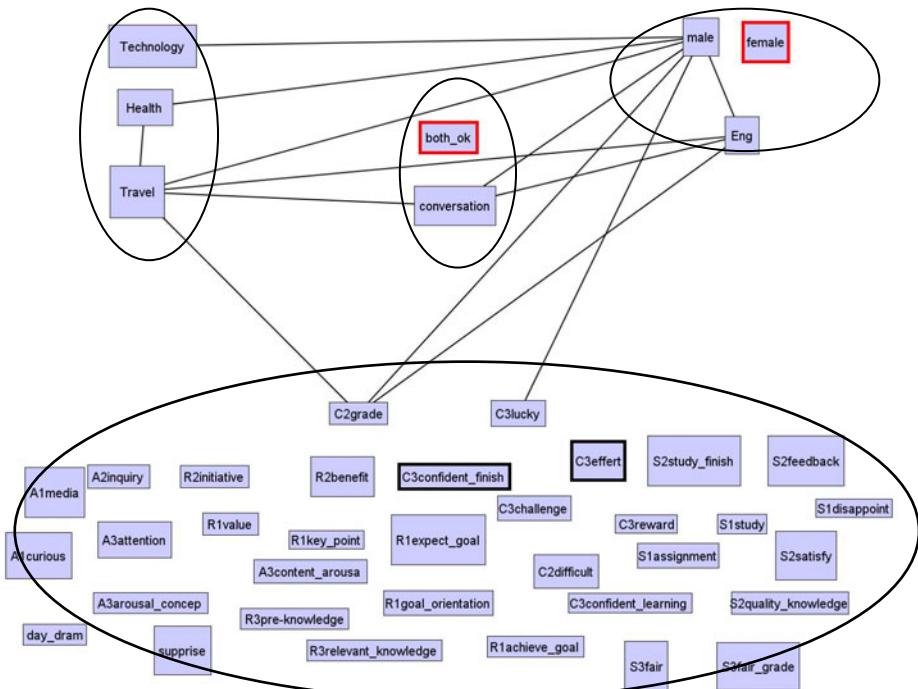


Fig. 3. The jaccard value

## 5 Results and Discussion

The result, namely the final scenario indicated that the attributes such as Confidence in ARCS, the school of engineer and male were generated in students' characters. Such attributes as technology, health, travel, conversation were generated in teaching characters. The important points, the black nodes, generated by KeyGraph were students' confidence in finishing the course and students efforts made in courses. Also, the KeyGraph generated the chance points, the red nodes, which were female and "either kind of material type will do" chosen by students.

The refined scenario was showed in figure 4.



**Fig. 4.** The scenario of students' and teaching' character interaction

### Discussion

In KeyGraph technology the black nodes indicated the high frequency which implied the important points and the red nodes indicated the low frequency which implied the chance points. Hence, the important points both belonged to ARCS factor which was the students' character, and the chance points belonged to either the students' character or the teaching character. In this study case, the word terms in ARCS were transformed from the questionnaire, of which the result made the word terms from ARCS appeared much higher than other word terms. Therefore, the high frequency word terms were all in ARCS model, so that all the ARCS were above the threshold, which made the data analyzed more difficult. The chance building model would not limit

the concept levels, yet the attributes level would be more various in order to provide innovated scenarios.

## 6 Conclusion and Suggestion

The text mining and KeyGraph technologies provide a new methodology for the field of knowledge representation. The black nodes and the red nodes stand for the important attributes and chance attributes. The links represent the relationship. With the nodes and links, the scenarios will be innovated. However, when the chance building model applies the technologies as well as the objectives, it makes the scenarios more meaningful.

The results of this study provide interaction between the students' characters and teaching characters. Each character cluster is composed of two main factors (1) the motivation and graduate school in students' character cluster, and (2) the material topics and types in teaching character cluster. Therefore, this study would suggest some principles to educators and some opinions for future studies.

According to the scenario, the remedial English e-learning course may focus on the confidence of students. The topics such as technology, health and travel may increase the students' motivation. However, there are chances for the difference to appear when female students and the short essay and conversation materials are included.

For the future studies, more various phases should be considered together to conduct the analysis. The intelligent thinking can be applied to the chance building model to provide intelligent scenarios. More research is needed for future study.

**Acknowledgments.** We would like to thank Dr. Hong, C. F. and Dr. Chiu, T. F. from Aletheia University for their support of computer system. The e-learning course was supported by Tamkang University.

## References

1. Hsu, C.L., Chang, Y.F.: Study of the Relationship with the Media Material and the Students' Learning motivation. *J. Educational Study* 116, 64–76 (2003)
2. Taylor, R.P.: *The computer in the school: Tutor, tool and tutee*. Teacher College Press, New York (1980)
3. Hsu, C.L.: E-CAI Case Study. *Educational Technology and Media* 33, 28–35 (1997)
4. Hsu, C.L., Kuo, C.H.: Study of e-Learning Material Technology. In: 2000 e-Learning Theory and Practice Conference, pp. 61–65. National Chiao Tung University, Shin-Chu (2000)
5. Ohsawa, Y., McBurney, P.: *Chance Discovery*. Springer, Germany (2003)
6. Ohsawa, Y., Tsumoto, S.: *Chance Discoveries in Real World Decision Making*. Springer, New York (2006)
7. Watts, D.: *Small Worlds: the Dynamics of networks begween order and randomness*. Princeton, New York (1999)
8. Hsu, C.C., Wang, L.H., Hong, C.F., Sung, M.Y., Tasi, P.H.: The KeyGraph Perspective in ARCS motivation model. In: The 6th IEEE International Conference on Advanced Learning Technologies, pp. 970–974. IEEE Press, New York (2006)

9. Hsu, C.C., Wang, L.H., Hong, C.F.: Understanding students' Conceptions and providing Scaffold Teaching Activities. In: International Conference of Teaching and Learning for Excellence, Tamsui, pp. 166–175 (2007)
10. Smith, P.L., Ragan, T.J.: Instructional design. Macmillan, New York (1993)
11. Michael, T., Marlon, M., Roberto, J.: The third dimension of ADDIE: A cultural embrace. *Tech. Trends* 46(12), 40–45 (2002)
12. Keller, J.M.: Motivation and Instructional Design. A Theoretical perspective. *J. Instructional Development*, 26–34 (1979)
13. Keller, J.M.: Motivational design of instruction. In: Reigeluth, C.M. (ed.) *Instructional Design Theories and Models: An Overview of their Current Status*. Erlbaum, Hillsdale (1983)
14. Visser, J., Keller, J.M.: The Clinical Use of Motivational Messages: an Inquiry into the Validity of the ARCS Model of Motivational Design. *J. Instructional Science* 19, 467–500 (1990)
15. Weiss, S.M., Indurkhy, N., Zhang, T., Damerau, F.: Text mining predictive methods for analyzing unstructured information. Spring Science-Business Media, Inc., New York (2005)
16. Hearst, M.A.: Untangling text data mining. In: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland (1999)
17. Grimes, S.: Mining text can boost research. *Information Outlook* 9(11), 20–21 (2005)
18. Ohsawa, Y., Benson, N.E., Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor. In: Proc. Advanced Digital Library Conference (IEEE ADL 1998), pp. 12–18. IEEE Press, New York (1998)
19. Montero, C.A.S., Araki, K.: Discovering Critically Self-Organized Chat. In: Proc. the Fourth IEEE International Workshop WSTS, pp. 532–542. IEEE Press, New York (2005)
20. Sakakibara, T., Ohsawa, Y.: Knowledge, Discovery Method by Gradual Increase of Target Baskets from Sparse Dataset. In: Proc. the Fourth IEEE International Workshop WSTS, pp. 480–489. IEEE Press, New York (2005)
21. Ohsawa, Y.: KeyGraph as Risk Explorer in Earthquake-Sequence. *J. of Contingencies and Crisis Management* 10, 119–128 (2002)
22. Wang, L.H., Hong, C.F., Hsu, C.L.: Closed - ended Questionnaire Data Analysis. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) *KES 2006. LNCS (LNAI)*, vol. 4253, pp. 1–7. Springer, Heidelberg (2006)
23. Huang, C.J., Tsai, P.H., Hsu, C.L.: Exploring Cognitive Difference in Instructional Outcomes Using Text Mining Technology. In: Proc. of 2006 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2116–2120. IEEE Press, New York (2006)
24. Tsai, P.H., Huang, C.J., Hong, C.F., Wang, L.H., Sung, M.Y., Hsu, C.L.: Discover Learner Comprehension and Potential Chances from Documents. In: The 11th IPMU International Conference, Paris, France (2006)

# A Web-Based E-Teaching System under Low Bit-Rate Constraint

Wei-Chih Hsu<sup>1</sup>, Cheng-Hsiu Li<sup>2</sup>, and Tzu-Hung Chuang<sup>1</sup>

<sup>1</sup> Department of Computer and Communication Engineering, National Kaohsiung First University of Science and Technology, 2 Jhuoyue Rd., Nanzih, Kaohsiung City, 811, Taiwan, R.O.C.

weichih@ccms.nkfust.edu.tw

<sup>2</sup> Graduate Institute of Engineering Science and Technology, National Kaohsiung First University of Science and Technology  
lcs@ntc.edu.tw

**Abstract.** The current teaching situation that developed from traditional teaching to long-distance teaching makes a breakthrough for time and space restriction, it provides the opportunity for learning in anywhere and anytime. In this research, the web-based teaching system will be proposed and implemented by a new concept of Command Script and a idea of Distributed-system into the network structure. By creating the Command Script, it could improve the efficiency more than 3.41 times, decrease the network bandwidth consumption, and increase the transmitting smoothness as well. According to the test results, the web-base e-teaching system could operate in the low bit-rate constraint network environment smoothly.

**Keywords:** Command script, web-based, long-distance teaching, media encoder, streaming technique.

## 1 Introduction

Nowadays, the transmitting speed of the knowledge is unimaginable fast. Nevertheless, the learners are confined from the space and time, because the limited environment can't provide the progressively effective teaching models. The long-distance teaching style has been developed with tremendous pace. By connecting the network systems, learning can be performed in anywhere and anytime[8].

The long-distance teaching can offer the convenience of learning environment. In order to serve different demands efficiently, the teaching system must focus on various network technologies, especially the network bandwidth restriction that must be conquered as the first step, and then the teaching process will be smoothly proceed[10].

The synchronous long-distance teaching model emphasizes on the interaction and instantaneity between the teacher and students, including the screens of the teaching frames, the voice, the teaching materials, and etc.[6]. Currently, ADSL is the most common adopted network, but the uploading bandwidth is less than the 1/8 times of

the downloading bandwidth[11]. This situation will affect the network bandwidth that teacher needed in the teaching process. When it combines with the teacher's images and sounds, the network bandwidth will be occupied a lot. Hence the teaching system won't work smoothly in the low bit-rate network, and the teaching and learning activities are unable to transmit at the same time. Once the insufficient bandwidth occurred, the process of the long-distance teaching will be influenced, and the interaction between the teachers and students will not come out with the expected quality. It will cause the teaching quality dropped obviously, and also led to the negative results in the learning. On the other hand, the system stability is not easy to build up, the management procedure requires long time, and the maintenance step is complicated all make the synchronous long-distance teaching beset with difficulties.

In comparison, developing the system of asynchronous long-distance teaching model is easier than synchronous model. Since it doesn't have the same network bandwidth problems, and the required network techniques are not high. So far there has lots of free software that can be obtained and used (i.e. moodle,xoops, and etc.), and it focus on the diverse function of the on-line learning system platforms and the abundance of the lively teaching materials. Students can learn and review lessons on line with flexible schedule[8]. This is the main purpose that asynchronous long-distance teaching model has been used commonly now.

Based on the above statement, the research will propose a new type of web-based teaching system by Command Script, provide the same quality level for low bit-rate bandwidth users, improve the convenience for operating, and other portion as well. Therefore the main ideas have been concluded as following:

- (1) Adopt the web-based structure as the develop principle, and then create the composite web-based teaching system which includes both synchronous and asynchronous models.
- (2) Apply the ActiveX components to simplify the tedious and complex installation procedure, and improve the user-friendly interface for system operation.
- (3) Embedding the network structure idea of Distributed-system into the systematic structure of this research.
- (4) Create the Command Script that transfers the electric white board drawing steps and the resource website linkages to text format. It could disperse and reduce the consumption of the network bandwidth efficiently, and allows the learners to use the web-based teaching platform under low bit-rate constraint.

## 2 System Design

### 2.1 Architecture of the System

In order to apply the long-distance teaching system on commonly used web browser (e.g. IE, Chrome, and etc.), the transmitted teaching images have been separated into four regions initially, such as Material-display region, Teacher's A/V (Audio/Video) region, Electric-painting region, and Text-message region[8].

Material-display region is functioned as a web browser, which can open teaching materials as browsing the web pages. Teacher's A/V region can record and encode the audio and video of educating procedure, and then upload to the Media Server. Beside, teachers can also preview the video in this region. In Electric-painting region, it provides the white board which embedded with the painting function. Teachers can set up the painting brush in Paint Panel component. Electric-painting is implemented by MFC that uses a transparent rectangle canvas to cover on web browser component, and record all mouse movements and drawing processes as teacher's references. Teachers can transmit and show the messages or the lesson notes in Text-message region. All four regions are embedded into the homepage by applying the web-based style. Whole system structures and the dividing work are showed at Fig. 1. These ActiveX control components will be downloaded and installed while opening the homepages, and they do not have to install the extra application software. The installation only needs to be executed once. Hence, the web-based learning will be much easier and convenient for learners to experience.

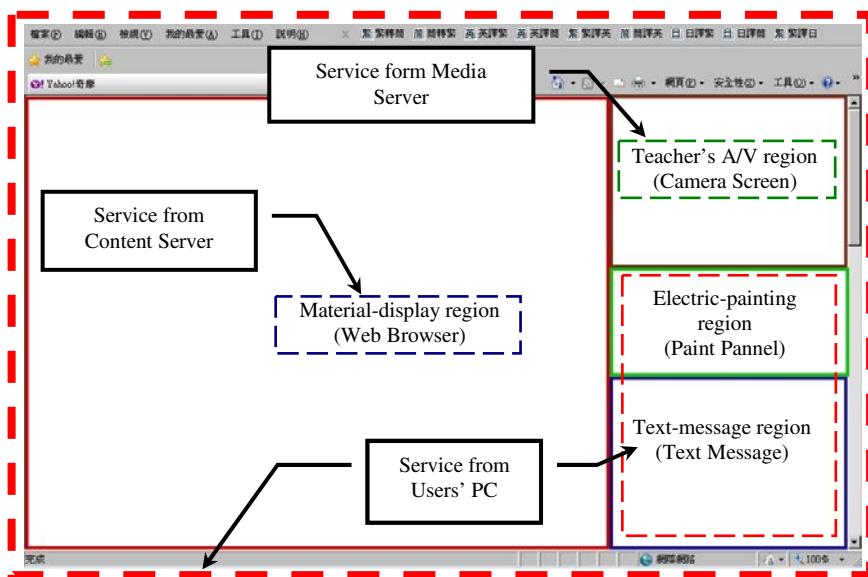


Fig. 1. Architecture and dividing work of the system

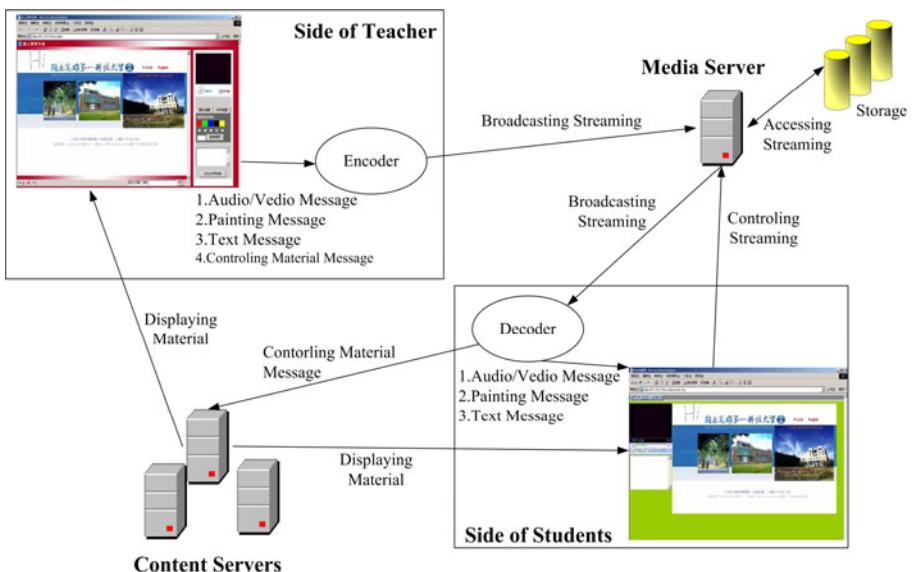
## 2.2 Analysis of the System

In this system, the users' roles have been defined to four sections as teachers, students, Media Server, and Content Servers. The teacher's sections mainly focus on web-based teaching system. The student's portions are the passive receiver of the information, and it can be received in various locations with plenty access points. Media Server is the repeater for processing the A/V data; it will receive the streaming from the teachers and store them as the multimedia files, and it also stand by for students' demands simultaneously. Content Server is the repeater for demonstrating the

teaching materials; it provides and displays the links of web-based teaching homepages, and it can be distributed in the different places. These four sections are embedded in the different domains. The teachers and students don't connect each other directly; however the repeaters of Media and Content Servers will deliver the teaching data. In order to reduce the occupancy rate of the online bandwidth, the system especially separates the teaching A/V and material providing servers, and then enhances the transmitting quality of the teaching A/V.

In students' operating interface, Microsoft IE (Internet Explore) has been adopted as the application platform, and then embeds the self developed ActiveX control components with the homepage.

When students connect to the Content Server and operate the web-based teaching interface, the learning interface will display and also receive the streaming that comes from Media Server. The streaming will be decoded by Media Player in student's section, and the teaching A/V and Command Script will also be separated from it. Afterwards the teaching A/V will be displayed by Media Player, and the functions of Command Script will start to execute, such as Electric-painting, Text-message, teaching material display, and etc.. Since they are encoded at the same timestamp, they will implement the commands and functions under the same timestamp in the decoding processes[4]. The Fig. 2. illustrates the flowing of the streaming in the teaching system.

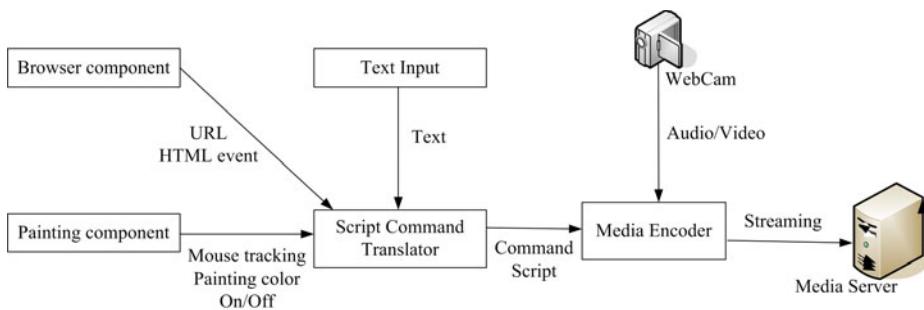


**Fig. 2.** Flowing of the streaming

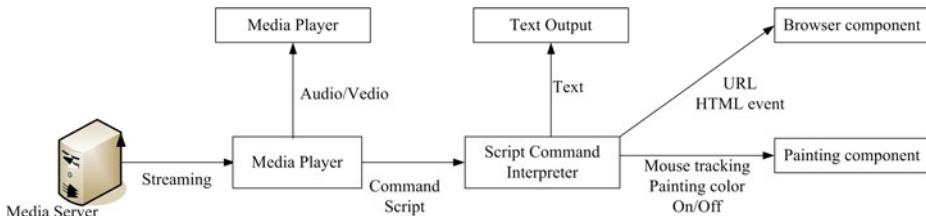
In order to control the activities of students' components by broadcasting method, it is necessary to transmit the pre-set event notices and the attribute coding through Translator which is on teacher's homepage, and then apply Media Encoder to release the coding. This method will make the synchronization between students' and teachers' components, and execute the same operating commands with them.

The teacher's Translator aims on users' controlling component data transmission, and also catch the exported events of the web browser and painting component. The Translator will transfer them into the pure text format, deliver them to the encoder component, and then control the web browser link activities at the same time.

In the functions of Interpreter, the Media Player will separate Command Script from the streaming, and then transfer them to the controlling command for handling the web browser components and Electric-painting of student's section. The component of Media Player Control that is structured by ActiveX control can be embedded in the homepage, and apply it to separate Command Script from the streaming[4][9]. The Fig. 3. and Fig. 4. demonstrate the operating situation between every component and Translator/Interpreter.



**Fig. 3.** Flowing of Translator



**Fig. 4.** Flowing of Interpreter

In order to solve the insufficient bandwidth problem while transmitting the teaching A/V on-line, the teachers' teaching A/V will be compressed by multimedia characters of Media Encoder and Media Player during the transmitting period, and utilize the pure text style of Command Script to control the display of the teaching materials, the teaching messages, and the Electric-painting. The streaming style has been adopted to transmit the multimedia data, and then using Windows Media Encoder to encode the other teaching contents. Meanwhile, the contents of teaching A/V and Command Script will be transmitted. Afterwards, the content will be decoded and separated at the received side. This method transfers the images to the text format will constrain the large occupancy bandwidth of the images.

When teachers operate the web-based interface during the teaching process, Media Encoder will encode the Command Script of teaching A/V, Electric-painting, text

message, teaching material display, and etc. into the streaming at the same timestamp, and then broadcasts it to Media Server, meanwhile it is not only waiting for students' requests, but also saving the streaming in the Media Server simultaneously for learners to review the lessons after class.

### 3 Experiments

The main purposes of this chapter will focus on the network testing and performance evaluation, and also provide the experiments for system implement and efficiency evaluation.

#### 3.1 Network Testing

The streaming time delay of web-based teaching system is measured by LAN (Local Area Network) and ADSL (Asymmetric Digital Subscriber Line), which adopts the NTP (Network Time Protocol) Clock from Bureau of Standard, Metrology & Inspection. According to the test results, the uploading and downloading bandwidth of ADSL were asymmetric, and the uploading bandwidth was less than downloading bandwidth about 1/8 times, however the system still can keep the streaming delay time within 17 seconds; and the transmitting performance even better in the LAN.

Other functions of web-based teaching system also operate precisely. For example, the Electric-painting function can correspond with teachers' demands to determine the line color, and then display on the web-based teaching system. On the other hand, the Text-message function is able to transmit teachers' additional messages to students; and the Material-display region can demonstrate the teaching materials immediately when teachers connect to Content Server and operate the teaching materials.

#### 3.2 Performance Evaluation

This chapter explains five evaluation aspects to assess the system performance: the installation method, system structure, long-distance teaching model design, bandwidth constraint, and operating interface. The details are described in the following statement.

##### 3.2.1 Evaluation of Installation Method

In order to provide the plentiful support for web-based teaching system, for instance Co-elearning[2], Tutor ABC, Global village, and etc., usually have to install the application software. Nevertheless the whole procedure from download program to complete installation may cause some problems for users who are not familiar with computer operating.

The innovative installation of the system is totally different than normal platforms. By applying the ActiveX control component to develop the web-based system; the A/V encoding, Material-display, and Electric-painting have been intergraded with the homepage browser. Users only need to browse the homepage, and the components will be downloaded and installed in client computer automatically. This method enhances the system operating convenience substantially; especially for inexperienced computer users to avoid the technical obstacle.

### 3.2.2 Evaluation of System Structure

The system adopts the Distributed System structure to locate the teaching materials and A/V media in the Content and Media Server individually, and provides their resource separately. When students browsed the teaching homepage, the web-based teaching interface and teaching materials were provided by Content Server, and the teacher's A/V were provided by Media Server.

This is a innovative idea that used Distributed System structure to separate the bandwidth and effective dispersedness, then could improve the operating performance of the system under the low bit-rate constraint.

### 3.2.3 Evaluation of Long-Distance Teaching Model Design

Generally speaking, the web-based teaching system includes synchronous and asynchronous mode. However, asynchronous is the most common type. Asynchronous mode usually does not restrict by bandwidth, but it can't display the content instantly.

The system combines synchronous and asynchronous mode for long-distance teaching platform. The system applies ActiveX control components for developing; the Command Scripts include painting information, interaction text messages, and etc., are encoded and broadcasted to client computers immediately by Media Encoder. Meanwhile the multimedia files will be stored in the Media Server. Therefore the advantages of synchronous and asynchronous mode will be integrated in the system simultaneously.

The features of this teaching system can provide the real-time teaching, after class practicing, and missed lesson catching up, so the teaching quality can be improved in a higher level.

### 3.2.4 Evaluation of Bandwidth Constraint

The system separates the original teaching images to Command Script and A/V streaming two parts. Since Command Script is pure text format with less than  $10^2$ Bytes size, the most bandwidth consumption comes from the A/V streaming, so the transmitted image is the only item to be compared with bandwidth performance.

For example, use a 300 thousand pixel webcam for A/V catching, adjust the image transmitting time to one minute, and then set 25 frames per second. Afterwards compare to the original full color image size of  $1024\times768$  pixels. Without compressing, the capacity can be reduced about 3.41 times.

In fact, the system is designed by A/V streaming which based on the Media Encoder. Besides, it will compress the A/V, and then display the A/V with streaming format. Hence both of the display and transmitting performance are much better than non-compressed conditions. The results verified the system can be operated smoothly under low bit-rate network constraint.

### 3.2.5 Evaluation of Operating Interface

Although the system utilizes Command Script to replace original image, it still can offer the full functions for web-based teaching classroom; including teacher's teaching A/V, teaching materials, electric painting, text message, and etc.. Therefore the virtual-reality learning can be imitated tremendously under the low bit-rate network constraint, and the effectiveness and efficiency of the long-distance teaching will be increased obviously.

## 4 Conclusions and Future Work

In this research, a new concept for implementing and enhancing the web-based teaching system by Command Script has been proposed. The system combines both advantages of synchronous and asynchronous modes simultaneously, and the features include real-time teaching, after class practicing, and absent lesson catch up. It makes “anyone can learn in anywhere and anytime” come true. Furthermore it can also be operated under the low bit-rate network constraint because of the low demand bandwidth.

The Distributed-system has been integrated into the network structure. The created Command Script transforms the electric white board drawing steps and the material website links to pure text formats, and then encoding them into the streaming with teachers’ video and audio. It can improve the efficiency more than 3.41 times of the network bandwidth consumption and the smooth transmission as well. The results confirmed the web-based teaching system can be installed and operated properly in low bit-rate network constraint without any interruption.

In the future, the system will be arranged to assist teaching, and it will be assessed by questionnaire and experimental evaluation. It can analysis teachers’ satisfaction of instructional proficiency. On the other hand, it can grade students’ effectualness before and after the learning. Overall, the influence for teaching and learning will be able to discover.

## References

1. Advanced Distributed Learning (ADL) Initiative (2004),  
<http://www.adlnet.gov/About/Pages/adlinitiative.aspx>
2. Collaborative eLearning website (2004), <http://www.coelearning.com/>
3. Davey, K.B.: Distance Learning Demystified. Phi. Kappa Phi. 79(1), 44–46 (1999)
4. Internet Engineering Task Force Internet Draft, draft-ietf-mmusic-rfc2326bis-06.txt (2003)
5. Pan, M.-J.: How many E-learning in the future (2003),  
[http://www.find.org.tw/0105/news/0105\\_news\\_disp.asp?news\\_id=2768](http://www.find.org.tw/0105/news/0105_news_disp.asp?news_id=2768)
6. Frank, M., Kurtz, G., Levin, N.: Implications of Presenting Pre-University Courses Using the Blended e-Learning Approach, Educational Technology & Society 5(4) (2002), ISSN 1436-4522, [http://www.ifets.info/journals/5\\_4/frank.html](http://www.ifets.info/journals/5_4/frank.html)
7. National Sun Yat-Sen Cyber University (2008), <http://cu1.nsysu.edu.tw/>
8. Power, D., Stevens, K., Boone, W., Barry, M.: VistaSchool District Digital Intranet: The Delivery of Advanced Placement Courses to Young Adult Learners in Rural Communities, ERIC Accession No. ED449933 (1999)
9. Deering, S.: Stanford University. RFC 1112 - Host extensions for IP multicasting (1989)
10. Yang, Z.-Y., Lee, L.-S., Tseng, L.-M.: The On-line Interactive and Off-line Self Learning Distance Learning System. In: National Computer Symposium, vol. 86, pp. 135–140 (1997)
11. Lee, Y.-P.: 6 How many Internet Subscriber (2009),  
<http://www.find.org.tw/find/home.aspx?page=many&id=243>

# Using Freeware to Construct a Game-Based Learning System

Yuh-Ming Cheng and Li-Hsiang Lu

**Abstract.** Game-Based Learning is being widely used in the field of education. To construct such a system needs considerable time and resources. The purpose of this research is to construct a Game-Based Learning system for junior and elementary school students by using game engine and network engine from freeware, combined with materials of questions designed by academic experts, characters and scenes designed by art staff. The experimental result can understand that this kind of learning is more active and effective for students to get online to learn in quantitative results, and the improvement of learning outcomes from students are also more obvious.

**Keywords:** Game-Based Learning, Integrated Science, E-Learning.

## 1 Introduction

Game-Based Learning is a widely discussed learning mode in recent years [1] [2] [3] [4] [5], since the game-based multi-media learning elements, such as characters, voice and images, are combined to enhance the attention of the students and foster their concentration, interest, creativity and community relationship [6]. Students are often addicted to computer games, especially for Internet-based games, as Internet-based games are characterized with the online operation mode, which produces a community relationship like a virtual world. Given the interaction, cooperation and competition amongst individuals in this virtual world, a unique coherence will be generated from such virtual/real interpersonal relationship. If the teaching materials are incorporated into online games, and every student gains an access to Internet for role-playing through cooperative learning and competition with others, it is possible to improve effectively their learning performance and educational value [7].

Game-Based Learning (GBL) is one of serious games with positive learning performance, so it is seldom discussed with respect to its effectiveness and education potential[8]. The Game-Based Learning is designed to reach a balance between Gameplay and teaching, and make the players capable of applying their knowledge into the real environment. Gameplay is set to connect single or multiple learning objectives to a series of stages for the purpose of natural learning in the games. With the help of Gameplay, the players are motivated to participate in the games and complete their missions more quickly. Despite of its sufficient acousto-optic effect, Gameplay will disappear quickly from the market if its educational significance is neglected. Therefore, it is required to pay equal attention to the educational purpose and Gameplay in Game-Based Learning [9].

As pointed out by Prensky, Game-Based Learning is conducted primarily owing to three factors [10]:

1. The combination of learning with games could draw the attention of the students and improve their learning desire, especially for the subjects disliked by them.
2. A variety of interactive learning processes could help realize the learning objectives.
3. Lots of approaches are used to combine the games with learning, of which highly contextual learning is preferred.

Due to implicit educational function, the games with narrative scenarios or stories enable the players to absorb knowledge unconsciously during the gaming process. The educational computer games can draw the attention of the students, allowing them to develop their cognition and experience along with the evolution of games [11]. In view of wide-ranging community relationships of different levels in existing games, the students can develop their own community relationships; for example, the children can express directly the learning problems encountered by themselves or their groups. Most of youth often join in Internet-based communities or groups during the games [12], and experience the interaction. Thus, online gaming environment with educational property provides the students with more in-depth and meaningful knowledge acquisition.

A successful GBL policy helps to improve the participation and interaction of students by using the gaming experience and user-friendly graphical interface as a learning tool. Such phenomenon already exists in the well-educated elite, who often indulge themselves into such games[3].

This study added the clue mechanism into the games as a prompt mechanism during the learning process. Through scaffolding learning process during Game-Based Learning, the students can acquire guidance and assistance by adding the clues [13], making it possible to develop more capabilities for the children. By putting focus on the students and laying emphasis on their attributes and potential, original limited cognitive capability can be developed into potential cognitive capability for the purpose of optimizing the learning performance [14].

At the same time, the challenges and degree of difficulty in the games are important control elements. For instance, psychologist Csikszentmihalyi initiated a Flow Theory in 1975 [15], showing that the people are prone to lose their consciousness or forget the time when they are interested in and fully involved in something. According to the definition of Csikszentmihalyi, flow experience means that “the users in a common empirical mode seem to be absorbed and their consciousness is limited collectively to a confined range. Thus, some irrelevant perceptions and ideas are filtered out without consciousness, and a perceived control is generated by manipulating the environment in response to specific objects and feedback. To make the students enter the intermediate state of the flow model and devote themselves to learning for a higher learning capability, this study applied this theory to Game-Based Learning. While designing the learning materials, the prerequisite knowledge and problem-solving capability of the students are taken into account.

The remainder of this paper is organized as follows: Section 2 describes the possible problems and solutions in development of GBL system; Section 3 explains the design methods of the system, the operating mechanism and system framework; Section 4 presents an actual test of the students, assess the students' satisfaction and the influences of the system on the learning performance; Section 5 gives the conclusion.

## 2 Development of Gaming and Learning System

Two problems are likely to occur during system development, which are mapping, and Internet connection and multithreads processing. Take Windows for example, textures must be processed in collaboration with the game programming interface (DirectX) developed by Microsoft, if the development is conducted without use of game engines, however, the establishment of entire window and the development of user's graphical interface must be considered. The current game engines of free software provide some necessary means, such as: 3D images, collision detection and user's graphical interface modules, helping to develop the games more quickly and conveniently with the functions offered by game engines.

### 2.1 Game Engines

Based on the support language C/C++ and Windows OS, three widely-used and well-known game engines are listed in Table 1:

**Table 1.** Comparison of Game Engines

Name	Graphical API	Support language	Efficacy	Stability	Ease-of-use	Support
OGRE	OpenGL, DirectX	C/C++	80	80	70	80
Irrlicht	OpenGL, DirectX, Software	C/C++,C#, VB.NET	80	80	90	80
Crystal Space	OpenGL, Software	C/C++	90	80	70	90

Devmaster[16] is an exchange platform of international game engines, onto which an extremely powerful retrieval system is established, and the game engines with top three open sources are compared. In spite of their advantages and disadvantages, game engines are developed with major purpose of saving the development time and cost, while the difficulty-of-use is evaluated and considered carefully. Owing to the limitation of expenditure, the computer classrooms of most schools are generally equipped with all-in-one embedded devices. In the absence of limitations of independent display cards and computer performance, the graphical API delivered by the game engines is of

utmost significance. Without this support, the games cannot run smoothly due to insufficient performance, inability of supporting Alpha Channel or lower FPS (Frames per Second).

## 2.2 Internet Engines

The games are often networked by 2 modes: Peer-to-Peer and Client/Server, and can be implemented by many methods irrespective of any mode (including RakNet network engine). Generally, the servers are equipped with computers of quickest processing and networking speed, and the clients with other kinds of computers. Today, the files are transferred by either UDP or TCP. TCP has excellent effect in transferring files, but often yields numerous transfer delays in games because of streaming (not packeting).

RakNet network engine plays a basic role of providing a complete UDP transfer environment and helping the developers in solving problems. In such case, the developers are only required to put their focus on the games. RakNet network engine can provide the following assistances in game development:

- Re-send automatically the packets not really reached.
- Send the packets automatically in sequence or in order to increase the transfer efficiency.
- Guarantee the entire sending process of packets, or notify automatically the programmer if tampered externally.
- Provide a quick and simply interface and restrict unauthorized transfer.
- Handle efficiency network problems such as: control and collection of streams.

It is time-consuming to enable the bit streams and packets in an efficient bandwidth and provide a number of network control functions. RakNet network engine guarantees easier and efficient networking with its many functions, such as remote functional calling, bit stream types, and automatic objects synchronization. Thus, network engine not only can save the development time, and also offer such functions as encryption, resources management, packet transfer and multithread management, making it possible to bring the functions of network elements into full play.

Based on the mapping and environment establishment functions offered by Irrlicht game engine and the networking control functions offered by RakNet network engine, this paper attempts to build an online competitive gaming and learning system to reduce the time and resources required for game development.

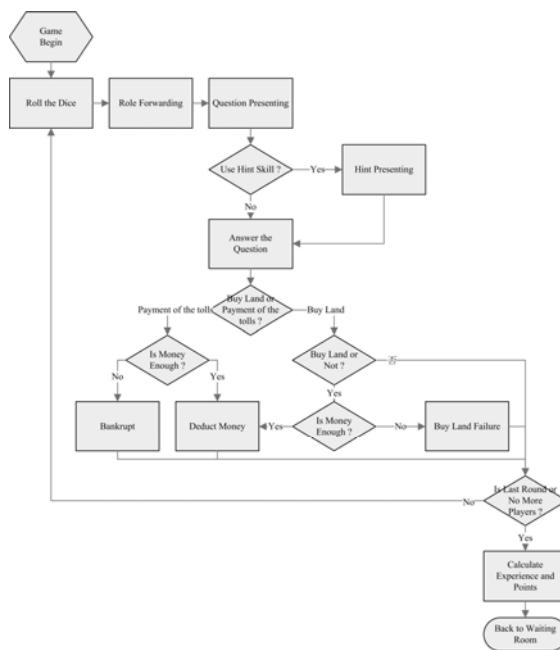
## 3 System Analysis and Framework

### 3.1 System Description

This system is an online multiplayer competitive gaming and learning system. After successful logging, players can enter into the gaming room, or activate a new game at their own discretion. Every game can be attended by 2 to 4 persons. For instance, in the Monopoly puzzle game, the chief of the gaming room can select a map, and every

participant can select the intended roles at the waiting room, then the game starts after the chief and all participants press OK button.

When the game begins, every participant rolls the dice in turn. All participants must answer a question in every round. And every player is provided with coins and points, and the coins are deducted only after answering the questions if he/she intends to buy land and build a house. The payment of coins depends on the deduction arising from the performance, and parameters can be set by the system administrator through a database. When walking to the premise of his competitor, the player must firstly answer the questions. The payment of tolls also requires parameter setting by the system administrator through a database. When answering the questions, users can use points to call the clues in order to obtain the prompts for the item, as shown in Figure 1.



**Fig. 1.** Flow Diagram of Games

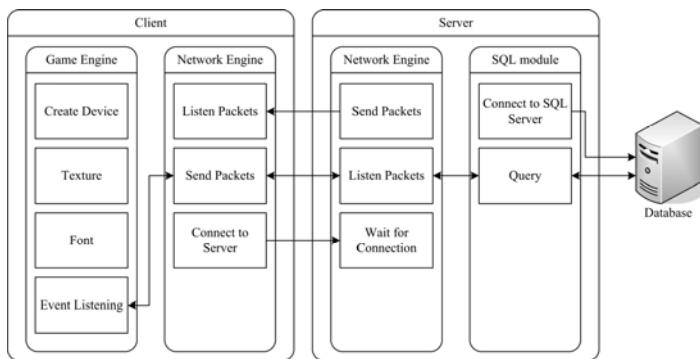
Finally, when the preset rounds are reached, or only a single person survives as the others are bankrupt (this round is terminated forcibly), the system will calculate the number of coins or the bankruptcy sequence, and then assign the points and experience. This game is characterized by:

- Research on learning materials: the teaching staff may forecast and analyze the teaching items and clues to establish a strong teaching library, and assist the arts and program designers in planning suitable learning materials.

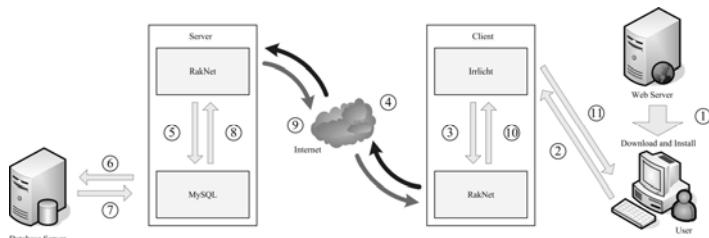
- Increase the interaction of players: use points to call the clues, thus increasing the interaction of players for the fun of “learning through play”.
- Personal training: the players may enter the personal study room if they fail to answer correctly the questions in the games. The items in the personal study room are designed by the teaching experts in order to guarantee the credibility and effectiveness.

### 3.2 System Framework

The system framework is shown in Figure 2, wherein Irrlicht game engine and Raknet network engine are combined integrally to build the entire system. At the user end, the game engine is responsible for creating the system window, font, texture and listening events at the user interface, the network engine responsible for connecting the user to the server and sending the packets. At the server, the network engine is responsible for transferring messages among the users. In addition, MySQL module is employed for connection to database server and query.



**Fig. 2.** System Framework



**Fig. 3.** Flow Diagram of System Operation

Figure 3 shows the flow diagram of game engine and network engine in the system. Firstly, the player downloads the user system (item ①) from the website delivered by the web server, then installs it and triggers various actions (item ②) through the graphical interface presented by Irrlicht game engine. If log-in verification is required, for instance, Irrlicht game engine will transfer the account and password entered by the players to RakNet network engine for packeting (item ③), and then send via Internet to the server (item ④).

The packets of the server are transferred by RakNet network engine by converting into corresponding format and calling the respective services. After RakNet network engine calls the log-in verification program, the program will call MySQL module, and send a quest for obtaining the player's account and password (item ⑤), and MySQL module is linked to the back-end database for acquiring the player's account and password (item ⑥, ⑦). After verifying and comparing the player's account and password obtained from the database with those entered by the player, the results are transferred to RakNet network engine for packeting (item ⑧), and RakNet network engine also sends the packets to the user (⑨) via Internet; after receiving the packets, RakNet network engine of the user also converts the packets into a corresponding format, and transfer the verification results to Irrlicht game engine (item ⑩), which will display the results at the graphical interface accessible to the players (item 11), thus finishing an entire verification process.

## 4 Performance Analysis

The students in middle and elementary schools were grouped into S classes, not just based on their learning capability. The research subjects were first-year students from two classes of a junior high school in Pingtung County, and the experimental period was 3 weeks. The teachers of the two classes were asked to teach the students to reach the same level, and a pretest was conducted. Then, the two classes were divided into a test group and a control group in S form, with students of high, medium and low learning capabilities distributed evenly. The students in the test group were allowed to learn and compete online at home with their classmates, while the control group received the homework and teaching materials.

To learn the performance of two groups of students subject to different teaching methods, SPSS software package was used for descriptive statistics, and then Paired-Samples T Test was conducted to analyze p value and t value by the significance level 0.05.

**Table 2.** Descriptive statistics of test and control groups

		Mean value	Standard deviation
Traditional teaching (Control group)	Pretest	53.625	10.697
	Post-test	70.125	14.966
Teaching with the system (Test group)	Pretest	53.030	8.932
	Post-test	86.273	13.531

Table 2 lists the mean value and standard deviation of two groups of students with respect to their performance of comprehensive subjects in pretest and post-test under different teaching methods.

**Table 3.** T statistics of test and control groups

Test and control groups	T-value	P value
Pretest	0.242	0.404
Post-test	-2.250	0.014
Rate of progress	-3.923	0.000

To prove the system performance, two groups of pretest scores were subject to t-test, as listed in Table 3. According to the statistics, p is 0.404, which is larger than the significance level if 0.05, indicating that there is insignificant difference between two groups in pretest. The t value is 0.242, indicating that both groups had insignificant difference with regard to the pretest scores, but the control group has a better score than test group.

In comparison of post-test scores, p value is 0.01, smaller than the significance level 0.05, showing significant difference between two groups. As mentioned above, t value is negative, indicating the test group has a bigger post-test score than the control group. In comparison of the rate of progress, p value is 0.00, smaller than the significance level 0.05, showing significant difference between two groups. Therefore, this system has significant effect in collaboration with traditional teaching methods.

## 5 Conclusions

As compared with other Monopoly computer games currently available, the games are set to strike a balance between amusement and education, but the game systems, such as: fortune, opportunity and trapping tools, are less abundant and diversified than Monopoly games. Given the fact of competition between computer AI and the players in a single-player mode, the current Monopoly game still maintains its competitiveness and fun. However, due to lack of competition between computer AI and the players in a single-player mode, the training room in this research is less attractive to the players as an important problem to be resolved in the next phase.

The actual test results indicate that, there is no significant difference between two groups of students in terms of pretest, while p value of post-test is 0.014, smaller than 0.05, indicating the significant effect of answering question mechanism on the students. As for the feedback mechanism, the students hope to gain awards or scores from the teachers by the game points. If the game points are changed into virtual currency similar to the current games, these currencies can be used to call the clues in the games,

and also enable the students to gain awards or scores from the teachers by the game points as another incentive for their proactive learning.

The quantified test data indicate that, the tested students have significant difference in traditional teaching and online game-based competitive learning. As mentioned in the literature review, there is a growing trend that free software will be incorporated into the teaching, so direct access to free software is now one of important topics in the current phase.

Future study will expand the samples and explore the effects of online game-based competitive learning by examples of grade 5 students of an elementary school in Kaohsiung County and six categories of Chinese characters. This will surely present the auxiliary effects of online Game-Based Learning.

## Acknowledgment

The authors wish to thank the project was supported by a grant from National Science Council, Executive Yuan ROC (98-2511-S-366-002-).

## References

1. Gendron, E., Carron, T., Marty, J.C.: Collaborative Indicators in Learning Games: an immersive factor. In: Proc. of the 2nd European Conference on Games Based Learning, Barcelona, Spain, pp. 16–17 (2008)
2. Jarmon, L., Traphagana, T., Mayratha, M., Trivedia, A.: Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life. *Computers & Education* 53(1), 169–182 (2009)
3. Johnson, L., Levine, A., Smith, R., Smythe, T.: The 2009 Horizon Report. The New Media Consortium, Austin (2009)
4. Proberta, E.: Information literacy skills: Teacher understandings and practice. *Computers & Education* 53(1), 24–33 (2009)
5. Şendağa, S., Odabaşb, H.F.: Effects of an online problem based learning course on content knowledge acquisition and critical thinking skills. *Computers & Education* 53(1), 132–141 (2009)
6. Raessens, J., Goldstein, J.: *Handbook of Computer Game Studies*. MIT Press, Cambridge (2003)
7. Tuzan, H.: *Motivating Learners in Educational Computer Games*. Indiana University, Bloomington (2004)
8. Kirriemur, J., McFarlane, A.: Literature review in games and learning. NESTA Futurelab series. NESTA Futurelab, Bristol (2004)
9. Kiili, K.: Digital game-based learning: Towards an experiential gaming model. *Internet and Higher Education* 8, 13–24 (2005)
10. Prensky, M.: *Digital Game-Based Learning*. McGraw-Hill, New York (2001)
11. Barab, S., Thomas, M., Dodge, T., Carteaux, R., Tuzun, H.: Making learning fun: Quest Atlantis, a game without guns. To appear in *Educational Technology Research and Development* (2005)

12. Lenhart, A., Kahne, J., Middaugh, E., Macgill, A., Evans, C., Vitak, J.: Teens, Video Games, and Civics: Teens' gaming experiences are diverse and include significant social interaction and civic engagement. Pew Internet and American Life Project, New York, Pew Research (2008)
13. Cheng, Y.: Building a General-Purpose Pedagogical Agent in a Web-Based Multimedia Clinical Simulation System for Medical Education. Preprints of IEEE Transactions on Learning Technologies: Accepted for Future Publication (2009)
14. Vygotsky, L.S.: *Mind and Society: The Development of Higher Psychological Processes*. Harvard University Press, MA (1978)
15. Csikszentmihalyi, M.: *Beyond Boredom and Anxiety*. Jossey-Bass, San Francisco (1975)
16. Devmaster, <http://www.devmaster.net/engines/>

# Three Kinds of Negations in Fuzzy Knowledge and Their Applications to Decision Making in Financial Investment

Zhenghua Pan, Cen Wang, and Lijuan Zhang

School of Science, Jiangnan University, Wuxi, China  
panzh@jiangnan.edu.cn

**Abstract.** The three kinds of negations of fuzzy information were proposed in this paper, which are contradictory negation, opposite negation, and fuzzy negation. Based on the medium predicate logic MF and the infinite-valued semantic interpretation  $\Phi$  of MF, the representation and reasoning on fuzzy information and its three kinds of negations were also investigated. To show applicability of the above results, an example which assists decision making in the financial investment was discussed. Concretely, the paper introduced a new Fuzzy Production Rules whose threshold value was related to  $\lambda$  in  $\Phi$ , and also discussed the reasoning and realization about fuzzy information and its three kinds of negations in the example.

**Keywords:** Fuzzy information, negation of information, medium predicate logic, assistant decision making, financial investment.

## 1 Introduction

Negation of concept plays a special role in information processing. Negative information has to be taken into account as important as positive information. Due to the fact of different negations in various information processing, especially fuzzy information processing, some scholars proposed that information processing need differentiate different negations [1-7]. Wagner et al considered that at least exist two kinds of negation: a weak negation expressing non-truth (in the sense of “she doesn’t like snow” or “he doesn’t trust you”) and a strong negation expressing explicit falsity (in the sense of “she dislikes snow” or “he distrusts you”) [1][2][3]. Kaneiwa proposed description logic ALC<sub>~</sub> with classical negation and strong negation, the classical negation  $\neg$  represents the negation of a statement, the strong negation  $\sim$  may be more suitable for expressing explicit negative information (or negative facts). In other worlds,  $\sim$  indicates information that is directly opposite and exclusive to a statement rather than its complementary negation [4]. Ferré introduce an epistemic extension for the concept of negation in Logical Concept Analysis and Natural Language, the extensional negation is traditional negation, for example, “young/not young” and “happy/not happy”. The intensional negation can be understood as opposition, for example, “hot/cold” and “tall/small”[5]. Pan consider that information processing should differentiate contradictory negation and opposite negation in information, and proposed the five kinds of negation relations in distinct information and fuzzy information, as well as a logic description to these relations [6][7].

## 2 Different Negation Relations in Fuzzy Concepts

Concept is base of information. A fuzzy concept means that its extension is vague. In Formal Logic, relation between the two concepts means that relation of extensions, and it includes consistent relation and inconsistent relation. Relation between concept  $A$  and  $B$  is inconsistent means that extension of  $A$  have nothing in common with extension of  $B$ , for instance, *white* and *nonwhite*, *young people* and *old people*, *conductor* and *nonconductor*.

A concept and its negation belong to a concept of genus, relation between a concept and its negation is inconsistent relation, and sum of extensions of a concept and its negation concept equates with extension of the concept of genus in formal logic. Since Aristotle, the inconsistent relation in concepts had distinguished into contradictory relation and opposite relation. Thus, the relation between a concept and its negation should include contradictory negative relation and opposite negative relation, and we consider that there exist three negative relations in fuzzy concept.

For the convenience of depiction, following  $F$  denotes a fuzzy concept,  $F_n$  denotes negation of  $F$ ,  $E(x)$  denotes extension of concept  $x$ .

### 1) Contradictory negation relation in Fuzzy Concept (CFC)

Character of CFC: “borderline between  $E(F)$  and  $E(F_n)$  is uncertain. Either  $E(F)$  or  $E(F_n)$  in extension of genus concept.”

For example, “*young people*” and its contradictory negation “*non-young people*” in the genus concept “*people*” are fuzzy concepts, the relation between *young people* and *non-young people* is CFC, the borderline between  $E(\text{young people})$  and  $E(\text{non-young people})$  is uncertain, and either  $E(\text{young people})$  or  $E(\text{non-young people})$  in  $E(\text{people})$ . “*quick*” and its contradictory negation “*not quick*” in the genus concept “*velocity*” are fuzzy concepts, the relation between “*quick*” and “*not quick*” is CFC, the borderline between  $E(\text{quick})$  and  $E(\text{not quick})$  is uncertain, and either  $E(\text{quick})$  or  $E(\text{not quick})$  in  $E(\text{velocity})$ , and so on.

### 2) Opposite negative relation in Fuzzy Concept (OFC)

Character of OFC: “ambit between  $E(F)$  and  $E(F_n)$  is uncertain, it is not  $E(F)$  neither  $E(F_n)$  in extension of the concept of genus.”

For example, “*young people*” and its opposite negation “*old people*” are fuzzy concepts in the concept of genus “*people*”, relation between *young people* and *old people* is OFC, ambit between  $E(\text{young people})$  and  $E(\text{old people})$  is uncertain, the ambit is not  $E(\text{young people})$  neither  $E(\text{old people})$  in  $E(\text{people})$ . “*quick*” and its opposite negation “*slow*” are fuzzy concepts in the concept of genus “*velocity*”, relation between “*quick*” and “*slow*” is OFC, ambit between  $E(\text{quick})$  and  $E(\text{slow})$  is uncertain, and this ambit is not  $E(\text{quick})$  neither  $E(\text{slow})$  in  $E(\text{velocity})$ , and so on.

For all of many opposite concepts, one of characteristics of them is that there exists “medium” concept in between the two opposite concepts. The medium concept represents intergradations from one to another between the two opposite concepts. For instance, “zero” between the positive number and the negative number, “middle” between left and right, “semiconductor” between the conductor and the nonconductor, “dawn” between daylight and night. We discovered that there is following characteristic between opposite fuzzy concepts after studying large numbers of instances,:

- If pair of opposite concepts is fuzzy concepts, there must exist medium fuzzy concept in between them. Conversely, if there is a medium fuzzy concept between the two opposite concepts, the two opposite concepts must be fuzzy concepts. In other words, pair of opposite concepts is fuzzy concepts if and only if there is a medium fuzzy concept between the two opposite concepts.

Therefore, we consider that medium fuzzy concept is a new negation of the two opposite fuzzy concepts, which is called “*medium negation*” of the opposite fuzzy concepts. Similarly, following  $F_m$  denotes the medium negation of fuzzy concept  $F$  in following depiction.

### 3) Medium negative relation in Fuzzy Concepts (MFC)

Character of MFC: “borderlines between  $E(Fm)$  and  $E(F)$  (or  $E(Fn)$ ) are uncertain. Not either this or that for  $E(Fm)$ ,  $E(F)$  and  $E(Fn)$  in extension of genus concept.”

For example, “*middleaged people*” in between the two opposite concepts “*young people*” and “*old people*” is a medium fuzzy concept in the concept of genus “*people*”. Relations between *middleaged people* and *young people* (or *old people*) are MFC, the borderlines between  $E(\text{middleaged people})$  and  $E(\text{young people})$  (or  $E(\text{old people})$ ) are uncertain, and not either this or that for  $E(\text{middleaged people})$ ,  $E(\text{young people})$  and  $E(\text{old people})$  in  $E(\text{day})$ . “*dawn*” between “*daylight*” and “*night*” is medium fuzzy concept in the concept of genus “*day*”, relations between *dawn* and *daylight* (or *night*) are MFC, the borderlines between  $E(\text{dawn})$  and  $E(\text{daylight})$  (or  $E(\text{night})$ ) are uncertain, and not either this or that for  $E(\text{dawn})$ ,  $E(\text{daylight})$  and  $E(\text{night})$  in  $E(\text{day})$ .

The above results show that there are three kinds of negative relations in fuzzy concepts, viz. contradictory negative relation CFC, opposite negative relation OFC and medium negative relation MFC. Therefore, we propose that negations of fuzzy information should include contradictory negation, opposite negation and medium negation.

## 3 Medium Predicate Logic and Infinite-Valued Interpretation

The medium predicate logic MF is a subsystem of the medium logic system ML, ML is a non-classical formal system[8][9]. In this paper, we adopt the formal symbols of MF.

In MF, the connectives  $\neg$ ,  $\bar{\neg}$  and  $\sim$  in the formal language  $L$  denoted the contradictory negation, the opposite negation and the fuzzy negation, respectively. Suppose  $P$  is a unary predicate, for any an individual variable  $x$ , if either  $x$  completely satisfy  $P$  or  $x$  completely does not satisfy  $P$ , then  $P$  is called *distinct predicate* and denoted as  $disP$ . If there is an individual variable  $x$ , which partially satisfy  $P$  and partially does not satisfy  $P$ ,  $P$  is called a *fuzzy predicate* and denoted as  $fuzP$ . In MF,  $\neg P$  represents the contradictory negation predicate of  $P$ ,  $\bar{\neg} P$  represents the opposite negation predicate of  $P$ , and unconditionally acknowledged that for some predicate  $P$ , there is individual variable  $x$  which partially satisfy  $P$  and partially satisfy  $\bar{\neg} P$ , such  $x$  is called a *medium object* and “ $x$  partially satisfies  $P$ ” is denoted as  $\sim P(x)$ .  $\sim P$  represents the fuzzy negation predicate of  $P$ . Moreover  $\neg P = \bar{\neg} P \vee \sim P$ .

In the semantic theories of MF, we consider that infinite-valued semantic interpretation  $\Phi$  for  $L$  [10] is optimum approach to describe the fuzzy concepts and contradictory negation, opposite negation and medium negation.

Let  $\Phi: \langle D, \varphi \rangle$ , where  $D$  is universe of individuals,  $\varphi$  is an infinite-valued evaluation function for formulas in  $L$ .  $\varphi$  is defined as follows.

**Definition 1:** Let  $\Gamma$  be set of formulas in  $L$ ,  $\lambda \in (0, 1)$ . Mapping

$$\varphi: \Gamma \rightarrow [0, 1]$$

is called a infinite-valued evaluation function for formula if satisfies following assignments. For any  $A, B \in \Gamma$ ,

- (1) assigning an object in  $D$  for each symbol of individual constant in formula;
- (2) assigning an mapping from  $D^n$  to  $D$  for each symbol of the n-variable function in formula;
- (3) assigning an mapping from  $D^n$  to  $[0, 1]$  for each symbol of the n-variable predicate in formula, and

(3.1)  $\varphi(A)$  only takes single value in  $[0, 1]$  when  $A$  is an atomic formula;

(3.2)  $\varphi(A) + \varphi(\neg A) = 1$ ,

(3.3)  $\varphi(\neg A) =$

$$\frac{2\lambda - 1}{1 - \lambda} (\varphi(A) - \lambda) + 1 - \lambda, \text{ when } \lambda \in [\frac{1}{2}, 1) \text{ and } \varphi(A) \in (\lambda, 1],$$

$$\frac{2\lambda - 1}{1 - \lambda} \varphi(A) + 1 - \lambda, \text{ when } \lambda \in [\frac{1}{2}, 1) \text{ and } \varphi(A) \in [0, 1 - \lambda),$$

$$\frac{1 - 2\lambda}{\lambda} \varphi(A) + \lambda, \text{ when } \lambda \in (0, \frac{1}{2}] \text{ and } \varphi(A) \in [0, \lambda),$$

$$\frac{1 - 2\lambda}{\lambda} (\varphi(A) + \lambda - 1) + \lambda, \text{ when } \lambda \in (0, \frac{1}{2}] \text{ and } \varphi(A) \in (1 - \lambda, 1],$$

$$1/2, \text{ when } \varphi(A) = 1/2,$$

(3.4)  $\varphi(A \rightarrow B) = \text{Max}(1 - \varphi(A), \varphi(B))$ ,

(3.5)  $\varphi(A \vee B) = \text{Max}(\varphi(A), \varphi(B))$ ,

(3.6)  $\varphi(A \wedge B) = \text{Min}(\varphi(A), \varphi(B))$ ,

(3.7)  $\varphi(\forall x P(x)) = \text{Min}_{x \in D} \{\varphi(P(x))\}$ ,  $\varphi(\exists x P(x)) = \text{Max}_{x \in D} \{\varphi(P(x))\}$ .

Based on MF and  $\Phi$ , we first introduce formal symbols “ $\sim^+$ ” and “ $\sim^-$ ” for the unary fuzzy predicate  $P$ .

$\sim^+P$  denotes that “ $\sim P$  verge on  $P$ ” for the fuzzy negation  $\sim P$  of  $P$ ,

$\sim^-P$  denotes that “ $\sim P$  verge on  $\neg P$ ” for the fuzzy negation  $\sim P$  of  $P$ .

**Definition 2:** Let  $P$  be a unary fuzzy predicate in the definition 1. For the truth-value  $\varphi(\sim^+P(x))$ ,  $\varphi(\sim^-P(x))$  of  $\sim^+P(x)$  and  $\sim^-P(x)$ , suppose  $\varphi(\sim^+P(x))$ ,  $\varphi(\sim^-P(x)) \in \{0, 1, 1/2\}$ . When  $\lambda \in (\frac{1}{2}, 1)$

$$\varphi(\sim^+P(x)) = \begin{cases} 0, & 1 - \lambda \leq \varphi(\sim P(x)) < \frac{1}{2}, \\ \frac{1}{2}, & \varphi(\sim P(x)) = \frac{1}{2}, \\ 1, & \frac{1}{2} < \varphi(\sim P(x)) \leq \lambda. \end{cases}$$

$$\varphi(\sim^-P(x)) = \begin{cases} 0, & \frac{1}{2} < \varphi(\sim P(x)) \leq \lambda, \\ \frac{1}{2}, & \varphi(\sim P(x)) = \frac{1}{2}, \\ 1, & 1 - \lambda \leq \varphi(\sim P(x)) < \frac{1}{2}. \end{cases}$$

*Remark:*  $\varphi(\sim P(x))$  always is belong to  $[1-\lambda, \lambda]$  i.e.  $[1-\lambda, \frac{1}{2}] \cup [\frac{1}{2}, \lambda]$  in the definition 1 and Fig.7 and Fig. 8 when  $\lambda \in (\frac{1}{2}, 1)$ .

## 4 Application in the Assistant Decision Making of Financial Investment

The aim of assistant decision making of financial investment is to assist consumer that deposited superfluous money in the bank or invest in the stock market. Suppose strategies of financial investment lie on the income and savings of investor, and according to following decision rules:

- If investor has not savings, investor should deposit his money in the bank whether how much income.
- If investor has much savings and much income, investor can consider stock investment, which takes a risk and profitable.
- If investor has much savings and little income, investor can consider that majority of money to buy stock, and the minority of money to deposit bank in superfluous income.
- If investor has little savings and much income, he can consider that majority of money to deposit bank, and the minority of money to buy stock in superfluous income.

*Remark:* If investor has little income, we think that he has no superfluous money.

### 4.1 An Example

In real life, people's viewpoints on the *much* (or *little*) *income* and *much* (or *little*) *savings* can be affected by many factors, and difference of areas is main factor.

**Table 1.** The people's viewpoints on the much (or little) income and savings in some areas

views	Area	<i>much income (month)</i>	<i>little income (month)</i>	<i>much savings</i>	<i>little savings</i>
1	Shanghai	$\geq 15,000$	$\leq 2000$	$\geq 200,000$	$\leq 100,000$
2	Pudong, Shanghai	$\geq 20,000$	$\leq 2500$	$\geq 250,000$	$\leq 150,000$
3	Xuhui, Shanghai	$\geq 10,000$	$\leq 2000$	$\geq 200,000$	$\leq 80,000$
2	Nanjing, Jiangsu	$\geq 10,000$	$\leq 1500$	$\geq 200,000$	$\leq 80,000$
3	Wuxi, Jiangsu	$\geq 12,000$	$\leq 1200$	$\geq 150,000$	$\leq 100,000$
4	Suzhou, Jiangsu	$\geq 15,000$	$\leq 1500$	$\geq 150,000$	$\leq 100,000$
5	Hefei, Anhui	$\geq 6,000$	$\leq 1000$	$\geq 100,000$	$\leq 80,000$
6	Fuyang, Anhui	$\geq 5,000$	$\leq 1000$	$\geq 100,000$	$\leq 50,000$
7	Tongning, Anhui	$\geq 4,000$	$\leq 800$	$\geq 100,000$	$\leq 50,000$
8	Jinan, Shandong	$\geq 7,000$	$\leq 1200$	$\geq 150,000$	$\leq 80,000$
9	Yantai, Shandong	$\geq 6,000$	$\leq 1000$	$\geq 120,000$	$\leq 50,000$
10	Weihai, Shandong	$\geq 10,000$	$\leq 1500$	$\geq 150,000$	$\leq 80,000$

For example, we investigated the people's viewpoints on the much (or little) income and much (or little) savings in some areas of China, obtained the result of random sample (Table 1) as follows (money unit: RMB).

In order to integrate investigation data in the same province, we compute the average of investigation data for each province, thus we can get integrative data in each province (Table 2).

**Table 2.** The integrative data table in each province

Area	Viewpoint of much income ( $\pm 500/\text{month}$ )	Viewpoint of little income ( $\pm 100/\text{month}$ )	Viewpoint of much savings ( $\pm 20,000$ )	Viewpoint of little savings ( $\pm 10,000$ )
Shanghai	$\geq 14400$	$\leq 2000$	$\geq 210,000$	$\leq 100,000$
Jiangsu	$\geq 11,000$	$\leq 1340$	$\geq 160,000$	$\leq 82,000$
Anhui	$\geq 5,000$	$\leq 920$	$\geq 100,000$	$\leq 56,000$
Shandong	$\geq 7,000$	$\leq 1100$	$\geq 124,000$	$\leq 68,000$

Based on table 2, we can establish a flexible interval for each data type, the flexible interval reflect veracity of integrative data. It can be see from table 2 that

- the viewpoint of much income is more than 5000 /month at least and 14400/month at most, thus, the corresponding flexible intervals are [4500, 5500] and [13900, 14900] respectively.
- the viewpoint of little income is less than 920 /month at least and 2000 /month at most, thus, the corresponding flexible intervals are [820, 1020] and [1900, 2100] respectively.
- the viewpoint of much savings is more than 100000 at least and 210000 at most, thus, the corresponding flexible intervals are [80000, 120000] and [190000, 230000] respectively.
- the viewpoint of little savings is less than 68000 at least and 100000 at most, thus, the corresponding flexible intervals are [58000, 78000] and [90000, 110000] respectively.

## 4.2 Representation of Fuzzy Predicate and Different Negations in Example

In the above decision rules and example, “*much savings*”, “*much income*”, “*little savings*” and “*little income*” are fuzzy predicates. “*savings*”, “*income*” and “*stock*” are functions. We need to point out:

*fuzzy predicate “little savings” is opposite negation of “much savings”, and “little income” is opposite negation of “much income”.*

In order to differentiate these different fuzzy predicates and their opposite negation in example, we express name of predicate using capital letters, name of function using small letters.

Let  $x$  be an investor. The representations of functions are as follows:

$\text{savings}(x)$ : denotes the savings of individual  $x$ .

$\text{income}(x)$ : denotes the income of  $x$ .

$mfsavings(x)$ : denotes the money which  $x$  need to deposit bank.

$mfstocks(x)$ : denotes the money which  $x$  need to buy stocks.

The representations of fuzzy predicates are as follows:

$MUCH(savings(x))$ : denotes that  $x$  has much savings.

$MUCH(income(x))$ : denotes that  $x$  has much income.

$INVESTMENT(x, stocks)$ : denotes that  $x$  buys stocks.

$INVESTMENT(x, savings)$ : denotes that  $x$  invests money in bank.

$MORE(mfsavings(x), mfstocks(x))$ : denotes that  $x$  deposited moneys in bank more than buy stocks.

According as the meaning of symbols  $\neg$ ,  $\sim$ ,  $\sim^+$  and  $\sim^-$ , also

$\neg MUCH(savings(x))$ : denotes that  $x$  has few savings,

$\sim MUCH(savings(x))$ : denotes that  $x$  has moderate savings,

$\sim^+ MUCH(savings(x))$ : denotes that  $x$  has biggish savings,

$\sim^- MUCH(savings(x))$ : denotes that  $x$  has lesser savings;

$\neg MUCH(income(x))$ : denotes that  $x$  has few income,

$\sim MUCH(income(x))$ : denotes that  $x$  has moderate income,

$\sim^+ MUCH(income(x))$ : denotes that  $x$  has biggish income,

$\sim^- MUCH(income(x))$ : denotes that  $x$  has lesser income,

The representation of decision rules as follows:

- decision rule (1):

$\neg MUCH(savings(x)) \rightarrow INVESTMENT(x, stocks)$ .

- decision rule (2):

$MUCH(savings(x)) \wedge MUCH(income(x)) \rightarrow INVESTMENT(x, stocks)$ .

- decision rule (3):

$\sim^+ MUCH(savings(x)) \wedge (MUCH(income(x)) \vee \sim MUCH(income(x))) \rightarrow (INVESTMENT(x, stocks) \wedge INVESTMENT(x, savings) \wedge MORE(mfstocks(X), mfsavings(X)))$ .

- decision rule (4):

$\sim^- MUCH(savings(x)) \wedge (MUCH(income(x)) \vee \sim MUCH(income(x))) \rightarrow (INVESTMENT(x, stocks) \wedge INVESTMENT(x, savings) \wedge MORE(mfsavings(x), mfstocks(x)))$ .

#### 4.3 Measurement of Truth Valued for Fuzzy Predicate Expressions in Example

In order to actualize the financial investment based on above example, we think that measurement of the truth-valued of fuzzy predicate expressions is focus.

In example,  $x$  is income number or savings number in fuzzy predicate expression  $MUCH(x)$ . It can also be found out from data in table 2 that if an income is high in Shanghai, it must be high in other area. If an income is low in Anhui, it must be low in other area. And savings is also. For the sake of this characteristic of the data, we adopt the Euclidean distance of one dimension, which is expressed as  $d(x, y)$ , viz.  $d(x, y) = |x - y|$ .

According to Definition 1, we can define the truth valued degree  $\varphi(MUCH(x))$  of fuzzy predicate expression  $MUCH(x)$  as follows:

$$\varphi(MUCH(x)) = \begin{cases} 0, & \text{when } x \leq \alpha_F + \varepsilon_F \\ \frac{d(x, \alpha_F + \varepsilon_F)}{d(\alpha_F + \varepsilon_F, \alpha_T - \varepsilon_T)}, & \text{when } \alpha_F + \varepsilon_F < x < \alpha_T - \varepsilon_T \\ 1, & \text{when } x \geq \alpha_T - \varepsilon_T \end{cases}$$

In the same way, the truth-valued degree  $\varphi(\neg MUCH(x))$  of fuzzy predicate expression  $\neg MUCH(x)$  can be given.

$$\varphi(\neg MUCH(x)) = \begin{cases} 0, & \text{when } x \geq \alpha_T - \varepsilon_T \\ \frac{d(x, \alpha_T - \varepsilon_T)}{d(\alpha_F + \varepsilon_F, \alpha_T - \varepsilon_T)}, & \text{when } \alpha_F + \varepsilon_F < x < \alpha_T - \varepsilon_T \\ 1, & \text{when } x \leq \alpha_F + \varepsilon_F \end{cases}$$

*Remark:*  $\alpha_T$  is the integrated datum regarded as the truest one for fuzzy predicate *MUCH* in table 2, and  $\varepsilon_T$  is its flexible extent.  $\alpha_F$  is the integrated datum regarded as the truest one for fuzzy predicate  $\neg MUCH$ , and  $\varepsilon_F$  is its flexible extent. Obviously,

$$\varphi(MUCH(x)) + \varphi(\neg MUCH(x)) = 1$$

which corresponds to the infinite valued semantic interpretation of medium predicate logic.

For the viewpoint of much income, the corresponding integrated datum of Shanghai is the highest, which is 14400. We regard it as the truest datum for *MUCH*, and its flexible extent  $\varepsilon_T$  is 500. For the viewpoint of a little income, the corresponding integrated datum of Anhui is the lowest, which is 920. We regard it as the truest datum for  $\neg MUCH$  or the falsest datum for *MUCH*, and its flexible extent  $\varepsilon_F$  is 100.

So, for income data  $x$ ,  $\varphi(MUCH(x)) =$

$$\begin{cases} 0, & \text{when } x \leq 1020 \\ \frac{d(x, 1020)}{d(1020, 13900)}, & \text{when } 1020 < x < 13900 \\ 1, & \text{when } x \geq 13900 \end{cases}$$

and  $\varphi(\neg MUCH(x)) = 1 - \varphi(MUCH(x))$ .

In the same way, for savings data  $x$ ,  $\varphi(MUCH(x)) =$

$$\begin{cases} 0, & \text{when } x \leq 6.6 \\ \frac{d(x, 6.6)}{d(6.6, 19)}, & \text{when } 6.6 < x < 19 \\ 1, & \text{when } x \geq 19 \end{cases}$$

and  $\varphi(\neg MUCH(x)) = 1 - \varphi(MUCH(x))$ .

And then, for any investigation data, we can compute truth-valued degree of fuzzy predicate expression.

#### 4.4 Reasoning of Fuzzy Information and Different Negations in the Example

The general form of a fuzzy production rule is:

$$Q \leftarrow P, CF, \tau \text{ or if } P \text{ then } Q, (CF, \tau)$$

which in  $P$  represents a set of premises and conditions,  $Q$  represents some results or actions. Both premises  $P$  and results  $Q$  can be fuzzy.  $CF$  ( $0 \leq CF \leq 1$ ) is called confidence level of rule,  $\tau$  ( $0 \leq \tau \leq 1$ ) is threshold value. This rule means that if premises  $P$  are contented in a way, results  $Q$  can be educed (or carry out actions in a way) in confidence level  $CF$ .

##### 4.4.1 Establishment of Threshold Value in Fuzzy Production Rule

Because this paper educes truth values of logic formulae according to infinite-valued model of MF, it must be related to  $\lambda$  in model.

For example, in the demonstration in part 3, suppose premise is a fuzzy logic expression “ $MUCH(\text{savings}(x))$ ”. If  $x$  lives in Jiangsu, so  $\lambda = 0.815$ . For people  $x$  investigated, truth value of  $MUCH(\text{savings}(x))$  must be larger than 0.815 when we say that  $x$  has much savings. So, characteristic of  $\lambda$  in the infinite valued model of MF is similar to threshold  $\tau$  in a way. When confidence level of rule is larger than value of  $\lambda$ , the value of  $\lambda$  can be regarded as threshold.

##### 4.4.2 Realization of Reasoning

We continue to discuss about the demonstration in part 3, for rule (1):

$$\neg MUCH(\text{savings}(x)) \rightarrow INVESTMENT(x, \text{stocks})$$

where  $INVESTMENT$  is a distinct predicate, but for  $x$  investigated, whether it is true or false depends on  $\neg MUCH(\text{savings}(x))$ . So this medium logic formula can be regarded as a production rule. The confidence level can also be fetched by random investigation and statistic. Here suppose the confidence level  $CF = 0.9 > \lambda$ . If people  $x$  investigated lives in Jiangsu, so threshold  $\tau$  can be endowed with value of  $\lambda$ , viz. 0.815. Therefore, the rule can be expressed as:

$$INVESTMENT(x, \text{stocks}) \leftarrow \neg MUCH(\text{savings}(x)), 0.9, 0.815.$$

In the same way, rest rules can also be converted into fuzzy production rules. Here suppose confidence levels of rule(1)-rule(4) are all 0.9 for convenience.

For instance, one people lives in Wuxi, Jiangsu. His income is 5000 yuan a month and his family savings is 120 thousand. How does he invest?

First, the people lives in Jiangsu, so ascertain the value of  $\lambda_1$ , viz.  $\lambda_1 = 0.875$ , when considering about income and the value of  $\lambda_2$ , viz.  $\lambda_2 = 0.815$ , when considering about savings. Obviously, both  $\lambda_1$  and  $\lambda_2$  are smaller than  $CF$ .

According to corresponding function of truth value and infinite-valued model of MF, educe results as follows:

$$\varphi(MUCH(\text{savings}(x))) = \frac{d(12, 6.6)}{d(6.6, 19)} = 0.435;$$

$$\varphi(\neg MUCH(\text{savings}(x))) = 0.565;$$

$$\varphi(MUCH(\text{income}(x))) = \frac{d(5000, 1020)}{d(1020, 13900)} = 0.309;$$

$$\varphi(\neg MUCH(\text{income}(x))) = 0.691.$$

For rule (1):

$$\begin{aligned} INVESTMENT(x, \text{stocks}) \leftarrow \\ \neg MUCH(\text{savings}(x)), 0.9, 0.815; \end{aligned}$$

Prescribe that  $t = \min\{0.565, 0.9\} = 0.565 < 0.815$ , so the rule can not act.

For rule (2):

$$INVESTMENT(x, \text{stocks}) \leftarrow MUCH(\text{savings}(x)) \wedge MUCH(\text{income}(x)), 0.9, 0.875;$$

*Remark:* here premises come down to savings and income, so threshold value  $\tau = \max\{0.875, 0.815\} = 0.875$ .

The truth value of premises:

$$t_1 = \min\{\varphi(MUCH(\text{savings}(x))), \varphi(MUCH(\text{income}(x)))\} = \min\{0.435, 0.309\} = 0.309,$$

$$t = \min\{0.309, 0.9\} = 0.309 < 0.875.$$

So, the rule can not act.

For rule (3):

$$INVESTMENT(x, \text{stocks}) \wedge INVESTMENT(x, \text{savings}) \wedge MORE(mfstocks(x), mfsavings(x)) \leftarrow \sim^+ MUCH(\text{savings}(x)) \wedge MUCH(\text{income}(x)) \vee \sim MUCH(\text{income}(x)), 0.9, 0.875$$

Because  $1 - 0.815 < \varphi(MUCH(\text{savings}(x))) < 0.5$ , according to definition 3,

$$\varphi(\sim^+ MUCH(\text{savings}(x))) = 0.$$

Obviously, the rule can not act.

For rule (4):

$$INVESTMENT(x, \text{stocks}) \wedge INVESTMENT(x, \text{savings}) \wedge MORE(mfsavings(x), mfstocks(x)) \leftarrow \sim^- MUCH(\text{savings}(x)) \wedge MUCH(\text{income}(x)) \vee \sim MUCH(\text{income}(x)), 0.9, 0.875$$

Because of  $1 - 0.815 < \varphi(MUCH(\text{savings}(x))) < 0.5$ , according to definition 3,  $\varphi(\sim^- MUCH(\text{savings}(x))) = 1$ , and  $1 - 0.875 < \varphi(MUCH(\text{income}(x))) < 0.875$ , so  $\varphi(\sim^- MUCH(\text{income}(x))) = 1$ .

So, truth value of premises is 1.

$$t = 0.9 > 0.875.$$

So, the rule can act.

In conclusion, this people can adopt rule (4) for financing.

## Acknowledgment

This work was supported by the National Natural Science Foundation of China (60973156) and the Program for Innovative Research Team of Jiangnan University.

## References

1. Wagner, G.: *Vivid Logic: Knowledge-Based Reasoning with Two Kinds of Negation*. Springer, New York (1994)
2. Wagner, G.: Partial Logics with Two Kinds of Negation as a Foundation for Knowledge-Based Reasoning. In: Gabbay, D., Wansing, H. (eds.) *What Is Negation?* Oxford University Press, Oxford (1999)
3. Wagner, G.: Web Rules Need Two Kinds of Negation. In: Bry, F., Henze, N., Małuszyński, J. (eds.) *PPSWR 2003. LNCS*, vol. 2901, pp. 33–50. Springer, Heidelberg (2003)
4. Kaneiwa, K.: Negations in Description Logic-Contraries, Contradicories, and Subcontraries. In: Dau, F., Mugnier, M.L., Stumme, G. (eds.) *Contributions to ICCS*, pp. 66–79 (2005)
5. Ferré, S.: Negation, Opposition, and Possibility in Logical Concept Analysis. In: Missaoui, R., Schmidt, J. (eds.) *ICFCA 2006. LNCS (LNAI)*, vol. 3874, pp. 130–145. Springer, Heidelberg (2006)
6. Pan, Z., Zhu, W.: A New Cognition and Processing on Contradictory Knowledge. In: Proceedings of 2006 International Conference on Machine Learning and Cybernetics, vols. 1–7, pp. 1532–1537. IEEE Press, New York (2006)
7. Pan, Z.: A Logic Description on Different Negation Relation in Knowledge. In: Huang, D.-S., Wunsch II, D.C., Levine, D.S., Jo, K.-H. (eds.) *ICIC 2008. LNCS (LNAI)*, vol. 5227, pp. 815–823. Springer, Heidelberg (2008)
8. Wujia, Z., Xian, X.: Proposition Calculus System of Medium Logic (I). *Nature Journal* 8(4), 315–316 (1985) (in Chinese)
9. Xian, X., Wujia, Z.: Predicate Calculus System of Medium Logic (I). *Nature Journal* 8(7), 540–542 (1985) (in Chinese)
10. Pan, Z., Zhu, W.: An Interpretation of Infinite Valued for Medium Proposition Logic. In: Proceedings of IEEE-Third International Conference on Machine Learning and Cybernetics, pp. 2495–2499. IEEE Press, New York (2004)

# Using Fuzzy Neural Network to Explore the Effect of Internet on Quality of Life

Jui-Chen Huang\*

Department of Health Business Administration, HUNGKUANG University, Taiwan, ROC  
34 Chung-Chie Rd, Sha Lu, Taichung, 443, Taiwan, ROC  
juichen@ms17.hinet.net

**Abstract.** The purpose of this research was to discuss users' opinions on the use of Internet, its effect on their quality of life, and the relationship between perceived use and willingness of use. This research adopted Fuzzy Neural Network (FNN) to propose a feasible and effective analysis method, which is different from the previous method. The results indicate that the FNN method is better than the regression analysis. In addition, the results showed the acquisition of information has the greatest effects on the perceived effects of Internet on their quality of life, followed by the health promotion, social relationship, and overall living quality. This finding may serve as a reference to future studies.

**Keywords:** Internet; Fuzzy Neural Network (FNN); Quality of life (QoL); Regression.

## 1 Introduction

Globalization and rapid advances in information technology offer us vast, unprecedented opportunities to improve life quality. Extensive qualitative and quantitative evidences also supported the Internet's potential that home Internet access enabled the informationally disadvantaged or low-income families to experience powerful emotional and psychological transformations in identity (self-perception), a new sense of confidence, and social standing or development of personal relationships on the Internet [1-4]. Besides, health information can increase individuals' knowledge of their disease and its treatments, reduce distress and anxiety and help individuals make informed decisions regarding their treatment. Increased access to the Internet has provided patients with a new source of information and the rapid growth of the Internet has triggered an information revolution [5-8].

Quality of life (QoL) is conceptualized as a generic, multidimensional construct that describes an individual's subjective perception of his or her physical and psychological health, as well as his or her social functioning, environment, and general life status [9-10]. The World Health Organization (WHO) defines QoL as an individual's perception of their position in life within the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards and concerns [11].

---

\* Corresponding author.

In 1991, the WHO initiated a cross-cultural project to develop a quality-of-life (QoL) questionnaire (WHOQOL); soon after this, the clinically applicable short form was developed and named WHOQOL-BREF, followed by a Taiwanese version (WHOQOL-BREF [TW]) [12]. The WHOQOL is a generic quality of life instrument that was designed to be applicable to people living under different circumstances, conditions and cultures [13-14]. The WHOQOL-BREF is a generic, transcultural and short instrument that represents good psychometric characteristics in general and clinical settings [13]. The WHOQOL-BREF (TW) is reliable and valid from various validation studies [15].

On the other hand, little research has been carried out to further explore the potential relationship between the Internet and QoL. For the time being, both theoretical and empirical researches on the impact of the Internet are still in their infancy [4].

Furthermore, most of previous studies tended to adopt simple statistic methods, regression-based multivariate analysis, or path analysis-based variable causation identification, all of which are linear models. Thus, few studies employed nonlinear structural models (such as the nonlinear neural network model). Therefore, this research adopted a Fuzzy Neural Network (FNN) and compared with the traditional regression analysis to propose a feasible and effective analysis method, which is different from the previous method.

The purpose of this research was to discuss users' opinions on the use of Internet, its effect on their quality of life, and the relationship between perceived use and willingness of use. The results are provided as reference for future studies, developers, and policy-makers.

## 2 Methodology

### 2.1 Data Collection

This study interviewed 369 samples (>15 years old) in Taiwan. A total of 369 valid copies of a questionnaire were obtained with females accounting for 53% of the respondents. Most of them were age group of 35-44 (amounting to 34%), followed by the groups of 45-54 and 55-64 (amounting to 28% and 27% respectively). In terms of the educational level, 53% of the subjects had completed university and graduate schools, and the average monthly income ranged between NT\$50,000 and NT\$80,000 (1 USD ≈ NT\$32.65).

### 2.2 Measures of the Constructs

This multidimensional instrument, the WHOQOL-100 [14], reflects the view that QoL is a broad-ranging concept that incorporates subjectively experienced QoL. It is a broad-ranging concept incorporating in a complex way the person's physical health, psychological state, level of independence, social relationships, personal beliefs and their relationship to salient features of the environment. The WHOQOL-100 was developed through an international collaboration of 15 culturally diverse countries. It has been expanded over the years to include more than 40 different language versions [16]. The WHOQOL-BREF, an abbreviated 26-item version of the WHOQOL-100, has been demonstrated to be a valid and reliable brief assessment of QoL.

The Taiwan version of the WHOQOL-BREF [17] contains the 26 original items of the WHOQOL-BREF, plus two national items for Taiwan.

This research was based on domains covered by the Taiwanese version of WHO-QOL-BREF. The factors used to measure perceived effects of the use of Internet on QoL are divided into 7 items: health promotion, safety, accessibility of medical care services, overall living quality, financial burden, social relationship, and acquisition of information. All evaluation items employ a five-point Likert-type scale for measurement, where 1, 2, 3, 4, and 5 indicate “strongly disagree,” “disagree,” “fair,” “agree,” and “strongly agree,” respectively.

### **2.3 Data Analysis Methods**

This research adopted a Fuzzy Neural Network (FNN) and compared with the traditional regression analysis to propose a feasible and effective analysis method, which is different from the previous method. The Artificial Neural Networks (ANN) and FNN are described as follows:

ANN was first introduced as a mathematical aid by McCulloch and Pitts (1943) [18]. It is one of recent artificial intelligence techniques that has gained widely acceptance beginning from the 1990s [19]. Artificial neural networks (ANN) are defined as complex systems created through connecting artificial neurons, which were developed with inspiration from neurons in human brain, with different connection geometries. An ANN is a computational system, comprising software and hardware. ANN has been employed to various problems because of their fascinating features of learning, fast computation and ease of implementation [20-21]. An activation function can be linear or nonlinear form depending on applications. It can be seen as a legitimate part of statistics that fits snugly in the niche between parametric and non-parametric methods [22]. ANN is a collective system with massively parallel interconnections of simple neurons, and consisted of many non-linear computational elements, called nodes, which are interconnected through direct links. One or more input values are combined into a single value, and transformed into an output value. It can be used in applications where a model of a system is required based on an input set of training data. During the iterative process of training, the error measurements between the desired output and the actual output (produced by the ANN) are very crucial. The adjustment of the synaptic weights is performed aiming in reducing these error values. ANN has been widely applied to examine the complex relationship between input variables and output variables [23]. The applications of ANN range from signal processing in communications to pattern recognition in business, engineering and medicine [19].

Although ANN is quite powerful for modeling various real world problems, it also has its shortcomings. If the input data are less accurate or ambiguous, ANN would be struggling to handle them and a fuzzy system might be a better option.

Fuzzy inference systems (FIS) are powerful tools for the simulation of nonlinear behaviors with the help of fuzzy logic (FL) and linguistic fuzzy rules [24]. The theory of fuzzy set, founded by Zadeh [25], is a set of mathematics theory based on linguistic analysis and intelligent behavior. Fuzzy logic is a superset of boolean logic. It can be used as a basis for constructing a set of fuzzy “If–Then” rules with appropriate membership functions in order to generate the preliminary stipulated input–output pairs.

On the other hand, the fuzzy control is very difficult to design and adjust automatically. So it is necessary to design the Fuzzy Neural Network (FNN) model, it can use the both advantages [26-27]. FNN combines fuzzy logic control (FLC) with artificial neural network (ANN) and realizes fuzzy logic by fuzzy neural network. With the expert knowledge of fuzzy and the learning capabilities of neural network, the controller could save the time of searching space to achieving optimal solution. For the FNN approaches, it should been determined by trial-and-error in advance for the reason that it is difficult to consider the balance between the rule number and the desired performance.

In this study, the input variables are normalized to Z scores to be within the interval of 1.96 times the standard deviation for eliminating the affection of some abnormal values; while the output variable is scaled to 0.2 to 0.8. Two third of the total samples (246 samples) were taken as the training data, and the one third of the samples (123 samples) were taken as the testing data.

Lastly, this study used t-test to discuss the effect of gender and marital status on the use of Internet and its influence on quality of life. One-Way ANOVA, Kruskal-Wallis test, and LSD post-hoc comparison are adopted to discuss the effect of monthly household income on the use of Internet and its influence on quality of life.

### 3 Results and Discussion

#### 3.1 Effect of Use of Internet on Quality of Life

Table 1 shows the effects of the perceived use of Internet on quality of life. As seen, the acquisition of information (mean and standard deviations (S.D.) are 4.17 and 0.63, respectively) has the greatest effect on QoL, followed by the health promotion (mean and S.D. are 4.08 and 0.73, respectively), social relationship (mean and S.D. are 4.04 and 0.99, respectively), and overall living quality (mean and S.D. are 4.00 and 0.62, respectively).

**Table 1.** Effect of use of Internet on quality of life

Categories	Mean	Standard deviation (S.D.)
Health promotion	4.08	0.73
Safety	3.66	0.84
Accessibility of medical care services	3.95	0.73
Overall living quality	4.00	0.62
Financial burden	2.91	0.98
Social relationship	4.04	0.99
Acquisition of information	4.17	0.63

#### 3.2 Comparison between the Fuzzy Neural Network and Traditional Regression Method

Table 2 shows the perceived use of Internet on living quality and the relationship between willingness to use. This research adopted a Fuzzy Neural Network (FNN) and compared with the traditional regression analysis.

As presented in the FNN, the training and testing data RMSE (root mean square error) are 0.076 and 0.258, respectively. As shown, the errors are acceptable. On the other hand, the regression model also shows the results of the RMSE. The training and testing data RMSE are 0.750 and 0.781, respectively.

In summation, this comparison clearly shows that the FNN method is much better than regression method. In addition, it is evident that this non-traditional modeling method is a useful approach to evaluate the effect of Internet. By using the FNN method, the data do not need to be restricted to linear or nonlinear forms.

**Table 2.** Comparison between the Fuzzy Neural Network (FNN) and regression method

Method	Training data RMSE	Testing data RMSE
FNN	0.076	0.258
Regression	0.750	0.781

### 3.3 Effect of Demographic Characteristics on Quality of Life

This study used t-test to discuss the effect of gender and marital status on the use of Internet and its influence on quality of life. One-Way ANOVA, Kruskal-Wallis test, and LSD post-hoc comparison are adopted to discuss the effect of monthly household income on the use of Internet and its influence on quality of life.

#### 3.3.1 Effect of Gender on Quality of Life

Table 3 shows the effect of gender on the use of Internet and its influence on quality of life. As seen, among the 7 categories of quality of life, accessibility of medical care services, social relationship, and acquisition of information did not reach a significant difference statistically. Thus, the categories that reached a significant difference ( $p<0.05$ ) included health promotion, safety, overall living quality, and financial burden.

**Table 3.** T test summary of the effects of gender on quality of life

Categories	Males		Females		t value
	Mean	S.D.	Mean	S.D.	
Health promotion	4.33	0.77	3.86	0.61	6.39*
Safety	3.76	0.66	3.57	0.97	2.18*
Accessibility of medical care services	4.02	0.71	3.88	0.75	1.84
Overall living quality	4.07	0.52	3.94	0.69	1.98*
Financial burden	3.14	0.89	2.69	0.99	4.53*
Social relationship	4.02	0.99	4.06	0.98	-0.37
Acquisition of information	4.22	0.63	4.13	0.64	1.39

\* $p<0.05$ .

Their t values are 6.39, 2.18, 1.98, and 4.53, respectively. Also, among the 4 items that reached a significant difference, the values of males were higher than those of females.

### 3.3.2 Effect of Marital Status on Quality of Life

Table 4 shows the effect of marital status on the use of Internet and its influence on quality of life. As seen, among the 7 categories of quality of life, only health promotion reached a significant difference ( $p<0.05$ ), which t value was 3.62; and the values of the married subjects were higher than those of other marital status. Other items did not reach a significant difference.

**Table 4.** T test summary of the effects of marital status on quality of life

Categories	Married		Other		t value
	Mean	S.D.	Mean	S.D.	
Health promotion	4.22	0.76	3.95	0.67	3.62*
Safety	3.74	0.89	3.58	0.79	1.76
Accessibility of medical care services	3.91	0.79	3.99	0.68	-1.06
Overall living quality	3.97	0.63	4.04	0.60	-1.09
Financial burden	2.92	0.99	2.89	0.98	0.27
Social relationship	4.09	0.99	3.98	0.99	1.03
Acquisition of information	4.13	0.64	4.21	0.63	-1.25

\* $p<0.05$ .

**Table 5.** One-Way ANOVA and Kruskal-Wallis test summary of the effects of monthly household income on quality of life

Categories	Monthly household income (Mean)					F value	Kruskal- Wallis Chi- Square	LSD post-hoc comparison
	(1)	(2)	(3)	(4)	(5)			
Health promotion	3.50	4.00	4.15	4.12	3.67	2.55*		(3)>(1)*, (4)>(1)*
Safety	3.13	3.71	3.62	3.72	4.00	1.34		
Accessibility of medical care services	3.75	4.01	3.93	3.93	4.00	0.37		
Overall living quality	4.00	4.00	3.98	4.07	4.17	0.33		
Financial burden	3.38	2.74	2.92	3.11	2.33	2.34		
Social relationship	3.38	4.02	4.13	3.90	3.67	1.75		
Acquisition of information	4.00	4.29	4.13	4.15	4.17		5.46	

Note 1: \* $p<0.05$ .

Note 2: (1) (< NT\$20000), (2) (NT\$20001~50000), (3) (NT\$50001~80000), (4) (NT\$80001~120000), (5) (>NT\$120001) (1 USD ≈ NT\$32.65).

### 3.3.3 Effect of Monthly Household Income on Quality of Life

This study conducted One-Way ANOVA on the 7 categories of quality of life with the monthly household income, in order to understand the effect of monthly household income on the use of Internet and its influence on quality of life. In the homogeneity test for variance, if  $p < 0.05$ , it means that the population variance represented by the samples was not equal, thus Kruskal-Wallis test is adopted. As shown in Table 5, only health promotion and the influence of Internet on quality of life reached a significant difference ( $F$  value is 2.55,  $p=0.039$ ). Based on the LSD post-hoc comparison, the value of the monthly household income of NT\$50,001~80,000 and NT\$80,001~120,000 is significantly higher than that of < NT\$20000. In other words, subjects with higher monthly household income perceived higher influence of the Internet on health promotion than those with lower monthly household income.

## 4 Conclusion

The purpose of this research was to discuss users' opinions on the use of Internet, its effect on their quality of life, and the relationship between perceived use and willingness of use. The results are provided as reference for future studies. This research adopted Fuzzy Neural Network (FNN) and compared with the traditional regression analysis to propose a feasible and effective analysis method, which is different from the previous method.

The results showed the acquisition of information has the greatest effects on the perceived effects of Internet on their quality of life, followed by the health promotion, social relationship, and overall living quality. Based on the subjects' opinions on the influence of Internet on quality of life, it is suggested that information acquisition could be enhanced; in other words, allowing individuals to acquire more information through the Internet could effectively enhance the quality of life. Also, the use of Internet could also enhance the information acquisition on health related knowledge, thus promoting individual health. The subjects also suggested that using the Internet could enhance interpersonal interaction, social support, and independent living opportunities, thus improving the overall living quality.

In terms of the effects of demographic characteristics on the influence of the Internet on quality of life, it is found that male subjects' rating of health promotion, safety, overall living quality, and financial burden is higher than females'. Married subjects' rating of health promotion is higher than those of other marital status. Subjects with higher monthly household income perceived higher influence of Internet on health promotion than those with lower monthly household income.

Moreover, the result indicates that the FNN methods of ANN are better than the regression analysis. It is a feasible and effective analysis method. This finding may serve as a reference to future studies.

## References

1. Bier, M., Gallo, M.: Personal empowerment in the study of home Internet use by low-income families. *Journal of Research on Computing in Education* 30(2), 107–121 (1997)
2. Anderson, B., Tracey, K.: Digital living: the impact (or otherwise) of the Internet on everyday life. *American Behavioral Scientist* 45(3), 456–475 (2001)

3. Henderson, C.: How the Internet is changing our lives. *Futurist* 35(4), 38–45 (2001)
4. Leung, L., Lee, P.S.N.: Multiple determinants of life quality: The roles of Internet activities, use of new media, social support, and leisure activities. *Telematics and Informatics* 22(3), 161–180 (2005)
5. Michie, S., Rosebert, C., Heaversedge, J., Madden, S., Parbhoo, S.: The effects of different kinds of information on women attending an out-patient breast clinic. *Psychol. Health Med.* 1, 285–296 (1996)
6. Jadad, A., Gagliardi, A.: Rating health information on the Internet: navigating to knowledge or to Babel? *J. Am. Med. Assoc.* 279, 611–614 (1998)
7. Humphris, G.M., Duncalf, M., Holt, D., Field, E.A.: The experimental evaluation of an oral cancer information leaflet. *Oral Oncol.* 35, 575–582 (1999)
8. Davison, B.J., Kirk, P., Degner, L.F., Hassard, T.H.: Information and patient participation in screening for prostate cancer. *Patient Educ. Couns.* 37, 255–263 (1999)
9. Jang, Y., Lin, H.C., Wang, Y., Wu, Y.H.: A validity study of the WHOQOL BREF assessment in persons with traumatic spinal cord injury. *Arch. Phys. Med. Rehabil.* 85, 1890–1895 (2004)
10. Kuehner, C., Buerger, C.: Determinants of subjective quality of life in depressed patients: the role of self-esteem, response styles, and social support. *J. Affect. Disord.* 86, 205–213 (2005)
11. The WHOQOL Group, The World Health Organization Quality of life assessment (WHOQOL): Position paper from the World Health Organization. *Soc. Sci. Med.* 41, 1403 (1995)
12. Yang, S.C., Kuo, P.W., Wang, J.D., Lin, M.I., Su, S.: Quality of Life and Its Determinants of Hemodialysis Patients in Taiwan Measured With WHOQOL-BREF(TW). *American Journal of Kidney Diseases* 46(4), 635–641 (2005)
13. The WHOQOL Group, Development of the World Health Organization WHOQOL-bref. Quality of life assessment instrument. *Psychol. Med.* 28, 551–558 (1998a)
14. The WHOQOL Group, The World Health Organization Quality of Life Assessment (WHOQOL): development and general psychometric properties. *Soc. Sci. Med.* 46, 1569–1585 (1998b)
15. Yao, G., Wang, J.D., Chung, C.W.: Cultural Adaptation of the WHOQOL Questionnaire for Taiwan. *Journal of the Formosan Medical Association* 106(7), 592–597 (2007)
16. The WHOQOL-Taiwan Group, Development and Manual of the Taiwanese Version of WHOQOL, 2nd edn. The WHOQOL-Taiwan Group, Taipei (2005) (Chinese)
17. Yao, G., Chung, C.W., Yu, C.F., Wang, J.D.: Development and verification of validity and reliability of the WHOQOL-BREF Taiwan version. *J. Formos. Med. Assoc.* 101, 342–351 (2002)
18. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133 (1943)
19. Razi, M.A., Athappilly, K.: A comparative analysis of neural network (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Application* 29(1), 65–74 (2005)
20. Haykin, S.: *Neural Networks: A Comprehensive Foundation*. Macmillan College Publishing Company, New York (1994)
21. Maren, A., Harston, C., Pap, R.: *Handbook of Neural Computing Applications*. Academic Press, London (1990)
22. Aldrich, C.: *Exploratory Analysis of Metallurgical Process Data with Neural Networks and Related Methods*, p. 5. Elsevier, Amsterdam (2002)

23. Nelson, M.M., Illingworth, W.T.: Practical guide to neural nets. Addison Wesley Publishing Company, USA (1994)
24. İnan, G., Göktepe, A.B., Ramyar, K., Sezer, A.: Prediction of sulfate expansion of PC mortar using adaptive neuro-fuzzy methodology. *Build. Environ.* 42, 1264–1269 (2007)
25. Zadeh, L.: Fuzzy sets. *Information Control* 8, 338–353 (1965)
26. Bongards, M.: Improving the efficiency of a wastewater treatment plant by fuzzy control and neural networks. *Water Science and Technology* 3(11), 189–196 (2001)
27. Chen, J.C., Chang, N.B.: Mining the fuzzy control rules of aeration in a submerged biofilm wastewater treatment process. *Engineering Applications of Artificial Intelligence* (2007)

# The Power Load Forecasting by Kernel PCA

Fang-Tsung Liu<sup>1</sup>, Chiung-Hsing Chen<sup>1</sup>, Shang-Jen Chuang, and Ting-Chia Ou<sup>2</sup>

<sup>1</sup> Department of Electronic Communication Engineering, National Kaohsiung Marine University, Kaohsiung 81157, Taiwan, R.O.C.

chiung@mail.nkmu.edu.tw

<sup>2</sup> Institute of Nuclear Energy Research, Atomic Energy Council, Taoyuan 32546, Taiwan, R.O.C.

tcou@iner.gov.tw

**Abstract.** We use one year's subset to train the Support Vector Machines (SVM) and the next year's data was used for testing with Kernel Principal Components Analysis (KPCA). This is clearly not optimal for a non-stationary time series such as we have here nevertheless the MAPE of peak load data set with back-propagation neural network [Chuang et al., 1998] is 3.0 and Support Vector Machine is 2.6.

**Keywords:** Support Vector Machines (SVM), Kernel Principal Components Analysis (KPCA).

## 1 Introduction

Support vector machines (SVM) can be used for pattern classification and nonlinear regression. The main idea of a support vector machine is to construct a hyper plane as the decision surface in such a way that the margin of separation between positive and negative examples is maximized. Support vector machines are an approximate implementation of the method of structural risk minimization. This induction principle is based on the fact that the error rate of a learning machine on test data is bound by the sum of the training-error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension.

In many real world applications appropriate pre-processing transformations of high dimensional input space can increase the overall performance of algorithms. In general, there exist some correlations among input variables; thus dimensionality reduction or so-called feature extraction allows us to restrict the entire input space to a sub-space of lower dimensionality.

A notion that is central to the construction of the support vector learning algorithm is the inner-product kernel between a “support vector”  $x_i$  and the vector  $x$  drawn from the input space. The support vectors consist of a small subset of the training data extracted by the algorithm. Depending on how this inner-product kernel is generated, we may construct different learning machine characterized by nonlinear decision surfaces.

## 2 Support Vector Machines

### 2.1 Kernel Methods

Kernel Methods use a mapping  $\varphi$  of an input vector  $x$  into a high-dimensional feature space in which we perform some linear operation. Performing the linear operation in feature space is equivalent to performing a nonlinear operation in input space. Consider first regression: let

$$\omega \cdot \varphi \quad \text{ith } \varphi \in \mathbb{R}^n \quad (1)$$

where  $x$  is an input vector,  $\omega$  is an adjustable weight vector, and  $b$  is a bias term. Then linear regression in high dimensional space corresponds to nonlinear regression in the low dimensional input space. The dot product  $\omega \cdot \varphi(x)$  would have to be computed in this high dimensional space, which is usually intractable. Since  $\varphi$  is fixed, we determine  $\omega$  from the data by minimizing the sum of the empirical risk  $R_{\text{emp}}[f]$  and a complexity term  $\|\omega\|^2$ , which enforces flatness in feature space

$$R_{\text{reg}}[f] = R_{\text{emp}}[f] + \lambda \|\omega\|^2 = \sum_{i=1}^n C(f(x_i) - y_i) + \lambda \|\omega\|^2 \quad (2)$$

where  $n$  denotes the sample size,  $C(\cdot)$  is a cost function and  $\lambda$  is a regularization constant. This idea is clearly related to the regularisation which we discussed when dealing with Radial Basis Networks.

For a large set of cost functions, equation (2) can be minimized by solving a quadratic programming problem, which is uniquely solvable. It can be shown that the vector  $\omega$  can be written in terms of data points in an easily calculable form

$$\omega = \sum_{i=1}^n (\alpha_i - \alpha'_i) \varphi(x_i) \quad (3)$$

with  $\alpha_i$  and  $\alpha'_i$  being the solution of the aforementioned quadratic programming problem. Taking (3) and (1) into account, we are able to rewrite the whole problem in term of dot products in the low dimensional input space.

$$\begin{aligned} f(x) &= \sum_{i=1}^n (\alpha_i - \alpha'_i) (\varphi(x_i) \cdot \varphi(x)) + b \\ &= \sum_{i=1}^n (\alpha_i - \alpha'_i) K(x_i, x) + b \end{aligned} \quad (4)$$

Where the kernel function  $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ . It can be shown that any symmetric kernel function  $K$  satisfying Mercer's [Scholkopf et al., 1999] condition corresponds

to a dot product in some feature space. Mercer's condition may be approximately viewed as stating that the kernel produced must be positive definite.

The cost function we use is the  $\varepsilon$ -insensitive loss function, which is optimal since we have previously used in this thesis and elsewhere [Chuang et al., 1998] Mean Absolute Percentage Error to define the goodness of our predictors. This measure is more robust than a least-squares estimator which is sensitive to presence of outliers. The cost function has the form

$$L_{\varepsilon}(d, y) = \begin{cases} |d - y| - \varepsilon & \text{for } |d - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where  $d$  is the desire response and  $y$  is the estimator output and  $\varepsilon$  is a prescribed parameter. We use one year's subset to train the Support Vector Machines and the next year's data was used for testing. This is clearly not optimal for a non-stationary time series such as we have here nevertheless the MAPE of peak load data set with back-propagation neural network [Chuang et al., 1998] is 3.0 and Support Vector Machine is 2.6. However it should be noted that this is at the expense of an intensive search for the optimal parameter set for  $C$  and  $\varepsilon$ . Also kernel methods are very dependent on number of data samples.

## 2.2 KPCA

However Kernel methods may also be used for unsupervised investigations of data sets. Thus we perform Kernel Principal Components Analysis (KPCA) [Schölkopf et al., 1998] [Smola et al., 1998] on the data set. With KPCA, PCA is performed in the feature space that is nonlinearly related to the input space. The operation of Kernel PCA may be described as follows; given the training examples  $\{x_i\}_{i=1}^N$ , computer the kernel matrix  $K = \{K(x_i, x_j)\}$ , where  $K(x_i, x_j) = T(x_i)^T T(x_j)$ . The eigenvalue problem  $K \mathbf{v} = \lambda \mathbf{v}$  is then solved where  $\lambda$  is the eigenvalue of the kernel matrix and  $\mathbf{v}$  is the associated eigenvector. The eigenvectors are then normalized  $T_k \mathbf{v}_k = 1/k$  ( $k=1, 2, \dots, p$ ) where  $p$  is the smallest nonzero eigenvalue of matrix  $K$ . This is to ensure that the corresponding eigenvectors in feature space are of unit length.

For inner-product kernels defined in accordance with Mercer's theorem, we are basically performing ordinary PCA in a  $m_1$ -dimensional feature space, where the dimension  $m_1$  is a design parameter. In particular, kernel PCA is linear in the feature space, but nonlinear in the input space. As such, it can be applied to all those domains where ordinary PCA has been used for feature extraction or data reduction, in which nonlinear extensions would make sense.

Power load data for the period of 1992 to 1996 from Taiwan Power Company Ltd.(Taiwan) has been used in these experiments. In the last chapter, we performed an Exploratory Projection Pursuit on this data set and showed that much of the structure in this data is day-of-the-week dependent. Therefore, in this experiment, we split the data set into seven disjoint subset (Monday, Tuesday...); each subset includes Max. and Min. temperature of the previous day, peak of energy on the previous day, peak of energy on the equivalent day one week before and the time on the previous day when the peak was reached.

This section shows that sample Principal Component Analysis (PCA) may be performed on the samples of a data set in a particular way which will be useful in the performance of PCA in the nonlinear feature space.

PCA finds the eigenvectors and corresponding eigenvalues of the covariance matrix of a data set. Let  $\chi = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$  be iid (independent, identically distributed) samples drawn from a data source. If each  $\mathbf{x}_i$  is  $n$ -dimensional,  $\exists$  at most  $n$  eigenvalues/eigenvectors. Let  $C$  be the covariance matrix of the data set; then  $C$  is  $n \times n$ . Then the eigenvectors,  $e_i$ , are  $n$  dimensional vectors which are found by solving

$$C\mathbf{e} = \lambda\mathbf{e} \quad (6)$$

where  $\lambda$  is the eigenvalue corresponding to  $e$ . We will assume the eigenvalues and eigenvectors are arranged in non-decreasing order of eigenvalues and each eigenvector is of length 1. We will use the sample covariance matrix as though it was the true covariance matrix and so

$$C \approx \frac{1}{M} \sum_{j=1}^M \mathbf{x}_j \mathbf{x}_j^T \quad (7)$$

Now each eigenvector lies in the span of  $\chi$ ; i.e. the set  $\chi = \{\mathbf{x}_1, \dots, \mathbf{x}_M\}$  forms a basis set (normally overcomplete since  $M > n$ ) for the eigenvectors. So each  $e_i$  can be expressed as

$$\mathbf{e}_i = \sum_j \alpha_j^i \mathbf{x}_j \quad (8)$$

If we wish to find the principal components of a new data point  $x$  we project it onto the eigenvectors previously found: the first principal component is  $(x, e_1)$ , the second is  $(x, e_2)$ , etc. These are the coordinates of  $x$  in the eigenvector basis. There are only  $n$  eigenvectors (at most) and so there can only be  $n$  coordinates in the new system: we have merely rotated the data set.

Now consider projecting one of the data points from  $\chi$  on the eigenvector  $e_1$ ; then

$$\mathbf{x}_k \mathbf{e}_1 = \mathbf{x}_k \cdot \sum_j \alpha_j^1 \mathbf{x}_j = \mathbf{a}_1 \cdot \sum_j \mathbf{x}_k \mathbf{x}_j \quad (9)$$

Now let  $K$  be the matrix of dot products. Then  $K_{ij} = \mathbf{x}_i \mathbf{x}_j$ . Multiplying both sides of (1) by  $\mathbf{x}_k$  we get

$$\mathbf{x}_k C \mathbf{e}_1 = \lambda \mathbf{e}_1 \cdot \mathbf{x}_k \quad (10)$$

and using the expansion for  $e_1$ , and the definition of the sample covariance matrix,  $C$ , gives

$$\frac{1}{M} \mathbf{K}^2 \mathbf{a}_1 = \lambda \mathbf{K} \mathbf{a}_1 \quad (11)$$

Now it may be shown [Scholkopf et al., 1999] that all interesting solutions of this equation are also solutions of

$$\mathbf{K}\mathbf{a}_1 = M\lambda_1 \mathbf{a}_1 \quad (12)$$

whose solution is that  $\alpha_1$  is the principal eigenvector of  $\mathbf{K}$ .

Now so far we have only found a rather different way of performing Principal Component Analysis. But now we preprocess the data using  $\Phi: \mathcal{X} \rightarrow \mathcal{F}$ . So  $\mathcal{F}$  is now the space spanned by  $\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}_M)$ . The above arguments all hold and the eigenvectors of the dot product matrix  $K_{ij} = (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j))$  are the equivalent vectors in the feature space. But now the Kernel Trick: provided we can calculate  $\mathbf{K}$  we don't need the individual terms  $\Phi(\mathbf{x}_i)$ .

As an example of how to create the Kernel matrix, we may use Gaussian kernels so that

$$\begin{aligned} K_{ij} &= (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^2 / (2\sigma^2)) \\ K_{ij} &= (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^2 / (2\sigma^2)) \end{aligned} \quad (13)$$

This kernel has been shown [Scholkopf, 1999] to satisfy the conditions of Mercer's theorem and so can be used as a kernel for some function  $\Phi(\cdot)$ . One issue that we must address in feature space is that the eigenvectors should be of unit length. Let  $\mathbf{v}_i$  be an eigenvector of  $\mathbf{C}$ . Then  $\mathbf{v}_i$  is a vector in the space  $\mathcal{F}$  spanned by  $\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}_M)$  and so can be expressed in terms of this basis. This is an at most  $M$ -dimensional subspace of a possibly infinite dimensional space which gives computational tractability to the kernel algorithms. Then

$$\mathbf{v}_i = \sum_{j=1}^M \alpha_j^i \Phi(\mathbf{x}_j) \quad (14a)$$

for eigenvectors  $\mathbf{v}_i$  corresponding to non-zero eigenvalues. Therefore

$$\begin{aligned} \mathbf{v}_i^T \mathbf{v}_i &= \sum_{j,k=1}^M \alpha_j^i \Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_k) \alpha_k^i \\ &= \sum_{j,k=1}^M \alpha_j^i K_{jk} \alpha_k^i \\ &= \mathbf{a}_i^i \cdot (\mathbf{K} \mathbf{a}_i^i) \\ &= \lambda_i \mathbf{a}_i^i \cdot \mathbf{a}_i^i \end{aligned} \quad (14b)$$

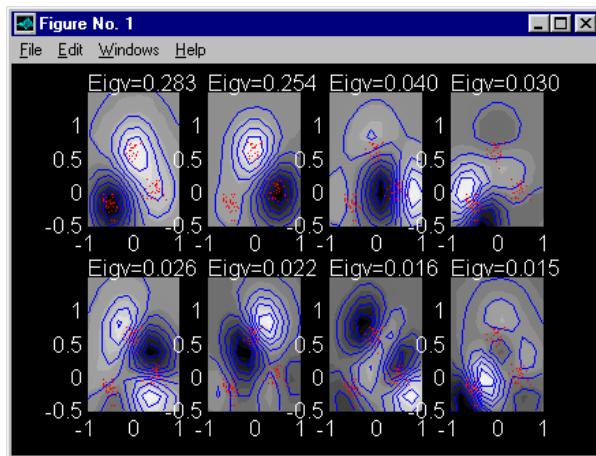
Now  $\alpha^i$  are (by definition of the eigenvectors of K) of unit magnitude. Therefore since we require the eigenvectors to be normalised in feature space, F, i.e.  $\mathbf{v}_i^T \mathbf{v}_i = 1$ , we must normalise the eigenvectors of K,  $\alpha^i$ , by dividing each by the square root of their corresponding eigenvalues.

Now we can simply perform a principal component projection of any new point x by finding its projection onto the principal components of the feature space, F. Thus

$$\begin{aligned} \mathbf{v}_i \cdot \Phi(\mathbf{x}) &= \sum_{j=1}^M \alpha_j^i \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}) \\ &= \sum_{j=1}^M \alpha_j^i K(\mathbf{x}_j, \mathbf{x}) \end{aligned} \quad (15)$$

### 3 KPCA for Exploratory Data Investigation Method

We use KPCA as an exploratory data investigation method. Fig. 1 shows the clustering ability of Kernel PCA with a Gaussian Kernel. The data set comprises 3 sets each of 30 points each of which is drawn from a Gaussian distribution. The centres of the three Gaussians are such that there is a clear separation between the clouds of points. The figure shows the contours of equal projection onto the first 8 KPCA directions. Note that linear PCA would only be able to extract 2 principal components; however because the kernel operation has moved us into a high dimensional space in a nonlinear manner, there may be up to 90 non-zero eigenvalues. The three clusters can be clearly identified by projecting the data points onto the first two eigenvectors. Subsequent Kernel Principal Components split the clusters into sections.

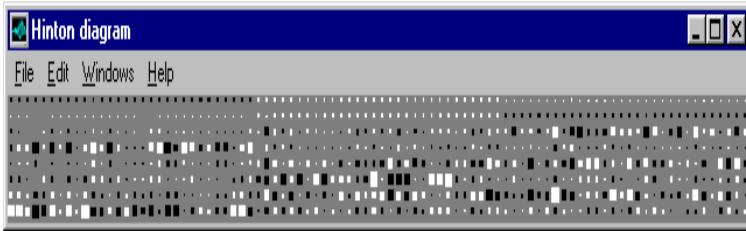


**Fig. 1.** The 3 clusters data set is shown as individual points

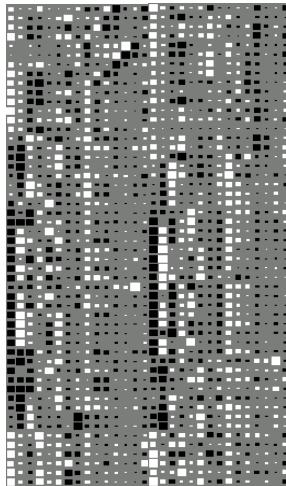
However Fig. 2 shows the components of the eigenvectors in feature space. We see why the first two projections were so successful at identifying the three clusters but we note that there is a drawback to the method if we were to use this method to identify

cases: each eigenvector is constructed with support from projections of very many points. What we really wish is to identify individual points in terms of their importance. This issue has previously been addressed in (Fyfe et al., 2000) using a number of heuristics. In this thesis, we use a novel sparsification of the Kernel PCA method.

The contours are contours of equal projection on the respective Principal Components. The first two principal components are sufficient to differentiate between the three clusters; the others slice the clusters internally and have much less variance associated with them.



**Fig. 2.** The first eight eigenvectors found (each vector is represented in a horizontal line) by Kernel PCA



**Fig. 3.** The first 15 columns show the first 15 eigenvectors of the 1995 data set. The second 15 column vectors show the projections of the 1996 data onto these eigenvectors.

Fig. 3 shows the results from a Kernel PCA on the 1995 data set. We see that the first KPC differentiates between summer and winter, the second differentiates between the change in the seasons and those constant seasons etc. It also illustrates how KPCA may be thought of as a clustering method in data space rather than a projection method in data space. This suggests that, in the context of forecasting, we use KPCA to cluster like data points and then use the similar points to forecast (just as the

nonlinear forecasters use similar points on an attractor). There is however one obvious difficulty with this method- the kernel methods scale computationally with the number of data points. We must look therefore for a method which cuts down the number of data points (which is what the original Support Vector algorithm does automatically).

## 4 Sparse Kernel Principal Component Analysis

It has been suggested [Smola et al., 2000] that we may reformulate the Kernel PCA problem as follows: let the set of permissible weight vectors be

$$\begin{aligned} V = & \left\{ \mathbf{w} : \mathbf{w} = \sum_{i=1}^M \alpha_i \Phi(\mathbf{x}_i), \text{ with } \|\mathbf{w}\|^2 \right. \\ & \left. = \sum_{i,j} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \leq 1 \right\} \end{aligned} \quad (16)$$

Then the first principal component is

$$\mathbf{v}_1 = \arg \max_{\mathbf{v} \in V} \frac{1}{M} \sum_{i=1}^M |\mathbf{v} \cdot \Phi(\mathbf{x}_i)|^2 \quad (17)$$

for centred data. This is the basic KPCA definition which we have used above. Now we may ask whether other sets of permissible vectors may also be found to be useful. Consider

$$V_{LP} = \left\{ \mathbf{w} : \mathbf{w} = \sum_{i=1}^M \alpha_i \Phi(\mathbf{x}_i), \text{ with } \sum_i |\alpha_i| \leq 1 \right\} \quad (18)$$

This is equivalent to a sparsity regulariser used in supervised learning and leads to a type of kernel feature analysis

$$\mathbf{v}_1 = \arg \max_{\mathbf{v} \in V_{LP}} \frac{1}{M} \sum_{i=1}^M |\mathbf{v} \cdot \Phi(\mathbf{x}_i)|^2 \quad (19)$$

We may think that subsequent "principal vectors" can be found by removing this vector from further consideration and ensuring that the subsequent solutions are all orthogonal to the previously found solutions. However as we shall see there are problems in this simple solution. [Smola et al., 2000] point out that this system may be generalised by considering the  $l_p$  norm to create permissible spaces

$$V_p = \left\{ \mathbf{w} : \mathbf{w} = \sum_{i=1}^M \alpha_i \Phi(\mathbf{x}_i), \text{ with } \sum_i |\alpha_i| \leq 1 \right\} \quad (20)$$

## 5 Solutions and Problems

[Smola et al. 2000] have shown that the solutions of

$$\mathbf{v}_1 = \arg \max_{\mathbf{v} \in V_p} \frac{1}{M} \sum_{i=1}^M |\mathbf{v} \cdot \Phi(\mathbf{x}_i)|^2 \quad (21)$$

are to be found at the corners of the hypercube determined by the basis vectors,  $\Phi(\mathbf{x}_i)$ . Therefore all we require to do is find the element  $\mathbf{x}_k$  defined by

$$\mathbf{x}_k = \arg \max_{\mathbf{x}_i \in \mathcal{X}} \sum_{i=1}^M |\Phi(\mathbf{x}_k) \cdot \Phi(\mathbf{x}_i)|^2 = \arg \max_{\mathbf{x}_i \in \mathcal{X}} \sum_{i=1}^M |K_{ki}|^2 \quad (22)$$

which again requires us only to evaluate the kernel matrix. So the solution to finding the "First Principal Component" using this method is exceedingly simple. However, subsequent PCs cause us more concern. Consider first the "naive" solution which is simply to remove the winner of the first competition from consideration and then repeat the experiment with the remainder of the data points. However these data points may not reveal interesting structure: typically indeed the same structure in input space (e.g. a cluster) may be found more than once. In the data set to be considered in this paper, this indeed happens. Indeed the first 10 Kernel Principal Components are in fact all from the same cluster of data and are highly redundant.

An alternative is to enforce orthogonality using a Gram Schmidt orthogonalisation in feature space. Let  $\mathbf{v}_1 = \Phi_1(\mathbf{x}_i)$  for some i. Then

$$\begin{aligned} \Phi_2(\mathbf{x}_j) &= \Phi_1(\mathbf{x}_j) - \frac{\mathbf{v}_1}{|\mathbf{v}_1|^2} (\Phi_1(\mathbf{x}_j) \cdot \mathbf{v}_1) \\ &= \Phi_1(\mathbf{x}_j) - \frac{\mathbf{v}_1}{|\mathbf{v}_1|^2} K(\mathbf{x}_j, \mathbf{x}_i) \end{aligned} \quad (23)$$

where we have used  $\Phi_1$  to denote the nonlinear function mapping the data into feature space and  $\Phi_2$  to denote the mapping after the orthogonalisation has been performed i.e. the mapping is now to that part of the feature space orthogonal to the first Principal Component. Using the same convention with the K matrices gives

$$\begin{aligned} &\Phi_2(\mathbf{x}_j) \Phi_2(\mathbf{x}_k) \\ &= \left( \Phi_1(\mathbf{x}_j) - \frac{\mathbf{v}_1}{|\mathbf{v}_1|^2} K_1(\mathbf{x}_j, \mathbf{x}_i) \right) \left( \Phi_1(\mathbf{x}_k) - \frac{\mathbf{v}_1}{|\mathbf{v}_1|^2} K_1(\mathbf{x}_k, \mathbf{x}_i) \right) \\ &= K_1(\mathbf{x}_j, \mathbf{x}_k) - \frac{2K_1(\mathbf{x}_j, \mathbf{x}_i)K_1(\mathbf{x}_k, \mathbf{x}_i)}{|\mathbf{v}_1|^2} + \frac{K_1(\mathbf{x}_j, \mathbf{x}_i)K_1(\mathbf{x}_k, \mathbf{x}_i)}{|\mathbf{v}_1|^2} \end{aligned} \quad (24a)$$

i.e.

$$K_2(\mathbf{x}_j, \mathbf{x}_k) = K_1(\mathbf{x}_j, \mathbf{x}_k) - \frac{K_1(\mathbf{x}_j, \mathbf{x}_i)K_1(\mathbf{x}_k, \mathbf{x}_i)}{K_1(\mathbf{x}_i, \mathbf{x}_i)} \quad (24b)$$

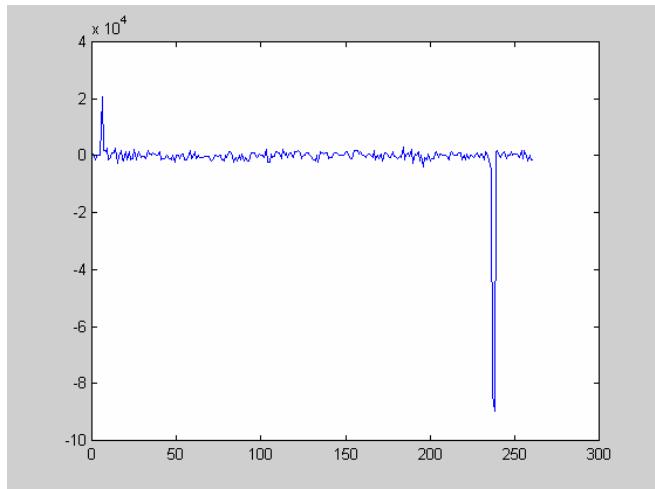
which can be searched for the optimal values. The method can clearly be applied recursively for any time instant  $i+1$ .

$$K_{i+1}(\mathbf{x}_j, \mathbf{x}_k) = K_i(\mathbf{x}_j, \mathbf{x}_k) - \frac{K_i(\mathbf{x}_j, \mathbf{x}_i)K_i(\mathbf{x}_k, \mathbf{x}_i)}{K_i(\mathbf{x}_i, \mathbf{x}_i)} \quad (25)$$

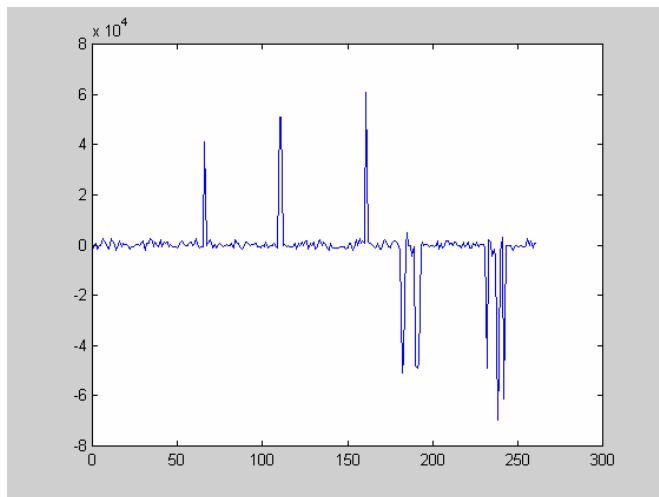
One difficulty with this method is that we can be (and typically will be) moving out of the space determined by the norm. [Smola et al, 2000] suggest renormalising this point to move it back into  $V_p$ . This can be easily done in feature space and both the orthogonalisation and renormalising can be combined into

$$\begin{aligned} K_{i+1}(\mathbf{x}_j, \mathbf{x}_k) \\ = \frac{K_i(\mathbf{x}_j, \mathbf{x}_k)K_i(\mathbf{x}_i, \mathbf{x}_i) - K_i(\mathbf{x}_j, \mathbf{x}_i)K_i(\mathbf{x}_k, \mathbf{x}_i)}{K_i^3(\mathbf{x}_i, \mathbf{x}_i)\{K_i(\mathbf{x}_i, \mathbf{x}_i) + K_i(\mathbf{x}_j, \mathbf{x}_k)\}\{K_i(\mathbf{x}_i, \mathbf{x}_i) - K_i(\mathbf{x}_j, \mathbf{x}_i)\}} \end{aligned} \quad (26)$$

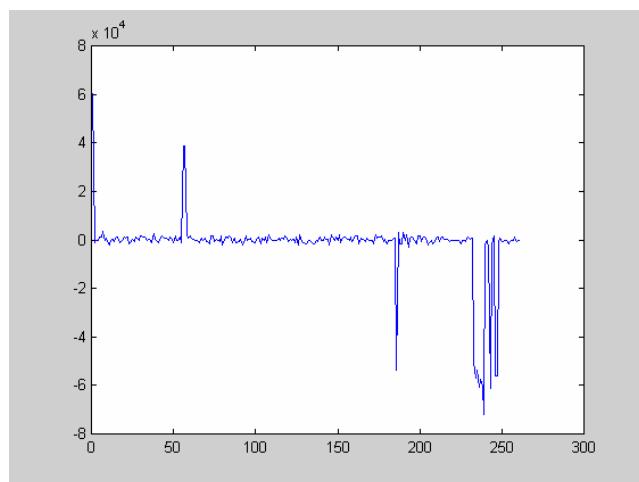
which is a somewhat cumbersome expression and must be proved to be a valid kernel. In this paper we do not perform this step having found it to be unnecessary. We will demonstrate that finding the maximal projection corner from the remainder after orthogonalisation is sufficient for our purposes. We use each day of week data set to train the Kernel PCA. Fig. 4 to Fig. 10 shows the results from a Kernel PCA on each day of week data sets.



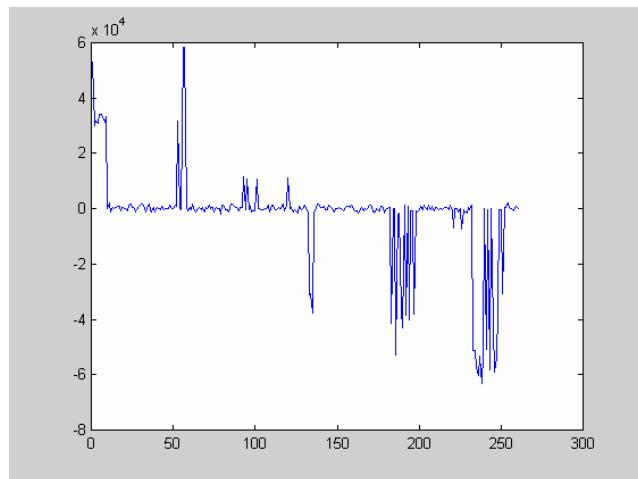
**Fig. 4.** Forecasting errors on Monday data set



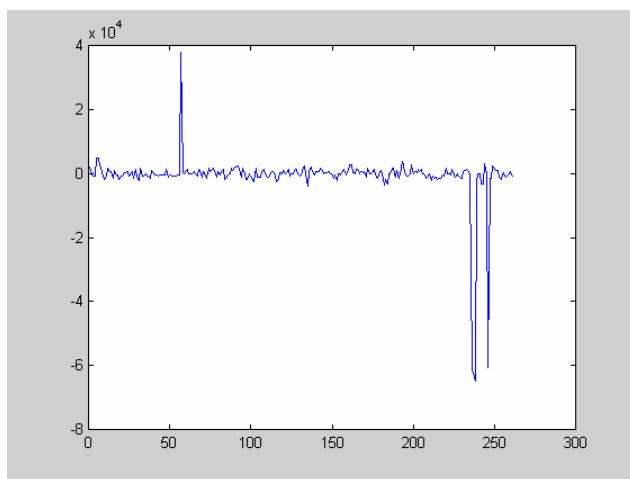
**Fig. 5.** Forecasting errors on Tuesday data set



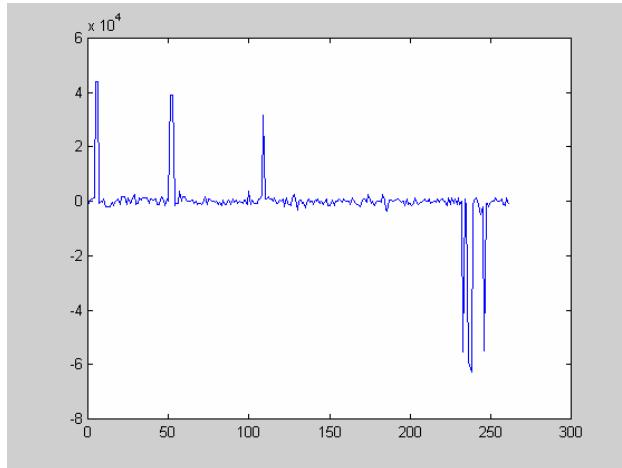
**Fig. 6.** Forecasting errors on Wednesday data set



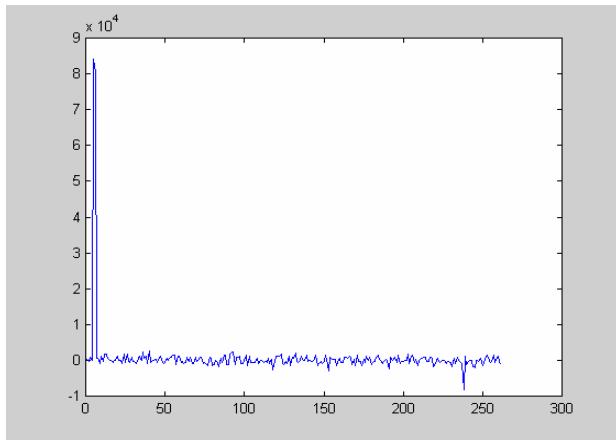
**Fig. 7.** Forecasting errors on Thursday data set



**Fig. 8.** Forecasting errors on Friday data set



**Fig. 9.** Forecasting errors on Saturday data set



**Fig. 10.** Forecasting errors on Sunday data set

We see that for each data set some points are very accurately predicted. Presumably these are the points which lie close to the points which the Sparse Kernel PCA has identified. However other points are very badly forecast. These points are those which are far from any of the points which the Sparse KPCA has used as important points when carrying out the Principal Component Analysis in feature space.

## 6 Conclusions

This chapter has investigated the use of kernel methods for the prediction of power load data. The standard SVM has shown a great deal of promise in this area but we should note that it is a computationally intensive method. We have also investigated

the data set using Kernel Principal Component Analysis; we have made the case that a sparsification of this method is necessary to get a computationally viable and easily comprehensible representation in feature space. However the Sparse Kernel PCA method which we have developed has been shown to perform well for only some data points. Points which are more distant from the data points identified by the method as contributing a great deal to the variance in the data set are less well forecast by this method. Future research will inevitably be into other methods of finding the optimal data points to use as well as which kernel is most appropriate for each problem and which methods are best in feature space for results in data space.

## References

1. Chuang, S.J., Fyfe, C., Hwang, R.C.: Fuzzy backpropagation for power load forecasting: a comparative study. *Soft Computing* (1998)
2. Chuang, S.J., Fyfe, C., Hwang, R.C.: Power load forecasting by neural network with weights-limited training algorithm. In: ESS 1998 (1998)
3. Lin, Y., Wahba, G., Zhang, H., Lee, Y.: Statistical Properties and Adaptive Tuning of Support Vector Machines. TR 1022 (September 2000)
4. Swee, E.G.T., Elangovan, S.: Applications of systems for denosing and load forecasting. In: Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics, 1999, pp. 165–169 (1999)

# A Study of USB 3 in Perspective Aspect

Yeh Wei-Ming

Department of Radio & Television, National Taiwan University of Arts, Taipei, Taiwan  
t0150@mail.ntua.edu.tw

**Abstract.** Since 1996, the first USB product was introduced (USB 1.0, and USB 2.0 specs was released in 2000. Many people have shared the new technology of Windows and flash memory card through USB devices. Meanwhile, another heavy weight high-speed communications and real-time data transfer: IEEE 1394 interface was announced one year ahead of USB1.0 (1995), and many professional consumers are happy to have both devices on PC or Mac. Recently, as soon as USB 3.0 has introduced in May 2010, the battle of new and old format are happened. This research took few years to trace the development of USB system. We collect more than 300 cases from the telephone survey during Jan., 2010 to May, 2010. Total of 212 cases comply with the conditions. To probe mainly into the relationship between new USB 3.0 confidence level and 3 different groups: IT industries, Multimedia specialist, and Consumers in Taiwan. The profiles of contingency table were used to explore the relationship between eBooks' confidence level and different peoples, and special model were used to confirm the relationship of each other. The result is an effective method that may help to improve management qualities and push media management forward.

**Keywords:** Asymmetric Design; FireWire; HANA; Peripheral Devices.

## 1 Introduction

USB (Universal Serial Bus) is a specification to establish communication between devices and a host controller (usually personal computers), developed and invented by Ajay Bhatt while working for Intel. USB may replace some serial and parallel ports, which can connect computer peripherals basics such as: mice, keyboards, digital cameras, printers, personal media players, flash drives, and external hard drives. Soon after, USB has become the standard connection method. In 2008, there are about 2 billion USB devices sold per year, and approximately 6 billion of them sold to date around the world. It began development in 1994 by a group of seven companies: Compaq, DEC, IBM, Intel, Microsoft, NEC and Nortel. The USB 1.0 specification was introduced in January 1996. The original USB 1.0 specification had a data transfer rate of 12 Mbit/s. The USB 2.0 was released in April 2000, and was standardized by the USB-IF at the end of 2001, was widely accepted by HP, Intel.. NEC and Philips jointly led the initiative to develop a higher data transfer rate, with the resulting specification achieving 480 Mbit/s. USB 3.0 (*SuperSpeed*) is announced in May 2010,

**Table 1.** The milestone of USB and IEEE 1394

Year	Content	Remark
2010	1.USB 3.0 is announced in Taiwan 2.Industrial Technology Research Institute (ITRI) has integrated the leading technology of USB 3.0 in press conference	* May, 2010 *June,2010
2009	1. NEC ships world's first USB 3.0 host silicon 2. SuperSpeed USB logo debuted 3. Linux begins native USB 3.0 support	
2008	USB 3.0 specs released	
2006	IEEE 1394c released	
2005	Wireless USB 1.0 specs released	
2002	1. Windows XP SP1 supports USB 2.0 natively 2. IEEE 1394b released	
2001	USB OTG specification released.	
2000	1.USB 2.0 specs released 2.USB started to gain reputation as a mainstream bus technology 3.IEEE 1394a released	
1998	1.Apple shipped iMac with USB ports only 2.USB 1.1 specification released	
1997	USB-IF membership increased to over 400 companies, Over 500 USB products were in development worldwide	
1996	1.USB 1.0 specs released 2.First USB product introduced 3.First USB plug compliance workshop held.	
1995	1.USB Implementers Forum (USB-IF) formed with an initial membership of 340 companies 2.Intel introduced the first USB silicon. 3.Apple created Fire Wire(IEEE 1394)	
1994	USB core companies assembled	

Source:<http://www.everythingsub.com/superspeed-usb.html>

which provides a fourth transfer mode at 5.0 Gbit/s. The raw throughput is 4 Gbit/s, and the specification considers it reasonable to achieve 3.2 Gbit/s (0.4 GByte/s or 400 MByte/s. All of implements connections for USB is a storage devices, which using a set of standards called the “*USB mass storage device class*”. This was initially intended for traditional magnetic and optical drives, but has been extended to support a wide variety of devices, particularly flash drives(flash memory), which is preferable

for many people all over the world. In addition, there are more external drives available, such as: IDE, ATA, SATA, PATA, ATAPI, or SCSI to a USB interface port. In fact, eSATA, ExpressCard (now at version 2.0), HDMI, and FireWire (IEEE 1394), are welcome too.

## 2 Literature Review

### 2.1 Different Aspects of USB and IEEE 1394

- Anderson, Don ., FireWire System Architecture. MindShare, Inc.1394ta.org. (1999),
- Teener, Michael J. What is Firewire,[http://www.teener.com/firewire\\_FAQ/.\(2008\)](http://www.teener.com/firewire_FAQ/.(2008))
- IEEE Std. 1394-2008, IEEE Standard for a High-Performance Serial Bus, [http://ieeexplore.ieee.org/servlet/opac?punumber=4659231\(2008\)](http://ieeexplore.ieee.org/servlet/opac?punumber=4659231(2008))
- Usb-ware.com. FireWire — USB Comparison, [\(2010\).](http://www.usb-ware.com/firewire-vs-usb.htm)
- USB Implementers Forum, About USB-IF, [http://www.usb.org/about.\( 2009\),](http://www.usb.org/about.( 2009),)
- Melissa J. Perenson, SuperSpeed USB 3.0: More Details Emerge s, [\(2009\).](http://www.pcworld.com/article/156494/superspeed_usb_30_more_details_emerge.html)

### 2.2 Technical Innovation

The world's best-known computer peripheral interface once again receives a major revamp (following Wireless USB) to stay current with modern demands for connectivity bandwidth. Dubbed "*SuperSpeed USB*", USB 3.0 promises a major leap forward in transfer speeds and capability, while maintaining backwards compatibility with USB 2.0 devices. More over, this new device has new challenge: 1. increased maximum bus power and increased device current draw to better accommodate power-hungry devices, 2. new power management features, 3. full-duplex data transfers and support for new transfer types, 4. new connectors and cables for higher speed data transfer...although they are backwards compatible with USB 2.0 devices and computers, 5. the most obvious change is an additional physical bus that is added in parallel with the existing USB 2.0 bus. Which indicates USB 2.0 previously had 4 wires (power, ground, and a pair for differential data), USB 3.0 adds 4 more for two pairs of differential signals (receive and transmit) for a combined total of 8 connections in the connectors and cabling. These extra two pairs were necessary to meet the requirement of bandwidth, because the two wire differential signals of USB 2.0 were not sufficient.

### 2.3 Phone Survey

In our experiment, we collect 300 cases from the telephone survey during March, 2010 and May, 2010. Total of 212 cases were effective. This telephone surveys were written by two specialists, and those non-typical cases were determined after further discussion of 3 groups: IT industries (IT), Multimedia specialist (MS), and Consumers (CS).

## 2.4 Associated Analysis of Video (Image) Effect Confidence Level

We took profile of contingency table proposed by Jobson (1992) to describe the video effect Confidence level associated with three groups. The study probes into confidence level and three specific groups, use chi-square test of independence to confirm the result.

## 2.5 The Different Groups in Confidence Level

The Moving Average Convergence and Divergence is attributed to Gerald Appel and Schisler of Signalert Corporation. Although the study's formula suggests the periods to be used for the two averages, the user can experiment with different inputs. The results are two exponentially smoothed moving averages that revolve above and below a zero line. The most useful signals are given when the shorter line crosses the slower line. A signal may occur, when the shorter line crosses above the longer and a sell signal when it crosses below the slower line.

The relation of Confidence level and three groups we present as  $3 \times 4$  contingency table (the Confidence level 0, 1, 2, 3 classified low, slight, moderate and high). Seldom have confidence level "3", so we amalgamate "2" and "3" to "2" (above moderate) and then utilize this data to find out conditional, and marginal probability.

**Table 2.** The conditional and marginal probability of USB 3.0 Confidence level

groups	Confidence level		
	level0	level1	above2
IT	0.2844	0.4587	0.2569
MS	0.7143	0.1429	0.1429
CS	0.5366	0.3171	0.1463
TOTAL	0.4682	0.3318	0.2000

We utilize table2 and draw profile of different groups with conditional and marginal probability of different groups Confidence level. It is for IT, MS and CS respectively in Fig. 1. Where full line denotes the marginal probability of the levels, the dotted lines the conditional probability. Find levels "low"(level 0) and "slight"(level 1), there is a maximum disparity from full line to others, nearly up to 0.2 in the B(MS), therefore can infer there is less confidence level in the MS relatively, and greater probability of having slight confidence level in the IT, hence the confidence level associates with USB 3.0. Add up confidence level "2" and "3" to "2", presented  $3 \times 3$  contingency table, we use chi-squared test for independence to confirm and associated with the Confidence level, the  $\chi^2$ -value would be 33.67, and P = 8.6E-7.

value  $\hat{\beta}_2 > 0$ , ascertaining the result is unanimous of this result by Table 3.

### 3 Proportional Odds Model

We made use of test for independence to confirm the different groups Confidence level and three groups in section 3: The different groups' Confidence level is related, but this test has not used the ordinal response level to confidence level. The regular collects the classification data with order in the media research. For example it mentioned in the article the response variables of 'low', 'slight', 'moderate' and 'high' reading, we hope to probe into the factor influencing its Confidence level, so we analyze utilizing proportional odds model. if only one covariate, model present straight line relation to the logarithm of the accumulated odds of level p and covariate variable x, because this model supposes that there is the same slope  $\beta$  to all level p, this c-l straight line is parallel each other. This model is based on McCullagh and Nelder theory (1989) and Diggle, Liang and Zeger (1994).

One predictable variable was included in the study, representing known or potential risk factors. They are kinds of variable status is a categorical variable with three levels. It is represented by two indicator variables x ( $x_1$  and  $x_2$ ), as follows:

**Table 3.** 3 groups of indicator variables

kinds	$x_1$	$x_2$
IT	1	0
MS	0	1
CS	0	0

Note the use of the indicator variables as just explained for the categorical variable. The primary purpose of the study was to assess the strength of the association between each of the predictable variables.

$$\text{Let } L_p(x_1, x_2) = \theta_p + \beta_1 x_1 + \beta_2 x_2, \quad p = 1, 2 \quad (1)$$

Use likelihood ratio tests for  $b_1 = b_2 = 0$  hypothesis, the explanation of likelihood ratio tests is as follows: given L be the log likelihood of models , then  $G2 = -2L$  From hypothesis  $b_1 = b_2 = 0$ , we have

$$L_p(x) = \theta_p, \quad p = 1, 2, \dots, c-1 \quad (2)$$

likelihood ratio tests use the difference of two deviances between model (1) and model (2) as reference value to test  $H_0 : b_1 = b_2 = 0$

If reject the hypothesis, furthermore, to test if  $b_1 = 0$  or  $b_2 = 0$ .

As fact, the confidence level "under moderate" to the confidence level "high", the odds ratio of SP compared with CS is also  $\exp(b_1)$ . Now according to formula,  $b_1$  is the logarithm of the estimated odds when the confidence level "under high" to the confidence level "high", the SP compared with CS.  $b_1 > 0$  means the confidence level presented by SP is less high than CS. Alternatively,  $b_1 < 0$  means the confidence level presented by SP is more high than CS.

Therefore,  $b_1 - b_2$  is the logarithm of the estimated odds when the confidence level "under moderate" to the confidence level "high", the SP compared with AP.  $b_1 - b_2 > 0$  means the confidence level presented by SP is less high than AP. Alternatively,  $b_1 - b_2 < 0$  means the specify confidence level presented by IT is more high than AP.

On the rear part in the article, we will utilize odds ratio to probe into association USB 3.0 confidence level.

### 3.1 Analysis of Different Groups Confidence Level

Combining confidence level "2" and "3" to "2"(above moderate), we utilize odds ratio to calculate, deviance equal 215.76, if  $b_1 = b_2 = 0$ , deviance = 229.51, the difference between two deviances is 13.75, the related  $\chi^2$ -critical value will be test  $H_0: \beta_1 = \beta_2 = 0$ , we find  $\chi^2_{0.05} = 13.75$ , and P=0.001. The null hypothesis is rejected and we conclude that  $\beta_1$  and  $\beta_2$  are not zero simultaneously. Furthermore, we analyze  $\beta_1 = 0$  or  $\beta_2 = 0$  or both not equal to zero. Table3 is the results by maximum likelihood estimates.

**Table 4.** Analysis of the video effect Confidence level

Effect	B	Coefficient	Std Error	z-statistic	Sig
Intercept	$\hat{\theta}_1$	0.137	0.303	0.452	0.652
	$\hat{\theta}_2$	1.799	0.331	5.428	0.0000
USB 3.0	$\hat{\beta}_1$	0.906	0.350	2.586	0.009
	$\hat{\beta}_2$	-0.675	0.402	-1.680	0.093

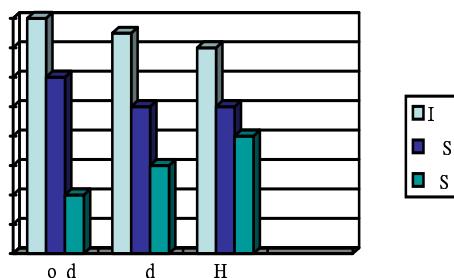
From Table4 we find the hypothesis  $\hat{\beta}_2 = 0$  is accepted,  $\hat{\beta}_1 = 0$  is rejected, P-value is 0.009 and 0.093 respectively,  $\hat{\beta}_2 = 0$  represents the logarithm of odds ratio for B(MS), thus the odds ratio would be estimated to be  $e^{\beta_2}$ .  $\hat{\beta}_1 > 0$  represents the logarithm of odds ratio of confidence level not high of the MS is higher than CS. This indicates that

the odds of USB 3.0 Confidence level not high is about 2.47 and 0.51 fold for IT and MS compared to CS customer. Confidence level of the CS is more high by its estimated value  $\hat{\beta}_2 > 0$ , ascertaining the result is unanimous of this result.

### 3.2 Preference of Various USB 3.0 in Different Groups

We integrated all possible devices, such as: USB host controllers, peripheral chipsets, hubs compliant, flash memory.., are widely accepted. Based on this data, IT (IT industries) may have more expectance than the rest of groups, we assumed USB 3.0 could bring bright “imagination” to replace the USB 2.0 system with a decade long life-span, especially after the Windows 7 has accepted by many consumers with a long struggle by Microsoft. It sounds naturally for IT industries to accept the new USB 3.0 system, and ready to prepare new mass production products for USB 3.0 appliance, and may have more profit than before. As to the MS (Multimedia specialist), may wonder the test report, and true credit of it, not simply judged by “beautiful” statistic report Consumers may express least intention on this new device (s), the reasons could be: 1. USB 2.0 is good enough, 2. every new system should have some hidden drawback, 3. the current price for new system is not practical.

Among many new products for USB 3.0, motherboard (Board), adapter card (Card), and mass storage HD (HD) may be accepted by many people at the first “round”, different group of them has perspective preference.



**Fig. 1.** Preference of USB 3.0 products in different groups

## 4 Conclusion and Suggestion

Based on the result of the section 3, 4, the confidence level associate with various groups, and analyze the possible meaning.

As to the statistical method use, we arrange discussion to adopt profile model compared with logarithm of odds. Adopt profile put emphasis on ascertaining and explaining conveniently mainly, because it can make very apt to find out less seldom of confidence level for the AP, further, IT is relatively comfortable with the probability of having slight confidence level. But profile model will often have different explanations with different persons in interpretation, so uses the proportion odds model to confirm its result in this profile.

Since a decade ago, USB 2.0 set its unique standard; many IT industries all over the world are glad to follow and got a huge profit from it, no matter more strong competitors such as: eSATA, FireWire 3200, and ExpressCard 2.0 are there, USB 2.0 has challenged them all, and survived.

As to the future, at least the next five years, we do not see the market for USB 2.0 devices of all types to dwindle. High-bandwidth devices, such as video cameras or storage devices will likely be the first to migrate to USB 3.0, in this particular industry are mainly driven by demand and volume, will restrict USB 3.0 implementation to higher-end products.

By 2010, computer motherboards should start to come equipped with USB 3.0 ports supplementing USB 2.0 ports. USB 3.0 adapter cards will likely play a important role in driving the installed base of USB 3.0 ports up, become standard on new PCs with operation system of Win 7 (or Win 8), device manufacturers will be further motivated to migrate to the this new system, which may appear new market for over millions dollar per year hopefully, IT industries in Taiwan, this time with key- technology of it, may enjoy the fruit of success as a early bird. However, some sad memories of Win Vista, and economic depression may stop the intention for many people to replace any existing reliable system and products, such as: XP, and USB 2.0

Consequently, USB 2.0 may be phased out as USB 1.1 did, but in foreseeable future, .USB 2.0 isn't going anywhere, and be stay with us quite a while.

## References

- [1] Chin, K.: Everythingusb, SuperSpeed USB 3 FAQ (2010),  
<http://www.everythingusb.com/superspeed-usb.html>
- [2] Sutter, J.D.: USB inventor is tech's unlikely 'rock star', CNN (2010)
- [3] Perenson, M.J.: SuperSpeed USB 3.0: More Details Emerge (2009),  
[http://www.pcworld.com/article/156494/superspeed\\_usb\\_30\\_more\\_details\\_emerge.html](http://www.pcworld.com/article/156494/superspeed_usb_30_more_details_emerge.html)
- [4] Teener, Michael, J.: What is Firewire (2008),  
[http://www.teener.com/firewire\\_FAQ/](http://www.teener.com/firewire_FAQ/)
- [5] IEEE Std. 1394-2008, IEEE Standard for a High-Performance Serial Bus (2008),  
<http://ieeexplore.ieee.org/servlet/opac?punumber=4659231>
- [6] Merritt, R.: 2007/09/18. USB 3.0 guns for Firewire, EE Times (2007)
- [7] Diggle, P.J., Liang, K.Y., Zeger, S.L.: Analysis of Longitudinal Data. Oxford Press, Oxford (1994)
- [8] Jobson, J.D.: Applied Multivariate Data Analysis Volume: Categorical and Multivariate Methods. Springer, New York (1992)
- [9] McCullagh, P., Nelder, J.A.: Generalized Linear Models. Chapman and Hall, New York (1989)

# A Study of CAPTCHA and Its Application to User Authentication

Albert B. Jeng<sup>1</sup>, Chien-Chen Tseng<sup>2</sup>, Der-Feng Tseng<sup>2</sup>, and Jiunn-Chin Wang<sup>3</sup>

<sup>1</sup> Dept. of Computer Science and Information Engineering,  
Jinwen University of Science and Technology, Taipei, Taiwan  
[albertjeng@hotmail.com](mailto:albertjeng@hotmail.com)

<sup>2</sup> Dept. of Electrical Engineering,  
National Taiwan University of Science and Technology, Taipei, Taiwan  
[joker913\\_0@hotmail.com](mailto:joker913_0@hotmail.com),  
[dtseeng@mail.ntust.edu.tw](mailto:dtseeng@mail.ntust.edu.tw)

<sup>3</sup> Dept. of Computer Science and Information Engineering,  
National Taiwan University of Science and Technology, Taipei, Taiwan  
[jcwang1209@hotmail.com](mailto:jcwang1209@hotmail.com)

**Abstract.** A CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a scheme used to determine whether the user is a human or a malicious computer program. It has become the most widely used standard security technology to prevent automated computer program attacks. In this paper, we first give an overview of CAPTCHA. Next, we discuss the pros and cons of various CAPTCHA techniques. Then, we present the common attacks and vulnerability analysis in CAPTCHA design. Subsequently, we suggest counter-measures and remedies for those attacks. Finally we propose a personalized CAPTCHA to replace the traditional password-based authentication system as possible further research in applying CAPTCHA to user authentication application.

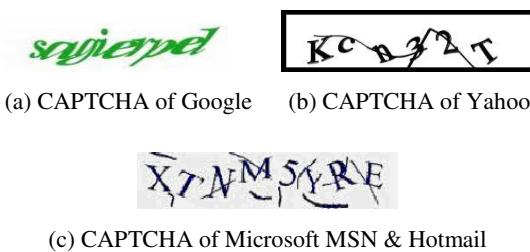
**Keywords:** Human Interactive Proof, Attack, Personalized CAPTCHA.

## 1 Introduction

CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) or HIP (Human Interactive Proof), as the name suggests, is an automatic security mechanism to distinguish between computer programs and humans [1]. It was first made by Luis Von Ahn, Manuel Blum, Nicholas J. Hopper and John Langford (all of Carnegie Mellon University) in 2002. With the advancement of information technology and network communication, a variety of malicious software (e.g.: Trojans, Worms, Botnets, etc.) can easily use Spam as a transmission medium. However, CAPTCHA can effectively prevent these attacks, because it is easy to construct and simple identification for humans. So CAPTCHA has been the standard security mechanism, which eliminates spam or unknown malicious botnet invasion in recent years. Most major websites (Google, Yahoo, Microsoft MSN and Hotmail, etc), blogs, message boards, online trading systems, public transit system and the online

game authentication systems use it extensively. CAPTCHA can be divided into four types of architectures:

1. **Text-based scheme:** It's the most widely used and highly acceptable form which combines alphabets and numbers to do variety treatments (such as: rotation, deformation, distortion, division, etc) [12]. Its aim is to make computers automatic character recognition program difficult to identify, but the human eyes can easily understand the message. However, over-distorted characters often lead users difficult to recognize in order to achieve better security. The designer must pay attention to whether the design lost its usability. Fig. 1 shows some examples used in Google, Yahoo, and Microsoft MSN & Hotmail.
2. **Image-based scheme:** It provides user multiple images to do the graphic recognition, namely the use of image content delivering challenge for the user to answer [14]. On one hand, the advantage of the picture implies a number of different messages can be used to carry many challenges in it. There is some research about image identification in [2]. On the other hand, the disadvantage is that each person's understanding of the picture is not necessarily the same. From the security point of view, database of pictures should be sufficiently large. It is better to automatically update pictures, so that it avoids hackers from cracking the same images. Fig. 2 shows an example of image CAPTCHA presented in [15].
3. **Audio-based scheme:** Users typically make voice recognition of the voice content, and it can be used for visually-impaired authentication. On the web, the voice CAPTCHAs [13] have been presented as an alternative for image-based CAPTCHAs. It is a convenient way for users with eyesight disabilities. However, for non-native users of the language, it may cause error of identification.
4. **Video-based scheme:** This is the newer design approach using animation or video mode. According to some studies [3] [4] [5], this approach may provide greater security than text-based and image-based CAPTCHA (i.e., hard to be broken by computer programs). However, video is also more complex and need more time to answer the challenge than other schemes.



**Fig. 1.** Text-based scheme

There are still many different innovative designs, such as [6] and [7]. CAPTCHA design is still evolving constantly now. We will first introduce the background of

CAPTCHA in Section 2. Then, we will investigate the attacks of CAPTCHA and vulnerability of CAPTCHA design in Section 3. And finally we will target at different types of CAPTCHA to analyze their security and provide some new ideas and comments.



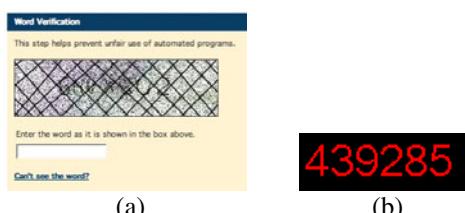
**Fig. 2.** Asirra: an example of Image-based CAPTCHA

## 2 Background

A good CAPTCHA must be designed to follow these two important criteria:

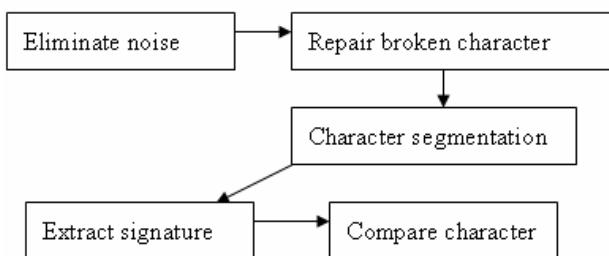
- 1. Robustness**
- 2. Usability (human friendly)**

As a popular standard security mechanism, CAPTCHA requires not only usability & human friendly (i.e., to make it easy for human to understand, and interpretation), but also the system robustness, in order to prevent destruction and invasion by interested parties or malicious software. However, in some cases, robustness and usability could be conflict goals. Using text-based scheme as an example, when the system has a better robustness, then the characters must be more distorted, so that the malicious computer program (e.g., a robot program) could not recognize them, and consequently the user was not able to identify them either. That is not a good design obviously. On the contrary, if the design is quite easy to be identified by the user, then the computer may also be able to easily identify and solve it as illustrated in Fig. 3.



**Fig. 3.** (a) Better robustness but lack of usability, (b) Better usability but lack of robustness

Actually, designing a CAPTCHA which balances robustness and human friendliness is very difficult. How to make a good tradeoff between these two criteria is the CAPTCHA designer's goal. Text-based scheme and Image-based scheme CAPTCHA are two most commonly used modes with Text-based mode at the top. Most attack methods on text-based scheme identify characters using the OCR (Optical Character Recognition) to reach the break point automatically. Before talking about CAPTCHA attacks and security analysis, we must first introduce how the OCR works. Simply, the working principle of OCR is to scan and process the document into character recognition and converted into electronic signals which computer understands. The workflow of OCR character recognition is illustrated in Fig. 4.



**Fig. 4.** Workflow of OCR

**Eliminate noise:** Since the surface of the input file is not clear or there is scan distortion, it will result in recognition trouble. Therefore, we must eliminate the noise before identification. In text-based and image-based CAPTCHA schemes, hackers often use similar method to eliminate background.

**Repair broken character:** Fix the discontinuity, sawtooth and damaged characters.

**Character segmentation:** Separate each character (including overlap part), so that computer can do single character recognition.

**Extract signature:** It's the most important and difficult technology. Extracting the signature in character and encode it. OCR usually uses several different signatures to indicate the character.

**Compare character:** Compare the encoded character with letter in the database, then output the recognized character.

OCR is widely used to break CAPTCHA the architecture of which is simple. Most of break methods are similar to OCR working processes. In next two sections, we will discuss how to break CAPTCHA and how to achieve better security.

### 3 CAPTCHA Attacks

CAPTCHA attacks are various and evolving with time. For different design of CAPTCHA, hackers also break it in different ways. There are no news at all that Yahoo, Microsoft and Google CAPTCHA are broken. In PWNtcha's web page [8], the author claims that he had broken all kinds of CAPTCHA. He does not mention the technology part, but his comments are quite valuable to CAPTCHA designers. In this section, we discuss several common attacks and vulnerabilities in CAPTCHA.

**Botnet attack:** CAPTCHA can be used to block botnet spam attacks. But recent new attacks used the victims of botnet to fill out registration list. After decoding, CAPTCHA will deliver the result to the remote server. When the decoded CAPTCHA is sent back to client, fake account will register successfully. The probability of such attack is about 12% to 20%, but hacker can try many times in a short period. The more powerful the attack technique becomes, the more resultant victims it will cause. Only maintaining good network browsing habits (i.e., do not visit the website or click on unknown unidentified pages), and using the anti-virus protection software could avoid becoming a victim of Botnet attack.

**Binarization:** This type of attack occurs because of CAPTCHA vulnerability. The general CAPTCHA includes foreground and background. Foreground usually design to hide the challenges (characters). Foreground and background color intensity are quite different. Following this feature, attacker can use *threshold* method: When the color intensity is higher than or below the threshold, it will convert the background into either white or black. As shows in Fig.5, the background can be easily removed, only leaving the necessary information to identify characters.



**Fig. 5.** (a) Original image, (b) Binarized image (Both Images are taken from Microsoft CAPTCHA)

**Vertical Segmentation:** This design has the same flaw as caused by the lack of security. According to K. Chellapilla in [9], computer has very powerful reorganization ability. Attackers just need to separate CAPTCHA characters, and give the divided character to the computer identification, then CAPTCHA will be broken easily. Hackers used vertical segmentation to divide up the picture. On the sum of each Column of pixel into a histogram to find the point no pixel exists. Refer to Fig. 6 for detail.

CAPTCHA attacks are various. CAPTCHA designer should focus on architecture, in next Section we will analyze security of different CAPTCHA, and provide suggestions to resist these attacks.



**Fig. 6.** (a) Vertical pixel histogram, (b) segment five blocks (Both Images are taken from [10])

## 4 Remedy

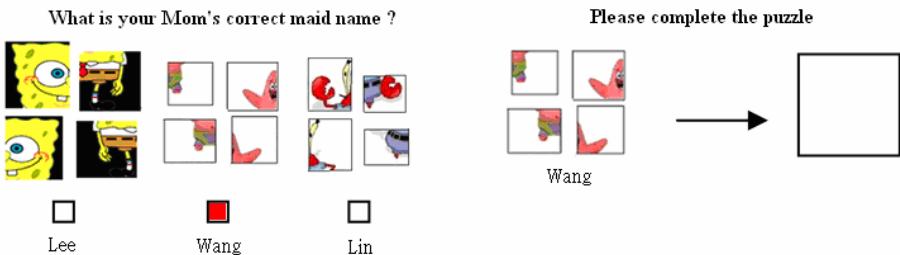
Most of CAPTCHA attacks are caused by various architecture vulnerabilities. Here we propose some useful suggestions to designer.

- **Letter overlap:** Letter overlap is a very powerful way to avoid vertical segmentation attacks. If all of the letters are connected, the OCR will be very hard to recognize each word.
- **Various fonts:** Every font must be different. Constant font (same font size, no rotation, no deformation) will be broken easily. We suggest that the font should be random combination of alphabet (e.g., capital or small letters, number or punctuation mark), and every letter has different height and width.
- **String length:** The longer the string is, the more difficult to break in CAPTCHA. If the probability of OCR's successful recognition at single character is  $n$ , the string length is  $l$ . Then, the correct recognition probability is  $n^l$ . So the longer string length will make better security.
- **Perturbative background:** This protection method adds one more step before segmenting because to remove noise hackers must find a specific property that distinguishes letters. Designer should establish a robust background that is difficult to remove. Simple background will be attacked easily (Binarization). The robust background can have colors, dots, lines, circles, and rectangles, which can be hard to break.

## 5 Personalized CAPTCHA

CAPTCHA is a challenge-response authentication mechanism. It is fast and convenient and widely used to distinguish humans from automatic program attacks. However it is not suitable as a personal authenticate scheme because of its lack of personalization. In [2] [11], they used CAPTCHA to replace traditional password based system. But they do not mention the CAPTCHA architecture. Hence we propose a personalized CAPTCHA that is not only fast and convenient but also suitable for personal authentication. First, a user must provide some privacy information used in a secure challenge-response question (e.g., it could be based on any personalized information such as the mother's maid name, pet's name, etc) to the server during the account registration. Suppose, the challenge question is "what is your Mom's correct maid name?". Then, we construct a puzzle-based CAPTCHA

which contains the answer to the challenge in image puzzle pieces. Assume the correct maid name of your Mom is Wang. Server shows several pictures which have different answers for the user to choose (e.g., Lin, Lee or Wang). After the user picks the picture puzzle pieces (say Wang), which has the right answer, he must also put all the image puzzle pieces together to get the correct final answer. Here is an example shown in Fig.7. This scheme has two advantages: 1. Avoid automatic program attacks. 2. Provide effective personalized authentication.



**Fig. 7.** Puzzle-based personalized CAPTCHA

## 6 Conclusion and Future Work

We've covered the most popular CAPTCHA approaches here but not all of them. Most users are used to the more commonly used graphical CAPTCHAs, so designers must have convincing reasons to migrate to another more difficult-to-break approach. Overall, the design of CAPTCHA is still an art, rather than a science. It requires considerable study to evolve the design of secure and usable CAPTCHAs into a science. We have proposed the first CAPTCHA that used a puzzle-based scheme to implement a personalized authentication. We will continue to analyze the security and usability of this new puzzle-based personalized CAPTCHA for other potential security applications. There's still a lot of interesting research could be done in this topic.

## References

1. von Ahn, L., Blum, M., Langford, J.: Telling Humans and Computer Apart (Automatically) or How Lazy Cryptographers do AI. Comm. of the ACM 47(2), 56–60 (2004)
2. Jameel, H., Shaikh, R.A., Lee, H., Lee, S.: Human Identification through Image Evaluation using Secret Predicates. In: Abe, M. (ed.) CT-RSA 2007. LNCS, vol. 4377, pp. 67–84. Springer, Heidelberg (2006)
3. Cui, J.-S., Mei, J.-T., Wang, X., Zhang, D., Zhang, W.-Z.: A CAPTCHA Implementation Based on 3D Animation. In: International Conference on Multimedia Information Networking and Security, MINES, vol. 2, pp. 179–182 (2009)
4. Ince, I.F., Salman, Y.B.: Execution Time Prediction for 3D Interactive CAPTCHA by Keystroke Level Model. In: Fourth International Conference on Computer Sciences and Convergence Information Technology, pp. 1057–1061 (2009)

5. Kluever, K.A., Zanibbi, R.: Balancing Usability and Security in a Video CAPTCHA. In: ACM International Conference Proceeding Series (2009)
6. Desai, A., Patadia, P.: Drag and Drop: A Better Approach to CAPTCHA. In: 2009 Annual IEEE India Conference, pp. 1–4. IEEE Press, New York (2009)
7. Vimina, E.R., Areekal, A.U.: Telling Computers and Humans Apart Automatically Using Activity Recognition. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 4906–4909. IEEE Press, New York (2009)
8. PWNTcha caca labs, <http://caca.zoy.org/wiki/PWNTcha>
9. Chellapilla, K., Larson, K.: Computers beat Humans at Single Character Recognition in Reading based Human Interaction Proofs (HIPs). In: Proceeding of the ACM Conference (2005)
10. Yan, J., EI Ahmad, A.S.: A Low-cost Attack on a Microsoft CAPTCHA. In: Proceedings of the 15th ACM Conference on Computer and Communications Security, pp. 543–554 (2008)
11. Le, X.H., Lee, S.: Secured WSN-integrated Cloud Computing for u-Life Care. In: Proceedings of the 7th IEEE Conference on Consumer Communications and Networking Conference, pp. 702–703. IEEE Press, New York (2010)
12. Chew, M., Baird, H.S.: Baffletext: A human interactive proof. In: Proceedings of SPIE-IS&T Electronic Imaging, Document Recognition and Retrieval X, pp. 305–316 (2003)
13. Markkola, A., Lindqvist, J.: Accessible Voice CAPTCHAs for Internet Telephony. In: The Symposium on Accessible Privacy and Security (SOAPS 2008) (2008)
14. Chew, M., Tygar, J.D.: Image recognition CAPTCHAs. In: Zhang, K., Zheng, Y. (eds.) ISC 2004. LNCS, vol. 3225, pp. 268–279. Springer, Heidelberg (2004)
15. Elson, J., Douceur, J., Howell, J., Saul, J.: Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. In: Proceedings of 14th ACM Conference on Computer and Communications Security, pp. 366–374 (2007)

# **TAIEX Forecasting Based on Fuzzy Time Series and the Automatically Generated Weights of Defuzzified Forecasted Fuzzy Variations of Multiple-Factors**

Shyi-Ming Chen and Huai-Ping Chu

Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan, R. O. C.

**Abstract.** This paper presents a new method to forecast the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) based on fuzzy time series and the automatic generated weights of defuzzified forecasted fuzzy variations of multiple-factors. The proposed method uses the variation magnitude of adjacent historical data to generate fuzzy variation groups of the main factor (i.e., the TAIEX) and the elementary secondary factors (i.e., the Dow Jones, the NASDAQ and the M1B), respectively. Based on the variation magnitudes of the main factor TAIEX and the elementary secondary factors of a particular trading day, it gets the forecasted variation of the TAIEX of the next trading day forecasted by each factor. Based on the correlation coefficients between the forecasted fuzzy variation of the main factor and the forecasted fuzzy variation of each elementary secondary factor, it automatically generates the weights of the defuzzified forecasted fuzzy variation of the main factor and the defuzzified forecasted fuzzy variation of each elementary secondary factor, respectively. Based on the closing index of the TAIEX of the trading day and the weighted forecasted variation, it generates the final forecasted value of the next trading day. The experimental results show that the proposed method gets higher average forecasting accuracy rates than the existing methods.

**Keywords:** Elementary Secondary Factors, Forecasted Fuzzy Variations, Fuzzy Time Series, Fuzzy Variation Groups, Main Factor, Secondary Factors, TAIEX, Variation Magnitude, Correlation Coefficients.

## **1 Introduction**

Since Song and Chissom proposed the concepts of fuzzy time series [8], some fuzzy forecasting methods based on fuzzy time series have been presented for different applications [1], [3], [4], [5], [6], [9], [10].

In this paper, we present a new method to deal with the forecasting of the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) [13] based on fuzzy time series and the automatic generated weights of defuzzified forecasted fuzzy variations of multiple-factors. The proposed method uses the variation magnitude of adjacent historical data to generate fuzzy variation groups of the main factor (i.e., the

TAIEX) and the elementary secondary factors (i.e., the Dow Jones, the NASDAQ and the M1B), respectively. Based on the variation magnitudes of the main factor and the elementary secondary factors of a particular trading day, it gets the forecasted variation of the TAIEX of the next trading day forecasted by each factor. Based on the correlation coefficients between the forecasted fuzzy variation of the main factor and the forecasted fuzzy variation of each elementary secondary factor, it automatically generates the weights of the defuzzified forecasted fuzzy variation of the main factor and the defuzzified forecasted fuzzy variation of each elementary secondary factor, respectively. Based on the closing index of the TAIEX of the trading day and the weighted forecasted variation, it generates the final forecasting value of the next trading day. The experimental results show that the proposed method gets higher average forecasted accuracy rates than the existing methods.

## 2 Fuzzy Time Series

In [8] and [9], Song and Chissom proposed the concept of fuzzy time series based on the fuzzy set theory [12]. Let  $U$  be the universe of discourse, where  $U = \{u_1, u_2, \dots, u_n\}$ . A fuzzy set  $A$  of  $U$  is defined as follows:

$$A = f_A(u_1)/u_1 + f_A(u_2)/u_2 + \dots + f_A(u_n)/u_n,$$

where  $f_A$  is the membership function of the fuzzy set  $A$ ,  $f_A(u_i)$  denotes the degree of membership of  $u_i$  belonging to  $A$ ,  $f_A(u_i) \in [0,1]$  and  $1 \leq i \leq n$ .

**Definition 2.1 [8]:** Let  $Y(t)$  ( $t = \dots, 0, 1, 2, \dots$ ) be the universe of discourse in which fuzzy sets  $f_i(t)$  ( $i = 1, 2, \dots$ ) are defined. Let  $F(t)$  be a collection of  $f_i(t)$  ( $i = 1, 2, \dots$ ). Then,  $F(t)$  is called a fuzzy time series on  $Y(t)$  ( $t = \dots, 0, 1, 2, \dots$ ).

**Definition 2.2 [8]:** Let  $F(t - 1) = A_i$  and  $F(t) = A_j$ , where  $A_i$  and  $A_j$  are fuzzy sets. The fuzzy logical relationship (FLR) between  $F(t-1)$  and  $F(t)$  can be denoted by  $A_i \rightarrow A_j$ , where  $A_i$  and  $A_j$  are called the left-hand side (LHS) and the right-hand side (RHS) of the fuzzy logical relationship, respectively.

**Definition 2.3 [1]:** Assume that there are fuzzy logical relationships with the same left-hand side “ $A_i$ ”, shown as follows:

$$\begin{aligned} A_i &\rightarrow A_{j1}, \\ A_i &\rightarrow A_{j2}, \\ &\vdots \\ A_i &\rightarrow A_{jn}, \end{aligned}$$

then these fuzzy logical relationships can be grouped into the fuzzy logical relationship group “Group  $A_i$ ”, shown as follows:

$$A_i \rightarrow A_{j1}, A_{j2}, \dots, A_{jn}.$$

### 3 A New Method for the TAIEX Forecasting Based on Fuzzy Time Series and the Automatically Generated Weights of Defuzzified Forecasted Fuzzy Variations of Multiple-Factors

In this section, we present a new method for forecasting the TAIEX [13] based on fuzzy time series and the automatic generated weights of defuzzified forecasted fuzzy variations of multiple-factors. We let the TAIEX be the main factor and let the Dow Jones, the NASDAQ and the M1B be the elementary secondary factors. For each year, the data from January to October are used as the training data set and the data from November to December are used as the testing data set. The proposed method is presented as follows:

**Step 1:** Calculate the variation magnitude of the closing indices between the adjacent historical training data of the main factor and calculate the variation magnitude of the closing indices between the adjacent historical training data of each elementary secondary factor, respectively. The variation magnitude  $VM_t$  of the closing indices between the adjacent historical training data of each factor on trading day  $t$  is calculated as follows:

$$VM_t = \frac{Close_t - Close_{t-1}}{Close_{t-1}} \times 100\% , \quad (1)$$

where  $Close_t$  and  $Close_{t-1}$  denote the closing indices on trading day  $t$  and trading day  $t - 1$ , respectively.

**Step 2:** Because the “maximum negative fluctuation” and the “maximum positive fluctuation” of the variation magnitude of the main factor TAIEX are -7% and 7%, respectively, we let the universe of discourse  $U$  of the main factor TAIEX be [-7%, 7%] and partition the universe of discourse  $U$  into fourteen intervals  $u_1, u_2, \dots$ , and  $u_{14}$ , where  $u_1 = [-7\%, -6\%]$ ,  $u_2 = [-6\%, -5\%]$ ,  $u_3 = [-5\%, -4\%]$ ,  $u_4 = [-4\%, -3\%]$ ,  $u_5 = [-3\%, -2\%]$ ,  $u_6 = [-2\%, -1\%]$ ,  $u_7 = [-1\%, 0\%]$ ,  $u_8 = [0\%, 1\%]$ ,  $u_9 = [1\%, 2\%]$ ,  $u_{10} = [2\%, 3\%]$ ,  $u_{11} = [3\%, 4\%]$ ,  $u_{12} = [4\%, 5\%]$ ,  $u_{13} = [5\%, 6\%]$  and  $u_{14} = [6\%, 7\%]$ . Because there are no limitations on the variation magnitude of the Dow Jones and because the variation magnitude of the Dow Jones is usually within a specific range, we let the universes of discourse  $V$  of the elementary secondary factor Dow Jones be  $(-\infty, \infty)$  and partition the universe of discourse  $V$  into fourteen intervals  $v_1, v_2, \dots$ , and  $v_{14}$ , where  $v_1 = (-\infty, -6\%)$ ,  $v_2 = [-6\%, -5\%]$ ,  $v_3 = [-5\%, -4\%]$ ,  $v_4 = [-4\%, -3\%]$ ,  $v_5 = [-3\%, -2\%]$ ,  $v_6 = [-2\%, -1\%]$ ,  $v_7 = [-1\%, 0\%]$ ,  $v_8 = [0\%, 1\%]$ ,  $v_9 = [1\%, 2\%]$ ,  $v_{10} = [2\%, 3\%]$ ,  $v_{11} = [3\%, 4\%]$ ,  $v_{12} = [4\%, 5\%]$ ,  $v_{13} = [5\%, 6\%]$  and  $v_{14} = [6\%, \infty]$ . Because there are no limitations on the variation magnitude of the NASDAQ and because the variation magnitude of the NASDAQ is usually within a specific range, we let the universes of discourse  $W$  of the elementary secondary factor NASDAQ be  $(-\infty, \infty)$  and partition the universe of discourse  $W$  into fourteen intervals  $w_1, w_2, \dots$ , and  $w_{14}$ , where  $w_1 = (-\infty, -6)$ ,  $w_2 = [-6\%, -5\%]$ ,  $w_3 = [-5\%, -4\%]$ ,  $w_4 = [-4\%, -3\%]$ ,  $w_5 = [-3\%, -2\%]$ ,  $w_6 = [-2\%, -1\%]$ ,  $w_7 = [-1\%, 0\%]$ ,  $w_8 = [0\%, 1\%]$ ,  $w_9 = [1\%, 2\%]$ ,  $w_{10} = [2\%, 3\%]$ ,  $w_{11} = [3\%, 4\%]$ ,  $w_{12} = [4\%, 5\%]$ ,  $w_{13} = [5\%, 6\%]$  and  $w_{14} = [6\%, \infty]$ . Because there are no limitations on the variation magnitude of the M1B and because the variation magnitude of the elementary

secondary factor M1B distributes irregularly, we let the universes of discourse  $Z$  of the elementary secondary factor M1B be  $(-\infty, \infty)$  and partition the universe of discourse  $Z$  into fourteen intervals  $z_1, z_2, \dots$ , and  $z_{14}$ . By analyzing the historical training data of the universe of discourse  $Z$  of the elementary secondary factor M1B, the length of each interval in the universe can be determined, described as follows. By taking the absolute value  $R$  of the largest variation magnitude as the range of possible changes in either the positive variation magnitude or the negative variation magnitude, we partition the universe of discourse  $Z$  into 14 intervals  $z_1, z_2, \dots$ , and  $z_{14}$ , where  $z_1 = (-\infty, -R\%)$ ,  $z_{14} = [R\%, \infty)$ . Besides the intervals  $z_1 = (-\infty, -R\%)$  and  $z_{14} = [R\%, \infty)$ , we partition the interval  $[-R\%, R\%]$  in the universe of discourse  $Z$  into 12 intervals  $z_2, z_3, \dots$ , and  $z_{13}$ , where the length  $l$  of each interval in the interval  $[-R\%, R\%]$  can be calculated, shown as follows:

$$l = \left\lceil \frac{2 \times R}{12} \right\rceil.$$

Thus, we can partition the universe of discourse  $Z$  of the elementary secondary factor M1B into fourteen intervals  $z_1, z_2, \dots$ , and  $z_{14}$ , where  $z_1 = (-\infty, -R\%)$ ,  $z_2 = [-R\%, -5l\%)$ ,  $z_3 = [-5l\%, -4l\%)$ ,  $z_4 = [-4l\%, -3l\%)$ ,  $z_5 = [-3l\%, -2l\%)$ ,  $z_6 = [-2l\%, -l\%)$ ,  $z_7 = [-l\%, 0\%)$ ,  $z_8 = [0\%, l\%)$ ,  $z_9 = [l\%, 2l\%)$ ,  $z_{10} = [2l\%, 3l\%)$ ,  $z_{11} = [3l\%, 4l\%)$ ,  $z_{12} = [4l\%, 5l\%)$ ,  $z_{13} = [5l\%, R\%)$  and  $z_{14} = [R\%, \infty)$ .

**Step 3:** Define the linguistic terms  $A_1, A_2, A_3, \dots, A_{13}$  and  $A_{14}$  of the main factor TAIEX represented by fuzzy sets, respectively, shown as follows:

$$\begin{aligned} A_1 &= 1/u_1 + 0.5/u_2 + 0/u_3 + \dots + 0/u_{12} + 0/u_{13} + 0/u_{14}, \\ A_2 &= 0.5/u_1 + 1/u_2 + 0.5/u_3 + \dots + 0/u_{12} + 0/u_{13} + 0/u_{14}, \\ &\vdots \\ A_{14} &= 0/u_1 + 0/u_2 + 0/u_3 + \dots + 0/u_{12} + 0.5/u_{13} + 1/u_{14}. \end{aligned}$$

Define the linguistic terms  $B_1, B_2, B_3, \dots, B_{13}$  and  $B_{14}$  of the elementary secondary factor Dow Jones represented by fuzzy sets, respectively, shown as follows:

$$\begin{aligned} B_1 &= 1/v_1 + 0.5/v_2 + 0/v_3 + \dots + 0/v_{12} + 0/v_{13} + 0/v_{14}, \\ B_2 &= 0.5/v_1 + 1/v_2 + 0.5/v_3 + \dots + 0/v_{12} + 0/v_{13} + 0/v_{14}, \\ &\vdots \\ B_{14} &= 0/v_1 + 0/v_2 + 0/v_3 + \dots + 0/v_{12} + 0.5/v_{13} + 1/v_{14}. \end{aligned}$$

Define the linguistic terms  $C_1, C_2, C_3, \dots, C_{13}$  and  $C_{14}$  of the elementary secondary factor NASDAQ represented by fuzzy sets, respectively, shown as follows:

$$\begin{aligned} C_1 &= 1/w_1 + 0.5/w_2 + 0/w_3 + \dots + 0/w_{12} + 0/w_{13} + 0/w_{14}, \\ C_2 &= 0.5/w_1 + 1/w_2 + 0.5/w_3 + \dots + 0/w_{12} + 0/w_{13} + 0/w_{14}, \\ &\vdots \\ C_{14} &= 0/w_1 + 0/w_2 + 0/w_3 + \dots + 0/w_{12} + 0.5/w_{13} + 1/w_{14}. \end{aligned}$$

Define the linguistic terms  $D_1, D_2, D_3, \dots, D_{13}$  and  $D_{14}$  of the elementary secondary factor M1B represented by fuzzy sets, respectively, shown as follows:

$$\begin{aligned} D_1 &= 1/z_1 + 0.5/z_2 + 0/z_3 + \dots + 0/z_{12} + 0/z_{13} + 0/z_{14}, \\ D_2 &= 0.5/z_1 + 1/z_2 + 0.5/z_3 + \dots + 0/z_{12} + 0/z_{13} + 0/z_{14}, \\ &\vdots \\ D_{14} &= 0/z_1 + 0/z_2 + 0/z_3 + \dots + 0/z_{12} + 0.5/z_{13} + 1/z_{14}. \end{aligned}$$

**Step 4:** Fuzzify the variation magnitude between the adjacent historical training data of the main factor and fuzzify the variation magnitude between the adjacent historical training data of the elementary secondary factors, respectively, to get fuzzy logical relationship groups of the main factor and the elementary secondary factors, respectively. The sub-steps of **Step 4** are shown as follows:

**Step 4.1:** Fuzzify the variation magnitude of the main factor on each trading day into a fuzzy set defined in **Step 3**. If the variation magnitude of the main factor on trading day  $t$  belongs to  $u_i$ , where  $1 \leq i \leq 14$ , then the variation magnitude of the main factor on trading day  $t$  is fuzzified into  $A_i$ , where  $1 \leq i \leq 14$ .

**Step 4.2:** Fuzzify the variation magnitude of each elementary secondary factor on each trading day into a fuzzy set defined in **Step 3**. For example, if the variation magnitude of the elementary secondary factor “Dow Jones” on trading day  $t$  belongs to  $v_j$ , where  $1 \leq j \leq 14$ , then the variation magnitude of the elementary secondary factor “Dow Jones” on trading day  $t$  is fuzzified into  $B_j$ , where  $1 \leq j \leq 14$ .

**Step 4.3:** Group the fuzzy variations of the main factor on the trading days having the same fuzzy variations of the main factor on the previous trading day into a fuzzy variation group. For example, assume that the fuzzy variation of the main factor on trading day  $t$  is  $A_2$  and assume that the fuzzy variation of the main factor on trading day  $t - 1$  is  $A_1$ ; assume that the fuzzy variation of the main factor on trading day  $t + m$  is  $A_3$  and assume that the fuzzy variation of the main factor on trading day  $t + m - 1$  is  $A_1$ . Then, the fuzzy variations  $A_2$  and  $A_3$  can be grouped into “Group  $A_1$ ” of the main factor, shown as follows:

$$\text{Group } A_1: A_2, A_3.$$

**Step 4.4:** Group the fuzzy variations of the main factor on the trading days having the same fuzzy variation of an elementary secondary factor on the previous trading day into a fuzzy variation group. For example, assume that the fuzzy variation of the main factor on trading day  $t$  is  $A_2$  and assume that the fuzzy variation of the elementary secondary factor on trading day  $t - 1$  is  $B_1$ ; assume that the fuzzy variation of the main factor on trading day  $t + n$  is  $A_3$  and assume that the fuzzy variation of the elementary secondary factor on trading day  $t + n - 1$  is  $B_1$ . Then, the fuzzy variations  $A_2$  and  $A_3$  can be grouped into “Group  $B_1$ ” of the elementary secondary factor, shown as follows:

$$\text{Group } B_1: A_2, A_3.$$

**Step 5:** Generate the forecasted fuzzy variation of the main factor and the forecasted fuzzy variations of the elementary secondary factors, respectively, described as follows. Assume that the fuzzy variation of the main factor on trading day  $t-1$  is  $A_m$  and assume that the fuzzy variation of the elementary secondary factor on trading day  $t-1$  is  $B_s$ , where  $m$  and  $s$  are positive integers,  $1 \leq m \leq 14$  and  $1 \leq s \leq 14$ . Let  $A_{m1}, A_{m2}, \dots$ , and  $A_{mp}$  be the fuzzy variations appearing in the fuzzy variation group “Group  $A_m$ ” of the main factor and let  $N_{m1}, N_{m2}, \dots$ , and  $N_{mp}$  denote the number of occurrences of fuzzy variations  $A_{m1}, A_{m2}, \dots$ , and  $A_{mp}$  in Group  $A_m$ , respectively; let  $A_{s(i)1}, A_{s(i)2}, \dots$ , and  $A_{s(i)q}$  be the fuzzy variations appearing in the fuzzy variation group “Group  $B_s$ ” of the  $i$ th elementary secondary factor and let  $N_{s(i)1}, N_{s(i)2}, \dots$ , and  $N_{s(i)q}$  denote the

number of occurrences of fuzzy variations  $A_{s(i)1}$ ,  $A_{s(i)2}$ , ..., and  $A_{s(i)q}$  in Group  $B_s$ , respectively, where  $p$  and  $q$  are positive integers,  $1 \leq p \leq 14$  and  $1 \leq q \leq 14$ . Then, we define two fuzzy sets  $V_M(t)$  and  $V_{S(i)}(t)$ , respectively, shown as follows:

$$V_M(t) = N_{m1}/A_{m1} + N_{m2}/A_{m2} + \dots + N_{mp}/A_{mp}, \\ V_{S(i)}(t) = N_{s(i)1}/A_{s(i)1} + N_{s(i)2}/A_{s(i)2} + \dots + N_{s(i)q}/A_{s(i)q},$$

where  $V_M(t)$  denotes the forecasted fuzzy variation of the main factor on trading day  $t$  derived from the fuzzy variation group “Group  $A_m$ ” of the main factor;  $V_{S(i)}(t)$  denotes the forecasted fuzzy variation of the main factor on trading day  $t$  derived from the fuzzy variation group “Group  $B_s$ ” of the  $i$ th elementary secondary factor. Then, we define the fuzzy set  $V_M(t)$  denoting the forecasted fuzzy variation of the main factor on trading day  $t$  and define the fuzzy set  $V_{S(i)}(t)$  denoting the forecasted fuzzy variation of the  $i$ th elementary secondary factor on trading day  $t$ , respectively.

**Step 6:** Based on the fuzzy variations appearing in the forecasted fuzzy variations  $V_M(t)$  and  $V_{S(i)}(t)$  on trading day  $t$ , respectively, construct the numerical data series  $T_{VM(t)}$  from the forecasted fuzzy variation  $V_M(t)$  and construct the numerical data series  $T_{VS(i)(t)}$  from the forecasted fuzzy variation  $V_{S(i)}(t)$ . If the number of occurrences of the fuzzy variation  $A_{mp}$  appearing in the forecasted fuzzy variation  $V_M(t)$  is  $N_{mp}$ , then set the corresponding value in  $T_{VM(t)}$  to  $N_{mp}$ , where  $p = 1, 2, \dots, 14$ . If the number of occurrences of the fuzzy variation  $A_{s(i)q}$  appearing in the forecasted fuzzy variation  $V_{S(i)}(t)$  is  $N_{s(i)q}$ , then set the corresponding value in  $T_{VS(i)(t)}$  to  $N_{s(i)q}$ , where  $q = 1, 2, \dots, 14$ . Therefore, the corresponding numerical data series  $T_{VM(t)}$  and  $T_{VS(i)(t)}$  for the forecasted fuzzy variations  $V_M(t)$  and  $V_{S(i)}(t)$  on trading day  $t$  can be constructed, respectively, shown as follows:

$$T_{VM(t)} = (N_{m1}, N_{m2}, N_{m3}, N_{m4}, N_{m5}, N_{m6}, N_{m7}, N_{m8}, N_{m9}, N_{m10}, N_{m11}, N_{m12}, N_{m13}, N_{m14}), \quad (2)$$

$$T_{VS(i)(t)} = (N_{s(i)1}, N_{s(i)2}, N_{s(i)3}, N_{s(i)4}, N_{s(i)5}, N_{s(i)6}, N_{s(i)7}, N_{s(i)8}, N_{s(i)9}, N_{s(i)10}, N_{s(i)11}, N_{s(i)12}, N_{s(i)13}, N_{s(i)14}), \quad (3)$$

where  $B_{mp} \in \{0, 1\}$ ,  $B_{s(i)q} \in \{0, 1\}$ ,  $1 \leq p \leq 14$  and  $1 \leq q \leq 14$ . Calculate the correlation coefficient  $r_i$  between the numerical data series  $T_{VM(t)} = (x_1, x_2, \dots, x_n)$  of the main factor and the numerical data series  $T_{VS(i)(t)} = (y_1, y_2, \dots, y_n)$  of the  $i$ th elementary secondary factor, respectively, shown as follows [7]:

$$r_i = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (4)$$

where  $\bar{x}$  denotes the mean value of the numerical data series  $T_{VM(t)} = (x_1, x_2, \dots, x_n)$  of the main factor,  $\bar{y}$  denotes the mean value of the numerical data series  $T_{VS(i)(t)} = (y_1, y_2, \dots, y_n)$  of the  $i$ th elementary secondary factor and  $r_i \in [-1, 1]$ . If  $r_i > 0$ , then we adopt the  $i$ th elementary secondary factor to assist the main factor in forecasting the variation of the TAIEX of the next trading day. Otherwise, we do not adopt the  $i$ th

elementary secondary factor to assist the main factor in forecasting the variation of the TAIEX of the next trading day.

**Step 7:** Calculate the weight  $W_{VM(t)}$  of the forecasted fuzzy variation  $V_M(t)$  of the main factor on trading day  $t$  and calculate the weight  $W_{VS(i)(t)}$  of the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$ , respectively, shown as follows:

$$W_{VM(t)} = \frac{1}{1 + \sum_{\substack{i=1,2,\dots,n \\ r_i > 1}} r_i}, \quad (5)$$

$$W_{VS(i)(t)} = \frac{r_i}{1 + \sum_{\substack{i=1,2,\dots,n \\ r_i > 1}} r_i}, \quad (6)$$

where  $n$  denotes the number of elementary secondary factors,  $r_i$  denotes the correlation coefficient between the numerical data series  $T_{VM(t)}$  (constructed according to the forecasted fuzzy variation  $V_M(t)$  of the main factor on trading day  $t$ ) and the numerical data series  $T_{VS(i)(t)}$  (constructed according to the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$ ), and  $1 \leq i \leq n$ . If  $r_i \leq 0$ , then it indicates that there is no correlation or a negative correlation between the numerical data series  $T_{VM(t)}$  (constructed according to the forecasted fuzzy variation  $V_M(t)$  of the main factor on trading day  $t$ ) and the numerical data series  $T_{VS(i)(t)}$  (constructed according to the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$ ). In this situation, the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor will be of no help for forecasting the fuzzy variation  $V_M(t)$  of the main factor and we do not adopt this elementary secondary factor to help the main factor for forecasting the variation  $V_M(t)$  of the main factor on trading day  $t$ . If  $r_i > 0$ , then according to the correlation coefficient between the numerical data series  $T_{VM(t)}$  (constructed according to the forecasted fuzzy variation  $V_M(t)$  of the main factor on trading day  $t$ ) and the numerical data series  $T_{VS(i)(t)}$  (constructed according to the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$ ), we can determine the weights of the forecasted variation of the main factor and each elementary secondary factor for the prediction, respectively.

**Step 8:** Defuzzify the forecasted fuzzy variation of the main factor on training day  $t$  and defuzzify the forecasted fuzzy variations of the elementary secondary factors on training day  $t$ , respectively, and then integrate them to generate the final forecasted fuzzy variation of the main factor on training day  $t$ . Assume that the forecasted fuzzy variation  $V_M(t)$  of the main factor and the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$  are as follows:

$$\begin{aligned} V_M(t) &= N_{m1}/A_{m1} + N_{m2}/A_{m2} + \dots + N_{mp}/A_{mp}, \\ V_{S(i)}(t) &= N_{s(i)1}/A_{s(i)1} + N_{s(i)2}/A_{s(i)2} + \dots + N_{s(i)q}/A_{s(i)q}, \end{aligned}$$

then the weight  $W_{Amh}$  of the fuzzy variation  $A_{mh}$  in  $V_M(t)$  and the weight  $W_{As(i)k}$  of the fuzzy variation  $A_{s(i)k}$  in  $V_{S(i)}(t)$  are calculated as follows:

$$W_{Amh} = \frac{N_{mh}}{\sum_{h=1}^p N_{mh}}, \quad (7)$$

$$W_{As(i)k} = \frac{N_{s(i)k}}{\sum_{k=1}^q N_{s(i)k}}, \quad (8)$$

where  $1 \leq h \leq p$  and  $1 \leq k \leq q$ . The defuzzified forecasted fuzzy variation  $FVar_M(t)$  of  $V_M(t)$  and the defuzzified forecasted fuzzy variation  $FVar_{S(i)}(t)$  of  $V_{S(i)}(t)$  are calculated as follows:

$$FVar_M(t) = \sum_{h=1}^p (W_{Amh} \times u_h^M), \quad (9)$$

$$FVar_{S(i)}(t) = \sum_{k=1}^q (W_{As(i)k} \times u_k^M), \quad (10)$$

where  $u_h^M$  denotes the midpoint of the interval  $u_h$ , the maximum membership value of linguistic term  $A_{mh}$  occurs at interval  $u_h$  and  $1 \leq h \leq p$ ;  $u_k^M$  denotes the midpoint of the interval  $u_k$ , the maximum membership value of linguistic term  $A_{S(i)k}$  occurs at interval  $u_k$  and  $1 \leq k \leq q$ . Then, based on the weight  $W_{VM(t)}$  of the forecasted fuzzy variation  $V_M(t)$  of the main factor on trading day  $t$ , based on the weight  $W_{VS(i)(t)}$  of the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$  obtained in **Step 7**, based on the defuzzified forecasted fuzzy variation  $FVar_M(t)$  of the main factor on trading day  $t$  and based on the defuzzified forecasted fuzzy variation  $FVar_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$ , where  $1 \leq i \leq n$  and  $n$  denotes the number of elementary secondary factors, calculate the final forecasted variation  $FVar(t)$  of the main factor on trading day  $t$ , shown as follows:

$$FVar_F(t) = W_{VM(t)} \times FVar_M(t) + \sum_{i=1}^n (W_{VS(i)(t)} \times FVar_{S(i)}(t)), \quad (11)$$

where  $W_{VM(t)}$  denotes the weight of the forecasted fuzzy variation  $V_M(t)$  of the main factor on trading day  $t$  and  $FVar_M(t)$  denotes the defuzzified forecasted fuzzy variation of the main factor on trading day  $t$ ;  $W_{VS(i)(t)}$  denotes the weight of the forecasted fuzzy variation  $V_{S(i)}(t)$  of the  $i$ th elementary secondary factor on trading day  $t$  and  $FVar_{S(i)}(t)$  denotes the defuzzified forecasted fuzzy variation of the  $i$ th elementary secondary factor on trading day  $t$ , where  $1 \leq i \leq n$  and  $n$  denotes the number of elementary secondary factors.

**Step 9:** Calculate the forecasted value  $F(t)$  of the main factor on trading day  $t$ , shown as follows:

$$F(t) = \begin{cases} Close_{t-1} \times (1 + FVar_F(t)), & \text{if } FVar_F(t) \neq 0 \\ Close_{t-1} \times (1 + \frac{V_M(t-1)}{3}), & \text{otherwise} \end{cases} \quad (12)$$

where  $Close_{t-1}$  denotes the closing index of the main factor on trading day  $t - 1$ .

## 4 Experimental Results

In this section, we apply the proposed method to forecast the TAIEX from 1999 to 2004, where we let the TAIEX be the main factor and let the combination of the Dow Jones, the NASDAQ and the M1B be the secondary factors, respectively. The data from January to October for each year are used as the training data and the data in November and December for each year are used as the testing data. We evaluate the performance of the proposed method using the root mean square error (RMSE), which is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (forecasted\ value_i - actual\ value_i)^2}{n}}, \quad (13)$$

where  $n$  denotes the number of dates needed to be forecasted, “*forecasted value<sub>i</sub>*” denotes the forecasted value of day  $i$ , “*actual value<sub>i</sub>*” denotes the actual value of day  $i$  and  $1 \leq i \leq n$ . From Table 1, we can see that the proposed method outperforms the existing methods for forecasting the TAIEX.

**Table 1.** A comparison of the RMSEs and the average RMSE for different methods

Methods	RMSE	Year						Average RMSE
		1999	2000	2001	2002	2003	2004	
Huang et al.’s Method [5] (Use NASDAQ)	N/A	158.7	136.49	95.15	65.51	73.57		105.88
Huang et al.’s Method [5] (Use Dow Jones)	N/A	165.8	138.25	93.73	72.95	73.49		108.84
Huang et al.’s Method [5] (Use M1B)	N/A	169.19	133.26	97.1	75.23	82.01		111.36
Huang et al.’s Method [5] (Use NASDAQ & Dow Jones)	N/A	157.64	131.98	93.48	65.51	73.49		104.42
Huang et al.’s Method [5] (Use NASDAQ & M1B)	N/A	155.51	128.44	97.15	70.76	73.48		105.07
Huang et al.’s Method [5] (Use NASDAQ & Dow Jones & M1B)	N/A	154.42	124.02	95.73	70.76	72.35		103.46
Chen’s Fuzzy Time Series Model (U_FTS Model) [1], [10], [11]	120	176	148	101	74	84		117.4
Univariate Conventional Regression Model (U_R Model) [10], [11]	164	420	1070	116	329	146		374.2
Univariate Neural Network Model (U_NN Model) [10], [11]	107	309	259	78	57	60		145.0
Univariate Neural Network-Based Fuzzy Time Series Model (U_NN_FTS Model) [4], [10], [11]	109	255	130	84	56	116		125.0
Univariate Neural Network-Based Fuzzy Time Series Model with Substitutes (U_NN_FTS_S Model) [4], [10], [11]	109	152	130	84	56	116		107.8
Bivariate Conventional Regression Model (B_R Model) [10], [11]	103	154	120	77	54	85		98.8
Bivariate Neural Network Model (B_NN Model) [10], [11]	112	274	131	69	52	61		116.4
Bivariate Neural Network-Based Fuzzy Time Series Model (B_NN_FTS model) [10], [11]	108	259	133	85	58	67		118.3
* Bivariate Neural Network-Based Fuzzy Time Series Model with Substitutes (B_NN_FTS_S Model) [10], [11]	112	131	130	80	58	67		96.4
Chen and Chen’s Method [2]	Use Dow Jones	N/A	127.51	121.98	74.65	66.02	58.89	89.81
	Use NASDAQ	N/A	129.87	123.12	71.01	65.14	61.94	90.22
	Use M1B	N/A	129.87	117.61	85.85	63.1	67.29	92.74
	Use Dow Jones & NASDAQ	N/A	124.06	125.12	72.25	57.14	56.95	87.10
	Use Dow Jones & M1B	N/A	127.75	115.64	79.45	60.41	65.86	89.82
The Proposed Method	Use NASDAQ & M1B	N/A	128.45	126.14	76.03	66.96	65.5	92.62
	Use NASDAQ & Dow Jones & M1B	N/A	129.57	119.66	73.25	66.8	65.41	90.94
	Use Dow Jones	99.87	122.75	117.18	68.45	53.96	52.55	85.79
	Use NASDAQ	102.6	119.98	114.81	69.07	53.16	53.57	85.53
	Use M1B	101.22	123.99	117.75	70.63	54.92	55.29	87.3
	Use Dow Jones & NASDAQ	101.33	121.27	114.48	67.18	52.72	52.27	84.88
	Use Dow Jones & M1B	100.59	124.1	116.28	68.11	53.5	53.33	85.99
	Use NASDAQ & M1B	102.25	122.47	115.02	68.51	52.82	53.99	85.84
	Use NASDAQ & Dow Jones & M1B	101.47	122.88	114.47	67.17	52.49	52.84	85.22

(\*Note: In [11], Yu and Huarng have corrected the typing errors appearing in Table 6 of [10]).

## 5 Conclusions

In this paper, we have presented a new method for forecasting the TAIEX based on fuzzy time series and the automatic generated weights of defuzzified forecasted fuzzy variations of multiple-factors. From the experimental results shown in Table 1, we can see that the proposed method gets higher average forecasting accuracy rates than the existing methods.

## Acknowledgements

This work was supported in part by the National Science Council, Republic of China, under Grant NSC 97-2221-E-011-107-MY3.

## References

1. Chen, S.M.: Forecasting Enrollments Based on Fuzzy Time Series. *Fuzzy Sets and Systems* 81, 311–319 (1996)
2. Chen, C.D., Chen, S.M.: A New Method to Forecast the TAIEX Based on Fuzzy Time Series. In: Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, San Antonio, Texas, pp. 3550–3555 (2009)
3. Chen, S.M., Hwang, J.R.: Temperature Prediction Using Fuzzy Time Series. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 30, 263–275 (2000)
4. Huarng, K., Yu, T.H.K.: The Application of Neural Networks to Forecast Fuzzy Time Series. *Physica A* 363, 481–491 (2006)
5. Huarng, K., Yu, H.K., Hsu, Y.W.: A Multivariate Heuristic Model for Fuzzy Time-Series Forecasting. *IEEE Transactions on Systems, Man, and Cybernetics Part-B: Cybernetics* 37, 836–846 (2007)
6. Lee, L.W., Wang, L.H., Chen, S.M., Leu, Y.H.: Handling Forecasting Problems Based on Two-Factors High-Order Fuzzy Time Series. *IEEE Transactions on Fuzzy Systems* 14, 468–477 (2006)
7. Lohninger, H.: Teach/Me Data Analysis. Springer, Germany (1999)
8. Song, Q., Chissom, B.S.: Fuzzy Time Series and Its Model. *Fuzzy Sets and Systems* 54, 269–277 (1993)
9. Song, Q., Chissom, B.S.: Forecasting Enrollments with Fuzzy Time Series - Part I. *Fuzzy Sets and Systems* 54, 1–9 (1993)
10. Yu, T.H.K., Huarng, K.H.: A Bivariate Fuzzy Time Series Model to Forecast the TAIEX. *Expert Systems with Applications* 34, 2945–2952 (2008)
11. Yu, T.H.K., Huarng, K.H.: Corrigendum to A Bivariate Fuzzy Time Series Model to Forecast the TAIEX. *Expert Systems with Applications* 34(4), 2945–2952 (2010)
12. Zadeh, L.A.: Fuzzy Sets Information and Control, vol. 8, pp. 338–353 (1965)
13. TAIEX,  
<http://www.twse.com.tw/en/products/indices/tsec/taiex.php>

# Concept Document Repository to Support Research of the Coal Industry Development Forecasting

Liudmila Takayshvili

Energy Systems Institute, SB RAS, Lermontov street, 130,  
664033 Irkutsk, Russia  
luci@isem.sei.irk.ru

**Abstract.** The paper shows the necessity of creating a Document Repository. The analysis of composition and structure of stored documents to be used for research of coal industry development is made. Main principles of Document Repository are formulated. Problems of document integration are outlined. Issues related to implementation of the Document Repository are touched upon.

**Keywords:** Document, Document Repository, Coal Industry Development, Documents Mart, Metadata, Concepts, Classifications.

## 1 Introduction

The forecasting the development and operation of Coal Industry as part of Fuel and Energy Complex includes several stages of research:

1. Analysis of the current condition of Coal Industry.
2. The development of economy scenarios (for country or region).
3. The forecast demand for electricity and heat.
4. The forecast demand for fuel, depending on the scenario accepted.
5. Calculate the volume of fuel demand, taking into account current trends in production, supply, consumption, available resources of fuel, the capacity of the transport infrastructure and other factors.
6. The Forecast of Coal Industry development on the basis of the balance of coal.

For stages 4-6 developed versions of software and dataware [1-3]. As for stages 1 and 2, these directions are not enough developed. Stage 1 is connected with analysis of the large volume of documents, presented in the form of articles, reports, statistical reviews and etc. Orientation in the information flow - one of the main problems in any areas of activities, including the research of Coal Industry development.

This article is devoted to consideration of the possible direction of the partial automation of Research on the first stage, taking into account that the process of research is low formalizable [4], and for the first stage – especially.

Commonly, documents, used in research, come from different sources, or have been created by a user in the process of activity. There is problems with choice of the most

suitable structure of keeping documents for orientation in a multitude of documents. Since time, different multitudes of documents begin to be difficult identifiable archive. Appear problems with searching not only fragment document, but also document themselves.

In have recently got development to the concept of data warehouse, workflow system and document management systems [5,6]. Current systems work with documents contain many functions not called-up when performing scientific research. Analysis of such systems has shown that it is not rational to adapt them for research. In these systems, organization work with the document is subordinated certain business-process, absent in the performance of scientific research. Scientific studies characterize other processes. A specific of the scientific studies necessitate the organization of certain procedures for working with documents that are unique to this type of activity. The heterogeneity of internal and external documents and forms of documents used in research requires the organizations very flexible and adequate system of storing and retrieval of documents and providing opportunities to create new documents based on existing ones. Herewith document templates do not exist. The task of Document Repository is several another, than the data warehouse, workflow system and document management systems, though in tasks of these systems are similar in than-that. Document Repository focuses on supporting scientific research, to facilitate the working with documents, required for analysis. Articles and reports should be as a result of manipulate with documents. Document Repository is a collection of methods, techniques, and tools used to support of Researcher to conduct analyse of contents of a document that help in making of articles, reports and etc.

## **2 The Function of Document Repository**

Document repository should be specially organised repository of files, with the ability to operate with groups of documents (Documents Mart), and with the metadata about documents and with fragments of documents. A Documents Mart is a subset or an aggregation of documents, grouped by certain signs or for a particular purpose, and stored in a primary Document Repository. It includes a set of information about stored documents (Metadata).

Goals of Document Repository are: automation of some formalizable procedures of stage 1; to reduce the time of a study, simplify the process for the preparation of reports, articles, etc. Concepts, which is expected to manipulate in a document repository, are "document" and "element of the document structure": fragment of the document, keywords, content (Section List), references to a document (article, book, etc.) and others.

Main tasks of Document Repository can be formulated as follows:

- organization of a structured document storage specific subject area;
- search for documents on the set parameters (on request);
- organization of basic (fixed) and workers (temporary), virtual, groups of documents in accordance with the needs of the researcher (Document Mart);
- quick access to groups of documents selected from the repository of documents for certain queries;

- choice links to sources of information on request;
- creating new documents, partly on the basis of the chosen;
- other more minor procedures to facilitate rapid clearance of the printed material when performing the scientific studies.

## 2.1 Analysis of the Composition and Structure of Stored Documents

Consider the composition of documents used in Research of the development and operation of Coal Industry in the energy industry. In analysing the current state of Coal Industry and describing the current state, following documents are used: forms of statistical reporting, statistical reviews, scientific reports, articles, articles from Internet, links to sites on Internet on themes, reference manual, literature on the direction of research, articles in journals (small format), large publications (in large format), other publications, references to the literature.

Analysis of the composition of electronic documents when carrying out scientific research in order to predict the development of industries of Fuel and Energy Complex allows conditionally select group of documents on various signs (Table 1). Also other ways of classification of documents are possible.

**Table 1.** Ways of classification of documents

Way of classification	the characteristic for the classification	The value of the characteristic, note
1) by the method of storage organisation	<ul style="list-style-type: none"> <li>- on electronic carrier</li> <li>- on paper</li> </ul>	file format: doc, pdf, xls, ppt and etc.
2) by origin (source information)	- documents received from outside	official source, Internet, organisation, from authors, other source
	- documents compiled, created by the user or user group	authors
3) in the form of internal organisation	<ul style="list-style-type: none"> <li>- tables data</li> <li>- an illustrative material (drawings, maps and etc.)</li> <li>- text documents</li> <li>- documents of mixed type</li> </ul>	
4) on substantive signs	<ul style="list-style-type: none"> <li>-Statistics reports and articles</li> <li>- Presentations</li> <li>- Links to publications and etc.</li> </ul>	name of the document
5) the internal organisation of storage	structured documents	elements of the structure: chapter headings, contents, presence heterogeneous objects (pictures, tables, diagrams, text)
	not structured	
6) on a importance and reliability	- degree of a confidence to document	official information, estimations of authors, opinions of journalists, etc.
7) frequency of receipt	<ul style="list-style-type: none"> <li>-1 per year</li> <li>- No periodicity</li> </ul>	

Documents may be as in electronic form (mostly), also and p-books.

The classification can be used to form attributes of stored document, as well as the organisation of Documents Marts, the creation of the classifier and a set of keywords included in the classifier.

Separately it is necessary to regard documents of large-format , as reports, statistical reporting forms and statistical reviews (Table 2). Statistical reviews have about same characteristics as statistical reporting forms, as well as some characteristics of reports (contents, pictures, tables, diagrams).The process of obtaining necessary data to work from a form of statistical reporting rather labourious and poorly formalised. To work of researchers are necessary not forms themselves, but chosen from these forms of individual indicators presented in the form of certain tables to analyse the current state of Coal Industry. Tables contain indicators of the development of facilities in dynamics, and estimates are not contained in the source documents. Forms of statistical reporting and statistical reviews, generally, contain redundant set of indicators for only one period (year).

**Table 2.** Characteristics of large-format documents

Characteristic	Document	
	Form of statistical reporting	Reports
Structure	Table	structured text
Elements of structure	stubs, headlines	content, subsections, lists of tables, figures, diagrams
Number of documents	1-20	1
Storage Format mainly: rarely:	*.doc, *.xls, *.txt, *.pdf, p-books	*.doc, *.pdf, p-books
Number of indicators	to 2-3 million in one table	It depends on number of tables in the report and quantity of indicators in tables
file size	1-30 MB	0,5-10 MB (30-500 pages of text)
extract indicators,% of total	1-10	1-80
Periodicity of output	1 per year	is absent
Stability of the structure of the document	is absent	is absent

One of problems in the processing of forms of statistical reports and statistical reviews is the diversity of forms of storage and structure of documents, even for the same documents. For example, in forms of statistical reports vary from year to year, not only formats of storage(text, MS Word, MS Excel), but also a set of indicators, except for key indicators. Composition of stubs and headlines of tables and codes of indicators in forms of statistical reporting are changing. Names of file can be the same in different years or changed. Browse forms rather difficult. For example, in one of tables of the statistical form "4-fuel" the number of indicators on a single administrative unit can be up to 500, respectively, over 2 million indicators across the whole table.

The heterogeneity of documents used determines the different approaches to storing and accessing to information in documents.

Statistical reporting forms and statistical reviews, for several reasons, it seems reasonable to keep as a primary source in the Document Repository while indicators needed to create tables to analyse - in a database. Large-format documents (reports, statistical reviews), easier to use in the presence of detailed metadata about documents.

Articles and other documents of small size, also require the organisation of a structured storage with the possibility of finding relevant documents by keyword and other characteristics .

In addition to electronic documents are used documents in paper form, references to which are necessary, for using these sources, and for inserting references to sources when creating new documents.

### **3 Basic Principles of Organisation Document Repository**

1. Problem-Object orientation: (Documents are combined in the category in line with some signs or characteristics).
2. Integrity: unites documents containing information about a particular subject area and sections of this subject area.
3. Non-updatable separate groups of documents: documents in certain groups (Documents Mart) document repository are not editable (permanent display of documents).
4. Documents from the Document Repository may be removed.
5. Documents can be created on the basis of documents contained in the Documents Repository.
6. Temporary Documents Mart are created and belong to updatable groups of documents.
7. Temporary Documents mart may change status and become a permanent Documents Mart.
8. Creating Metadata two levels:
  - a. general ( formalistic data about the document);
  - b. profound informative aspect;
9. Development of Classifiers.

Documents in the Document Repository come from different sources, relate to different types of documents and used in the study for different purposes: to study and analyse information, data extraction, extraction of text fragments and etc. In the information space allocated to certain areas - Documents Mart, or specified in advance - permanent (statistics, articles, etc.), or temporary, created for a particular study. In the process of uploading documents are created Metadata about documents and connection of documents to a specific Documents Mart.

#### **3.1 Composition of Document Repository**

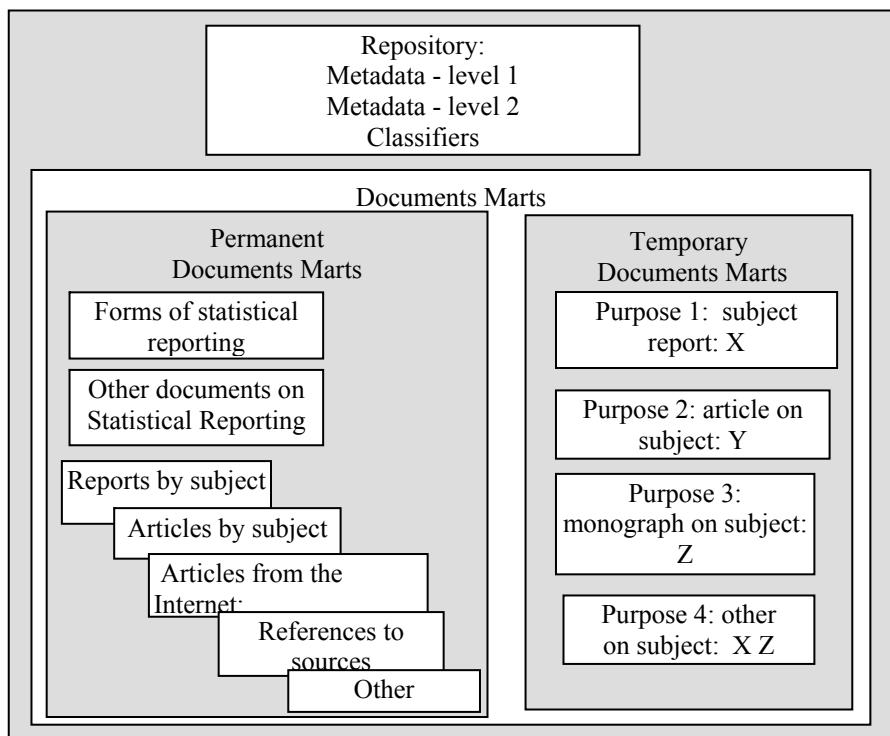
Components of Document Repository (fig. 1):

1. Documents Mart: Permanent; Temporary.
2. Repository: Metadata - level 1; Metadata - level 2; Classifiers.
3. References to sources and other composed of Permanent Documents Mart.

Under of Classifiers is understood as the hierarchy of set of concepts used in the subject area, which consist of sets of characteristics of objects. For example, an administrative division of the country and etc.

Documents Marts are organised in accordance with the need of end-user and the specific subject area. Every Documents Mart may consist of smaller ones. For example, Documents Mart of statistical reporting forms may contain Documents Mart for every form, etc.

Metadata (repository, "data about documents"). They play the role of a handbook containing information about sources of primary documents, their structure, content, etc. Metadata should be presented in two levels: general (formal) information about the document and data reflecting contents of the document (substantive aspect).



**Fig. 1.** Composition of Document Repository

Formal details about the document contain information: description of the document structure, source of the document, authors, date received, date of the document, to which type of documents include (statistics, reports, articles, etc.), the estimation of the validity, the reliability of the source document and etc.

The substantive aspect includes information about the internal content of the document: the document structure (for example, for the report: name, contents, list of tables, list of figures, list of applications names), keywords, headlines and stubs of tables of statistical reporting forms, etc.

- Classifiers include dictionaries of keywords, a hierarchy of key terms of domain objects, relationship of concepts and are intended for use for: formation requests for searching for document in the Document Repository;
- the establishment of metadata about the document;
- realising the connection with the database "Perspective of the development Coal Industry".

Major Classifiers for the Coal Industry may be the following: administrative division of the country, the classification of coals by different characteristics, classification of company-producers and etc. Classifiers associated with basic concepts and terminology of the subject area [7].

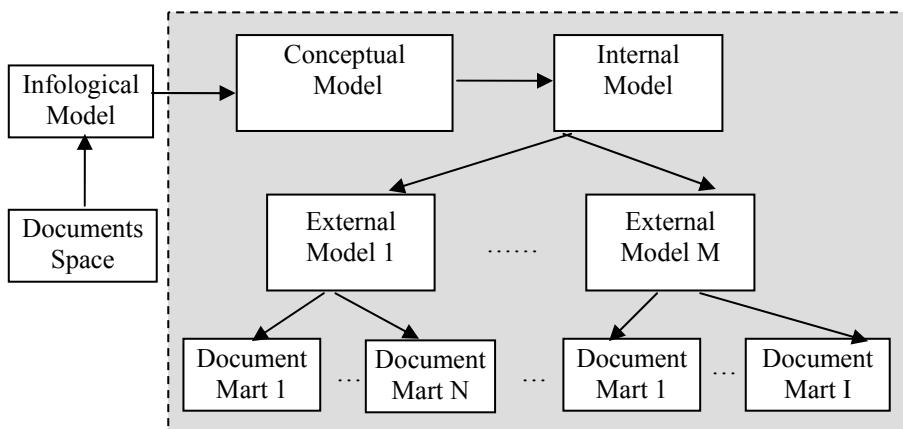
One of main conditions for successful implementation of Document Repository is availability of developed Metadata and means of providing Metadata to end-user. For example, before the request will be formulate to a system, user should know what documents there are, are its relevance, should he trust them or not, contents of documents, etc.

Data on possible sources of information in the subject area contain references to books and websites with a brief description of content (annotation) and, possibly, a set of keywords for this source. Possible content of links to the source: the organisation, authors, year of establishment of the source of information or date of copying information from a Internet, a reference to the literature, etc. Data on sources of information should be decorated for placement in a list of references.

## 4 The Approach to Implementation of Document Repository

The process of implementation of Document Repository should to include engineering and building of different level models by analogy with database development technology [8].

In our situation, database design theory will be operating not on data and data sets, but on Documents and set of Documents. Structuring of Document set implies its



**Fig. 2.** Generalised scheme of Document Repository design

division into subsets, which can be called Documents Marts. The Documents Mart is regarded as a virtual document collection, singled out on some characteristic or for particular purpose.

The scheme of Document Repository was build (Fig. 2) by analogy with database development technology, where document space goes instead of informational space, and external data models are amplified with Documents Mart.

#### **4.1 Building Infological Model of Document Repository**

During infological model building there main information objects of the application environment, their attributes (characteristics or features) and object relationships are selected [9]. The Infological model is user-oriented and does not depend on DBMS type.

The Infological model is represented in Chen's notation: model "entity-relationship", or ER-model. The model describes all essences, attributes and connections. The Model encloses by certain rules associated essences which are data domain imposed: «Document Repository», «Document Mart», «Constant Document mart», «Temporary Document mart», «Metadata level\_1», «Metadata level\_2», «Reference books», «Articles», «Reports», «Statistical reporting», «Material from Internet», «References Bibliographies», «Statistical Collection», and «Others».

#### **4.2 Procedures of Document Repository**

Main procedures of Document Repository can be divided into some groups:

1. Forming the Documents Mart;
  - 1.1. Forming Temporary and Permanent Documents Marts: Procedures for working with lists of Documents;
  - 1.2. Adding Documents to Documents Mart and removed from it;
2. Procedures for working with Documents
  - 2.1. Initial documents loading in Document Repository: Formation of Metadata -level 1; Formation of metadata -level 2;
  - 2.2. Formation of new Documents: Procedures for working with the content of Documents;
3. Working procedures with Document Repository configuration (generation, deleting constant Documents Mart, working with keywords and Classifiers).

#### **4.3 Algorithms of Procedures**

Algorithm input document to the Document Repository:

1. Open the document;
2. Entering metadata level 1, type / copy / to choose from existing \*: document title; annotation; authors \*; comment; source documents; year \*; degree of reliability of the information; keywords \*;
3. Entering metadata level 1, select from classifiers;
4. Entering metadata level 2, copy, or generated automatically in the presence of procedures: content; list of names of tables; list of names of graphs; list of names of applications;

5. Choosing documents mart to save the document in the Document Repository;

Metadata are entered at the discretion of the user, completely or only main.

An algorithm for the formation of a Temporary Documents Mart in the Documents Repository:

1. Search for documents in Document Repository, specify your search, choose from the list / to type: authors; source of documents; year; keywords; classifiers;
2. Get list of documents satisfying the search parameters;
3. View the metadata retrieved documents;
4. Creating the list of documents for the establishment of a Temporary Documents Mart;
5. Entering metadata about Mart: name; comments (target, title, etc.).
6. Saving Mart of documents.

#### **4.4 Selection of Tools**

In the view of the foregoing in article Document Repository engineering approach it expected the implementation of Document Repository. Firebird is selected as DBMS, due to its compactness and functionality. IBExpert is selected as manager of Document Repository engineering, since it was developed for DBMS InterBase and Firebird and completely grants our demands. NetBeans (Java language) is selected as development framework and language.

### **5 Conclusion**

In this article is made attempt statement of problem to develop Document Repository, does not pretend to be complete coverage of the problem. This direction of development of information technology is promising and topical. Experience in developing tools for a given subject area allow to analyse the possibility of the presented direction of the development of dataware and, subsequently, to develop a more versatile tool.

The field of application of Document Repository systems is not only restricted to Coal Industry, but it also may use in any science and education.

### **Acknowledgment**

The research described in this article was partially supported by grant RFBR № 07-07-00264. The results will be used in projects supported by this grant.

I deeply thank Dr. Prof. Massel L.V. for help and supporting this direction of research.

### **References**

1. Orekhova, L.N.: Software and Information engineering for forecasting coal industry development of country / Abstract - Irkutsk. In: ERS, p. 18. SB RAS Press (1991) (in Russian)
2. Sokolov, A.D., Takayshvili, L.N.: Tools for study coal industry. In: Proceedings of Russian Conference on Information Technology in Science and Education, pp. 116–121. ISEM SB RAS Press, Irkutsk (2002) (in Russian)

3. Sokolov, A.D., Takaishvili, L.N.: Modeling and optimization of the coal industry in market conditions. In: Proceedings of Materials Science and Practical Conference Fifth Melentevskie theoretical reading, December 8-9, pp. 281–291. INEI RAN Press, Moscow (2004)
4. Takayshvili, L.N.: Features computational experiment study of the coal industry within the framework of fuel and energy complex. J. Modern Technology. Systems Analysis. Modeling, Special Edition, IrGUPS, Irkutsk, pp. 64–69 (2008) (in Russian)
5. Golfarelli, M., Rizzi, S.: Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, New York (2009)
6. Document management system,  
[http://en.wikipedia.org/wiki/Document\\_management\\_system](http://en.wikipedia.org/wiki/Document_management_system)
7. Takayshvili, L.N.: Terms classification for formalised description of the Coal industry as part of Fuel and Energy Complex. In: Proceedings of International Workshop on Computer Science and Information Technologies (CSIT 2009), Crete, Greece, October 5-8, vol. 1, pp. 91–95. USATU, Ufa (2009)
8. Massel, L.V., Boldyrev, E.A., et al.: The Integration information technology in system study of energy. Science Press, Novosibirsk (2003)
9. Takayshvili, L.N., Osama El Sayed Ahmed Mohamed Seth.: Projecting document repository for studies of the coal industry. In: Proc. of XIVth Baikal Russian Conference Information Technologies in Mathematics and Science, Technology and Education, vol. 3, pp. 208–213. ISEM SB RAS, Irkutsk (2009)

# Forecasting Coal and Rock Dynamic Disaster Based on Adaptive Neuro-Fuzzy Inference System

Jianying Zhang, Jian Cheng, and Leida Li

School of Information and Electrical Engineering,  
China University of Mining & Technology, Xuzhou 221116, China  
zjycumt@126.com

**Abstract.** Forecasting model of coal rock electromagnetic radiation was built with combining time series analyze and adaptive neuro-fuzzy inference system (ANFIS). In the first, coal rock electromagnetic radiation phase space was reconstructed through Takens theory, and time delay and embedding dimension are determined by mutual information method and false nearest neighbor method respectively. Then, the forecasting model of coal rock electromagnetic radiation was constructed via ANFIS in the reconstruction phase space, and the parameters of ANFIS are tuned by hybrid learning algorithm. Finally, the simulation results and comparison analysis are presented, the training and checking root mean squared error are 0.0248 and 0.0286 respectively, which indicates that the ANFIS has better learning ability and generalization performance, thus, the model is creditable and feasible.

**Keywords:** Coal rock electromagnetic radiation, time series, phase space reconstruction, ANFIS.

## 1 Introduction

Coal rock dynamical disasters in coal mining, such as coal and gas outburst [1] and rock burst [2], are causing severe threat on the coal mine safety. With the increase of the mining depth, the frequencies of such disasters increase accordingly, which bring great threat to the hewers. Electromagnetic radiation is a phenomenon of energy dissipation caused by excessive load on the rocks, and it is closely related to the disaster processes. The coal rock electromagnetic radiation signal can be used to achieve non-contact, no-destruction and real-time monitoring of coal rock dynamical disasters [3]. However, the current methods only use the strength and number of pulses to predict the possibility of disaster, causing in-correct forecasting. Datum obtained from both labs and on-site environments show that electromagnetic radiations produced during the cracking of coal rocks is a complex system changes dynamically with the time. It is usually affected nonlinearly by the natural factors and the mining techniques. It is not easy to forecast using traditional methods. It has been proved that it is more feasible to achieve accurate forecasting using parts of the time series datum obtained on site. The reason is that the time series in complex dynamic systems contain much richer information, and include all the traces of all other parameters.

Recently, neural networks and fuzzy inference systems have been widely applied in modeling of nonlinear regression. While the neural network can obtain accurate solutions, the priori knowledge cannot be used. On the contrary, fuzzy inference system can deal with priori knowledge, but fails to produce accurate solutions. Adaptive neuro-fuzzy inference system (ANFIS) is produced by combining the above two techniques [4]. Fuzzy inference system is widely used for fussy control, while neural network is characterized by self-adaptive learning. Therefore, ANFIS makes good use of both of their advantages, which enable it to be used in fussy control and pattern recognition. As a new kind of neural network, ANFIS has the ability to approximating any linear or nonlinear functions with any required precision. Furthermore, it is featured by fast convergence, little error and few training samples. Therefore, it is useful for self-adaptive signal processing.

Based on the researches on the chaos properties of time series of coal rock electromagnetic radiation, this paper determines the time delay of reconstruction phase space and embedding dimensions using mutual information algorithm and pseudo nearest neighbor method. In the reconstruction phase space of coal rock electromagnetic radiation, the prediction model of the self-adaptive neural fussy inference system is established and the electromagnetic radiation signal is analyzed accordingly. The trend of future electromagnetic radiation signal is forecasted. Simulation results show that the proposed scheme is applicable to coal mines.

## 2 Analysis of Coal Rock Electromagnetic Radiation Time Series

### 2.1 Phase Space Reconstruction

In the 1980s, Takens proposed the well-known Takens theorem based on Whitney's early work on topology [5]. This theorem is the basis of phase space reconstruction, and reveals the dynamic mechanism of some non-linear systems. Phase space reconstruction theorem is the basis of chaos time series prediction. Packard et al. and Takens et al. propose to achieve phase space reconstruction of the chaos time series  $\{x(t)\}$  using delay coordinate method [5, 6]. The points in the phase space are denoted by:

$$X(t) = [x(t), x(t - \tau), \dots, x(t - (m-1)\tau)] \quad (1)$$

where  $m$  is the embedding dimension and  $\tau$  is the time delay.

The Takens theorem proves that if the embedding dimension  $m \geq 2d + 1$  ( $d$  is the dimension of system dynamics), then the reconstructed dynamic system and the original system are equal in topology meaning of entropy, and the chaos attractors from the two phase spaces have diffeomorphisms. Therefore, the next status of a system can be obtained from the current status, obtaining the prediction value of the next moment. This also provides a basis for the prediction of time series signal. The nature of time series prediction is an inverse problem of a system, namely reconstructing the system dynamic model  $F(\cdot)$  through the status of the dynamic system.

$$x(t + T) = F(X(t)) \quad (2)$$

where  $T$  ( $T > 0$ ) is the step of forward prediction.

Many methods can be used to approximating  $F(\cdot)$  by constructing a nonlinear function  $f(\cdot)$ , and ANFIS is employed in this paper to achieve time series prediction.

## 2.2 Selection of Time Delay

Traditional time delay selection methods include autocorrelation function method and minimum mutual information method [7], among which the latter has been widely used. The minimum mutual information method employs the first minimum mutual information value as the optimal time delay.

For time series  $\{x_i : t = 1, \dots, N\}$  and its  $\tau$ -delayed series  $\{x_{i+\tau} : t = 1, \dots, N\}$ , assume the probability that  $x_i$  appears in  $\{x_t : t = 1, \dots, N\}$  is  $P(x_i)$  and the probability that  $x_{i+\tau}$  appears in  $\{x_{i+\tau} : t = 1, \dots, N\}$  is  $P(x_{i+\tau})$ . The joint probability that  $x_i$  and  $x_{i+\tau}$  appear in the two series is  $P(x_i, x_{i+\tau})$ . The probabilities  $P(x_i)$  and  $P(x_{i+\tau})$  can be computed by the frequencies they appear in the corresponding series, while the joint probability  $P(x_i, x_{i+\tau})$  can be obtained by the corresponding lattices on the plane  $(x_i, x_{i+\tau})$ . Therefore, the corresponding mutual information is a function of the time delay  $\tau$ :

$$I(\tau) = \sum_{i=1 \rightarrow N} P(x_i, x_{i+\tau}) \ln \frac{P(x_i, x_{i+\tau})}{P(x_i)P(x_{i+\tau})} \quad (3)$$

This function measures the dependence of the successive test results. The time delay can be determined using the value  $\tau$  when  $I(\tau)$  takes the first minimum.

## 2.3 Selection of Embedding Dimension

In one-dimension time series based phase space reconstruction, it is of great importance to determine the optimal embedding dimension. According to the Takens theorem, if  $m$  is very small, the attractors may intersect in some area due to folding, and it may contain the points from different parts of the attractors in a small region of the intersecting area. If  $m$  is too big, although feasible in theory, the computation cost will increase significantly when  $m$  increases in practical application. What is more, the noise and rounding error will increase accordingly. The main algorithms for computing the embedding dimension include pseudo nearest neighbor method and the method by computing some geometric invariants of the attractors etc [8].

During the reconstruction of phase space, when the dimension is  $m$  and  $X_{i'}$  is the nearest neighboring point of  $X_i$ , the distance between them is  $\|X_{i'} - X_i\|^{(m)}$ . When the dimension increases to  $m+1$ , the distance is denoted by  $\|X_{i'} - X_i\|^{(m+1)}$ . If

$$\frac{\|X_{i'} - X_i\|^{(m+1)} - \|X_{i'} - X_i\|^{(m)}}{\|X_{i'} - X_i\|^{(m)}} > R_T \quad 10 \leq R_T \leq 50 \quad (4)$$

Then  $X_{t'}$  is called the false nearest point, and  $R_T$  is the threshold. This method of computing the minimum embedding dimension by the concept of false nearest point is called False Nearest Neighbor (FNN). The computation cost of the method is low.

When the dimension changes from  $m$  to  $m+1$ , we check if there are false nearest point in the trajectory of  $X_n$ . If no, then the geometric structure is regarded to be open. In implementation, we start to compute from  $m=2$  with a fixed threshold. Then the proportion of false nearest point is determined. Then  $m$  is increased until the proportion of false nearest point is less than the threshold or the number of nearest points does not decrease with increasing  $m$  values, producing the minimum embedding dimension.

### 3 Adaptive Neural-Fuzzy Inference System

#### 3.1 Structure of ANFIS

Adaptive Neural-Fuzzy Inference System (ANFIS) is the product combining fuzzy inference system and neural network. As fuzzy inference system makes good use of experience and inspired knowledge, while the neural network has the ability of adaptive self-learning ability, ANFIS combines the advantages of the above two techniques [4,9], which have lead to its application in fussy control and pattern recognition [4,10,11]. As a new kind of neural network, ANFIS has the ability to approximating any linear or nonlinear functions with any required precision. Furthermore, it is featured by fast convergence, little error and few training samples. Therefore, it is useful for self-adaptive signal processing.

For simplicity, assume that there are two input parameters ( $x$  and  $y$ ) and one output parameter  $f$ , as illustrated in Fig.3. The rule base contains two *if-then* rules of Takagi-Sugeno fuzzy model, with the same function in the same layer.

**Layer one:** Each node is a self-adaptive node with node function, with Eq.(5) as the commonly used generalized bell-shaped function.

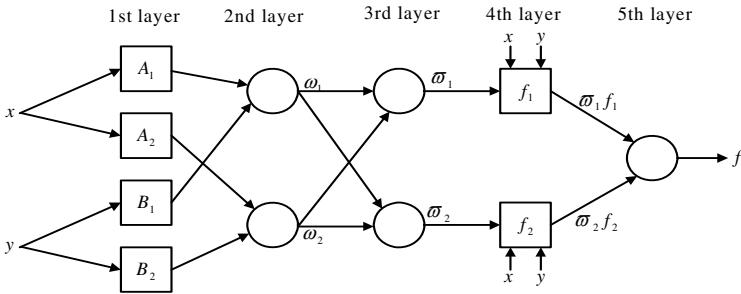
**Layer two:** The output of each node is the multiplication of the input signals.

**Layer three:** The computation of the impulse strength from the second layer.

**Layer four:** Applying the *if-then* rule of Takagi-Sugeno type. Each node has the linear parameter set  $\{p_i, q_i, r_i\}$  called the conclusion parameter.

$$\mu_{A_i}(x) = \frac{1}{1 + \left[ \left( \frac{x - c_i}{a_i} \right)^2 \right]^{b_i}}, \quad i=1, 2 \quad (5)$$

where  $\{a_i \ b_i \ c_i\}$  is called the parameter set of the premise parameters, which inflect the different forms of membership function.



**Fig. 1.** ANFIS structure with two input variables and one output variable

**Layer five:** The total output is the sum of all the input signals computed from a single point. When the premise parameter values are given, the output of ANFIS can be denoted by the linear combination of the conclusion parameters.

### 3.2 Learning Algorithm of ANFIS

Although the parameters of self-adaptive networks can be selected using the gradient method, it is slow in convergence and it often falls to local minimum. Therefore, the parameters in ANFIS are recognized combining gradient method and least square method, and it can be achieved by the following steps:

**Step 1:** The premise parameter is initialized, and the conclusion parameters are computed using the least square estimation method.

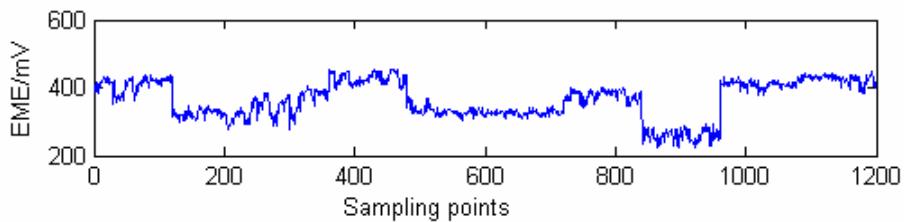
**Step 2:** Based on the conclusion parameters, the error is computed using the back-propagation algorithm of the feed forward network. In this way, the shape of the membership function will change accordingly.

## 4 Simulations

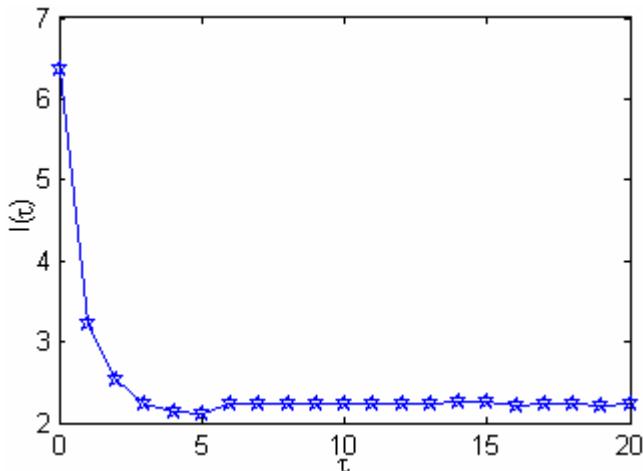
The coal rock electromagnetic radiation signal is collected from a China coal mine, denoted by  $\{x(t) : t = 1, \dots, 1200\}$ , as shown in Fig.2.

The phase space  $X(t) = [x(t), x(t - \tau), \dots, x(t - (m-1)\tau)]$  is reconstructed using the method described in Section 2. The optimal time delay is determined using the minimum mutual information estimation, which is shown in Fig.3. It can be seen from the figure that the optimal time delay is  $\tau = 5$ . Fig.4 shows the change of the proportion of false nearest point with the embedding dimension. Based on the analysis of section 2.3, we know that the dimension of reconstructed phase space is  $m = 5$ .

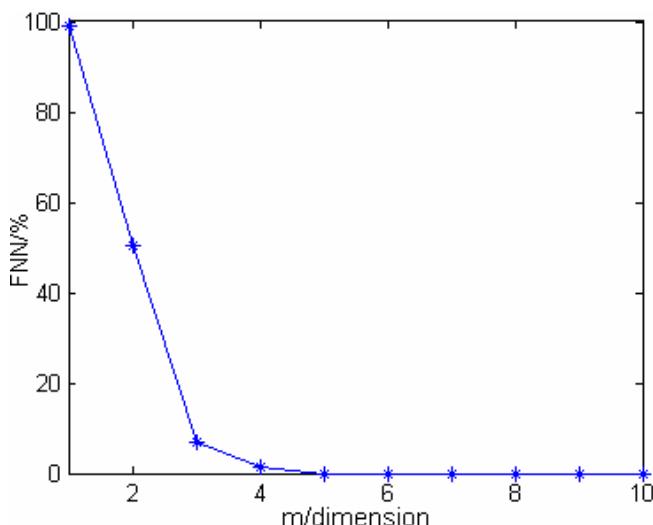
It should be noted that the time delay and embedding dimension will not be the best whatever kind of method is employed. In practice, close optimal is often used instead. By experiments, the embedding dimension is 5 and the time delay is 5.



**Fig. 2.** Electromagnetic radiation data



**Fig. 3.** Determine delay time through mutual information



**Fig. 4.** Determine embedding dimension based on false nearest neighbor

In the reconstructed phase space of the coal rock electromagnetic radiation, the self-adaptive fuzzy reference system is built, where the number of input parameters is  $m-1=4$ . Two membership functions are used for each parameter. The type of membership function is shown in Fig.5. The input space is divided to lattices, and distributed averagely and intersect adequately in the input domain, producing 16 fuzzy *if then* rules. The number of rules is not big so that dimension disaster can be avoided. The parameters of ANFIS are recognized using the hybrid learning algorithm.

The first 1000 groups of datum are used as the training samples for the model, and the first 100 groups of the remaining datum are used to check the prediction ability of the model. The prediction ability of the model is evaluate by the root-mean-square error (RMSE) and non-dimensional-error-index (NDEI):

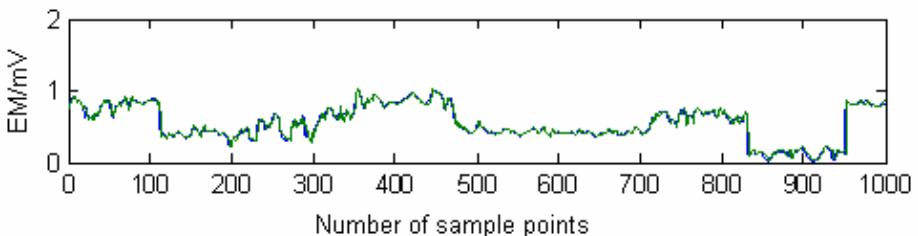
$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$NDEI = \frac{1}{\delta} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad \delta = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

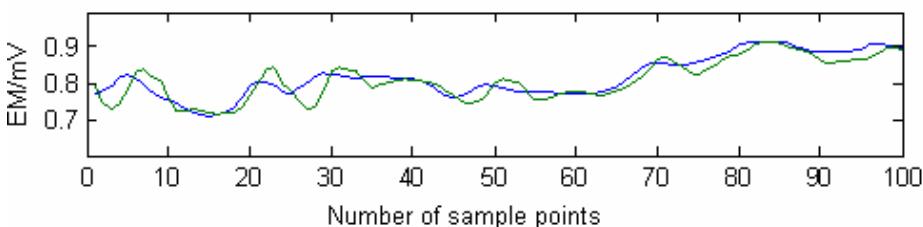
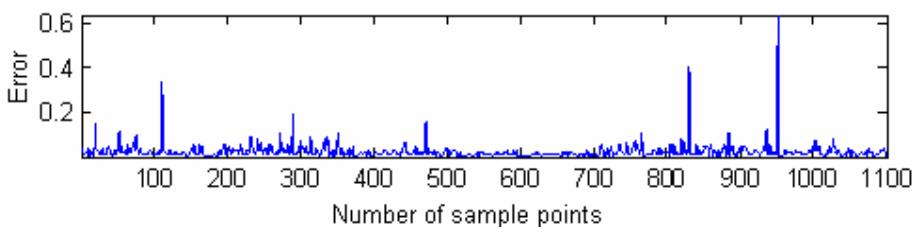
where  $n$  is the number of samples,  $y_i$  and  $\hat{y}_i$  denotes the actual value and prediction value,  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ .

Fig.6 and Fig.7 show the actual values and the prediction values of the training samples and testing samples respectively. It can be seen that the model proposed in this paper reflects in nature the dynamic characteristics of the coal rock electromagnetic radiation time series. The prediction error of the model is shown in Fig.8, and the error evaluation of the prediction is listed in table 1. It is observed that most of the prediction errors are within a small range except for some singular values, which also demonstrates that it is applicable to on site requirements.

For the same training samples and test samples, a three-layer 4-10-1 BP Neural Network is constructed. The average error of 20 simulations is listed in table 1. It is known that the error of BPNN is bigger than ANFIS. In regard if the computation time, the average converging time of ANFIS is 7.3210s, while 23.702s for BPNN. As a result, the proposed model is more suitable for practical applications.



**Fig. 6.** Model output of training samples

**Fig. 7.** Model output of checking samples**Fig. 8.** The absolute value of prediction errors**Table 1.** Error analysis of simulation

Model	Training error		Prediction error	
	RMSE	NDEI	RMSE	NDEI
ANFIS	0.0248	0.8735	0.0286	0.9011
BPNN	0.0311	0.8903	0.0564	1.2836

## 5 Conclusion

The characteristic of coal rock electromagnetic radiation signal is its non-linear property. Based on the analysis of its time series, this paper proposes to determine the time delay and dimension of the reconstructed phase space using mutual information and false nearest neighbor method. Then the prediction model is established in reconstructed phase space by self-adaptive neural-fuzzy inference system.

1) The proposed model combines the individual advantage of time series analysis, neural network and fuzzy inference system, so that the ability to deal with complex dynamic system.

2) The proposed model is featured by less parameter, fast convergence. ANFIS can obtain high nonlinear mapping. Although priori knowledge is not needed, rational initialization parameter can be obtained and covers the whole input space. In this way, ANFIS can converge quickly to the parameter reflecting the dynamic properties. The proposed ANFIS model here needs much fewer parameters than the BPNN model. In our simulations, the average convergence time is 7.3210s, which is much shorter than 23.702s of the BPNN model.

3) The proposed model can obtain higher prediction accuracy and it has better generalization ability. Experiments show that the training RMSE of ANFIS model is 0.0248. The training NDEI obtained is 0.8735, RMSE of the prediction error is 0.0286, and the prediction NDEI is 0.9011. The above errors are all smaller than those of BPNN model.

Combining time series and adaptive neural fuzzy inference system for modeling is not only suitable for coal rock electromagnetic radiation prediction, but also applicable to other nonlinear signal processing.

## References

1. Yu, Q.X.: Coal Mine Gas Control. China University of Mining and Technology Press, Xuzhou (1992)
2. Dou, L.M., He, X.Q., Xueqiu, H.E.: Rock Burst Control Theory and Technology. China University of Mining and Technology Press, Xuzhou (2001)
3. He, X.Q., Wang, E.Y., Nie, B.S.: Coal Electric Magnetic Dynamics. Science Press, Beijing (2003)
4. Jang, J.S.R.: ANFIS: Adaptive-Network-based Fuzzy Inference System. IEEE Transactions on System, Man and Cybernetics 23, 665–685 (1993)
5. Takens, F.: Detecting Strange Attractors in Turbulence. Lecture Notes in Mathematics, vol. 898, pp. 361–381 (1981)
6. Packard, N.H., Crutchfied, J.P., Farmer, J.D., Shaw, R.S.: Geometry from a Time Series. Physical Review Letters 45, 712–716 (1980)
7. Wang, H.Y., Sheng, Z.H.: Choice of the Parameters for the Phase Space Reconstruction of Chaotic Time Series. Journal of Southeast University: Natural Science Edition 30, 113–117 (2000)
8. Wang, Y., Xu, W., Qu, J.S.: The Algorithm and Chick of Phase-Space Reconstruction Based on The Time Series. Journal of Shandong University: Engineering Science 35, 109–114 (2005)
9. Lee, S.J., Ouyang, C.S.: A Neuro-Fuzzy System Modeling With Self-Constructing Rule Generation and Hybrid SVD-based Learning. IEEE Transactions on Fuzzy Systems 11, 341–353 (2003)
10. Cheng, J., Guo, Y.N., Qian, J.S.: Estimation of Loose Status of Jigging bed based on Adaptive Neuro-Fuzzy Inference System. Journal of China University of Mining & Technology: English Edition 16, 270–274 (2006)
11. Panella, M., Gallo, A.S.: An Input-Output Clustering Approach to The Synthesis of ANFIS Networks. IEEE Transactions on Fuzzy Systems 13, 69–81 (2005)

# Context-Aware Workflow Management Based on Formal Knowledge Representation Models

Fu-Shiung Hsieh

Department of Computer Science and Information Engineering,  
Chaoyang University of Technology,  
41349 Taichung County, Taiwan  
fshsieh@cyut.edu.tw

**Abstract.** In existing literature, different methodologies for the design of context-aware systems have been proposed. However, workflow models have not been considered in these methodologies. Despite the fact that many context-awareness projects have been launched, few address the issues of context-aware user interface generation and automated resource allocation. Our interests in this paper are to propose a framework to develop workflow driven context-aware human computer interaction to effectively control resource allocation. To achieve the objective, a model for knowledge management is required. In existing literature, one formal knowledge representation and reasoning model is Petri net. We first propose Petri net models to describe the workflows in the systems. Next, we propose models to capture resource activities in the systems. Finally, the interactions between workflows and resources are combined to obtain a complete model for the systems. Based on the aforementioned model, we propose architecture to automatically generate context-aware graphical user interface to guide the users and control resource allocation.

**Keywords:** Context aware, human computer interaction, workflow, knowledge representation.

## 1 Introduction

Workers in a real business environment are highly mobile; they change location and are assigned different tasks from time to time to perform their work. The information required by these workers is highly dependent on their location, role, time, workflow and activity involved. With the wide acceptance of pervasive computing devices, how to streamline the processes by providing timely information to workers and exploiting the advantage of handheld devices is an important research issue. Our research aims to propose an effective model for the above task force assignment problem based on context-aware computing technologies.

Context-aware computing [1]-[8] is becoming more and more important with the wide deployment of ubiquitous/pervasive computing infrastructure. In this paper, we focus on application of context-aware computing technology in task force automation [9]. The task force automation process is driven by the arrival of a potential order.

Depending on the requirements of the order, a task workflow will be formed to determine whether the order should be committed. Typical business workflows usually require the formation of a collaborative network to solve the problem or process the task. Although collaborative network formation has been studied in literature, how to achieve effective collaborative network formation by taking the advantage of context-aware computing technologies is an important issue. The goal of this research is to propose a formal model for modeling and verification of the collaborative networks formed in a context-aware computing environment. To achieve the objective, a model for workflow knowledge management is required. In existing literature, one formal workflow knowledge representation and reasoning model is Petri net [10]. For examples, modeling and analysis of workflow processes have been extensively studied in [11]-[15]. Knowledge representation and reasoning based on Petri nets have also been studied in [16]-[18]. To describe the context of a task, we model its contextual information and knowledge with Petri nets. Our previous works [20]-[27] on modeling of workflows and processes in multi-agent systems paves way for the development of this paper. With the proposed Petri net models, we propose a methodology to facilitate the design and analysis of context-aware task force automation applications.

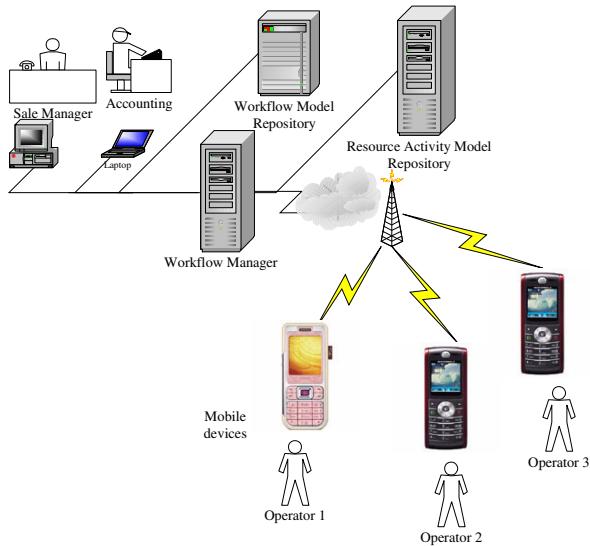
The remainder of this paper is organized as follows. In Section 2, we introduce the workflow management problem. In Section 3, we propose resource activity models and workflow models in Petri nets. In Section 4, we propose a complete workflow management Petri net model and a resource allocation scheme based on reachability graph of the workflow management Petri net. We conclude this paper in Section 5.

## 2 Workflow Management Problem

Task force automation is concerned with allocation of resources to achieve a certain business objective in a timely and efficient manner. Typically, a variety of resources are involved in the processes to accomplish a task. Different steps throughout the lifecycle of handling an order include sale, accounting, receiving, manufacturing, quality assurance, packaging and shipping. Each step requires distinct resources for processing. For example, in sale step, sale staffs are involved in the negotiation as well as contracting. Costing of the order requires the assistance from the accountants. Therefore, a message is forwarded to the accountants for costing. Once the order contract is established, it will be released to the shop floor for production. In the production process, different operators and machines are assigned to perform the relevant operations. Each final product is then forwarded to the quality control staff for examination. Once the required qualified products are available, the operators package and then ship the products to the customer. The above scenario exhibits a very complicated process flow and intensive messaging and interactions between different workers in the team. How to effectively manage the concurrent workflows as well as messages in the system is a challenge.

We assume each worker is equipped with a mobile phone that is used for communication with other team members. One of the design issues is to deliver the required information only to avoid overloading the mobile phone as well as the users. Another issue is to effectively enact the execution of operations to accomplish the task. Fig. 1 shows our proposed architecture for context-aware task force automation. In Fig. 1, the contextual model repository is used to capture the context of interest to different types

of team members and that of task workflows. The information to be delivered to different members in the system varies depending on the role of the member involved in carrying out the task. The workflow manager aims to provide a model to monitor the progress of tasks that are needed to fulfill an order. A team member usually needs to participate in different parts of the task workflows. Depending on the role of a team member, the communication devices accessible vary. How to generate the context-aware information to guide a team member using a variety of different types of communication devices to collaboratively accomplish the task is a significant design issue.



**Fig. 1.** Architecture for workflow task force automation

### 3 Resource Activity and Workflow Models

To facilitate automated generation of context-aware guidance information for each team member, an appropriate contextual modeling approach must be adopted in this paper to capture the interactions between the individual contexts of the team members, the contexts of different task workflows as well as the contexts of the communication devices accessible. There are two requirements in selection of models. First, the model must have well-established formal mechanisms for modeling as well as analysis. Second, the model must be accompanied with a standardized interchange format. Petri net [10] is a model that meets these two requirements. Therefore, we adopt Petri nets to model the contexts. The advantages of applying Petri nets formalism to model and analysis of a context-aware system can be summarized as follows. First of all, the graphical nature of Petri nets can visualize sequences of firing via token passing over the net. Second, Petri nets have well-established formal mechanisms for modeling and property checking of systems with concurrent, synchronous and/or asynchronous structures. Third, the mathematical foundation of Petri nets can analyze structural and dynamic behaviors of a system. These advantages make Petri

nets a suitable modeling and analysis tool. Furthermore, the emerging Petri Net Markup Language (PNML) [19] is a preliminary proposal of an XML-based interchange format for Petri nets. PNML makes it possible for individual companies to exchange their process models in Petri net. In this paper, PNML is used for sharing and exchanging the process models of participants in a team.

A Petri Net (PN) [10]  $G = (P, T, I, O, m_0)$ , where  $P$  is a finite set of places,  $T$  is a finite set of transitions,  $I \subseteq P \times T$  is a set of transition input arcs,  $O \subseteq T \times P$  is a set of transition output arcs, and  $m_0 : P \rightarrow \mathbb{Z}^{|P|}$  is the initial marking of the PN with  $\mathbb{Z}$  as the set of nonnegative integers. A marking of  $G$  is a vector  $m \in \mathbb{Z}^{|P|}$  that indicates the number of tokens in each place under a state.  $\bullet t$  denotes the set of input places of transition  $t$ . A transition  $t$  is enabled and can be fired under  $m$  iff  $m(p) \geq I(p, t) \quad \forall p \in \bullet t$ . Firing a transition removes one token from each of its input places and adds one token to each of its output places. A marking  $m'$  is reachable from  $m$  if there exists a firing sequence  $s$  bringing  $m$  to  $m'$ . The reachability set  $R(m_0)$  denotes all the markings reachable from  $m_0$ . A Petri net  $G = (P, T, I, O, m_0)$  is live if, no matter what marking has been reached from  $m_0$ , it is possible to ultimately fire any transition of  $G$  by progressing through some further firing sequence.

To model a task workflow in Petri net, we use a place to represent a state in the workflow while a transition to represent an event or operation that brings the workflow from one state to another one. The workflow of a task  $w_n$  is modeled by an acyclic sequential marked graph  $W_n = (P_n, T_n, I_n, O_n, m_{n0})$ , where the set  $P_n$  of places denotes the production states whereas the set  $T_n$  of transitions denotes the operations.

**Definition 3.1:** The workflow of  $w_n$  is an acyclic sequential marked graph  $W_n = (P_n, T_n, I_n, O_n, m_{n0})$ .

As each transition represents a distinct operation in a task,  $T_j \cap T_k = \Phi$  for  $j \neq k$ . Each  $W_n$  has a final transition  $t_n^f$  whose firing terminating the task.

A workflow only specifies the sequence of operations to complete a task. Each operation in a workflow consumes a number of different types of resources. In task force automation systems, the set of resources includes sale staffs, accountants, operators, etc. In addition to the workflow specified by  $W_n$ , the activities for each type of resources must also be specified. Each resource may take part in some of the operations in different workflows. A resource may visit different states throughout the lifetime of its usage in performing various operations in the workflows. For example, a sale staff may be busy with a number of orders. The sale staff will return to idle state after the tasks are finished. For another example, a shipping operator may be in either busy state or idle state. Depending on the location, the shipping operator may be in different busy states. A machine may be either in idle state or in busy state. A quality assurance operator may also be in different states depending on the time, schedule, location and the products to be inspected. To facilitate monitoring and control of resource allocation, a model is required to capture the activities of each type of resources.

An activity is a sequence of operations to be performed and states to be visited by a certain type of resources. We use a place in Petri net to represent a state in the resource activity. Each resource has an idle state. Each resource activity starts and ends with an idle state. A resource activity is described by a circuit in Petri net. A circuit indicates that the resource activity includes resource allocation and de-allocation.

Let  $\mathbf{R}_n$  denote the set of resource types required to perform the operations in  $W_n$ . The Petri net model for the  $k-th$  activity for a type- $r$  resources, where  $r \in \mathbf{R}_n$ , is described by a Petri net  $A_r^k$  defined as follows.

**Definition 3.2:** Petri net  $A_r^k = (P_r^k, T_r^k, I_r^k, O_r^k, m_r^k)$  denotes the  $k-th$  activity for a type- $r$  resources, where  $r \in \mathbf{R}_n$ . Remark that  $T_r^k \cap T_r^{k'} = \Phi$  for  $k \neq k'$ .

Let  $K_r$  be the number of activities of a type- $r$  resources. Let  $\Omega_n^r \subseteq \{1, 2, 3, \dots, K_r\}$  denote the set of type- $r$  activity IDs in  $W_n$ . The initial marking  $m_r^k$  is determined based on the set of resource tokens allocated to the  $k-th$  activity. More specifically,  $m_r^k(p_r)$  is the number of resources allocated to place  $p_r$ , where  $p_r$  is the idle place of type- $r$  resources.

## 4 Workflow Management

The complete Petri net model of a process is constructed by combining the resource activity models with the workflow model to take into account the interactions between resources and the workflow. To combine resource activity models with the workflow model, we define the operator “ $\parallel$ ” as follows to merge two Petri net models with common transitions, places and/or arcs. The operator “ $\parallel$ ” can be easily implemented based on the incidence matrices of the Petri nets. By applying the “ $\parallel$ ” operator, resources involved in an operation are synchronized so that the operation can be executed.

**Definition 4.1:** The complete Petri net model to process workflow  $W_n$  is modeled by  $G_n = \parallel_{r \in \mathbf{R}_n} A_r \parallel W_n$ , where  $A_r = \parallel_{k \in \Omega_n^r} A_r^k$ .

Based on the composition operation defined in Definition 4.1, we may construct the complete Petri net model. Fig.2 shows  $G_n =$

$$W_n \parallel A_{r1}^1 \parallel A_{r1}^2 \parallel A_{r1}^3 \parallel A_{r1}^4 \parallel A_{r2}^1 \parallel A_{r3}^1 \parallel A_{r4}^1 \parallel A_{r5}^1 \parallel A_{r6}^1 \parallel A_{r7}^1 \parallel A_{r8}^1 \parallel A_{r9}^1 \parallel A_{r10}^1 \\ \parallel A_{r11}^1 \parallel A_{r12}^1 \parallel A_{r13}^1 \parallel A_{r14}^1$$

**Definition 4.2:**  $M_n^*$  denotes the set of initial markings of  $G_n$  with minimal resources for the existence of a control policy to keep it live. The set of resources under  $m_n^* \in M_n^*$  is called a minimal resource requirement (MRR) of medical process  $G_n$ .

**Property 4.1:** Given  $G_n$  with marking  $m \in R(m_0)$ , there exists a control policy  $u$  such that  $G_n$  is live under  $m$  if and only if there exists a sequence of control actions that bring  $m$  to a marking  $m' \in M_0$  with  $m'(p_{r0}) \geq m_n^*(p_{r0}) \forall r \in R$ , where  $m_n^* \in M_n^*$ .

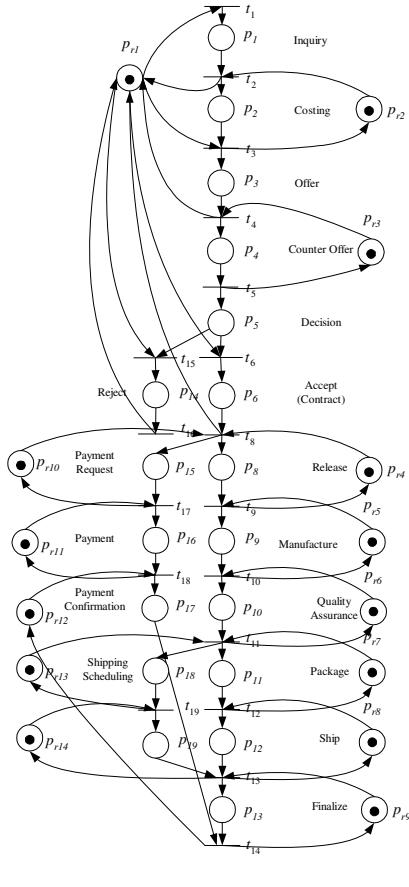


Fig. 2. Workflow management model

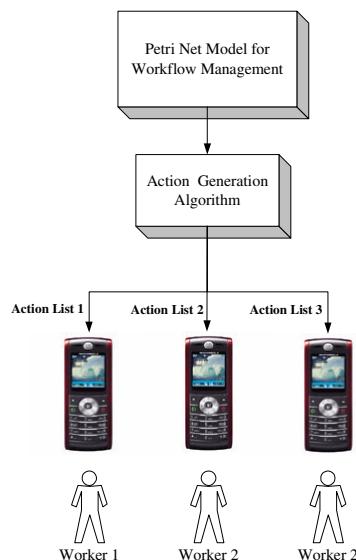


Fig. 3. Workflow management architecture

Application of Property 4.1 requires computation of  $m_n^*$  and testing reachability of a marking  $m'$  that covers  $m_n^*$ . Marking  $m_n^*$  can be found by firing the transitions in  $G_n$  once for each transition. With the complete Petri net model, the state of the system can be captured, represented and predicted conveniently. For example, we may apply the reachability tree method to predict the dynamic behaviors of the workflow system.

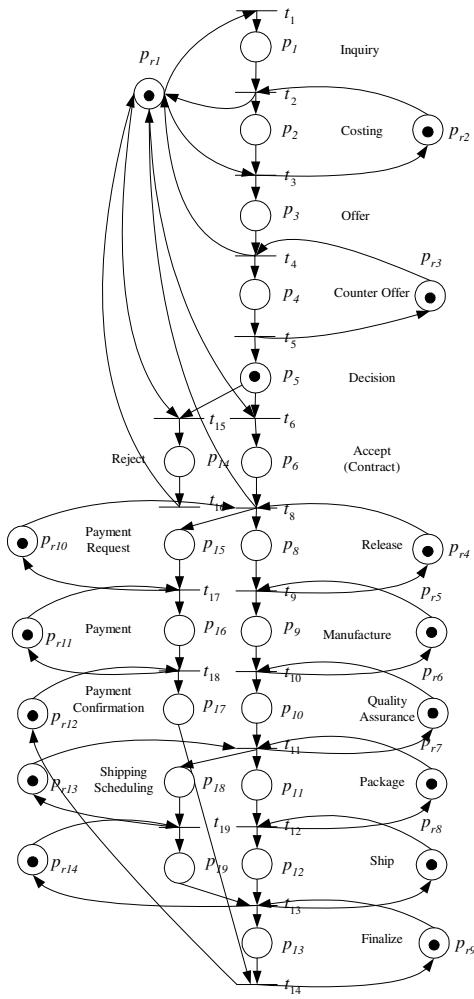


Fig. 4. A system state



Fig. 5. An action list for the state of Fig. 4

Given  $G = (P, T, I, O, m_0)$  with an initial marking  $m_0$ , the reachability graph can be generated with the following algorithm.

#### Reachability Graph Construction Algorithm $RGC(m)$

**Input:**  $G, m$

**Output:**  $RG = (N, A)$

**Step 0:**  $N \leftarrow \Phi$

$$A \leftarrow \Phi$$

Find the incidence matrix  $K$  of  $G$

Step 1: If  $m \notin N$

```

 $N \leftarrow N \cup \{m\}$ 
For each  $t_j \in T_n$ 
  If  $m'(i) = K(i, j) + m(i) \geq 0 \forall i$ 
    Create arc  $a(m, m')$ 
     $A \leftarrow A \cup \{a(m, m')\}$ 
     $RGC(m')$ 
  End If
End For
End If

```

Generally, the reachability graph may either contain safe markings as well as unsafe markings. Evolution to unsafe markings is undesirable. Therefore, we must identify the subgraph of a reachability graph that contains safe markings only. Construction of the safe subgraph of a reachability graph can be efficiently achieved by applying the strongly connected component from the reachability graph.

The model we propose previously can be used to monitor and control the resource allocation in a healthcare institution. Fig. 3 shows our proposed architecture for the implementation of a monitoring & supervisory control system based on our proposed Petri Net model. We assume each worker is equipped with a handheld device to receive the action list for execution. The complete Petri net model maintains the state of the medical processes based on the up-to-date state information from the state update algorithm. With the most up-to-date information of the medical processes, the action generation algorithm generates the action list for each worker. Different workers select their actions from the action lists for execution. By executing an action, a message is forwarded to the state update algorithm to update the state of the complete Petri Net model. The action generation algorithm then generates the action lists based on the new state. Fig. 4 shows a state in which a token reaches place  $p_5$ , which represents a decision needs to be made. In this case, our system will respond with the action list shown in Fig. 5 to instruct the user to either accept or reject the order.

## 5 Conclusion

This paper focuses on application of context-aware technology in workflow management. Consider the task force assignment problem that takes place from time to time in a company. The task force automation process is driven by the arrival of a potential order. Depending on the requirements of the order, a collaborative network needs to be formed to determine whether the order should be committed. We propose formal Petri nets for modeling individual contexts of the collaborative networks in a context-aware computing environment. Based on composition of the individual context models, workflows can be monitored. Our future research directions include theoretical development of supervisory control algorithms for the class of Petri nets proposed in this paper, implementation of the model composition mechanism and prototype design based on our proposed system architecture to facilitate the generation of user interface for context-aware applications. Context-aware computing refers to an application's ability to adapt to changing circumstances and respond based on the context

of use. The wide scale deployment of wireless networks improve communication among team members of a company as well as enable the delivery of accurate information anytime anywhere, thereby reducing errors and improving access. How to design a system with state of the art context-aware computing technologies is a significant issue. In existing literature, different methodologies for the design of context-aware systems have been proposed. Existing methodologies or prototype systems do not address the issues of workflow resource allocation, monitoring and generation of context-aware user interfaces. This paper aims to propose a framework to automate resource allocation and monitoring effectively. To achieve this goal, we first propose workflow models and resource activity models in Petri net to describe the processes in systems. Next, we propose a complete Petri net model to model resource contention and interactions between workflow and resources. The complete Petri net models serve to monitor and control the allocation of resources. To dynamically generate user interface, we propose a reachability graph construction to generate the list of actions for a resource in a given system state to avoid deadlocks. We develop a context-aware user interface generation program to generate the user interface based on the list of actions. We also propose a prototype system that implements the workflow driven context-aware application based on the complete Petri net models.

## Acknowledgement

This paper is supported in part by National Science Council of Taiwan under Grant NSC97-2410-H-324-017-MY3.

## References

1. Cheverst, K., et al.: Experiences of developing and deploying a context-aware tourist guide: The GUIDE project. In: 6th International Conference on Mobile Computing and Networking, Boston, pp. 20–31 (August 2000)
2. Patel, S.N., Abowd, G.D.: The ContextCam: Automated point of capture video annotation. In: Davies, N., Mynatt, E.D., Siio, I. (eds.) UbiComp 2004. LNCS, vol. 3205, pp. 301–318. Springer, Heidelberg (2004)
3. Shi, Y., et al.: The smart classroom: Merging technologies for seamless tele-education. IEEE Pervasive Computing 2(2), 47–55 (2003)
4. Chen, H., et al.: Intelligent agents meet semantic web in a smart meeting room. In: 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 854–861 (July 2004)
5. Helal, S., et al.: Enabling location-aware pervasive computing applications for the elderly. In: 1st IEEE Conference on Pervasive Computing and Communications, Fort Worth (March 2003)
6. Davis, F.D., Venkatesh, V.: Measuring user acceptance of emerging information technologies: An assessment of possible method biases. In: Proc. 28th Hawaii Int. Conf. System Sciences, pp. 729–736 (1995)
7. Schilit, B.N., Theimer, M.M.: Disseminating active map information to mobile hosts. IEEE Network 8(5), 22–32 (1994)
8. Hsieh, F.-S.: Context-aware Workflow Driven Resource Allocation for e-Healthcare. In: The Ninth International Conference on e-Health Networking, Application & Services (Healthcom 2007), June 19–22, pp. 34–39 (2007)

9. Croci, F., Perona, M., Pozzetti, A.: Work-force management in automated assembly systems. *International Journal of Production Economics* 64(1-3), 243–255 (2000)
10. Murata, T.: Petri Nets: Properties, Analysis and Applications. *Proceedings of the IEEE* 77(4), 541–580 (1989)
11. Hsieh, F.-S.: Robustness of deadlock avoidance algorithms for sequential processes. *Automatica* (39), 1695–1706 (2003)
12. Hsieh, F.-S.: Fault Tolerant Deadlock Avoidance Algorithm for Assembly Processes. *IEEE Transaction on System, Man and Cybernetics, Part A* 34(1), 65–79 (2004)
13. Hsieh, F.-S.: Robustness analysis of Petri nets for assembly/disassembly processes with unreliable resources. *Automatica* 42(7), 1159–1166 (2006)
14. Hsieh, F.-S.: Analysis of Flexible Assembly Processes based on Structural Decomposition of Petri Nets. *IEEE Transaction on System, Man and Cybernetics, Part A* 37(5), 792–803 (2007)
15. Hsieh, F.-S.: Robustness analysis of holonic assembly/disassembly processes with Petri nets. *Automatica* 44(10), 2538–2548 (2008)
16. Yu, S., Hsu, W., Pung, H.-K.: KPN: a Petri net model for general knowledge representation and reasoning. In: *Proceedings of the 1998 IEEE International Conference on System, Man and Cybernetics*, pp. 184–189 (1998)
17. Lianxiang, J., Huawang, L., Genqing, Y., Qingrong, Y.: An Improved Fault Petri Net for Knowledge Representation. In: *Proceedings of the 2009 Computational Intelligence and Software Engineering*, pp. 1–4 (2009)
18. Zhang, Z., Yang, Z., Liu, Q.: Modeling Knowledge Flow Using Petri Net. In: *2008 International Symposium on Knowledge Acquisition and Modeling*, pp. 142–146 (2008)
19. Weber1, M., Kindler, E.: The Petri Net Markup Language (2002),  
[http://www2.informatik.hu-berlin.de/top/pnml/  
download/about/PNML\\_LNCS.pdf](http://www2.informatik.hu-berlin.de/top/pnml/download/about/PNML_LNCS.pdf)
20. Hsieh, F.-S.: Model And Control Holonic Manufacturing Systems Based On Fusion Of Contract Nets And Petri Nets. *Automatica* (40), 51–57 (2004)
21. Hsieh, F.-S.: Analysis of contract net in multi-agent systems. *Automatica* 42(5), 733–740 (2006)
22. Hsieh, F.-S.: Hierarchy Formation and Optimization in Holonic Manufacturing Systems with Contract net. *Automatica* 44(4), 959–970 (2008)
23. Hsieh, F.-S.: Collaborative reconfiguration mechanism for holonic manufacturing systems. *Automatica* 45(11), 2563–2569 (2009)
24. Hsieh, F.-S.: Dynamic composition of holonic processes to satisfy timing constraints with minimal costs. *Engineering Applications of Artificial Intelligence* 22(7), 1117–1126 (2009)
25. Hsieh, F.-S.: Developing cooperation mechanism for multi-agent systems with Petri nets. *Engineering Applications of Artificial Intelligence* 22(4-5), 616–627 (2009)
26. Hsieh, F.-S.: Design of Reconfiguration Mechanism for Holonic Manufacturing Systems based on Formal Models. *Engineering Applications of Artificial Intelligence* (2010) doi: 10.1016/j.engappai.2010.05.008
27. Hsieh, F.-S., Chiang, C.Y.: Collaborative composition of processes in holonic manufacturing systems. *Computer In Industry* (2010), doi:10.1016/j.compind.2010.05.012

# A Consensus-Based Method for Fuzzy Ontology Integration\*

Ngoc Thanh Nguyen<sup>1</sup> and Hai Bang Truong<sup>2</sup>

<sup>1</sup> Institute of Informatics, Wroclaw University of Technology, Poland

ngoc-thanh.nguyen@pwr.edu.pl

<sup>2</sup> University of Information Technology, Ho Chi Minh City, Vietnam

bangth@uit.edu.vn

**Abstract.** Ontology can be treated as the background of a knowledge-based system. Fuzzy ontologies in many cases seem to be more useful than non-fuzzy ontologies because of the possibility for distinguishing the degrees to which concepts describe a real world, or relations between them. This paper includes a framework of consensus-based method for fuzzy ontology integration. For this aim a conception for fuzzy ontology definition is proposed and three problems for fuzzy ontology integration on concept and relation levels are presented. For these problems several algorithms have been proposed.

## 1 Introduction

Fuzzy ontologies have been more and more popular in playing an important role for knowledge systems [15]. In the difference with non-fuzzy ontologies, fuzzy ontologies enable among others distinguishing the degrees to which concepts describe a real world. Fuzzy ontologies are useful in such fields as Information Retrieval or in describing domains (fuzzy domain ontologies) [16]. In Information Retrieval ontology is often used to representing user preference. Fuzzy ontology is here very useful since owing to it degrees of user interests in terms can be introduced. Such structure better reflexes the user profile. In domain ontologies fuzzy elements can be helpful in representing the importance degrees of some attributes in describing a concept, or in relations between concepts [17].

Most often an (non-fuzzy) ontology is defined by the following elements:

- $C$  – a set of concepts (classes);
- $R$  – set of binary relations defined on  $C$ ;
- $Z$  – set of axioms, which formulas of the first order logic and can be interpreted as integrity constraints or relationships between instances and concepts, and which can not be expressed by the relations in set  $R$ .

In general, on the basis of the literature, we can state that an ontology is fuzzy if one of the above mentioned elements is fuzzy. The problems of non-fuzzy ontology integration have been solved in many works in the literature. In this paper we present

---

\* This paper was partially supported by Polish Ministry of Science and Higher Education under grant no. N N519 407437.

a definition of fuzzy ontology and a consensus-based approach for fuzzy ontology integration. Using consensus methods for integrating fuzzy ontologies is novel since, to the best knowledge of the authors, this approach is missing in the literature. The remaining part of this paper is organized as follows: In the next section an overview of fuzzy ontology definitions is presented. Section 3 includes a new definition of fuzzy ontology and the problems of integration on levels of concepts and relations. Section 4 contains several algorithms for integration. Some future works will be defined in the last section.

## 2 Related Works

In general, the problem of ontology integration can be formulated as follows: For given ontologies  $O_1, \dots, O_n$  one should determine one ontology which could replace them. Ontology integration is useful when there is a need to make fusion (or merge) of systems in which ontologies  $O_1, \dots, O_n$  are used. For integrating non-fuzzy ontologies many works have been done and presented, among others in [4]-[7], [15]. The main subject of these works is strongly related to using ontology in semantic web.

Fuzzy ontology conception is younger and not so many researchers have been dealing with this subject. For this kind of ontologies one can distinguish two group of works. The first of them consists of logical-based approaches, where several works tried to couple both fuzzy and distributed features using description and fuzzy logics [3], [8]. The main contribution of these papers is to propose a discrete tableau algorithm to achieve reasoning within this new logical system. In the non-logic approach Abulaish and Dey [1] proposed a fuzzy ontology generation framework in which a concept descriptor is represented as a fuzzy relation which encodes the degree of a property value using a fuzzy membership function. The fuzzy ontology framework provides appropriate support for application integration by identifying the most likely location of a particular term in the ontology. In [2] Blanco et al. proposed a flexible ontology system is proposed to enable to store fuzzy information in databases. This proposal allows users to manage imprecise and classic information.

The largest disadvantage in the approaches for ontology integration is that there is lack of clear criteria for integration. Most often proposed algorithms refer to concrete situations and their justification is rather intuitive than formal.

In this paper we propose to use consensus theory to fuzzy ontology integration. The advantages of such approach are based on the fact that consensus methods are very useful in solving many kinds of conflicts or inconsistencies which very often appear in integration tasks. Besides, consensus methods possess well-defined criteria. As it is well known, ontologies even describing the same real world often contain many inconsistencies because they have been created by autonomous systems. Using consensus methods for integrating fuzzy ontologies is novel since, to the best knowledge of the authors, this approach is missing in the literature.

## 3 A Proposal for Fuzzy Ontology Integration

### 3.1 Definition of Fuzzy Ontology

We assume a real world  $(A, V)$  where  $A$  is a finite set of attributes describing it and  $V$  – the domain of  $A$ , that is  $V$  is a set of attribute values, and  $V = \bigcup_{a \in A} V_a$  ( $V_a$  is the

domain of an attribute  $a$ ). We consider domain ontologies referring to the real world ( $A, V$ ), such ontologies are called  $(A, V)$ -based. As a fuzzy  $(A, V)$ -based ontology we understand a triple:

$$\text{Fuzzy ontology} = (\mathbf{C}, \mathbf{R}, \mathbf{Z})$$

where:

1.  $\mathbf{C}$  is the finite set of concepts. A concept of a fuzzy ontology is defined as a triple:

$$\text{concept} = (c, A^c, V^c, f^c)$$

where  $c$  is the unique name of the concept,  $A^c \subseteq A$  is a set of attributes describing the concept and  $V^c \subseteq V$  is the attributes' domain:  $V^c = \bigcup_{a \in A^c} V_a$  and  $f^c$  is a fuzzy function:

$$f^c: A^c \rightarrow [0,1]$$

representing the degrees to which concept  $c$  is described by attributes. Triple  $(A^c, V^c, f^c)$  is called the *fuzzy structure* of concept  $c$ .

2.  $\mathbf{R}$  is a set of fuzzy relations between concepts,  $\mathbf{R} = \{R_1, R_2, \dots, R_m\}$  where

$$R_i \subseteq \mathbf{C} \times \mathbf{C} \times (0, 1]$$

for  $i = 1, 2, \dots, m$ . A relation is then a set of pairs of concepts with a weight representing the degree to which the relationship should be. We assume that within a relation  $R_i$  in an ontology a relationship can appear between two concepts only with one value of weight, that is if  $(c, c', v) \in R_i$  and  $(c, c', v') \in R_i$  then  $v = v'$ .

3.  $\mathbf{Z}$  is a set of axioms. In this paper we will not deal with them. This will be a subject of another work in the future.

Such defined structure can be useful in representing a document preference profile of a user in information retrieval tasks or in using weights for distinguishing the importance degrees of attributes in describing a real world.

### 3.2 General Tasks of Fuzzy Ontology Integration

In general, an integration task most often refers to a set of elements (objects) with the same kind of structures, the aim of which is based on determining an element best representing the given. The kinds of structures mean for example relational, hierarchical, table etc. The words “best representation” mentioned above mean the following criteria:

- All data included in the elements to be integrated should be in the result of integration. This criterion guarantees the completeness, that is all information included in the component elements will appear in the integration result.
- All conflicts appearing among elements to be integrated should be solved. It often happens that referring to the same subject different elements contain inconsistent information. Such situation is called a conflict. The integration result should not contain inconsistency, so the conflicts should be solved.
- The kind of structure of the integration result should be the same as of the given elements.

Integration tasks are very often realized for database integration or knowledge integration. Ontology integration is a special case of the second case. Ontologies have well-defined structure and it is assumed that the result of ontology integration is also an ontology, therefore usually the first and second criteria are used.

It seems that satisfying the first criterion is simple since one can make the sum of all sets of concepts, relations and axioms from component ontologies in the integration process. However, it is not always possible because of the following reasons:

- Appearance of all elements in the integration result may contain inconsistency in the sense that some of the component ontologies may be in conflict and this conflict will be moved to the integration result.
- Summing all elements may cause lose of the ontology structure.

Satisfying the second criterion is based on solving conflicts, for example, by using consensus methods.

Conflicts between fuzzy ontologies may be considered on the following levels:

- Conflicts on concept level: The same concept has different fuzzy structures in different ontologies.
- Conflicts on relation level: The relations for the same concepts are different in different ontologies.

Conflicts mentioned in the above way are very general and inaccurate. In the following sub-sections we will provide with the concrete definitions of them.

### 3.3 Integration on Concept Level

On this level we assume that two ontologies differ from each other in the structures of their concepts. That means these ontologies can contain the same concept but its structure is different in each ontology. The reason of this phenomenon is that these ontologies come from different autonomous systems. Therefore, although they refer to the same real world, they can have different structures.

**Definition 1.** Let  $O_1$  and  $O_2$  be  $(A, V)$ -based ontologies. Let concept  $(c_1, A^{c^1}, V^{c^1}, f^{c^1})$  belongs to  $O_1$  and let concept  $(c_2, A^{c^2}, V^{c^2}, f^{c^2})$  belongs to  $O_2$ . We say that a conflict takes place on concept level if  $c_1 = c_2$  but  $A^{c^1} \neq A^{c^2}$  or  $V^{c^1} \neq V^{c^2}$  or  $f^{c^1} \neq f^{c^2}$ .

Definition 1 specifies such situations in which two ontologies define the same concept in different ways. For example concept *person* in one ontology may be defined by attributes with weights: *Name*(1.0), *Age*(0.8), *Address*(0.3), *Gender*(0.9), *Job*(0.4), while in other system it is defined by attributes: *Name*(0.9), *Age*(0.7), *Address*(0.3), *Gender*(0.9), *Job*(0.4), *Taxpayer's identification number*(1.0), *Occupation*(0.2). Note that the same attributes can occur in both ontologies for the same concept, but the weights may be different.

Thus on concept level the problem of fuzzy ontology integration is the following:

#### Problem FOI-1:

For given a set of fuzzy structures of the same concept

$$X = \{(A^i, V^i, f^i) : (A^i, V^i, f^i) \text{ is the fuzzy structure of concept } c \text{ in ontology } O_i \text{ for } i=1, \dots, n\}$$

it is needed to determine a triple  $c^* = (A^*, V^*, f^*)$  which best represents the given structures.

Criterion “best represents” mean that one or more postulates should be satisfied by  $c^* = (A^*, V^*, f^*)$ . In general, we would like to determine such  $A^*$  of final attributes that all attributes which appear in sets  $A^i$  ( $i=1,\dots,n$ ) are taken into account in this set. However, we cannot simply make the sum of  $A^i$ . According to the general criteria defined in Section 3.2, we should analyze the following postulates:

- All attributes from sets  $A^i$  for  $i = 1, \dots, n$  should appear in  $c^*$ .
- For each attribute appearing in  $c^*$  its domain should include all its domains in component ontologies, if any.
- Any inconsistency referring to attributes should be solved. Inconsistency may refer to attribute occurrences in component ontologies as well as their domains and fuzzy functions.

As stated above, in general it is impossible to satisfy all postulates simultaneously.

### 3.4 Integration on Relation Level

In works [11], [14] the author defined conflicts on relation level between non-fuzzy ontologies: Two ontologies are in inconsistency on relation level if referring to a relation a pair of concepts is in this relation in one ontology but in the second ontology it is not. For fuzzy ontologies the definition is similar with taking into account the fuzzy functions of concepts and the weights of relationships between them.

Similarly as in [11] we investigate two kinds of ontology inconsistency on relation level. The first kind of inconsistency refers to situations where between the same concepts  $c$  and  $c'$  different ontologies assign different relations. As an example let's consider concepts *Man* and *Woman*, in one ontology they are in relation *Marriage*, in another ontology they are in relation *Kinship*. Notice also that within the same ontology two concepts may be in more than one relation. Besides, the same pair of concepts can belong to the same relation, but with different weights in different ontologies. Let us denote by  $R_{ij}(c, c')$  the relationships between concepts  $c$  and  $c'$  within relation  $R_i$  and ontology  $O_j$  for  $i = 1, \dots, m, j = 1, \dots, n$ , i.e.

$$R_{ij}(c, c') = \langle c, c' \rangle \in R_i$$

where  $c$  and  $c'$  belong to ontology  $O_j$ .

We assume also that the set of concepts is the same for all ontologies to be integrated.

**Definition 2.** Let  $O_1$  and  $O_2$  be  $(A, V)$ -based ontologies. Let concepts  $c$  and  $c'$  belong to both ontologies. We say that an inconsistency takes place on relation level if

$$R_{i1}(c, c') \neq R_{i2}(c, c')$$

for some  $i \in \{1, \dots, m\}$ .

As an example let's consider 3 relations in two ontologies with concepts  $a$ ,  $b$  and  $c$  represented by the following table:

	$R_1$	$R_2$	$R_3$
$O_1$	$\langle a, b, 0.5 \rangle$	$\langle a, c, 1 \rangle$	$\langle a, b, 0.2 \rangle$
	$\langle a, c, 0.7 \rangle$	$\langle b, c, 0.3 \rangle$	$\langle b, a, 0.5 \rangle$
	$\langle b, c, 0.2 \rangle$	$\langle b, a, 0.9 \rangle$	$\langle c, a, 0.7 \rangle$
$O_2$	$\langle b, a, 0.5 \rangle$	$\langle c, a, 0.5 \rangle$	$\langle c, a, 0.5 \rangle$
	$\langle a, b, 0.7 \rangle$	$\langle c, b, 0.8 \rangle$	$\langle a, b, 0.4 \rangle$
	$\langle c, a, 0.2 \rangle$	$\langle a, c, 0.1 \rangle$	

We can note that an inconsistency appears because, for example,

$$R_{11}(a,b) = \langle a, b, 0.5 \rangle,$$

while

$$R_{12}(a,b) = \langle a, b, 0.7 \rangle.$$

Another inconsistency takes place because  $R_{31}(b,c) = \langle b, c, 0.7 \rangle$  while  $R_{31}(b,c)$  does not exist. In this case we can assume  $R_{31}(b,c) = \langle b, c, 0 \rangle$ .

Based on the defined kinds of inconsistencies between ontologies, we have the following problems of fuzzy ontology integration:

#### Problem FOI-2:

For given  $i \in \{1, \dots, m\}$  and set  $X = \{R_{ij}(c, c'): j = 1, \dots, n\}$  of relationships between two concepts  $c$  and  $c'$  in  $n$  ontologies it is needed to determine  $R_i(c, c')$  of final relationship between  $c$  and  $c'$  which best represents the given relationships.

In this problem all relationships are treated as independent on each other. However, it is always possible because some relation can be, for example transitive, that is if pair  $\langle c, c' \rangle$  belongs to  $R_i$  with some weight  $v_1$  and pair  $\langle c', c'' \rangle$  belongs to  $R_i$  with weight  $v_2$  then pair  $\langle c, c'' \rangle$  belongs to  $R_i$  with weight  $v = \min\{v_1, v_2\}$ . In this case each pair of concepts cannot be treated separately.

In more general case where there are known the characteristics of relations, for example symmetric, transitive, the problem of fuzzy ontology integration is more complex. The second kind of ontology inconsistency is related to situations in which the same relationship is defined differently in different ontologies. For example, for the same set of concepts {Student, Lecturer, Hand-book} relation *Preference* can be transitive, that is if a student prefers a lecture and this lecture prefers a handbook, then the student also prefers the handbook to some degree. Besides, in one ontology this relation occurs only between concepts *Student* and *Hand-book*, while in other ontology it occurs between concepts *Lecturer* and *Hand-book*. The ontology integration problem referring to this case is formulated as follows:

#### Problem FOI-3:

For given  $i \in \{1, \dots, m\}$  and set of relations  $X = \{R_{ij} \subseteq C \times C \times (0, 1]: j = 1, \dots, n\}$  it is needed to determine relation  $R_i \subseteq C \times C \times (0, 1]$  which best represents the given relations.

## 4 Consensus-Based Approaches for Fuzzy Ontology Integration

In this section we present a consensus-based approach for integrating fuzzy ontologies on concept and relation levels.

### 4.1 Outline of Consensus Methods

In short consensus methods are useful in a process of solving conflict or data inconsistency. The scheme of using a consensus method can be presented as follows [10]:

1. Defining the set of potential versions of data
2. Defining the distance function between these versions
3. Selecting a consensus choice criterion
4. Working out an algorithm for consensus choice

Defining distance functions for potential versions of data is very important task and is different for different structures of data. For selecting a consensus choice criterion there are known two consensus choice functions. The first of them is the median function defined by Kemeny [9], which minimizes the sum of distances between the consensus and given inconsistent versions of data. The second function minimizes the sum of squared distances from consensus to given elements [13]. As the analysis has shown, the first function gives the closest representative of the given versions while the second gives a consensus being a good compromise of them [10]. Let's denote the consensuses chosen by these functions by  $O_1$ -consensus and  $O_2$ -consensus, respectively. The choice of a consensus function should be dependent on the conflict situation. If we assume that the consensus represents a unknown solution of some problem then there two cases [12]:

- In the first case the solution is independent on the given versions of data. Thus the consensus should be the closest representative to the conflict versions of data. For this case the criterion for minimizing the sum of distances between the consensus and the conflict versions should be used, that is  $O_1$ -consensus should be determined.
- In the second case the solution is dependent on the given versions of data. Then the consensus should be a compromise which neither “harm” nor “prefer” any of the given versions of data. For this case a  $O_2$ -consensus should be determined.

We will now try to use these consensus functions for fuzzy ontology integration.

### 4.2 Algorithm for Problem FOI-1

**Algorithm 1.** Determining integration of fuzzy concepts.

*Input:* Given set of fuzzy structures of a concept in  $n$  ontologies

$$X = \{(A^i, V^i, f^i) : (A^i, V^i, f^i) \text{ is the fuzzy structure of concept } c \text{ in ontology } O_i \text{ for } i=1, \dots, n\}$$

*Output:* Triple  $c^* = (A^*, V^*, f^*)$  which best represents the elements from  $X$ .

*Procedure:*

```

BEGIN
  Set  $A^* = \bigcup_{i=1}^n A^i$  ;
  Set  $V^* = \bigcup_{i=1}^n V^i$  ;
  For each  $a \in A^*$  do
    Begin
      Determine multi-set  $X_a = \{f^i(a) : \text{if } f^i(a) \text{ exists and } i=1,\dots,n\}$ ;
      Calculate  $f^*(a) = \frac{1}{\text{card}(X_a)} \sum_{v \in X_a} v$  ;
    End.
  END.

```

The computation complexity of Algorithm 1 is  $O(n^2)$ . It is possible to prove that this algorithm determines an integration satisfying  $O_2$ -consensus function.

### 4.3 Algorithm for Problem FOI-2

**Algorithm 2.** Determining integration of fuzzy relations.

*Input:* Given set of  $X = \{R_{ij}(c, c') : j = 1, \dots, n\}$  of relationships between two concepts  $c$  and  $c'$  in  $n$  ontologies

*Output:* Triple  $R_i(c, c') = (c, c', v)$  which best represents the elements from  $X$ .

*Procedure:*

```

BEGIN
  Order set  $X$  in increasing order giving  $X = \{x_1, x_2, \dots, x_n\}$ ;
  Set interval  $\left( x_{\left\lfloor \frac{n+1}{2} \right\rfloor}, x_{\left\lfloor \frac{n+2}{2} \right\rfloor} \right)$ ;
  Set  $v$  as a value belonging to the above defined interval;
END.

```

The computation complexity of Algorithm 2 is  $O(n^2)$ . It is possible to prove that this algorithm determines an integration satisfying  $O_1$ -consensus function.

### 4.4 Algorithm for Problem FOI-3

**Algorithm 3.** Determining integration of fuzzy relations.

*Input:* - Given set  $X = \{R_{ij} \subseteq C \times C \times (0, 1) : j = 1, \dots, n\}$  of relations of the same kind between concepts in  $n$  ontologies.

- These relations are transitive.

*Output:* Relation  $R_i \subseteq C \times C \times (0, 1]$  which best represents the elements from  $X$ .

*Procedure:*

```

BEGIN
  Set  $R_i = \emptyset$ ;

```

For each pair  $(c, c') \in \mathbf{C} \times \mathbf{C}$  do  
 Begin  
   Determine multi-set  $X_{(c,c')} = \{v: \langle c, c', v \rangle \in R_{ij} \text{ for } j = 1, \dots, n\}$ ;  
   Order set  $X_{(c,c')}$  in increasing order giving  $X = \{x_1, x_2, \dots, x_k\}$ ;  
   Set interval  $\left(x_{\left[\frac{k+1}{2}\right]}, x_{\left[\frac{k+2}{2}\right]}\right)$ ;  
   Set  $v$  as a value belonging to the above defined interval;  
   Set  $R_i := R_i \cup \{\langle c, c', v \rangle\}$   
 End;  
 For each  $(c, c', c'') \in \mathbf{C} \times \mathbf{C} \times \mathbf{C}$  do  
 Begin  
   If  $\langle c, c', v_1 \rangle \in R_i, \langle c, c', v_2 \rangle \in R_i$  and  $\langle c, c', v_3 \rangle \in R_i$  then  
     change  $v_3 = \min \{v_1, v_2\}$ ;  
     If only  $\langle c, c', v_1 \rangle \in R_i$  and  $\langle c, c', v_2 \rangle \in R_i$  then set  $R_i := R_i \cup \{\langle c, c', v_3 \rangle\}$   
     where  $v_3 = \min \{v_1, v_2\}$ ;  
 End  
 END.

The computation complexity of Algorithm 3 is  $O(n^3)$ . It is possible to prove that in general this algorithm determines an integration satisfying  $O_1$ -consensus function. Because of the limited space for the paper, the proof will be included in an extended work.

## 5 Conclusions

In this paper a consensus-based approach for integration fuzzy ontologies on levels of concepts and relations is presented. The worked out algorithms require deeper analysis for stating their applications. The future work should concern testing these algorithms and performing experiments justifying their usefulness.

## References

1. Abulaish, M., Dey, A.: A Fuzzy Ontology Generation Framework for Handling Uncertainties and Non-uniformity in Domain Knowledge Description. In: Proceedings of the International Conference on Computing: Theory and Applications, pp. 287–293. IEEE, Los Alamitos (2007)
2. Blanco, I.J., Vila, M.A., Martinez-Cruz, C.: The Use of Ontologies for Representing Database Schemas of Fuzzy Information. International Journal of Intelligent Systems 23(4), 419–445 (2008)
3. Calegari, S., Ciucci, D.: Fuzzy Ontology, Fuzzy Description Logics and Fuzzy-OWL. In: Masulli, F., Mitra, S., Pasi, G. (eds.) WILF 2007. LNCS (LNAI), vol. 4578, pp. 118–126. Springer, Heidelberg (2007)
4. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Method for Integration across Text Corpus and WordNet-based Ontologies. In: IEEE/ACM/WI/IAT 2008 Workshops Proceedings, pp. 1–4. IEEE Computer Society, Los Alamitos (2008)

5. Duong, T.H., Jo, G.S., Jung, J.J., Nguyen, N.T.: Complexity Analysis of Ontology Integration Methodologies: A Comparative Study. *Journal of Universal Computer Science* 15(4), 877–897 (2009)
6. Duong, T.H., Nguyen, N.T., Jo, G.S.: A Method for Integrating Multiple Ontologies. *Cybernetics and Systems* 40(2), 123–145 (2009)
7. Fernandez-Breis, J.T., Martinez-Bejar, R.: A Cooperative Framework for Integrating Ontologies. *Int. J. Human-Computer Studies* 56, 665–720 (2002)
8. Jianjiang, L., et al.: Distributed Reasoning with Fuzzy Description Logics. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) *ICCS 2007. LNCS*, vol. 4487, pp. 196–203. Springer, Heidelberg (2007)
9. Kemeny, J.G.: Mathematics without Numbers". *Daedalus* 88, 577–591 (1959)
10. Nguyen, N.T.: Using Distance Functions to Solve Representation Choice Problems. *Fundamenta Informaticae* 48, 295–314 (2001)
11. Nguyen, N.T.: A Method for Ontology Conflict Resolution and Integration on Relation Level. *Cybernetics and Systems* 38(8), 781–797 (2007)
12. Nguyen, N.T.: Advanced methods for inconsistent knowledge management. Springer, London (2008)
13. Nguyen, N.T.: Consensus system for solving conflicts in distributed systems. *Journal of Information Sciences* 147, 91–122 (2002)
14. Nguyen, N.T.: Conflicts of Ontologies – Classification and Consensus-based Methods for Resolving. In: Gabrys, B., Howlett, R.J., Jain, L.C. (eds.) *KES 2006. LNCS (LNAI)*, vol. 4252, pp. 267–274. Springer, Heidelberg (2006)
15. Noy, N.F., Musen, M.A.: SMART: Automated Support for Ontology Merging and Alignment. In: Proc. of the 12th Workshop on Knowledge Acquisition, Modelling and Management (KAW 1999), Banff, Canada, pp. 1–20 (1999)
16. Pinto, H.S., Martins, J.P.: A Methodology for Ontology Integration. In: Proceedings of the First International Conference on Knowledge Capture, pp. 131–138. ACM Press, New York (2001)
17. Reimer, U.: Knowledge Integration for Building Organizational Memories. In: Proceedings of the 11th Banff Knowledge Acquisition for Knowledge Based Systems Workshop, vol. 2, pp. KM-6.1–KM-6.20 (1998)

# Contextual Information Search Based on Ontological User Profile

Nazim udin Mohammed, Trong Hai Duong, and Geun Sik Jo

School of Computer and Information Engineering,

Inha University, Korea

nazim@eslab.inha.ac.kr, haiduongtrong@gmail.com, gsjo@inha.ac.kr

**Abstract.** Internet users use the web to search for information they need. Every user has some particular interests and preferences when he/she searches information on the web. It is challenging to trace the exact interests of a user by a system to provide the information he/she wants. Personalization is a popular technique in information retrieval to present information based on an individual user's needs. The main challenges of effective personalization are to model the users and identify the users' context for accessing information. In this paper, we propose a framework to model the user details and context for personalized web search. We construct an ontological user profile describing the users preferences based on the users context. Finally, we use a semantic analysis of the log files approach for the initial construction of the ontological users profile and learn the profile over time. Web information can be accessed based on the ontological user profiles, re-ranking the searched results considering the users' context. Experiments show that our ontological approach to modeling users and context enables us to tailor the web search results for users based on users' interests and preferences.

**Keywords.** Ontology, Personalization, User Profile, Context, Ontology-based User Profile.

## 1 Introduction

Information retrieval deals with extracting necessary information from huge data repositories on the current web. Extraction of information takes place by submitting a user query to the search engine. Not all the search results returned by traditional search engines are relevant to the user query or user needs due to the heavy information overload. Numerous methods and techniques have been developed to alleviate this problem. A common technique in information retrieval is personalization to address the issue of information overloading. Personalization is a user-centric method to model the user's preferences in the form of a user profile or personal profile. Personalization is a successful technique in the domain of news recommendation [4] to web search [10]. A user profile and category hierarchy is created and mapped to improve retrieval effectiveness in personalized web search[10] based on the user's search histories. Creation of a user profile collects the necessary information about individual users. User profiles can be constructed by collecting information of users explicitly or implicitly. User references can be traced automatically by the user 's click history[11] and analyzing the

log files [9]. Semantic web personalization [1],[6],[7],[2] has drawn much attention for web search and recommender systems. The ontological approach is a proven technique to model users and context in the field of information retrieval. User preferences can be described as ontology for personalized semantic search[8]. The user profile can be represented as an ontology describing users' details in the form of a concept hierarchy. A new ontology can be created to represent a user's general information, such as name, age, birth date, educational background, to more specific information describing the user's interests. The ontology can be created manually collecting user's data [14] or automatically[3] based on an existing knowledge base. The users profile can be considered also as an instance of a pre-existing reference domain ontology [8],[5]. Additionally, User context is an important element to identify user's interests in personalization. Ontology-driven approach has been undertaken to enrich the representational ground for content meaning , users' interests, and contextual conditions for personalized information retrieval [6].

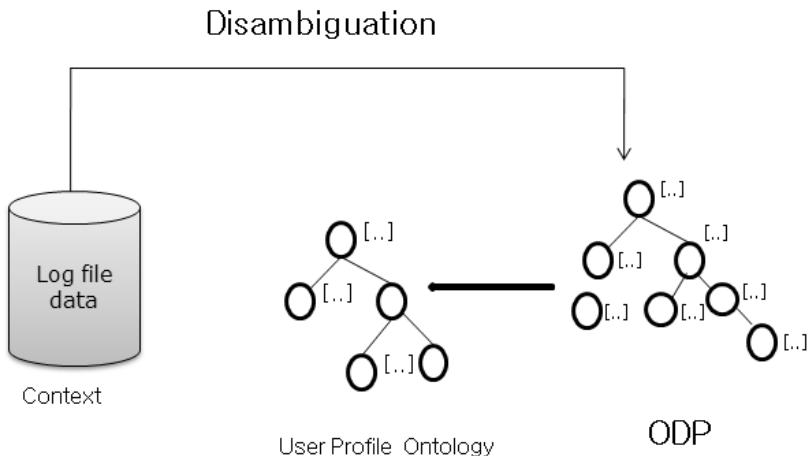
In this paper, we propose a personalized framework for personalized information search on the web. Our personalized framework is based on modeling users in the form of a user profile. We consider the ontological approach to create user profiles to describe and manage the user's preferences semantically. Most of the user's search is accomplished by entering a keyword query into search engines. This keyword query is treated as the current context of a user at a particular time. Modeling the current context is also necessary to identify the user's intention in personalized search. We use the WordNet [15] based approach to model the context to expand the meaning of the context in the form of a concept hierarchy. The remainder of the paper is organized as follows. Section 2 defines the detail of building the ontological personal profile. Based on this profile personalized searching and re-ranking are explained with our framework in section 3. Section 4 shows the experimental evaluation of our approach and finally, we conclude our approach in section 5.

## 2 Ontology-Based Personalization

Ontologies are used by a number of researchers to improve the web navigation, as well as for personalized web search and browsing. Ontology is used in [18] to focus on encoding semantics into web pages to describe their content. User contexts are utilized to personalize search results by re-ranking the search results returned from a search engine for a given query. We investigate the techniques to create user profiles automatically using the ontological approach for personalized semantic web search. In our approach, the user profile is considered as an instance of reference ontology, where each concept represents the user interests. User profile ontology is generated by mapping between the reference ontology and user context. Details about profile construction and context modeling are described in the following subsections. Figure 1 depicts the approach to construct the initial user profile.

Ontology specifies a conceptualization of a domain in terms of concepts, attributes and relationships. Concepts are typically organized into a tree structure based on the subsumption relationship among concepts. Here, we consider ontology is a tuple

$$O = (C, R, S, I) \quad (1)$$



**Fig. 1.** Construction of Initial Profile

where,  $C$  is the set of classes,  $R$  is the set of relations between classes ,  $S$  represents the set of relationship and  $I$  is for instances of the class.

The ontological users' profiles are created by the alignment of user interest to reference ontology . Reference ontology is a pre-existing hierarchical ontology scheme to describe a knowledge base in a specific domain. In our approach, users' profiles are created based on reference ontology. We use ODP (Open Directory Project) as reference ontology to build the profiles. ODP [16] is an open content directory of web pages maintained by a community of volunteer editors. ODP uses a hierarchical ontology schema to represent topics and web pages that belong to these topics. The personal profile is created with the user's contextual information based on this reference ontology. The profile is maintained and updated with the user's context and his/her ongoing behavior.

## 2.1 Contextual Information Collection

In our approach, we use the user's search histories for a particular period of time to construct an initial user profile. The log file provides the details of a user's search history, considering the time a user spends on a particular page, frequency of the visited page and click path analysis of switching from one page to another page. The contents of search histories are classified to a pre-existing hierarchy of domain concepts to construct an ontological user's profile. Additionally, we consider the user's short-term information needs as a current context in the form of a search query. A query consists of textual format of one or more keywords to represents the user's needs. Query context is utilized at the time of the information search for incorporation with the user's profile.

## 2.2 Reference Ontology

We implement the reference ontology using ODP in the following way . We investigate the web pages as documents to represent the reference ontology. Each concept in the

reference ontology is associated with related documents. Textual data are extracted from the related documents of each concept, considered as index terms with feature vectors of related concepts. Feature vectors are made by considering the importance of terms in the documents belonging to the concepts. All the documents belonging to a concept are merged to create a collection of documents for each concept in the reference ontology to produce the index terms with feature vectors. The Porter stemmer is applied to preprocess the documents to remove stop words and common suffixes. Then, the tf.idf approaches are applied to extract the related terms representing the document [25].

**Feature vector of Document.** Let  $T^d = (t_1, t_2, \dots, t_n)$  be the collection of all of key-words (or index terms) of the document d. Term frequency  $tf(d, t)$  is defined as the number of occurrences of term t in document d. A set of term frequency pairs,  $P^d = \{(t, f) | t \in T^d, f > threshold\}$ , is called the pattern of a document.

Given a pattern  $P^d = \{(t_1, f_1), (t_2, f_2), \dots, (t_m, f_m)\}$ , let  $\vec{d}$  be the feature vector of document d and let  $td$  be the collection of corresponding terms to the pattern, we have:

$$\vec{d} = (w_1, w_2, \dots, w_m) \quad (2)$$

$$td = (t_1, t_2, \dots, t_m) \quad (3)$$

where

$$w_i = \frac{f_i}{\sum_{j=1}^m f_j} * \log \frac{|D|}{|d : t_i \in d|} \quad (4)$$

**Concept vector generation.** There is leaf concept and non-leaf concept in the reference ontology. Non-leaf concepts are those that have sub-concepts.

- For each leaf concept, the feature vector is calculated as the feature vector of a set of documents:

$$\vec{S^c} = \vec{ds_c} \quad (5)$$

- For each non-leaf concept, the feature vector is calculated by taking into consideration contributions from the documents that have been assigned to it ( $D^c$ ), its direct sub concepts ( $D^{c'}$ , for any  $c'$  is a direct sub concept of  $c$ ):

$$\vec{S^c} = \alpha \vec{ds_c} + \beta \vec{ds_{c'}} \quad (6)$$

where  $\vec{ds_c}$  and  $\vec{ds_{c'}}$  correspond to the feature vectors of the sets of documents  $D^c$  and  $D^{c'}$ ,  $0 \leq \alpha, \beta \leq 1$  and  $\alpha + \beta = 1$ .

### 2.3 User Profile Ontology

In our approach, automatic creation of the user profile ontology is accomplished by alignment of user context to the reference ontology constructed in the previous subsection. User context, representing the user's long-term interests, are collected by monitoring the search histories for a particular period of time and the collection of bookmarks.

In this approach, we have analyzed the user's search histories for a period of one month. The user's past click histories provide the content related metadata of the web pages that the user has browsed or searched. Content related metadata of user's extracted web pages provide the semantic information of the user's interests. We consider the click path of the web site the user navigated, the time consumed on the web pages to read them, and the frequency of web pages users used to identify the user interests. We used Gerald's[9] methods and techniques to analyze the log files to collect the meta information of the user's past history of retrieved web documents. Finally, a concept list is created with term vectors based on the search history. We term this the context vectors. These context vectors are mapped with the reference ontology to create an instance of the reference ontology that will be treated as the ontological user profile. This profile will further be incorporated with the query context for personalized information searching and re-ranking. The similarity between context and reference ontology is calculated by the cosine measure between the two representative feature vectors. Let two feature vectors for concept  $l$  and  $m$  respectively, have a length of  $n$ . The cosine similarity is calculated as:

$$\text{sim}(l, m) = \text{sim}(c^l, c^m) = \frac{\sum_{i=0}^n (C_i^l * C_i^m)}{\sqrt{\sum_{i=1}^n (C_i^l)^2} * \sqrt{\sum_{i=1}^n (C_i^m)^2}} \quad (7)$$

Where,

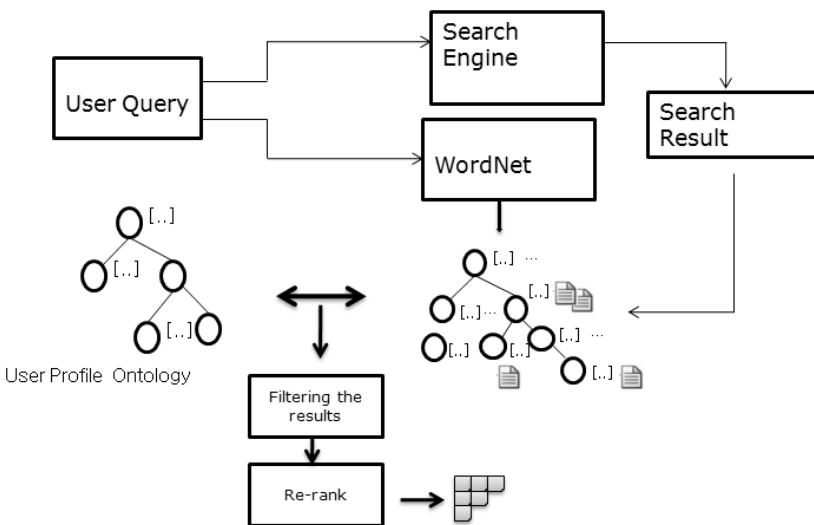
- $C^l$  and  $C^m$  are feature vectors for concepts  $l$  and  $m$ , respectively
- $n$  is length of feature vectors

A threshold value is maintained to select the similarity between two concepts. If the similarity value of a concept pair of less than the threshold, the pair will be considered as unrelated concepts.

The user's profile should be learned after a specified time to represent the user's updated preferences and interests. In our approach, we conveyed two ways to learn the user's profile. First, when a new query is generated and the user's profile did not contain any term related to the query, then an instance of the ODP ontology is created matching to the query ontology described in previous section and added to the user's profile ontology. Second, the user's profile could be learned with the collection of the user's previous search histories over one or two month(s). In our approach, we analyze the user's log file records, which provide semantics of the user's retrieved documents, to learn the user profile. We assume that learning the user's profile using these techniques makes it stable after a certain period with his/her updated interests and preferences.

### 3 Search Based on Ontological Personal Profile

In this section, we describe the semantic search based on our ontological personal profile approach. Figure 2 depicts the framework for personalized search and re-ranking the search results based on the personal profile ontology.



**Fig. 2.** Framework for Ontology Based Personalization

### 3.1 Query

The user's short-term information needs are determined by a search query. A search query consists of one or more search keyword in the form of text. A query is passed to the search engine and generates a list of search results. We consider that a query is the user's current context of information sought. We pass the query context to WordNet to generate a lexicographic hierarchy of query content. Web search results are mapped with this hierarchy to improve retrieval effectiveness.

### 3.2 WordNet

WordNet organizes lexical information in terms of word meanings called senses, rather than the word forms. Synsets are connected to each other through the external semantic relationships that are defined in WordNet. These relationships only connect word senses belonging to the same part of speech. Noun synsets are connected to each other through hypernym, hyponym, meronym, and holonym relationships. WordNet has been employed as a useful resource for automatic text analysis in information retrieval [17][19]. Since a user query is nothing but one or more keywords with a textual form, query disambiguation with WordNet can provide semantic matching between the user interest and search results. We model the user query with WordNet semantics to extend the meaning of a given query that can identify the actual user interests. The WordNet hypernym is used to produce a semantic hierarchy of query keywords to expand the meaning of the query. The WordNet hierarchy of the query will be treated as query ontology.

### 3.3 Filtering and Re-ranking the Search Results

Personalized search results are filtered and re-ranked based on the ontological user profile utilizing the user's context. In this context, we assume that the user's profile is relatively

stable with sufficient collection of information to represents the user's interests. Filtering determines which documents in the result lists are more relevant to the user's interests and preserves it by removing the irrelevant documents. The search result returned by the search engine for a given query is filtered by matching to the user profile ontology and query ontology. Documents are filtered by measuring the cosine distance between documents in the query ontology and concepts in the user profile. Details of the filtering process are described in Algorithm 1. In the worst case, if the user's profile is insufficient to determine the user's intention when a new query is encountered, then the ODP hierarchies can be utilized to map to the query ontology to filter the search results.

For a given query, re-ranking is done by modifying the order of ranking results returned by a filtering stage. We have designed a new algorithm to re-rank the search results based on the document-document similarity score considering a threshold value. We make a keyword pair  $\langle K_i, K_j \rangle$  for the search (filtered) results where  $i \neq j$  and measure the similarity scores. For example, if the word 'computer' appears several times in many documents, the score of the concept 'computer' will increase with the respective score value. Based on similarity scores, documents are re-ranked and returned to the user. The details of re-ranking methods are described in algorithm 2.

```

input : Results that are relevant to user's keyword
output: Results that are relevant to user's keyword and profile
1 Begin;
2 Generate set of Keywords representing results  $i$  that denotes  $r_i = k_1^i,..k_m^i$  and its
corresponding feature vectors are  $w_i = w_1^i,..w_m^i$  where  $k_j^i$  is a keyword with weight
 $w_j^i, j = 1...m$ ;
3 Generate set of Keywords representing concepts  $i$  in user profile, this denotes
 $p_i = p_1^i, p_2^i..p_m^i$  and its corresponding feature vectors are  $E_i = e_1^i, e_2^i..e_m^i$  where  $p_j^i$ 
is a keyword with weight  $e_j^i, j = 1...m$ ;
4  $R \leftarrow U_{i=1}^n R_i$ ;
5  $P \leftarrow U_{i=1}^n P_i$ ;
6 For each  $R_i$  in  $R$  do  $D_i(R_i, P) = \sum_{p_j \in k} d(R_i, P_j)$ , Where,  $d(R_i, P_j) = d(\vec{W_i}, \vec{E_j}) =$ 
 $1 - \frac{\sum_{(w_t^i \in R_i, e_h^j \in P_j)} (w_t^i * e_h^j)}{\sqrt{\sum_{w_t^i \in R_i} (w_t^i)^2} * \sqrt{\sum_{e_h^j \in P_j} (e_h^j)^2}}$ ;
7 chose  $N \subseteq R$ ;
8 Retrun(N);

```

**Algorithm 1.** Filtering the results

## 4 Experimental Evaluations

In this section, we discuss our methodology to perform the experiments to evaluate the effectiveness of the semantic search based on our framework.

### 4.1 Metrics

We measure effectiveness of re-ranking in terms of two widely used statistical methods in information retrieval, Recall and Precision. Recall is the ability of the search to find

all of the relevant documents in the corpus. Precision refers to the retrieval of top-ranked documents that are mostly related to users. Precision and Recall can be defined as

$$\text{precision} = \frac{\# \text{ of relevant documents retrieved}}{\# \text{ of documents retrieved}} \quad (8)$$

$$\text{Recall} = \frac{\# \text{ of relevant documents retrieved}}{\text{Total } \# \text{ of relevant documents retrieved}} \quad (9)$$

## 4.2 Data Sets

We use a three level ODP hierarchy, with a minimum of five documents indexed under each concept, to conduct experiments. We added some extra documents (web pages) related to ODP concepts if there are sufficient documents available in ODP. Every concept of ODP contains feature vectors from related documents. Feature vectors are measured by the *tf-idf* method, described in section 4. We write a program using Java to disambiguate the keyword from the query and make a concept hierarchy using WordNet. When the user generates a query to search information, the concept hierarchy is generated and mapped with the search results. We condense one or two keyword(s) query

```

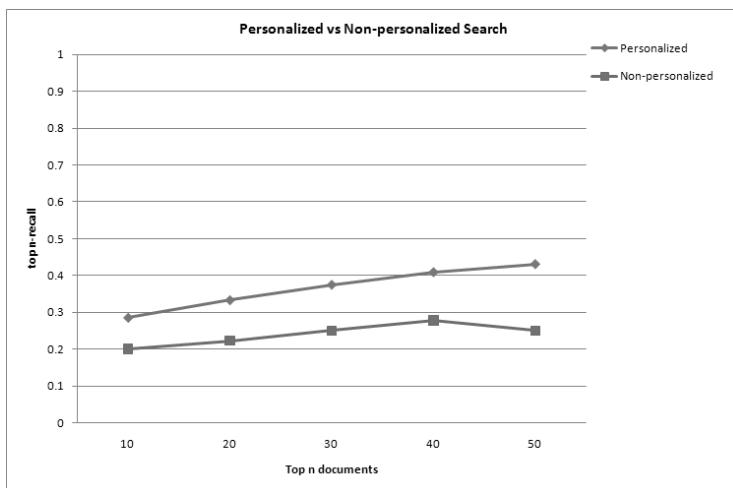
input : Results relevant to user's keyword and Profile (N)
output: Re-rank The results
1 begin;
2 Generate set of Keywords representing result  $i$ , which denotes  $R = K_1^i, K_2^i \dots K_m^i$ , its
corresponding feature vectors are  $W_i = w_1^i, w_2^i \dots w_m^i$ , where  $K_j^i$  is Keyword with
weight  $w_j^i$ ,  $j = 1..m$ ;
3  $R \leftarrow U_{i=1}^n R_i$ ;
4  $R \leftarrow \varphi$  ;
5  $b_i = 0$ ;
6 for each pair of Keywords  $< K_t^i, K_h^j >$  do
7   if  $K_t^i \leftrightarrow K_h^j$  and  $< K_t^i, K_h^j >$  exists in  $M$  then
8      $M[< K_t^i, K_h^j >] = +f(k_t^i, k_h^j)$ 
9   end
10  else
11     $M[< K_t^i, K_h^j >] = f(k_t^i, k_h^j)$ 
12  end
13 end
14 Sort  $M$  in Descending Ordered  $M[< k, k' >]$  ;
15 for each  $k_t^i$  in  $R_i$ ,  $i = 1..m$  do
16   for each pair of keyword  $< K_t^i, K_h^j >$  in  $M$  do
17      $b_i = +[< K_t^i, K_h^j >]$ 
18   end
19 end
20 Rank  $B_i$ ,  $i = 1..n$  with Descending order of  $b_i$ 
```

**Algorithm 2.** Re-ranking the searched results

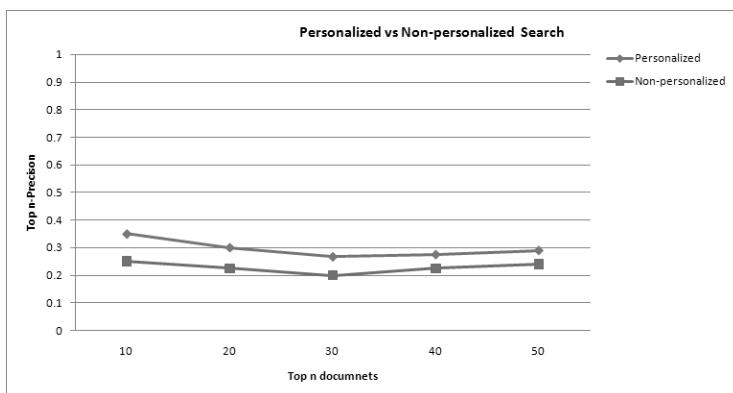
to search the information. Search documents are assigned to the query concepts based on similarity. Finally, matching of personal profile ontology and query ontology is undertaken to filter and re-rank the search results with similarity scores. We use our web crawl program to collect the top 50 results from the search engine related to a given query.

#### 4.3 Evaluation Measures

We use the Google and yahoo search engines to perform a search for a query. The top 50 results are fetched each time for a given query and create a feature vector for each result with the page title. Feature vectors are calculated using tf.idf for result page



**Fig. 3.** Recall



**Fig. 4.** Precision

and indexed under page title. The similarity map is measured with the result pages and concepts of the query ontology. The query ontology and personal profile ontology are matched to filter the search document for a given query. Finally, our ranking algorithm is applied to re-rank the searched results in descending order. Ten users participated in the experiments to test the system with different queries and test the relevancy of the re-ranked results to their opinions. We evaluated the result of the re-ranking with user relevancy in terms of recall and precision. We compared the personalized results with non-personalized search results based on re-ranking, using an interval of ten. Figure 3 and 4 show the average top n-recall and n-precision for personalized and non-personalized search results.

## 5 Conclusions

The main intention of personalization is to present search results to a particular user based on his/her interests. In this paper, we investigated an ontological approach to model the user, as well as context, to make an effective personalization. An ontological user profile is built with reference to a widely organized domain concept hierarchy of ODP. The web search history of a particular user is collected implicitly, as a user's context, and modeled with the reference ontology to build an initial user profile. The user's search query is represented as a current context of the user in a particular period of web search. The query is submitted to the search engine by the users in the form of one or more text keyword(s). The user may not use the exact words when looking for the documents; thus, by chance, may miss some relevant documents. We use WordNet hypernyms to extend the query content. We term this the query ontology. The match between the search results and the query ontology is estimated to categorize the search documents into related concepts of the query ontology. Moreover, the query ontology and personal profile ontology are matched to filter and re-rank the search results based on our new ranking algorithm that improves the effectiveness of accurately identifying the user's specific goal and intent for the search. Experimental results indicate that our approach boosts the personalized search with re-ranking to efficiently adapt the search results based on the users' context. In future, we plan to investigate more techniques and methods to learn the ontological user profile to accurately identify the users' context. Since user needs change over time, the profile stability and convergence will be considered to enrich the ontological user profile representing the user's updated interests and preferences.

## References

1. Sarabjot, S.A., Patricia, K., Mary, S.: Generating Semantically Enriched User Profiles for web Personalization. ACM Transaction on Internet Technology 7(4) Article 22 (2007)
2. Carsten, F., Markus, L.: Ontology-Based User Profile. In: Abramowicz, W. (ed.) BIS 2007. LNCS, vol. 4439, pp. 314–327. Springer, Heidelberg (2007)
3. Hyunjang, K., Myunggwon, H., Pankoo, k.: Design of Automatic Ontology Building System about the Specific Domain Knowledge. In: ICACT, pp. 20–22 (2006)
4. Abhinandan, D., Mayur, D., Ashutosh, G.: Google News Personalization: Schable Online Collaborative Filtering. In: WWW 2007, pp. 8–12 (2007)

5. Trajkova, J., Gauch, S.: Improving Ontology-Based User Profiles. In: RIAO (2004)
6. Mylonas, P., Vallet, D., Castells, P., Fernández, M., Avirthis, Y.: Personalized information retrieval based on context and ontological knowledge. The Knowledge Engineering Review 23, 73–100 (2008)
7. Guha, R., McCool, R., Miller, E.: semantic search. In: WWW 2003, pp. 20–24 (May 2003)
8. Sieg, A., Mobasher, B., Burke, R.: Learning Ontology-Based User Profiles: A Semantic Approach to Personalized Web Search. IEEE Intelligent Informatics Bulletin 8(1) (2007)
9. Gerald, S., Strembeck, M., Neumann, G.: A User Profile Derivation Approach based on Log-File Analysis, pp. 258–264. IKE, Las Vegas (June 2007)
10. Liu, F., Yu, C., Member, S., Meng, W.: Personalized Web Search For Improving Retrieval Effectiveness. IEEE Computer Society, Los Alamitos (2004)
11. Qiu, F., Cho, J.: Automatic identification of user interest for personalized search. In: WWW 2006, Edinburg, UK (May 2006)
12. Hearst, M.A.: Automated Discovery of WordNet Relations. MIT Press, Cambridge
13. Vaclav, S., Pavel, M., Jaroslav, P.: Wordnet Ontology Based Model for Web Retrieval. In: WIRI 2005. IEEE, Los Alamitos (2005)
14. Maria, G., Akrivi, K., Costas, V., George, L., Constantin, H.: Creating an Ontology for the User Profile: Method and Applications. In: Proceedings AI\*AI Workshop RCIS 2007, Italy (2002)
15. <http://wordnet.princeton.edu/>
16. <http://www.dmoz.org/>
17. Trong, H.D., Uddin, M.N., Delong, L., Jo, G.S.: A collaborative Ontology-based Profiles system. In: Nguyen, N.T., Kowalczyk, R., Chen, S.-M. (eds.) ICCCI 2009. LNCS (LNAI), vol. 5796, pp. 540–552. Springer, Heidelberg (2009)
18. Susan, G., Jason, C., Alaxander, P.: Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1(3-4), 219–334 (2003)

# Rough Sets Based Association Rules Application for Knowledge-Based System Design

Shu-Hsien Liao<sup>1</sup> and Yin-Ju Chen<sup>2</sup>

<sup>1</sup> Department of Management Sciences and Decision Making, Tamkang University,  
No. 151, Yingjuan Road, Danshuei Jen, 251 Taipei, Taiwan, ROC

<sup>2</sup> Department of Management Sciences, Tamkang University,  
No. 151, Yingjuan Road, Danshuei Jen, 251 Taipei, Taiwan, ROC  
michael@mail.tku.edu.tw, s5515124@ms18.hinet.net

**Abstract.** The Internet has emerged as the primary database, and technological platform for electronic business (EB), including the emergence of online retail concerns. Knowledge collection, verification, distribution, storage, and re-use are all essential elements in retail. They are required for decision-making or problem solving by expert consultants, as well as for the accumulation of customers and market knowledge for use by managers in their attempts to increase sales. Previous data mining algorithms usually assumed that input data was precise and clean, this assumes would be eliminated if the best rule for each particular situation. The Algorithm we used in this study however, proved to function even when the input data was vague and unclean. We provided an assessment model of brand trust as an example, to show that the algorithm was able to provide decision makers additional reliable information, in the hope of building a rough set theoretical model and base of resources that would better suit user demand.

**Keywords:** Machine Learning, Knowledge Representation, Knowledge-Based Systems, Rough sets, Association rules.

## 1 Introduction

In recent years, there has been a growing use of data mining and machine learning outside the computer science community. Such methods are being integrated into decision support systems for fields such as finance, marketing, insurance, and medicine [8]. Machine learning methods are well known for knowledge discovery. They can help to elicit knowledge (explicit and tacit) [9, 10] from data and generalize that knowledge to new, previously unseen cases [11]. Data with missing values may also be helpful if we need to guess at the correct classification of a condition using an “incomplete” table, for example by looking for rows that almost cover a condition [6]. One of the main tools of data mining is rule induction from raw data represented by a database. Real-life data are frequently imperfect: erroneous, incomplete, uncertain and vague [12]. Of the various data mining algorithms, this analysis uses the rough set algorithm due to its ability to deal with incomplete, imprecise or inconsistent information, which is typical in credit assessment analyses [4, 5]. Rough set theory is different from the fuzzy theory

and neural network. In addition, the majority of scholars mentioned that using association rules that need to face uncertainty or information inaccurate information. In the research, we further through the use of rough set theory to improve the thorny issue of the encounter for using association rules.

We can be to judge the decision of consumers relying on rules of thumb, when consumer choice factors taken into account are simple. But when the variety of choices as well as the growing number of factors to consider, how a simple analysis and consumer rule of thumb helping to determine the shopping behavior of consumers has become an important issue. At this time we may need more rigorous approach to help us to determine future consumer decision-making, and to find a complex combination of factors, and these effects of factors are tangible or intangible. The remainder of this paper is organized as follows. Section 2 reviews relevant literature correlate with the research and problem statement. Section 3 Mathematical models for new algorithm. Section 4 is presented an illustrative example. Closing remarks and future work are presented in Sect. 5.

## 2 Literature Review and Problem Statement

Machine learning can extract desired knowledge from existing training examples and ease the development bottleneck in building expert systems [9]. Knowledge-Based Systems are interactive computer programs that mimic and automate the decision-making and reasoning processes of human experts. Brand familiarity reflects the “share of mind” of a given consumer attained to the particular brand and the extent of a consumer's direct and indirect experience with a brand argue that brand familiarity is determined by strength of associations that the brand name evokes in consumer memory, and in this way it captures the consumer's brand attitude schemata [2]. The high levels of prior experience with a brand lead to the retention of stronger advertisement- brand links, making the attributes of previously familiar brands easier to recall [3]. In view of this, we hope to proceed to that the brand or product from the consumer's subjective or objective point of view preferences, according to the above-mentioned scholars, through establishing the brand image of the trust evaluation model by the ratio scale algorithms, combined with rough set theory, then find the influence degree of the consumer preferences variables between each other for the marketing decision-makers used.

## 3 Incorporation of Rough Set for Classification Processing

The knowledge in this Knowledge-Based System (KBS) consists of descriptions of domain classes and class hierarchies. We propose a classification of processed to describe these hierarchies, which could provide decision makers with addition information.

**Definition 1:** The questionnaire is  $X = \{x_1, x_2, \dots, x_n\}$ , where  $x_i \in X$ . The questionnaire answer item is  $A_{ij} = \{(a_{11}, a_{12} \dots a_{1j}), (a_{21}, a_{22} \dots a_{2j}), \dots, (a_{i1}, a_{i2} \dots a_{ij})\}$ . The decision attribute is  $D = \{d_1, d_2, \dots, d_m\}$ .

**Example 1:** Consumers of beer "brand recall" who sort and respondents related to what answer information are shown in Table 1 and 2.

**Table 1.** The original questionnaire answer

No.	Questionnaire $X$			Decision-making $D$
	Item1 $x_1$	Item2 $x_2$	Item3 $x_3$	Item4
	Gender $A_1$	Age $A_2$	Income $A_3$	Beer brand recall
1	Male $a_{11}$	30 $a_{21}$	35,000 $a_{31}$	As shown in Table 2.
2	Male $a_{11}$	40 $a_{22}$	60,000 $a_{32}$	As shown in Table 2.
3	Male $a_{11}$	45 $a_{23}$	80,000 $a_{34}$	As shown in Table 2.
4	Female $a_{12}$	30 $a_{21}$	35,000 $a_{31}$	As shown in Table 2.
5	Male $a_{11}$	40 $a_{22}$	70,000 $a_{33}$	As shown in Table 2.

**Table 2.** Beer brand recall ranking table

No.	First	Second	Third	Fourth
	Fifth	Sixth	Seventh	Eighth
1	Heineken	Miller	Taiwan light beer	Taiwan beer
—	Taiwan draft beer	Tsingtao	Kirin	Budweiser
2	Taiwan beer	Heineken	Taiwan light beer	Kirin
—	Taiwan draft beer	Tsingtao	Miller	Budweiser
3	Taiwan beer	Miller	Taiwan light beer	Heineken
—	Taiwan draft beer	Tsingtao	Kirin	Budweiser
4	Heineken	Miller	Taiwan beer	Tsingtao
—	Taiwan draft beer	Taiwan light beer	Budweiser	Kirin
5	Taiwan beer	Miller	Heineken	Tsingtao
—	Taiwan draft beer	Taiwan light beer	Budweiser	Kirin

**Definition 2:** All the respondents, whose answers were the same, are denoted as  $R_k^l = \bigcap_{i=1}^n A_{ij}(x_i)$ , where  $k = 1 \dots n$  is rule number and  $l = 1 \dots c$  is the number of accumulated customers.

**Example 2:** According to Table 1,  $R_1^1 = \{a_{11}, a_{21}, a_{31}\}$  indicated that male gender, age 30, income of 35,000.

**Definition 3:** We further discussion between the items and the items are there hidden relationships exist, that is,  $R_k^l = \bigcap_{i=1}^n A_{ij}(x_i)$  cross-comparison of the relationship between the ratio of  $x_i$

$$X_{A_{ij}} = \frac{\sum (A_{ij} - \bar{A}_{ij})}{\sqrt{\sum (A_{ij} - \bar{A}_{ij})^2}} = \frac{X_{A_i} - \bar{X}_{A_j}}{\sqrt{(X_{A_i} - \bar{X}_{A_j})(X_{A_j} - \bar{X}_{A_i})}},$$

$-1 < X_{A_{ij}} < 1$ , if  $X_{A_{ij}} < 0$  indicated a negative correlation between the two attributes;  $X_{A_{ij}} > 0$  indicated a positive correlation between the two attributes;  $X_{A_{ij}} = 0$  that is no related between the two attributes. If they are in categories (nominal) variables and numerical variables are mixed,

$$X_{A_{ij}} = \frac{\sum (\overline{A_{ij}} - \overline{A_{ij}})}{\sqrt{\sum (\overline{A_{ij}} - \overline{A_{ij}})^2}} \sqrt{P_{ij}} = \frac{\overline{X_{A_i}} - \overline{X_{A_j}}}{\sqrt{(\overline{X_{A_i}} - \overline{X_{A_j}})(\overline{X_{A_i}} - \overline{X_{A_j}})}} \sqrt{P_i P_j}, \text{ where } \overline{A_{ij}} \text{ is}$$

denoted as mean of each category and  $P_{ij}$  is denoted as percentage of each category.

**Example 3:** Taking the ratio of the relationship between gender, age, and income as an example, as shown in Table 1, we find  $X_{A_{12}} = 0$  indicates that there was no relation between sex and age;  $X_{A_{23}} > 0$  indicates a positive correlation between age and income. An increase in age accompanied an increase in income; with lower age, income was reduced. It also indicated that there was no relation between gender and income. The ratios of the relationship between gender, age and income are shown in Table 3, where  $A_{ij}^+$  denotes a positive correlation between  $i$  and  $j$ .  $A_{ij}^-$  denotes a negative correlation between  $i$  and  $j$ .  $\Delta_{ij}$  denotes no different between the two categories (nominal).

**Table 3.** The ratio of the relationship between gender, age and income

$Y_R$	Gender and age $A_1 A_2$	Age and income $A_2 A_3$	Income and gender $A_3 A_1$
$y_{12}$	$(a_{11}^+, +)$	$(+, +)$	$(+, a_{11}^+)$
$y_{23}$	$(a_{11}^+, +)$	$(+, +)$	$(+, a_{11}^+)$
$y_{34}$	$(\Delta_{1112}, -)$	$(-, -)$	$(-, \Delta_{1112})$
$y_{45}$	$(\Delta_{1112}, +)$	$(+, +)$	$(+, \Delta_{1112})$
$y_{51}$	$(a_{11}^-, -)$	$(-, -)$	$(-, a_{11}^+)$
$X_{A_{ij}}$	$X_{A_{12}} = 0$	$X_{A_{23}} > 0$	$X_{A_{31}} = 0$

**Definition 4:** We wanted to discover the hidden relationships between items, and whether they indicated another relationship with the set of ordinal attributes ( $D_c^{order}$ ). Where  $D_c^{order} = \{(d_{11}, d_{12}, \dots, d_{1j}), (d_{21}, d_{22}, \dots, d_{2j}), \dots, (d_{il}, d_{i2}, \dots, d_{ij})\}$  is a set of ordinal attributes that could be divided into several subsets.

**Example 4:** The correlation between the recall of Beer brand sequence with age and income, as shown in Table 4.

**Table 4.** The relationship between beer brand recall sequence with age and income

$Y_R$	$X_{A_{23}}$	Decision-making attributes set ( $D$ ) Beer brand recall sequence
$y_{12}$	(+,+)	$D_1^{order} = \{d_2^1, d_4^2, d_3^3, d_1^4, d_5^5, d_6^6, d_7^7, d_8^8\}$
$y_{23}$	(+,+)	$D_2^{order} = \{d_1^1, d_2^2, d_3^3, d_7^4, d_5^5, d_6^6, d_4^7, d_8^8\}$
$y_{34}$	(-, -)	$D_3^{order} = \{d_1^1, d_2^2, d_3^3, d_2^4, d_5^5, d_6^6, d_7^7, d_8^8\}$
$y_{45}$	(+,+)	$D_4^{order} = \{d_2^1, d_4^2, d_1^3, d_6^4, d_5^5, d_3^6, d_8^7, d_7^8\}$
$y_{51}$	(-, -)	$D_5^{order} = \{d_1^1, d_2^2, d_2^3, d_6^4, d_5^5, d_3^6, d_8^7, d_7^8\}$

**Definition 5:** We want to discover the hidden relationship between the decision-making attribute (Beer brand recall sequence) and the ratio of the definition3 found  $X_{A_{23}} > 0$  (Age and income have a positive ratio of inter-relationships). The hidden relationship is denoted as  $\delta_D^A = \wedge X_{A_{ij}} \wedge_{R_k^l} f(y_{ij}^D, y_{ij}^A)$  representing all of the hidden relationships, where  $f(y_{ij}^D, y_{ij}^A)$  is defined as:

$$f(y_{ij}^D) = \left\{ \begin{array}{l} y_{D_{ij}}^+ = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{D_c^{order}}{D_c^{order}} > 0 \forall y_{ij}^D \right\} \\ y_{D_{ij}}^- = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{D_c^{order}}{D_c^{order}} < 0 \forall y_{ij}^D \right\} \\ y_{D_{ij}}^0 = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{D_c^{order}}{D_c^{order}} = 1 \forall y_{ij}^D \right\} \end{array} \right\}$$

$$f(y_{ij}^A) = \left\{ \begin{array}{l} y_{A_{ij}}^+ = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{y_{ij}^A}{y_{ij}^D} > 0 \right\} \\ y_{A_{ij}}^- = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{y_{ij}^A}{y_{ij}^D} < 0 \right\} \\ y_{A_{ij}}^0 = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{y_{ij}^A}{y_{ij}^D} = 0 \right\} \\ y_{A_{ij}}^{D_i^{order}} = \left\{ y_{ij}^A \in A_{ij}, y_{ij}^D \in D_{ij} : \frac{y_{ij}^A}{y_{ij}^D} = 1 \right\} \end{array} \right\}$$

$[Y_R]_{cord}$  indicates a set defining the same hidden relationship between the decision-making attributes and the ratio found in definition3.

**Example 5:** We take the Customer No. 1 and No. 2 as an example of the relationship between beer brand recall sequence of decision-making attribute of Taiwan Beer  $d_1$ , as shown in table 5, is denoted as  $y_{d_1}^+$  and according to definition3, the relationship between the ratio of  $x_i$  is shown  $y_{A_{23}}^+$ , then is mean that decision attribute and items with the same orientation relationship ( $\delta_{d_1}^{A_{23}} = +$ ). We can be found from the above, as the "age and income" increases, the rank of Taiwan Beer, Tsingtao, and Budweiser are also rise, while with the "age and income" increases, the rank of Heineken and Kirin are also decreases, then  $[Y_R]_{cord} = \{(d_1, d_6, d_8) \cap (d_2, d_7)\}$ . In addition, Taiwan draft beer is always ranked fifth in the table.

**Table 5.** The relationship between beer brand recall

$y_{ij}^A$	$X_{A_{23}} > 0$					$f(y_{ij}^A)$	$\delta$
	$y_{12}$ (+,-)	$y_{23}$ (+,-)	$y_{34}$ (-,+)	$y_{45}$ (+,-)	$y_{51}$ (-,+)		
Taiwan beer	$y_{d_1}^+$	$y_{d_1}^0$	$y_{d_1}^-$	$y_{d_1}^+$	$y_{d_1}^+$	same $y_{A_{23}}^+$	$\delta_{d_1}^{A_{23}} = +$
Heineken	$y_{d_2}^-$	$y_{d_2}^-$	$y_{d_2}^+$	$y_{d_2}^-$	$y_{d_2}^+$	opposite $y_{A_{23}}^-$	$\delta_{d_2}^{A_{23}} = -$
Miller	$y_{d_4}^+$	$y_{d_4}^-$	$y_{d_4}^+$	$y_{d_4}^+$	$y_{d_4}^-$	inconsistent $y_{A_{23}}^0$	$\delta_{d_4}^{A_{23}} = 0$
Taiwan draft beer	$y_{d_5}^0$	$y_{d_5}^0$	$y_{d_5}^0$	$y_{d_5}^0$	$y_{d_5}^0$	ranked fifth $y_{A_{23}}^{d_5}$	$\delta_{d_5}^{A_{23}} = 5$

**Definition 6:** We use the rough set theory concept of approximation in the algorithm. According to Definition3, the lower estimate, denoted as  $Lower_{Y_R}$ , is defined as the union of all these elementary sets, which contained in  $R_k^l$ . More formally:

$$Lower_{Y_R} = \left\{ R_k^l \in A_{ij}(x_i) \mid [Y_R]_{core} \subset R_k^l \right\}$$

The upper estimate, denoted as  $Upper_{Y_R}$ , is the union of these elementary sets, which have a non-empty intersection with  $R_k^l$ .

$$Upper_{Y_R} = \left\{ R_k^l \in A_{ij}(x_i) \mid [Y_R]_{core} \cap R_k^l \neq \emptyset \right\}$$

The difference:  $Boundary_{Y_R} = Upper_{Y_R} - Lower_{Y_R}$  is called a boundary of  $R_k^l$ .

**Example 6:** According to Definition3, we know

$$\begin{aligned} Lower_{y_{di}} &= \{X_{A_2}, X_{A_3}\} = \{\text{age}, \text{income}\} \text{ and} \\ Upper_{y_{di}} &= \{(X_{A_1}, X_{A_2}), (X_{A_1}, X_{A_3}), \dots, (X_{a_{12}}, X_{a_{35}})\} \\ &= \{(\text{sex}, \text{age}), (\text{sex}, \text{income}), \dots, (\text{sex-female}, \text{income} - 80,000)\} \end{aligned}$$

**Definition 7:** Let us define an assessment model to establish brand trust  $E_i$ , as follows:  $E_c = \alpha \times \beta (a_{ij} / x_i / d_{ij})$ , where  $\alpha$  refers to the weight of assessment model, using the upper and lower bounds established by Definition6, while the attribute is included in  $Lower_{Y_R}$ , depending on attribute ranking computing with double and while the attribute is included in  $Upper_{Y_R}$ , depending on attribute ranking. And  $\beta = [n+1] - D^{order}$  is mean the rankings of the scores, where  $n$  is denoted the number of ordinal Decision-making attributes set and  $D^{order}$  is after that sort of ranking.

**Example 7:**  $E_1^{x_2x_3} = 2 \times 8 \times \text{Heineken}(d_2) = 16$ , is meaning that the assessment model to establish brand trust of Customer No. 1 in the condition in considering the age and income is 16.  $E_3^{x_2x_3} = 2 \times 5 \times \text{Heineken}(d_2) = 10$ , is mean that assessment model to establish brand trust of Customer No. 3 in the condition in considering the age and income is 10.

**Definition 8:** Finally,  $\frac{\sum E_c}{c}$  is the total trust value found in Definition7 that can provide marketing decision-maker as the basis for brand preference.

**Example 8:** We can calculate the total trust value of the Heineken brand for all customers  $\frac{\sum E_c}{c} = 13.5$ , and the higher the total value of the trust, the higher the degree of trust. The degree of brand trust with the brand recall ranking of decision-making attribute of Taiwan Beer, as Table 6.

**Table 6.** The brand trust level

Customer number $D_{ij}$	$E_c^{x_2x_3}$					Total trust value
	1	2	3	4	5	
Taiwan beer	10	16	16	12	16	14
Heineken	16	14	10	16	12	13.6
Taiwan light beer	12	12	12	6	6	9.6
Miller	14	4	14	14	14	12

## 4 Example Application

### 4.1 Rough Set Implementation

In the research, one thousand questionnaires were distributed; 900 were returned, of which 115 were disqualified as incomplete or invalid. This left 785 valid questionnaires, yielding a valid completion rate of 78.5%. First, we tried to determine the potential relationships through the regression, as shown in Table 7.

**Table 7.** Potential relationships through the regression (Gender, Age, and Income)

		Linear combination
1	Age * 0.531+Income * (-0.847)	

Second, we used the concept of redaction of rough set theory to find the core value. The redaction set is shown in Table 8. Pos. Reg. indicates the positive region of the reduction table and SC indicates the reduction of the stability coefficient.

**Table 8.** Redact set (Gender, Age, Income and Brand recall ranking)

Set	Pos. Reg.	SC	Redacts
1	0.997	1	{ Income, Beer brand recall ranking2, Beer brand recall ranking3, Beer brand recall ranking4, Beer brand recall ranking5, Beer brand recall ranking6, Beer brand recall ranking7 }
2	0.997	1	{ Age , Beer brand recall ranking1, Beer brand recall ranking2, Beer brand recall ranking3, Beer brand recall ranking4, Beer brand recall ranking5, Beer brand recall ranking6, Beer brand recall ranking7 }

Table 7 shows a negative correlation between age and income. According to Table 10, we find that { Income, Beer brand recall ranking2, Beer brand recall ranking3, Beer brand recall ranking4, Beer brand recall ranking5, Beer brand recall ranking6 and Beer brand recall ranking7 } are core attributes. The rule set is shown in Table 9.

**Table 9.** Rule set (Age, Income and Brand recall ranking)

Match			Decision Rules
1	12		{(Income= Below NT\$5,000)&(Beer brand recall ranking 2= Taiwan draft beer)&( Beer brand recall ranking 4= Budweiser) =>( Beer brand recall ranking 7= Heineken)}
2	11		{(Income= Below NT\$5,000)&(Beer brand recall ranking 3= Taiwan light beer)&( Beer brand recall ranking 4= Budweiser)=>( Beer brand recall ranking 7= Heineken) }

Third, we tried to determine the potential relationship through regression, as shown in Table 10. This indicates a positive correlation between income and Heineken, and a negative correlation between income and Miller. A positive relationship existed between age and Taiwan\_Beer, as well as between age and Taiwan\_Light\_Beer .

**Table 10.** Potential relationships through regression (Age, Income and Brand recall ranking)

Linear combination	
1	Income * 0.346+Heineken * 0.938
2	Income * 0.049+Miller * (-0.665)+Tsingtao * 0.744

According to the potential relationships found in Table 7 and Table 10, we established the degree of Brand Trust, as shown in Table 11.

**Table 11.** Total Brand trust

Brand	Brand trust	Brand	Brand trust
Taiwan beer $d_1$	12.77	Taiwan draft beer $d_5$	8.63
Heineken $d_2$	12.09	Tsingtao $d_6$	8.21
Taiwan light beer $d_3$	9.52	Kirin $d_7$	8.71
Miller $d_4$	6.51	Budweiser $d_8$	5.55

## 4.2 Association Rule Generation

Finally, we took the general attributes including Channels, Consumer Behavior, Product Features, and Medium into account, with Heineken as a decision-making variable. The rule sets are shown below.

**Table 12.** Rule set (Channels, Consumer Behavior, Product Feature, and Medium)

Match		Decision Rules
1	177	{(Channels= Convenience Stores)&( Product Feature= Price)&(Medium= Advertising)=>( Heineken)}
2	128	{(Consumer Behavior= Purchase by promotions )&(Channels= Hypermarkets)&( Product Feature= Price)&(Medium= Advertising)=>( Heineken)}
3	76	{(Consumer Behavior= Purchase by promotions )& ( Product Feature= Flavor)&(Medium= Advertising)=>( Heineken)}

## 5 Conclusion and Future Works

Traditional association rules should be adjusted to avoid retaining only trivial rules, or discarding interesting rules. In fact, situations using relative comparisons were more complete than those using absolute comparisons. In this paper, a new approach was used to determine rules of association possessing the ability to handle uncertainty in the process of classification suitable for ratio scale data. We established a brand trust evaluation model  $E_c = \alpha \times \beta (a_{ij} / x_i / d_{ij})$  in the first step of data processing, weighing them according to the upper and lower bounds set for the attributes. The system had to be readjusted to optimize the new rules, while the conditions of the traditional association rules changed. In the study, an extension of the concept of utility function used to establish the demand for users to adjust the brand image with brand trust evaluation model. The purpose and benefits for adjusting rules:

- Traditional rules of association can only generate rules, no function to amend rules. Through adjusted by the weight  $\alpha$  can increase the convenience for using association rules.
- Expert information can be used to adjust the weighting, in order to increase the credibility of the rules.

Knowledge engineers can acquire and represent knowledge in the form of decision tables, which are easily transformed to rules for use in knowledge-based systems [6]. It is our hope that in the future, we will be able to build a decision-making resources system based on rough set theory, which more closely matches the needs of users.

**Acknowledgements.** This research was funded by the National Science Council, Taiwan, Republic of China, under contract No. NSC 98-2410-H-032 -038-MY2.

## References

1. Liao, S.H., Ho, H.H., Yang, F.C.: Ontology-based data mining approach implemented on exploring product and brand spectrum. *Expert Systems with Applications* 36, 11730–11744 (2009)
2. Mikhailichenko, A., Javalgi Rajshekhar (Raj) G., Mikhailichenko, G., Laroche, M.: Cross-cultural advertising communication: Visual imagery, brand familiarity, and brand recall. *Journal of Business Research* 62, 931–938 (2009)
3. Kent, R.J., Kellaris, J.J.: Competitive interference effects in memory for advertising: are familiar brands exempt? *J. Mark Commun.* 7, 59–69 (2001)
4. Liu, G., Zhu, Y.: Credit Assessment of Contractors: A Rough Set Method. *Tsinghua Science & Technology* 11, 357–363 (2006)
5. Jiayuan, G., Chankong, V.: Rough set-based approach to rule generation and rule induction. *International Journal of General Systems* 31, 601–617 (2002)
6. Hewett, R., Leuchner, J.: Restructuring decision tables for elucidation of knowledge. *Data and Knowledge Engineering* 46, 271–290 (2003)
7. Hong, T.-P., Tseng, L.-H., Chien, B.-C.: Mining from incomplete quantitative data by fuzzy rough sets. *Expert Systems with Applications* 37, 2644–2653 (2010)
8. Štrumbelj, E., Kononenko, I., Robnik Šikonja, M.: Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* 68, 886–904 (2009)
9. Brohman, M.K.: Knowledge creation opportunities in the data mining process. In: *HICSS 2006: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, p. 170. IEE Computer Society, Washington (2006)
10. Nemati, H.R., Steiger, D.M., Iyer, L.S., Herschel, R.T.: Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. *Decision Support Systems* 33(2), 143–161 (2002)
11. Chi, C.-L., Street, W.N., Ward, M.M.: Building a hospital referral expert system with a Prediction and Optimization-Based Decision Support System algorithm. *Journal of Biomedical Informatics* 41, 371–386 (2008)
12. Fortes, I., Mora-López, L., Morales, R., Triguero, F.: Inductive learning models with missing values. *Mathematical and Computer Modelling* 44, 790–806 (2006)

# Author Index

- Abdul Raheem, Abdul Azeez I-499  
Adigun, Matthew O. III-122  
Ahmad, Mohd Sharifuddin I-296  
Ahmed, Moamin I-296  
Al-Saffar, Aymen I-178  
Alsaffar, Aymen Abdullah III-282, III-292  
Arcelli Fontana, Francesca II-352  
  
Bai, Huihui III-47  
Barbucha, Dariusz I-403  
Boryczka, Urszula I-363, I-373  
Borzemski, Leszek I-20  
Bossé, Tibor I-306  
  
Ceglarek, Dariusz I-162  
Cha, Jeong-Won II-22  
Chang, Bao Rong II-172, II-334  
Chang, Chia-Wei II-278  
Chang, Chuan-Yu III-1  
Chang, Chun-Chih I-491  
Chang, Chung C. II-85  
Chang, Hsuan-Ting I-64, I-74, I-81  
Chang, Jui-Fang I-136  
Chang, Shu-Han III-1  
Chang, Yao-Lang I-433, I-468  
Chao, Shu-Jun III-354  
Chen, Ching-I III-92  
Chen, Chiung-Hsing II-411  
Chen, Chun-Hao II-224  
Chen, Heng-Sheng II-324  
Chen, Liang-Ho III-317  
Chen, Lih-Shyang I-117  
Chen, Rung-Ching II-249  
Chen, Shao-Hsien I-491  
Chen, Shao-Jer III-1  
Chen, Shyi-Ming II-441  
Chen, Ya-Ning I-205  
Chen, Yen-Sheng I-491  
Chen, Yi-Huei III-333  
Chen, Yi-Ting I-109  
Chen, Yin-Ju II-501  
Chen, Ying-Hao II-249  
Cheng, Chih-Hsiang III-92  
Cheng, Jian II-461  
  
Cheng, Wen-Lin I-117  
Cheng, Yu-Huei III-448  
Cheng, Yuh-Ming II-381  
Chien, Jong-Chih III-200  
Chien, Li-Hsiang I-520  
Chien, Wei-Hsien III-367  
Chiu, Tzu-Fu I-152  
Chiu, Yu-Ting I-152  
Choi, Sang-Min II-22  
Choroś, Kazimierz I-11  
Chou, Jyh-Woei III-317  
Chouhuang, Wayne I-230  
Chu, Huai-Ping II-441  
Chu, Shu-Chuan I-109, III-71, III-174  
Chuang, Li-Yeh III-448  
Chuang, Shang-Jen II-411  
Chuang, Tzu-Hung II-373  
Chung, Hung-Chien I-482  
Chung, Kyung-Yong I-54  
Chynał, Piotr I-30  
Czarnowski, Ireneusz I-353  
  
Deb, Kaushik III-184  
Deng, Guang-Feng III-406  
Ding, Ing-Jr II-288  
Diwold, Konrad III-426  
Doskocz, Piotr II-11  
Drwal, Maciej I-20  
Dung, Nguyen Tien III-282  
Duong, Trong Hai II-490  
  
Encheva, Sylvia III-133  
  
Fan, Chia-Yu III-398  
Feng, Chong II-113  
Formato, Ferrante II-352  
  
Goczyła, Krzysztof III-102  
Górczyńska-Kosiorz, Sylwia I-320  
  
Haarslev, Volker III-457  
Haider, Mohammad II-153  
Han, Yo-Sub II-22  
Haniewicz, Konstanty I-162  
He, Jun-Hua I-172  
Hendriks, Monique I-330

- Hera, Lukasz I-320  
 Ho, Chengter I-413  
 Hong, Chao-Fu I-152  
 Hong, Tzung-Pei II-224, II-344  
 Hoogendoorn, Mark I-306  
 Horng, Ming-Huwi III-438  
 Horng, Mong-Fong I-109, III-63  
 Horng, Wen-Bing II-95  
 Hsiao, Huey-Der I-265  
 Hsiao, Kou-Chan II-85  
 Hsieh, Fu-Shiung II-470  
 Hsu, Chia-ling II-363  
 Hsu, Chien-Chang II-268, III-398  
 Hsu, Chiou-Ping III-342  
 Hsu, Chun-Liang III-142  
 Hsu, Jia-Lien III-367  
 Hsu, Li-Fu I-188  
 Hsu, Tsung-Shin III-327  
 Hsu, Wei-Chih II-373  
 Hu, Wu-Chih III-11, III-92  
 Huang, Chien-Feng II-172, II-334  
 Huang, Chien-Hsien II-61  
 Huang, Ching-Lien III-327  
 Huang, Deng-Yuan III-11, III-92  
 Huang, Heyan II-113  
 Huang, His-Chung II-172  
 Huang, Hong-Chu I-205  
 Huang, Hui-Hsin III-311  
 Huang, Jui-Chen II-402  
 Huang, Shih-Hao III-210  
 Huang, Shu-Chien II-183  
 Huang, Su-Yi III-333  
 Huang, Tien-Tsai III-317, III-333  
 Huang, Ying-Fung I-444  
 Huh, Eui-Nam I-178, I-195,  
     III-282, III-292  
 Hung, Kuo-Chen I-243  
 Hung, Mao-Hsiung III-174  
 Hung, Yi-Tsan II-203  
 Huy, Phan Trung III-252  
 Hwang, Chein-Shung II-104  
 Hwang, Hone-Ene I-74  
 Hwang, Wen-Jyi II-203  
 Islam, Md. Motaharul I-178,  
     III-282, III-292  
 Jain, Lakhmi C. III-71  
 Jan, Yee-Jee II-239  
 Jędrzejowicz, Joanna I-343  
 Jędrzejowicz, Piotr I-343, I-353, I-383,  
     I-393  
 Jembere, Edgar III-122  
 Jeng, Albert B. II-433  
 Jhan, Ci-Fong III-21  
 Jhu, Jia-Jie III-11  
 Jian, Jeng-Chih II-249  
 Jiang, Ting-Wei III-438  
 Jo, Geun Sik II-490  
 Jo, Kang-Hyun III-184  
 Juang, Jih-Gau I-520  
 Jung, Jason J. III-154  
 Juszczuk, Przemyslaw I-363  
 Kajdanowicz, Tomasz II-11  
 Katarzyniak, Radosław III-112  
 Kawamura, Takahiro II-163  
 Kazienko, Przemysław II-11  
 Kim, Jung-Won III-184  
 Kim, Youngsoo II-193  
 Klein, Michel C.A. I-306  
 Korff, R. I-90  
 Kornatowski, Eugeniusz II-298  
 Kozak, Jan I-373  
 Kozielski, Stanisław I-320  
 Krawczyk, M.J. I-90  
 Kułakowski, K. I-90  
 Kumar, T.V. Vijay II-153  
 Kuntanapreeda, Suwat III-242  
 Kuo, Bo-Lin II-316  
 Kuo, Jong-Yih III-376  
 Lai, Chih-Chin III-21  
 Lai, Shu-Chin I-468  
 Lay, Young-Jinn I-117  
 Lee, An-Chen III-29  
 Lee, Chien-Pang II-68  
 Lee, Chung-Nan II-344  
 Lee, Dong-Liang III-142  
 Lee, Huey-Ming II-51, II-61, II-324  
 Lee, Jung-Hyun I-54  
 Lee, Mn-Ta I-64  
 Lee, Tsang-Yean II-324  
 Lee, Yeong-Chyi II-224  
 Leu, Yungho II-68  
 Li, Che-Hao II-232  
 Li, Cheng-Hsiu II-373  
 Li, Cheng-Yi II-239  
 Li, Jen-Hsing III-81  
 Li, Leida II-307, II-461

- Liang, Bin III-236  
 Liao, Bin-Yih I-109, II-316, III-63  
 Liao, Shu-Hsien I-205, II-501  
 Lin, Cheng-Pin III-11  
 Lin, Chih-Hung II-278  
 Lin, Hsin-Hung III-387  
 Lin, Huan-wei III-302  
 Lin, Jennifer Shu-Jen I-252  
 Lin, Kuo-Ping I-243  
 Lin, Lian-Yong I-117  
 Lin, Lily II-51  
 Lin, Ruei-Tang III-81  
 Lin, Shiow-Jyu II-203  
 Lin, Shu-Chuan II-239  
 Lin, Wen-Ching I-444  
 Lin, Woo-Tsong III-406  
 Lin, Yi-Sin III-63  
 Lin, Yu-Jen I-117  
 Lin, Yuh-chang II-363  
 Liu, Chao-Yi III-163  
 Liu, Chen-Yi II-1  
 Liu, Chi-Hua II-316  
 Liu, Fang-Tsung II-411  
 Liu, Feng II-213  
 Liu, Feng-Jung III-210, III-227  
 Liu, Hsiang-Chuan I-509  
 Liu, Jing-Sin I-482  
 Liu, Li-Chang III-200  
 Liu, Yi-Hua III-63  
 Lo, Chih-Cheng II-316  
 Lu, Hoang-Yang III-200  
 Lu, Jonathan Chun-Hsien III-387  
 Lu, Li-Hsiang II-381  
 Lu, Song-Yun II-258  
 Lu, Wan-Chin III-210  
 Lu, Zhaolin II-307  
 Lu, Zhe-Ming III-56  
 Lv, Jing-Yuan I-172  
 Ma, Wei-Ming III-218  
 Malarz, K. I-90  
 Małysiak-Mrozek, Bożena I-320  
 Mei, Hsing II-258  
 Memic, Haris II-31, II-41  
 Meng, Lili III-47  
 Middendorf, Martin III-426  
 Mikolajczak, Grzegorz III-194  
 Minami, Toshiro I-274  
 Mohammed, Nazim uddin II-490  
 Momot, Alina I-320  
 Momot, Michał I-320  
 Mrozek, Dariusz I-320  
 Na, Sang-Ho III-282  
 Nam, Vu Thanh III-252  
 Nattee, Cholwich II-132  
 Nguyen, Hoang-Nam III-29  
 Nguyen, Ngoc Thanh II-480  
 Nguyen, The-Minh II-163  
 Nguyen, Tien-Dung I-178, III-292  
 Ni, Rongrong I-128  
 Nielek, Radoslaw II-122  
 Ohsuga, Akihiko II-163  
 Olatunji, Sunday Olusanya I-499  
 Ou, C.R. III-416  
 Ou, Chung-Ming III-416  
 Ou, Ting-Chia II-411  
 Pan, Jeng-Shyang I-109, I-128, II-316,  
     III-47, III-56, III-71, III-174  
 Pan, Zhenghua II-391  
 Pao, Cho-Tsan II-249  
 Pareschi, Remo II-352  
 Park, Junyoung I-195  
 Park, Jun-Young I-178  
 Park, Namje II-142, II-193  
 Pęksiński, Jakub III-194  
 Piotrowski, Piotr III-102  
 Qian, Jiansheng II-307  
 Ratajczak-Ropel, Ewa I-393  
 Rim, Kee-Wook I-54  
 Roddick, John F. III-174  
 Ruhnke, Thomas III-426  
 Rutkowski, Wojciech I-162  
 Selamat, Ali I-499  
 Shaban-Nejad, Arash III-457  
 Sharpanskykh, Alexei I-39, I-284  
 Shi, Guodong II-213  
 Shieh, Chin-Shiu I-174  
 Shieh, Wen-Gong II-95  
 Shih, Ming-Haur I-457  
 Shih, Teng-San II-51  
 Shih, Tsung-Ting I-433  
 Shin, Young-Rok III-292  
 Skakovski, Aleksander I-383  
 Skorupa, Grzegorz III-112  
 Sobecki, Janusz I-30  
 Song, Biao I-195

- Song, Chang-Woo I-54  
 Song, Youjin II-142  
 Strzalka, Krzysztof I-145  
 Su, Jin-Shieh II-51  
 Su, Yi-Ching II-104
- Tahara, Yasuyuki II-163  
 Tai, Shao-Kuo II-239  
 Takayshvili, Liudmila II-451  
 Tang, Jing-Jou I-117  
 Tang, Lin-Lin III-56  
 TeCho, Jakkrit II-132  
 Teng, Hsi-Che I-457  
 Teng, S.J. Jerome III-37  
 Theeramunkong, Thanaruk II-132  
 Tian, Huawei I-128  
 Tian, Yuan I-195  
 Treur, Jan I-39, I-284, I-306, I-330  
 Truong, Hai Bang II-480  
 Tsai, August III-342  
 Tsai, Hsiu Fen II-172  
 Tsai, Hui-Yi II-68  
 Tsai, Y.-C. I-243  
 Tseng, Chien-Chen II-433  
 Tseng, Chun-Wei III-210, III-227  
 Tseng, Der-Feng II-433  
 Tseng, Kuo-Cheng II-104  
 Tseng, Wen-Chang III-227  
 Tseng, Wen-Hsien II-258  
 Tuan, Chiu-Ching III-354  
 Tumin, Sharil III-133  
 Tung, Kei-Chen III-387
- Vavilin, Andrey III-184  
 Vinh, Ho Ngoc III-252
- Waloszek, Aleksander III-102  
 Waloszek, Wojciech III-102  
 Wang, Ai-ling II-363  
 Wang, Anhong III-47  
 Wang, Cen II-391  
 Wang, Chia-Nan I-421, I-444, I-468  
 Wang, Chieh-Hsuan II-232  
 Wang, Chih-Hong I-433  
 Wang, Jiunn-Chin II-433  
 Wang, Mu-Liang I-64, III-63  
 Wang, Ping-Tsung III-81  
 Wang, Shin-Jung I-457
- Wang, Shuozhong I-100  
 Wang, Wei-Shun III-21  
 Wang, Wei-Ting III-376  
 Wang, Wei-Yi III-398  
 Wang, Y. III-163  
 Wang, Yen-Hui I-421  
 Wang, Yen-Wen III-342  
 Wang, Ying-Wei III-92  
 Wei, Ching-Chuan II-232  
 Wei, Kexin III-236  
 Wei-Ming, Yeh II-425  
 Weng, Shaowei III-71  
 Wierzbicki, Adam II-122  
 Wong, Ray-Hwa III-163  
 Wou, Yu-Wen I-260  
 Wu, Bang Ye II-1  
 Wu, Jianchun II-213  
 Wu, Ming-Fang III-81  
 Wu, Min-Thai II-344  
 Wu, Qiumin I-100  
 Wu, Quincy III-302  
 Wu, Yi-Sheng III-200  
 Wu, Zong-Yu I-81
- Xulu, Sibusiso S. III-122
- Yang, Cheng-Hong III-448  
 Yang, Chih-Te III-342  
 Yang, Ching-Yu II-278, III-11  
 Yang, Chin-Ping II-1  
 Yang, Gino K. I-215, I-230, I-243  
 Yang, Ming-Fen I-265  
 Yang, Sheng-Yuan III-142  
 Yu, Chih-Min III-263, III-272  
 Yu, Jie II-268  
 Yu, Ping II-75  
 Yue, Youjun III-236  
 Yusoff, Mohd Zaliman M. I-296
- Zatwarnicki, Krzysztof I-1  
 Zawadzka, Teresa III-102  
 Zgrzywa, Aleksander I-145  
 Zhang, Jia-Ming III-354  
 Zhang, Jianying II-461  
 Zhang, Lijuan II-391  
 Zhang, Xiaofei II-113  
 Zhang, Xinpeng I-100  
 Zhao, Yao I-128, III-47