

Ford Lumban Gaol  
*Editor*

# Recent Progress in Data Engineering and Internet Technology

Volume 1



Ford Lumban Gaol (Ed.)

# Recent Progress in Data Engineering and Internet Technology

Volume 1

*Editor*

Ford Lumban Gaol  
Bina Nusantara University  
Perumahan Menteng  
Jakarta  
Indonesia

ISSN 1876-1100

e-ISSN 1876-1119

ISBN 978-3-642-28806-7

e-ISBN 978-3-642-28807-4

DOI 10.1007/978-3-642-28807-4

Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012933326

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Foreword

In the recent years we are being faced with flooding of data whose handling has promoted rapid advancements in internet technology. This book is a collection of selected papers, all of which were presented at the International Conference on Data Engineering and Internet Technology (DEIT 2011).

This conference, which took place in Bali Dynasty Resort, Bali, Indonesia on 15th–17th March 2011, brought together researchers and scientists from academia, industry, and government laboratories to share their results and identify future research directions in data engineering and internet technology. Topics of interest include, among others: computational algorithms and tools, database management and database technologies, intelligent information systems, data engineering applications, internet security, internet data management, web search, data grids, cloud computing, as well as web-based application.

The book makes fascinating reading and will be not only of great interest to researchers involved in all aspects of Data engineering and Internet Technology, but also it will surely attract more practitioners and academics towards these research areas.

Dr. Ford Lumban Gaol

# Contents

<b>Design of Fault Detection Filter for T-S Fuzzy Time-Delay Systems . . . . .</b>	<b>1</b>
<i>Wu Di, Hou Jiuyang, Li Qiang</i>	
<b>Autonomous and Pervasive Computing-Based Knowledge Service . . . . .</b>	<b>9</b>
<i>Keedong Yoo</i>	
<b>Short-Term Traffic Flow Forecasting Based on Periodicity Similarity Characteristics . . . . .</b>	<b>15</b>
<i>Chunjiao Dong, Chunfu Shao, Dan Zhao, Yinhong Liu</i>	
<b>An Eigenvector-Based Kernel Clustering Approach to Detecting Communities in Complex Networks . . . . .</b>	<b>23</b>
<i>Lidong Fu, Lin Gao</i>	
<b>Distributed Gaussian Mixture Learning Based on Variational Approximations . . . . .</b>	<b>29</b>
<i>Behrouz Safarinejadian</i>	
<b>The Function and Relationship of Verbal and Nonverbal Cues in IM Interpersonal Communication . . . . .</b>	<b>35</b>
<i>Lu Xiaoyan, Yao Jinyun</i>	
<b>Using Version Control System to Construct Ownership Architecture Documentations . . . . .</b>	<b>41</b>
<i>Po-Han Huang, Dowming Yeh, Wen-Tin Lee</i>	
<b>DF-RealL2Boost: A Hybrid Decision Forest with Real L2Boosted Decision Stumps . . . . .</b>	<b>47</b>
<i>Zaman Md. Faisal, Sumi S. Monira, Hideo Hirose</i>	
<b>Magazine Image Retrieval with Camera-Phone . . . . .</b>	<b>55</b>
<i>Cheng Yang, Jie Yang, Deying Feng</i>	

<b>Statistical Clustering and Times Series Analysis for Bridge Monitoring Data</b> .....	61
<i>Man Nguyen, Tan Tran, Doan Phan</i>	
<b>Towards Smart Advisor's Framework Based on Multi Agent Systems and Data Mining Methods</b> .....	73
<i>Madjid Khalilian</i>	
<b>Automatic LSA-Based Retrieval of Synonyms (for Search Space Extension)</b> .....	79
<i>Kamil Ekštejn, Lubomír Krčmář</i>	
<b>Research on Method of Wavelet Function Selection to Vibration Signal Filtering</b> .....	87
<i>Xiangzhong Meng, Jianghong Wang</i>	
<b>Time Series Subsequence Matching Based on Middle Points and Clipping</b> .....	93
<i>Nguyen Thanh Son, Duong Tuan Anh</i>	
<b>Effects of Spatial Scale in Cellular Automata Model for Land Use Change</b> .....	101
<i>Guanwei Zhao</i>	
<b>Prediction Model Based on PCA - DRKM-RBF</b> .....	107
<i>Wang Zhe, Sun Wen-wen, Zhou Tong, Zhou Chun-guang</i>	
<b>A Topic Detection and Tracking System with TF-Density</b> .....	115
<i>Shu-Wei Liu, Hsien-Tsung Chang</i>	
<b>Community Identification of Financial Market Based on Affinity Propagation</b> .....	121
<i>Lei Hong, Shi-Min Cai, Zhong-Qian Fu, Pei-Ling Zhou</i>	
<b>Research on Optimization of Target Oil Wells for CO<sub>2</sub> Huff and Puff in Yushulin Oil Well</b> .....	129
<i>Erlong Yang, Huaidong He, Lei Wang</i>	
<b>A Secret Embedding Scheme by Means of Re-indexing VQ Codebook upon Image Processing</b> .....	135
<i>Cheng-Ta Huang, Wei-Jen Wang, Shiuh-Jeng Wang, Jonathan C.L. Liu</i>	
<b>Knowledge Management System: Combination of Experts' Knowledge and Automatic Improvement</b> .....	143
<i>Hiroshi Sugimura, Kazunori Matsumoto</i>	
<b>Evaluating Recommender System Using Multiagent-Based Simulator: Case Study of Collaborative Filtering Simulation</b> .....	155
<i>Ryosuke Saga, Kouki Okamoto, Hiroshi Tsuji, Kazunori Matsumoto</i>	

<b>On the Constraint Satisfaction Method for University Personal Course Scheduling</b> .....	163
<i>Yukio Hori, Takashi Nakayama, Yoshiro Imai</i>	
<b>Factors Affecting Productivity of Fractured Horizontal Wells</b> .....	175
<i>Li Tiejun, Guo Dali, Tang Zhihao, Ke Xijun</i>	
<b>Informal Lightweight Knowledge Extraction from Documents</b> .....	181
<i>Francesco Colace, Massimo De Santo, Paolo Napoletano</i>	
<b>A Classification Model Using Emerging Patterns Incorporating Item Taxonomy</b> .....	187
<i>Hiroyuki Morita, Yukinobu Hamuro</i>	
<b>Synchronised Data Logging, Processing and Visualisation Framework for Heterogeneous Sensor Networks</b> .....	193
<i>Matthias Vodel, Rene Bergelt, Matthias Glockner, Wolfram Hardt</i>	
<b>Design of SENIOR: A Case Study Using NoGap</b> .....	199
<i>Per Karlström, Wenbiao Zhou, Dake Liu</i>	
<b>GWDL: A Graphical Workflow Definition Language for Business Workflows</b> .....	205
<i>M. Sultan Mahmud, Saad Abdullah, Shazzad Hosain</i>	
<b>A Feature Representation and Extraction Method for Malicious Code Detection Based on LZW Compression Algorithm</b> .....	211
<i>Yingxu Lai, Hongnan Liu, Zhen Yang</i>	
<b>CASM: Coherent Automated Schema Matcher</b> .....	219
<i>Rudraneel Chakraborty, Faiyaz Ahmed, Shazzad Hosain</i>	
<b>Research on Constructing 3D Geological Model of the Construction Layers in Daxing New City Area of Beijing City</b> .....	225
<i>Xiao Huang, Yingshuang Wang, Naiqi Shen, Yufeng Liu, Gang Chen</i>	
<b>Preliminary Study of HGML-Based Virtual Scene Construction and Interaction Display</b> .....	231
<i>Yang Wenhui, Xie Yan, Miao Fang</i>	
<b>The Next-Generation Search Engine: Challenges and Key Technologies</b> .....	239
<i>Bo Lang, Xianglong Liu, Wei Li</i>	
<b>An Efficient k-Anonymization Algorithm with Low Information Loss</b> . . .	249
<i>Md. Nurul Huda, Shigeki Yamada, Noboru Sonehara</i>	



<b>Design and Realization of Data Gateway for Large-Scale Alternately Supervisory System</b> . . . . .	255
<i>Yong Bai, Zhangli Lan, Yi Zhou</i>	
<b>Research on Mass Geospatial Raster Data Processing Based on Map/Reduce Model</b> . . . . .	261
<i>Fang Yin, Min Feng, Jia Song</i>	
<b>A Parallel-Enabled Frequency Offset Estimation Scheme for Optical Coherent Receivers</b> . . . . .	269
<i>Zhi-Yu Li, Xiao-Bo Meng, Xian Zhou, Xue Chen, Yang-Yang Fan, Hai Zhu</i>	
<b>Process-Driven Integrated Product Design Platform</b> . . . . .	275
<i>Bo Zhao, Jun Luo, Rongmei Nie, Xiaojun Hu, Hui Zhao</i>	
<b>Segmentation of Audio Visual Malay Digit Utterances Using Endpoint Detection</b> . . . . .	281
<i>Mohd Ridzuwary Mohd Zainal, Aini Hussain, Salina Abdul Samad</i>	
<b>Data Mining in Clinical Decision Support Systems</b> . . . . .	287
<i>Liljana Aleksovska-Stojkovska, Suzana Loskovska</i>	
<b>Spatial Influence Index Effects in the Analysis of Epidemic Data</b> . . . . .	295
<i>Fatima F. Santos, Nelson F.F. Ebecken</i>	
<b>EPON's Research on Transmission of Sampling Value in Process Level of Digital Substation</b> . . . . .	307
<i>Qiu Ming-yi, Xu Xi-dong</i>	
<b>Intelligent Clinical Information Systems for the Cardiovascular Diseases</b> . . . . .	315
<i>Nan-Chen Hsieh, Huan-Chao Keh, Chung-Yi Chang, Chien-Hui Chan</i>	
<b>Forecast the Foreign Exchange Rate between Rupiah and US Dollar by Applying Grey Method</b> . . . . .	321
<i>Tien-Chin Wang, Su-Hui Kuo, Truong Ngoc Anh, Li Li</i>	
<b>Simulating Patent Knowledge Contexts</b> . . . . .	329
<i>Jussi Kantola, Aviv Segev</i>	
<b>Apply Data Mining on Location-Based Healthcare System</b> . . . . .	337
<i>Chen-Yang Cheng, Ja-Hao Chen, Tsuo-Hung Lan, Chin-Hong Chan, Yu-Hsun Huang</i>	
<b>Using Fuzzy Logic in Virus-Mediated Gene Therapy</b> . . . . .	343
<i>Arie H. Tan, Fabian H. Tan</i>	

<b>Towards Using Cached Data Mining for Large Scale Recommender Systems</b> .....	349
<i>Swapneel Sheth, Gail Kaiser</i>	
<b>BAX-SET PLUS: A Taxonomic Navigation Model to Categorize, Search and Retrieve Semantic Web Services</b> .....	359
<i>Jaime Alberto Guzmán Luna, Ingrid Durley Torres Pardo, Jovani Alberto Jiménez Builes</i>	
<b>Text-Transformed Image Classification Based on Data Compression</b> . . . .	365
<i>Nuo Zhang, Toshinori Watanabe</i>	
<b>Numerical Analysis of Touch Mode Capacitive Pressure Sensor Using Graphical User Interface</b> .....	371
<i>Moon Kyu Lee, Jeongho Eom, Bumkyoo Choi</i>	
<b>Multithreading Embedded Multimedia Application for Vehicle Blackbox</b> .....	379
<i>Jajin Koo, Jeongheon Choi, Youjip Won, Seongjin Lee</i>	
<b>An EMD-Oriented Hiding Scheme for Reversible Data Embedding in Images</b> .....	385
<i>Chi-Yao Weng, Hung-Min Sun, Cheng-Hsing Yang, Shiu-H-Jeng Wang</i>	
<b>Gait Control for Guide Dog Robot to Walk and Climb a Step</b> .....	393
<i>Manabu Kosaka</i>	
<b>Artificial Neural Network-Based Lot Number Recognition for Cadastral Map</b> .....	401
<i>Dave E. Marcial, Ed Darcy Dy, Silvin Federic Maceren, Erivn Rygl Sarno</i>	
<b>The Novel Virtual Reality Fixture Design and Assembly System (VFDAS)</b> .....	407
<i>Qian Cao, Qiang Li</i>	
<b>Multi-phenomenon Oriented Embedded Software Fault Diagnosis</b> . . . .	413
<i>Shunkun Yang, Ge Lin, Lu Minyan</i>	
<b>Asymptotic Behavior of Random Walks with Resting State in Ergodic Environments</b> .....	421
<i>Liu Xiang-dong, Zhu Li-ye</i>	
<b>Research on the Spatio-temporal Pass Index Model in Emergency Evacuation</b> .....	429
<i>Liu Yi, Zhu Haiguo, Huang Quanyi, Zhao Yunxiu, Yuan Shengcheng, Zhang Hui</i>	

<b>On Issues of Multi-path Routing in Overlay-Networks Using Optimization Algorithms</b> .....	435
<i>Sameer Qazi, Tim Moors</i>	
<b>Reviews on Intelligent Building Systems in China</b> .....	441
<i>Feng Jiang, Min Gao, Zhijun Wang</i>	
<b>XBeGene: Scalable XML Documents Generator by Example Based on Real Data</b> .....	449
<i>Manami Harazaki, Joe Tekli, Shohei Yokoyama, Naoki Fukuta, Richard Chbeir, Hiroshi Ishikawa</i>	
<b>A Preliminary Activity Recognition of WSN Data on Ubiquitous Health Care for Physical Therapy</b> .....	461
<i>S.-Y. Chiang, Y.-C. Kan, Y.-C. Tu, H.-C. Lin</i>	
<b>Development and Evaluation of Question Templates for Text Mining</b> ...	469
<i>Norio Ishii, Yuri Suzuki, Takashi Fujii, Hironobu Fujiyoshi</i>	
<b>Segmentation of Fibro-Glandular Discs in Digital Mammograms Using Log-Normal Distribution</b> .....	475
<i>Y.A. Reyad, A. El-Zaart, H. Mathkour, M. Al-Zuair, H. Al-Salman</i>	
<b>An OOAD Model of Program Understanding System's Parser</b> .....	481
<i>Norazimah Rosidi, Nor Fazlida Mohd Sani, Abdul Azim Abd Ghani</i>	
<b>The Unified Modeling Language Model of an Object Oriented Debugger System</b> .....	489
<i>Noor Afiza Mohd Ariffin, Nor Fazlida Mohd Sani, Rodziah Atan</i>	
<b>A FPGA-Based Real Time QRS Complex Detection System Using Adaptive Lifting Scheme</b> .....	497
<i>Hang Yu, Lixiao Ma, Ru Wang, Lai Jiang, Yan Li, Zhen Ji, Yan Pingkun, Wang Fei</i>	
<b>A Novel Temporal-Spatial Color Descriptor Representation for Video Analysis</b> .....	505
<i>Li Xiang-wei, Zheng Gang, Zhao Kai</i>	
<b>Author Index</b> .....	511

# Design of Fault Detection Filter for T-S Fuzzy Time-Delay Systems

Wu Di, Hou Jiuyang, and Li Qiang

**Abstract.** A Fault Detection scheme for T-S fuzzy systems with unknown inputs is discussed. Based on  $H_\infty$  control theory, the proposed filter design provides sufficient conditions for the existence of a solution to the detection of faults. By means of the Projection Lemma, a quasi-convex formulation of the problem is obtained via LMI. The filter can guarantee the prescribed performance index, which minimizes the error between the residual and the real residual signal and has robust performance to unknown inputs. The effectiveness of the design technique is illustrated via a numerical example.

## 1 Introduction

Due to an increasing demand for higher performance and higher safety and reliability standards, the model-based approaches to fault detection and isolation (FDI) for dynamic systems have received more and more attention during last two decades. The fundamental purpose of an FDI scheme is to generate an alarm when a fault occurs (detection) and also to identify the nature and location of the fault (isolation)[1]. The most commonly used FDI methods are observer based where the measured plant output is compared to the output of an observer designed from a model of the system, and the discrepancy is used to form a residual.[1] Using this residual signal, a decision is made as to whether a fault condition is present and also an attempt is made to determine its location.

T-S fuzzy systems are nonlinear systems described by a set of IF-THEN rules which gives a local linear representation of an underlying system. T-S fuzzy control has become one of the most popular and promising research platform in the model-based fuzzy control and the theoretic researches on the issue have been

---

Wu Di · Hou Jiuyang

Department of Computer and Information Engineering, Heilongjiang Institute of science and technology, Harbin of China

e-mail: wudi55dd@sina.com, houjy2000@163.com

conducted actively by many fuzzy control theorists. Such models can approximate a wide class of nonlinear systems. Feng et al. [2] and Cao et al. [3] have proved that the T–S fuzzy system can approximate any continuous functions in at any preciseness and applied the method based on linear uncertain system theory to convert the stability.

Analysis of a fuzzy control system to the stability analysis of linear time-varying “extreme” subsystems. In this paper a model-based approach is developed for solving FD problems in T–S fuzzy systems with time-delay.

## 2 Problem Formulation

The continuous fuzzy dynamic model, proposed by Takagi and Sugeno, is described by fuzzy IF-THEN rules, which represented local linear input–output relations of nonlinear system.

Then the T-S fuzzy systems are defined as follows:

$$\begin{aligned} \dot{x} &= \sum_{i=1}^r \lambda_i(z) ((A_i + \Delta A_i)x + (A_{di} + \Delta A_{di})x_d \\ &\quad + (B_{ui} + \Delta B_{ui})u + B_{di}d + B_{fi}f) \\ y &= \sum_{i=1}^r \lambda_i(z) (C_i x + D_{ui}u + D_{di}d + D_{fi}f), \end{aligned} \quad (1)$$

where  $\lambda_i(z) = \beta_i(z) / \sum_{j=1}^r \beta_j(z)$ ,  $\beta_j(z) = \prod_{k=1}^p M_{kj}(z)$ ,  $t$  is omitted,  $x_d = x(t-d)$  and  $M_{kj}(z)$  is the membership function of fuzzy set  $M_{kj}$ . In the following we always assume that

$$\sum_{i=1}^r \lambda_i(z) = 1, \quad \lambda_k(z) \geq 0, \quad k = 1, 2, \dots, r,$$

In this paper, a residual generator based on an observer has the following general structure is proposed. Filter Rule i:

IF  $z_1(t)$  is  $M_{1i}$  and ... and  $z_p(t)$  is  $M_{1p}$ , THEN

$$F_i : \begin{cases} \dot{\hat{x}} = A_i \hat{x} + A_{di} \hat{x}_d + B_{ui} u + L_i (y - \hat{y}) \\ \hat{y} = C_i \hat{x} + D_{ui} u \\ r = V_i (y - \hat{y}) \end{cases} \quad (2)$$

Where  $\hat{x}$  is the state estimated vector,  $\hat{y}$  is the output estimated vector,  $L_i$  is the observer gain,  $V_i$  is the weighted matrix and  $r$  be the residual vector that depends on disturbances, fault signals and inputs. So the ideal residual is imported and represented as follows:

IF  $z_1(t)$  is  $M_{1i}$  and ... and  $z_p(t)$  is  $M_{1p}$ , THEN

$$W_i : \begin{cases} \dot{\hat{x}}_F = A_{Fi} \hat{x}_F + A_{Fdi} \hat{x}_{Fd} + B_{Fi} f \\ \hat{r} = C_{Fi} \hat{x}_F \end{cases} \quad (3)$$

where  $x_F \in R^{q_1}$  is the state variable,  $x_{Fd} = x_F(t-d)$ ,  $\hat{r} \in R^h$  is the ideal residual and  $A_{Fi}$ ,  $A_{Fdi}$ ,  $B_{Fi}$ ,  $C_{Fi}$ ,  $i = 1, 2, \dots, r$  are real matrices of appropriate dimensions.

Accordingly, the system in the augmented state space

$$\begin{aligned}\dot{\xi} &= \bar{A}_{cl}\xi + \bar{A}_{dcl}\xi_d + \bar{B}_{ucl}u + B_{dcl}d + B_{fcl}f \\ r &= C_{cl}\xi + D_{dcl}d + D_{fcl}f \\ r_e &= C_{cl1}\xi + D_{dcl}d + D_{fcl}f\end{aligned}\quad (4)$$

where  $\xi = [x^T \quad x_F^T \quad \tilde{x}^T]^T$ ,  $\xi_d = \xi(t-d)$ ,  $\tilde{x} = x - \hat{x}$ ,

$$\begin{aligned}r_e &= r - \hat{r}, \quad \bar{A}_{cl} = A_{cl} + \Delta A_{cl}, \quad \bar{A}_{dcl} = A_{dcl} + \Delta A_{dcl}, \quad \bar{B}_{ucl} = B_{ucl} + \Delta B_{ucl}, \\ A_{cl} &:= \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \begin{bmatrix} A_i & 0 & 0 \\ 0 & A_{Fi} & 0 \\ 0 & 0 & A_i - L_i C_j \end{bmatrix}, \quad \Delta A_{cl} := \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \begin{bmatrix} \Delta A_i & 0 & 0 \\ 0 & 0 & 0 \\ \Delta A_i & 0 & 0 \end{bmatrix} = \sum_{i=1}^r \lambda_i \bar{H} F \bar{E}_{li}, \\ A_{dcl} &:= \sum_{i=1}^r \lambda_i \begin{bmatrix} A_{di} & 0 & 0 \\ 0 & A_{Fi} & 0 \\ 0 & 0 & A_{di} \end{bmatrix}, \quad \Delta A_{dcl} := \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \begin{bmatrix} \Delta A_{di} & 0 & 0 \\ 0 & 0 & 0 \\ \Delta A_{di} & 0 & 0 \end{bmatrix} = \sum_{i=1}^r \lambda_i \bar{H} F \bar{E}_{di}, \\ B_{ucl} &:= \sum_{i=1}^r \lambda_i \begin{bmatrix} B_{ui} \\ 0 \\ 0 \end{bmatrix}, \quad \Delta B_{ucl} := \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \begin{bmatrix} \Delta B_{ui} \\ 0 \\ \Delta B_{ui} \end{bmatrix} = \sum_{i=1}^r \lambda_i \bar{H} F E_{ui}, \\ B_{dcl} &:= \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \begin{bmatrix} B_{di} \\ 0 \\ B_{di} - L_i D_{dj} \end{bmatrix}, \quad B_{fcl} := \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \begin{bmatrix} B_{fi} \\ 0 \\ B_{fi} - L_i D_{fj} \end{bmatrix}, \\ C_{cl} &:= \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j [0 \quad 0 \quad V_i C_j], \quad C_{cl1} := \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j [0 \quad -C_{Fi} \quad V_i C_j], \\ D_{dcl} &:= \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j V_i D_{dj}, \quad D_{fcl} := \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j V_i D_{fj}, \quad \bar{H} = [H^T \quad 0 \quad H^T]^T, \\ \bar{E}_{li} &= [E_{li} \quad 0 \quad 0], \quad \bar{E}_{di} = [E_{di} \quad 0 \quad 0]\end{aligned}$$

The objective of robust residual generation are the minimization of the effects of the disturbance and reference input on the residual and the maximization of the residual sensitivity to faults. The first leads to the minimization of the  $H_\infty$ -norms(5-7):

$$\min_{F_i} \sup_{d \in L_2-(0)} \frac{\|r\|_{L_2}}{\|d\|_{L_2}} \leq \gamma_d, \quad \min_{F_i} \sup_{u \in L_2-(0)} \frac{\|r\|_{L_2}}{\|u\|_{L_2}} \leq \gamma_u, \quad \min_{F_i} \sup_{f \in L_2-(0)} \frac{\|r_e\|_{L_2}}{\|f\|_{L_2}} \leq \gamma_f \quad (5-7)$$

The objectives is conflicting each other. In fact, there exists a trade-off between them, The detection problems can be recast into the following multi-objective  $H_\infty$  optimization problem.

Optimal FD design problem : Given positive real  $a$ ,  $b$  and  $c$ , find a filter realization  $F$  (equation (2)) such that

$$\min_{F_i} a\gamma_d + b\gamma_u + c\gamma_f \quad \text{s.t.} \quad (5) (6) \text{ and } (7) \quad (8)$$

### 3 FD Filter Design

The next step is to formulate the FD design problem and solve it via an LMI formulation. The following lemma is needed.

**Lemma 1.** If there are matrices  $Q_{ii} = Q_{ii}^T$ ,  $Q_{ij} = Q_{ji}^T (i \neq j = 1, 2, \dots, r)$ , such that

$$\Lambda_{ii} \leq Q_{ii}, \quad i = 1, 2, \dots, r, \quad (9)$$

$$\Lambda_{ij} + \Lambda_{ji} \leq Q_{ij} + Q_{ij}^T, \quad j < i \quad (10)$$

$$[Q_{ij}]_{r \times r} < 0, \quad (11)$$

hold, then

$$\sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \Lambda_{ij} < 0, \text{ where } \sum_{i=1}^r \lambda_i(z) = 1, \quad \lambda_k(z) \geq 0, \quad k = 1, 2, \dots, r$$

It can be established easily from reference [4] and hence its proof omitted.

**Lemma 2.** Let  $D, E$  and  $F(t)$  be real matrices of appropriate dimensions, and  $F(t)$  satisfying  $F^T(t)F(t) \leq I$ , Then, the following inequalities hold for any  $\varepsilon > 0$ ,

$$DF(t)E + (DF(t)E)^T \leq \varepsilon DD^T + \varepsilon^{-1} E^T E \quad (9)$$

**Lemma 3**[5]. Consider the system

$$\begin{aligned} \dot{x}(t) &= Ax(t) + A_d x(t-d) + Bw(t) \\ z(t) &= Cx(t) + Dw(t) \end{aligned}$$

If there exist matrices  $P = P^T > 0$  and  $S = S^T > 0$ , and positive scalar  $\gamma$  and

satisfying the inequality

$$\begin{bmatrix} A^T P + PA + S & PB & C^T & PA_d \\ * & -\gamma I & D^T & 0 \\ * & * & -\gamma I & 0 \\ * & * & * & -S \end{bmatrix} < 0$$

then the corresponding closed-loop system is globally exponentially stable and achieves L2 gain  $\gamma$ -performance for  $w$ .

Now, by exploiting the Lemma 3 and Lemma 2, conditions are satisfied, if there exist filter matrices  $L_i$ , and  $V_i$ , and an auxiliary matrix  $P = P^T > 0$ , and  $S = S^T > 0$ , By partitioning  $P$  and  $S$  as  $3 \times 3$  block-matrices,  $P = \text{diag}[P_1 \ P_2 \ P_3]$ ,  $S = \text{diag}[S_1 \ S_2 \ S_3]$ , such that the following matrix inequality

$$\begin{bmatrix} \Gamma_1 & PB_{ucl} & C_{cl}^T & PA_{dcl} & P\bar{H} & P\bar{H} & P\bar{H} \\ * & \Gamma_3 & 0 & 0 & 0 & 0 & 0 \\ * & * & -\gamma_u I & 0 & 0 & 0 & 0 \\ * & * & * & \Gamma_2 & 0 & 0 & 0 \\ * & * & * & * & -\varepsilon_1 I & 0 & 0 \\ * & * & * & * & * & -\varepsilon_d I & 0 \\ * & * & * & * & * & * & -\varepsilon_u I \end{bmatrix} < 0 \quad (12)$$

$$\begin{bmatrix} \bar{A}_{cl}^T P + P\bar{A}_{cl} + S & P\bar{B}_{ucl} & C_{cl}^T & P\bar{A}_{dcl} \\ * & -\gamma_u I & 0 & 0 \\ * & * & -\gamma_u I & 0 \\ * & * & * & -S \end{bmatrix} < 0 \quad (13)$$

$$\begin{bmatrix} \bar{A}_{cl}^T P + P\bar{A}_{cl} + S & P\bar{B}_{fcl} & C_{cl1}^T & P\bar{A}_{dcl} \\ * & -\gamma_f I & D_{fcl}^T & 0 \\ * & * & -\gamma_f I & 0 \\ * & * & * & -S \end{bmatrix} < 0 \quad (14)$$

where  $\varepsilon_{li} > \varepsilon_1 > 0$ ,  $\varepsilon_{ui} > \varepsilon_u > 0$ ,  $\varepsilon_{dli} > \varepsilon_{d1} > 0$ ,  $\Gamma_2 = -S + \sum_{i=1}^r \lambda_i \varepsilon_{di} \bar{E}_{di}^T \bar{E}_{di}$

$\Gamma_1 = A_{cl}^T P + P A_{cl} + S + \sum_{i=1}^r \lambda_i \varepsilon_{li} \bar{E}_{li}^T \bar{E}_{li}$ ,  $\Gamma_3 = -\gamma_u I + \sum_{i=1}^r \lambda_i \varepsilon_{ui} E_{ui}^T E_{ui}$

$$\sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \Omega_{ij} < 0, \quad \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \Xi_{ij} < 0, \quad \sum_{i=1}^r \sum_{j=1}^r \lambda_i \lambda_j \Psi_{ij} < 0 \quad (15-17)$$

Based on lemma 1, the inequality (15) (16) and (17) are linear in  $P_1, P_2, P_3, S_1, S_2, S_3, \hat{L}_i, V_i, \varepsilon_{li}, \varepsilon_{ui}, \varepsilon_{dli}, \varepsilon_1, \varepsilon_u, \varepsilon_d$ . The next main result summarizes the above discussion and provides a procedure for solving FD filter for T-S Fuzzy Time-delay Systems.

**Theorem 1.** if there are matrices  $Q_{ii} = Q_{ii}^T$ ,  $Q_{ij} = Q_{ji}^T$  ( $i \neq j = 1, 2, \dots, r$ ),  $T_{ii} = T_{ii}^T$ ,  $T_{ij} = T_{ji}^T$  ( $i \neq j = 1, 2, \dots, r$ ),  $U_{ii} = U_{ii}^T$ ,  $U_{ij} = U_{ji}^T$  ( $i \neq j = 1, 2, \dots, r$ ), a feasible solution to the FD problem is obtained by solving a sequence of optimization problems.



$$\min_{P_1, P_2, P_3, S_1, S_2, S_3, \tilde{L}_1, V_1, \epsilon_{i1}, \epsilon_{d1}, \epsilon_{i2}, \epsilon_{d2}, \epsilon_1, \epsilon_u, \epsilon_d} a\gamma_d + b\gamma_u + c\gamma_f, \quad \text{s.t. } P > 0, S > 0$$

$$\begin{aligned} \Omega_{ii} &\leq Q_{ii}, \quad \Omega_{ij} + \Omega_{ji} \leq Q_{ij} + Q_{ij}^T, (j < i), \quad [Q_{ij}]_{r \times r} < 0, \quad \Xi_{ii} \leq T_{ii}, \quad [T_{ij}]_{r \times r} < 0, \\ \Xi_{ij} + \Xi_{ji} &\leq T_{ij} + T_{ij}^T (j < i), \quad \Psi_{ii} \leq U_{ii}, \quad \Psi_{ij} + \Psi_{ji} \leq U_{ij} + U_{ij}^T (j < i), \quad [U_{ij}]_{r \times r} < 0, \end{aligned}$$

**Proof.** By collecting all the previous discussion and lemma 1.

## 4 Simulation Examples

To illustrate the proposed results, Consider the following truck-trailer model proposed in [2].

$$\begin{aligned} A_1 &= \begin{bmatrix} -3.6825 & 11.0149 & -1.3144 \\ -0.5091 & 0 & 0 \\ 0.5091 & -4.000 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -3.7104 & 11.2883 & -1.3411 \\ -0.5091 & 0 & 0 \\ 0.8102 & -6.3662 & 0 \end{bmatrix}, \quad A_{d1} = \begin{bmatrix} 0.2182 & 0 & 0 \\ -0.2182 & 0 & 0 \\ 0.2182 & 0 & 0 \end{bmatrix} \\ A_{d2} &= \begin{bmatrix} 0.2182 & 0 & 0 \\ -0.2182 & 0 & 0 \\ 0.3472 & 0 & 0 \end{bmatrix}, \quad B_{u1} = \begin{bmatrix} -1.4286 \\ 0 \\ 0 \end{bmatrix}, \quad B_{u2} = \begin{bmatrix} -1.4286 \\ 0 \\ 0 \end{bmatrix}, \quad B_{d1} = B_{d2} = \begin{bmatrix} 0.1 \\ 0 \\ 0.2 \end{bmatrix}, \quad B_{f1} = B_{f2} = \begin{bmatrix} 1 \\ 2 \\ 0.5 \end{bmatrix}, \\ C_1 = C_2 &= [1 \quad 0 \quad 0], \quad D_{u1} = D_{u2} = 0, \quad D_{d1} = 0.1, \quad D_{d2} = 0.2, \quad D_{f1} = 1, \quad D_{f2} = 2, \\ H &= [0.15 \quad 0 \quad 0.1]^T, \quad E_{i1} = E_{i2} = [0.1 \quad -0.15 \quad 0.15], \quad E_{d1} = E_{d2} = [0.15 \quad -0.15 \quad 0], \\ E_{u1} = E_{u2} &= [0.1 \quad 0.1 \quad 0.1], \quad F(t) = \sin(10t) \end{aligned}$$

The  $M_1(x_1)$  and  $M_2(x_1)$  represent the membership functions of fuzzy sets about 0 and  $\frac{\pi}{2}$  as follows:

$$\begin{cases} M_1(x_1) = \left( 1 - \frac{1}{1 + \exp(-3(\theta(t) - 0.5\pi))} \right) \left( \frac{1}{1 + \exp(-3(\theta(t) - 0.5\pi))} \right) \\ M_2(x_1) = 1 - M_1(x_1) \end{cases}$$

$$\theta(t) = -0.1273x_1 - 0.0545x_{d1} + x_2$$

To maintain the spacecraft in its upward vertical position, we design a state feedback controller that stabilizes the nominal dynamics by PDC

$$u = [-1676.8 \quad -303.8]x.$$

In simulation,  $w$  is assumed to be unitary white noises. Here the residual responses to the following fault signal

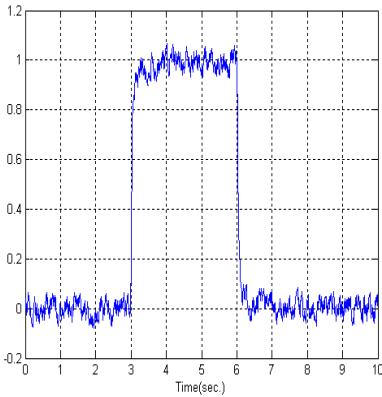
$$f(t) = \begin{cases} 0 & t < 3 \\ 1 & 3 \leq t \leq 6 \\ 0 & t > 6 \end{cases}$$

The following weighting function of fault  $W(s) = \frac{5}{s+1}$  has been selected.

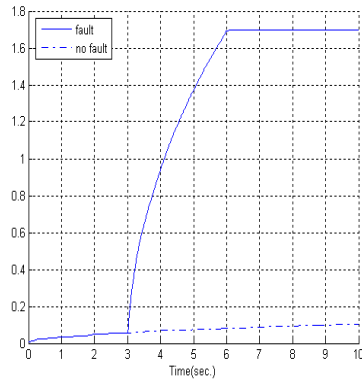
$$a = 100, b = 10, c = 1$$

Next, we design a robust filter capable to track fault signals. FD filter has been computed by means of Theorem 1. The filter is represented as follows:

$$L_1 = \begin{bmatrix} 0.0363 \\ 0.0897 \\ -0.0033 \end{bmatrix}, L_2 = \begin{bmatrix} -0.0138 \\ 0.1508 \\ -0.0120 \end{bmatrix}, V_1 = 0.0028, V_2 = 0.0017, \gamma_d = 0.0047, \gamma_u = 0.0144$$



**Fig. 1** Residual response



**Fig. 2** Evolution of residual evaluation

The simulation time is 10 sec. The threshold results  $J_{th}(t) := \sup_{d, u \in L_2, f=0} \|r(t)\|_{2,30} = 0.0024$ . The result  $J(0, 10.5) = 0.004 > J_{th}(t)$

$0.642 > J_{th}(t)$ , so the fault can  $J(r) = \sqrt{\int_0^{10.5} r^T(\tau)r(\tau)d\tau} = 0.004 > J_{th}$  successfully be detected at the 0.5 sec after fault has appeared.

From Figure 1, Residual can track the fault signal and restrain the disturbance.

## 5 Conclusions

In this paper, a novel solution for Robust Fault Detection for T-S fuzzy systems has been proposed. Sufficient conditions for the existence of a solution to the detection of faults is provided by the proposed filter design. The scheme focuses on the fault indicating signal to be an optimal estimate, in the  $H^\infty$ -norm sense, of the fault signal. By taking advantage of the Projection Lemma, the FD problem has been converted into a quasi-LMI optimization problem. Experiment results from a spacecraft system are presented to demonstrate the effectiveness of the proposed method.

**Acknowledgements.** The paper is supported by Heilongjiang Province Office of Education fund [11533056] and [11533060].

## References

- [1] Nobrega, E.G., Abdalla, M.O., Grigoriadis, K.M.: LMI-based filter design for fault detection and isolation. In: Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, Australia, pp. 4329–4334 (December 2000)
- [2] Bai, L.-S., Tian, Z.-H., Shi, S.-J.: Robust Fault Detection for Uncertain Time-delay Systems Based on  $H^\infty$  Filter. *Information and Control* 34(2), 163–166 (2005)
- [3] Zhong, M.-Y., Ye, H., Chen, G.-Y., Wang, G.-Z.: An ILMI Approach to RFDF for Uncertain Linear Systems with Nonlinear Perturbations. *Acta Automatica Sinica* 31(2), 297–300 (2005)
- [4] Feng, G., Cao, S.G., Rees, N.W., Chak, C.K.: Design of fuzzy control systems with guaranteed stability. *Fuzzy Sets Syst.* 85, 1–10 (1997)
- [5] Liu, X., Zhang, Q.: New approaches to  $H^\infty$  controller designs based on fuzzy observers for T-S fuzzy systems via LMI. *Automatica* 39, 1571–1582 (2003)
- [6] Cao, S.G., Rees, N.W., Feng, G.: Stability analysis and design for a class of continuous-time fuzzy control systems. *Int. J. Control* 64(6), 1069–1087 (1996)
- [7] Wang, H.-R., Wang, C., Gao, H.: Robust fault detection for uncertain systems via LMIs. *Electric Machines and Control* 9(5), 461–465 (2005)
- [8] Cao, Y.Y., Sun, Y.X.: Robust stabilization of uncertain systems with time-varying multi-state delay. *IEEE Trans. On Automatic Control* 43(10), 1484–1488 (1998)
- [9] Frank, P.M., Ding, X.: Survey of robust residual generation and evaluation methods in observer-based fault detection systems. *Journal of Process Control* 7, 403–424 (1997)
- [10] Yu, L.: Robust control—a LMI method. Tsinghua University Press, Beijing (2002)

# Autonomous and Pervasive Computing-Based Knowledge Service

Keedong Yoo

**Abstract.** Users, nowadays, are highly educated and trained, therefore, they are eager to solve ordinary problems by themselves without waiting expert's assistance. To fulfill these needs, the service of knowledge support must be provided in accurate and timely manners. The knowledge service needs to be not only preceded in itself without users' efforts to search what they want to know, but also delivered pervasively whenever and wherever the users are. Therefore, this paper proposes a framework of autonomous knowledge service, which autonomously provides proper knowledge to users by automatically identifying users' needs and situations. To provide knowledge services in a pervasive manner, a mobile network technology is additionally applied to the proposed framework. Autonomous as well as pervasive knowledge service can fulfill users' needs for computing services in the future, and hence underpin the foundation of intelligent, however more human-centered knowledge-based living environments.

## 1 Introduction

User's requirements towards the future computing services can be summarized as two categories; proactive and autonomous services. Proactive services can be provided by foreseeing the future events resulted from the past and present events, while autonomous services can be generated by the accumulated rules among events that have the relationship of cause and effect. Difference between autonomy-based computing and proactiveness-based one can be concluded that the former can be attained by the latter well-organized and hence clearly refined, however, both of them may share the origin: the context. Context is defined as the secondary information describing user-surrounding situations. Analyzing and

---

Keedong Yoo

Dept. of Management, Dankook University, Cheonan, Republic of Korea

e-mail: kdyoo@dankook.ac.kr

identifying users' context can give answers to the question of what users want to do, therefore construct the starting point of proactive and autonomous computing services. Context awareness-related techniques, such as voice and action recognition methods, machine learning algorithms, and context ontology, etc., can be applied to provide more accurate and effective autonomous computing services. Through user context-based autonomous computing, users can experience more replenished computing environments, as well as human life finally can be set as the well-being one.

Among various computing services, what users usually need to be supported are knowledge services. Users nowadays are highly educated and trained, therefore, they are eager to solve ordinary problems by themselves without waiting expert's assistance. To fulfill these needs, the service of knowledge supports must be provided in accurate and timely manners. To interpret these requirements from the technology viewpoint, the knowledge service must be ready for not only the immediate response but also the exact diagnosis-based recommendation. In other words, the knowledge service needs to be proceeded in itself without users' efforts to search what they want to know. Furthermore, knowledge must be distributed in a pervasive manner using portable devices carried by users.

## **2 Autonomous Computing-Based Knowledge Service**

To constitute a set of autonomous computing-based knowledge service, various theories and methods from various research areas must be put together so that the knowledge service can yield the best capabilities with holding the major objectives of proactiveness and autonomy in computing activities. Computer science, cognitive science, linguistics, ergonomics, industrial engineering, and telecommunication are typical areas to be aggregated in an interdisciplinary manner to achieve those objectives, and the resultant expected capabilities are summarized as follows;

### ***2.1 Capabilities***

#### *A. Autonomous knowledge acquisition (using sensing and text mining technologies)*

The first step to initiate knowledge service is to acquire knowledge from various sources, such as existing references, experts, knowledge workers, etc. Conventional ways to acquire knowledge depend on the manual operations, however this type of manual registration cannot guarantee the quantity and quality of knowledge. Therefore, using sensors invisibly embedded in workplaces and environments, knowledge within documents, discussions, and actions can be automatically scanned, dictated, and recorded, and finally converted into the text-based electronic documents. Applying text mining technologies to the documented pieces of knowledge, the topic of each piece can be automatically identified. Storing each document which describes each type of knowledge according to the each topic, autonomous knowledge acquisition is completed.

### *B. Knowledge part extraction (using linguistic and semantic analysis)*

Storing every part of acquired knowledge diminishes the efficiency of a knowledge-Base because whole part of knowledge containing documents, discussions, and actions cannot be regarded as knowledge itself. A piece of document, discussion, and action is composed of knowledge parts and background parts, therefore concentrating on and extracting the knowledge part only is necessary. To extract the knowledge part from the electronically converted documents of knowledge, the semantic analysis of the linguistic approach must be performed by analyzing each sentence within documents. Discriminating meaningful phrases from non-meaningful ones can deliver the unique knowledge part of the given documents, and eventually the exact part of knowledge can be extracted and stored with guaranteeing the efficient use of a knowledge-base.

### *C. Context-specific knowledge navigation (using a knowledge map with contextual semantics)*

To enhance the effectiveness of knowledge use, knowledge must be arranged according to the context of use as well as the relationship of cause and effect. So-called 'context-rich knowledge' explains both of the condition that the given knowledge must be used or executed and the resultant situation that the given knowledge has been used or executed. By arranging this cause and effect relationships surrounding a piece of knowledge, a knowledge map is to be developed, which describes the orders and conditions of knowledge application in the form of knowledge network. The node of the map denotes a piece of knowledge, and the arc illustrates when and where the related knowledge is to be fired. Based on the knowledge map explains the relationships between knowledge, context-specific knowledge navigation can be achieved, and consequently autonomous knowledge execution is enabled.

### *D. Automated Knowledge-Base expansion (using ontology technologies)*

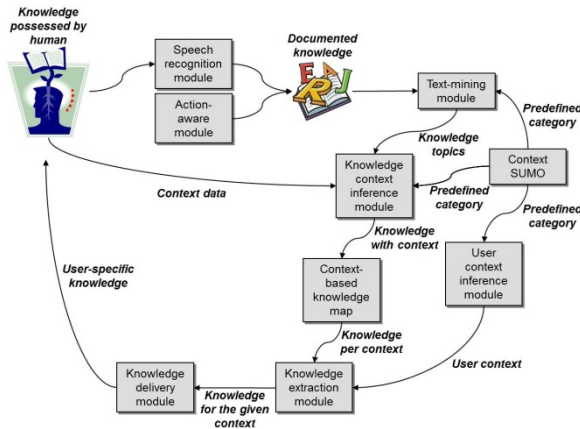
To implement the knowledge map with contextual semantics, knowledge and context must be modeled together with respect to their relationships and properties. An ontology is able to be applied to very properly describe the concepts and domains of knowledge and context, because an ontology provides a shared vocabulary, which can be used to model a domain, that is, the type of objects and/or concepts that exist, and their properties and relations. Although some limitations exist around the standard upper ontology which perfectly includes all properties and relations of objects and/or concepts yet, a semi-generalized upper ontology can be defined and implemented using existing ontologies by properly modifying them to fit given situations using satisfactorily formalized tools such as WordNet. An ontology-based knowledge-base can gather and store knowledge with respect to its context as well as reason another knowledge using the relations between knowledge and context in an automated manner. The self-expanding knowledge-base can also provide very effective knowledge query service in turn; therefore more plentiful and realistic knowledge search activities are generated.

### E. Ubiquitous knowledge service (using context inference and pervasive network)

Federated sensors and users' mobile devices can be applied to acquire users' context as well as to distribute resultant knowledge by combining them with the pervasive network technologies. Once user-surrounding data such as time, place, schedule, and identity, etc. are captured by sensors and mobile devices, they are combined with the topic which is retrieved by the knowledge acquisition subsystem and explains the job category of the user. The combination of the topic a user currently concerns, the time when an event takes place, the place where the user locates, the schedule the user plans to do, and the identity the user is whom can uniquely depict one fact what the user currently want to be serviced. To extract the resultant fact what a system must provide, either of machine learning algorithms such as neural networks, case-based reasoning, and nearest neighborhood method, etc., or ontology-based rule inference method can be applied. Resultant knowledge to be serviced can be distributed pervasively via mobile network protocols such as WiFi-based wireless LAN, HSDPA-based mobile Internet, and Mobile WiMax-based portable Internet by simply applying socket programming.

## 2.2 Conceptual Frameworks

Conceptual framework of the autonomous computing-based knowledge service, especially focused on verbal knowledge that resides in ordinary dialogues, can be abstractly diagrammed as "Fig.1." The core parts of the framework are knowledge context inference module, context-based knowledge map, and user context inference module, therefore they must be developed in advance.



**Fig. 1** Conceptual framework of the autonomous computing-based knowledge service system

Knowledge context inference module reads, analyzes, and interprets the transmitted text-based dialogues, and hence concludes what the dialogues are about. By extracting keywords of the transmitted text-based data, each dialogue can be classified in terms of the resulted keywords [10, 11]. The dialogue can be regarded as knowledge itself, because it contains various kinds of knowledge applicable to a similar situation. Because the keyword can be regarded as the topic of the dialogue, participants' knowledge can be categorized according to the topic. Besides the topic, context data which have been identified and transmitted from the context acquisition module must be applied to make the relationship between the knowledge and the situation of the dialogue. Context data, such as participant's identity, location, time, and schedule, can play the role of the meta-knowledge by uniquely depicting the knowledge used in the given situation [3, 2, 6, 5].

Context-based knowledge map plays the role of categories to store knowledge. This module stores knowledge according to its topics and displays knowledge when users request. The ontology technology is reasonably applicable to store various categories of knowledge according to various kinds of context [12, 11]. Because the number of context and its relationships are too many to count, the ontology is proper to manage them with, so called, the context ontology [9, 4, 1]. By storing knowledge with its context using the ontology, more autonomous expansion of knowledge as well as accurate application of knowledge can be expected.

User context inference module enables the proposed system to proactively generate and provide knowledge which is the most suitable to handle user's situation by automatically inferring user's context data. User's context data can be initially identified by user's hand-held mobile devices, and they are inputted to finally conclude what user's current situation is. To infer user's situation, not only some of machine learning algorithms can be applied, but the rule inference functionality of the ontology-based knowledge-base is also applicable [7, 8].

### 3 Conclusions and Research Vision

Autonomous computing, comparing to the concept of autonomic computing proposed by IBM, concerns self-generating features of computer-based services, providing adequate computer services in a proactive manner by continuously monitoring and identifying users' context data which explain users' current situations. An autonomous system, therefore, not only makes decisions on its own, but also communicates with users continually in a real time basis using another high-level policies; it will continuously monitor and identify users' situations and automatically adapt its services to changing situations surrounding users.

This paper tries to apply the capability of autonomous computing to the conventional approaches of knowledge service. Although the concept of autonomous computing is not widely known yet, its tremendous potentials elicit practitioners to pay more attentions for delivering human-like intelligent systems. By combining the technologies of pervasive computing, autonomous computing-based service of knowledge can facilitate utilization of knowledge at anytime and anywhere in more convenient ways, which is the main objective of this research.



Computers with intelligence can support human activities to be more convenient, and hence result in more fertilized human life. Computers cannot smartly think and act as human do, of course, because the way they work must be designated and coordinated by human, however they might be clever as much as a human being can be someday in the future. Studies and researches about knowledge services by intelligent computers are indispensable not only because computers must be trained to be correctly reactive, but also because they must be designed to be rightly reactive. Reliable, however more trustworthy knowledge services by computers are to be pursued and realized through this research.

**Acknowledgments.** 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.H00021).

## References

1. Henricksen, K., Indulska, J.: Developing context-aware pervasive computing applications: Models and approach. *Pervasive and Mobile Computing* 2(1), 37–64 (2006)
2. Kang, H., Suh, E., Yoo, K.: Packet-based context aware system to determine information system user's context. *Expert Systems with Applications* 35(1-2), 286–300 (2008)
3. Kim, S., Suh, E., Yoo, K.: A study of context inference on the Web-based information systems. *Electronic Commerce Research and Applications* 6, 146–158 (2007)
4. Korpipää, P., Mäntyjärvi, J., Kela, J., Keränen, H., Malm, E.: Managing Context Information in Mobile Devices. *IEEE Pervasive Computing* 2(3), 42–51 (2003)
5. Kwon, O., Yoo, K., Suh, E.: ubiDSS: A Proactive Intelligent Decision Support System as an Expert System Deploying Ubiquitous Computing Technologies. *Expert Systems with Applications* 28(1), 149–161 (2005)
6. Kwon, O., Yoo, K., Suh, E.: ubiES: Applying Ubiquitous Computing Technologies to An Expert System for Context-Aware Proactive Services. *Electronic Commerce Research and Applications* 5(2), 209–219 (2006)
7. Ranganathan, A., Al-Muhtadi, J., Campbell, R.H.: Reasoning about uncertain contexts in pervasive computing environments. *IEEE Pervasive Computing* 3(2), 62–70 (2004)
8. Ranganathan, A., Campbell, R.H., Ravi, A., Mahajan, A.: ConChat: A context aware chat program. *IEEE Pervasive Computing* 1(3), 51–57 (2002)
9. Wang, X., Dong, J.S., Chin, C.Y., Hettiachchi, S.R.: Semantic space: An infrastructure for smart spaces. *IEEE Pervasive Computing* 3(3), 32–39 (2004)
10. Yoo, K.: Autonomous knowledge acquisition methodology using knowledge workers' context information: Focused on the acquisition of dialogue-based knowledge for the next generation knowledge management systems. *Journal of Korea Knowledge Management Society* 9(4), 65–75 (2008)
11. Yoo, K., Kwon, O.: Vision and research challenges of the next generation knowledge management systems: A pervasive computing technology perspective. *Journal of Korea Knowledge Management Society* 10(1), 1–15 (2009)
12. Yoo, K., Suh, E., Kim, K.Y.: Knowledge flow-based business process redesign: Applying a knowledge map to redesign a business process. *Journal of Knowledge Management* 11(3), 104–125 (2007)

# Short-Term Traffic Flow Forecasting Based on Periodicity Similarity Characteristics

Chunjiao Dong, Chunfu Shao, Dan Zhao, and Yinhong Liu

**Abstract.** The methodology has been putted forward that the periodicity similarity should be consideration when using Elman neural network (ENN) to forecast short-term traffic flow, which is not only helpful to save training time, reduce training sample size, but also enhance forecasting efficiency. Firstly, training sample of ENN has been designed based on the periodicity similarity of traffic flow and network structure has been established aiming at improving ENN global stability. Secondly, short-term traffic flow forecasting method based on ENN have been established by taking daily, weekly and monthly periodicity similarity into account respectively. Finally, forecasting results have been evaluated by four error statistics from two aspects: forecasting effect and efficiency. The conclusion has been summarized to three aspects.

## 1 Introduction

Traffic jam is getting worse every day along with the development of economy and urban traffic flow increases rapidly. Since the 1980s, developed countries studied on the technology of transportation management and control, and Intelligent Transportation Systems (ITS) emerged responsively. Traffic control and route guidance is one of the important research domains in ITS, while precise short-term traffic flow forecasting is the key of traffic control and route guidance.

Since 1960s, some scholar started to research short-term traffic flow forecasting problems by applying mature reliable models of other field, produced a variety of

---

Chunjiao Dong

Center of Transportation Research, 309 Conference Center Building, 600 Henley Street,  
Knoxville, 37996, Tennessee, USA  
e-mail: CDONG5@UTK.EDU

Chunjiao Dong · Chunfu Shao · Dan Zhao · Yinhong Liu

MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology  
School of Traffic & Transportation, Beijing Jiaotong University, Beijing 100044, China

forecasting models and methods. Formally, these methods can be classified into two groups. One is based on certain mathematic model, and the other is intelligent algorithm without model. The former contains multivariate regression model, adaptive weight integrating model, Kalman filtering model, reference function-exponential smoothing model, UTS-2(3) model and combined model[1][2]. The latter includes nonparametric regression method, spectra analysis, state space reconstruction model, wavelet neural network, multi-fractal method, method based on wavelet decomposition and reconstruction, and various composite forecasting models related to neural network[3][4]. In fact, transportation system is complex, time-varying, comprehensive and subjected to random disturbance, its variation rules are nonlinear too[5], so the method based on linear theory cannot explain transportation system variation accurately. Besides, for its disability on adaption and self-learning, the linear forecasting system is not very robust. During the 20th century, Artificial neural network is really hot due to its nonlinear mapping. Nonetheless, neural network is not perfect too, for example, the difference between training data and forecasting data is wider, and the forecasting error will be larger. This paper analyses the relationship between periodicity similarity of traffic flow and forecasting effect as well as the efficiency evaluation index. The methodology is put forward that periodicity similarity of traffic flow must be taken into consideration when using ENN to forecast short-term traffic flow.

## 2 Periodicity Similarity of Traffic Flow

According to statistical time, periodicity similarity of traffic flow can be classified into annual, monthly, weekly and daily variation. Researching time-variety of traffic flow is conducive to master the variation regularity of traffic flow, thus establish valid forecasting method.

The periodicity similarity of traffic flow is researched with traffic flow in consecutive two days, one week and one month, because the weekly traffic flow curve had the highest similarity and the strongest regularity, what means the trends of traffic flow is very close every week, and this can be called as weekly similarity. On that basis, the periodicity similarity of traffic flow was further analyses at the same working day of four weeks in one month. Some researchers have entirely analyzed and approved that traffic flow has periodicity similarity, the relevant evidences are as follows [6]:

(1) Similarity Coefficient (SC for short). Supposing  $F=[f_{i1}, f_{i2}, \dots, f_{in}]_{k \times n}$  is a matrix constituted by the traffic flow, speed and density parameter of  $n$  days data at a special position, and  $k$  is the number of daily data, hence similarity coefficient is defined:

$$S = \frac{\sum_{n \geq i > j \geq 1} R(i, j)}{n(n-1)/2} \quad |S| \leq 1 \quad (1)$$

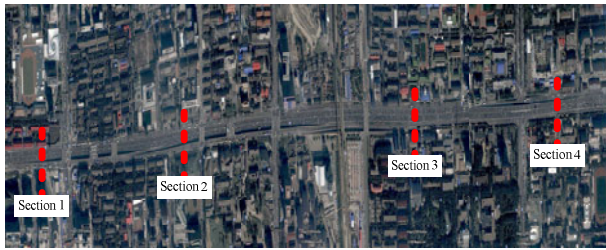
Where  $R$  is the correlation matrix of  $F$ . The periodicity similarity of traffic flow improves as value of  $S$  increases. When  $S$  is greater than 0.92, it can be estimated that traffic flow have periodicity similarity. Because the effect of traffic flow absolute value has been eliminated, the value of  $S$  only represents the trends

of single-point traffic flow variety similarly with the time axis. Therefore, some researchers redefined variation coefficient  $\delta$  to determine whether the traffic flow has absolute similarity.

(2) Variation Coefficient (VC for short). The column vector mean of matrix  $F$  is noted as  $M=[E(fi1), E(fi2), \dots, E(fin)]$ , the variation coefficient  $\delta$  is described as follows:

$$\delta = \frac{\sqrt{D(M)}}{E(M)} \tag{2}$$

Where  $D(M)$  is variance of matrix  $M$ , and  $E(M)$  is mathematical expectation of matrix  $M$ . The periodicity similarity of traffic flow increases as value of  $\delta$  decreases, and if  $\delta$  is less than 0.05, the traffic flow will have absolute periodicity similarity.



**Fig. 1** Section position on the urban expressway network

In order to satisfy the requirement of traffic control and guidance the data used in this paper is collected from four sections on urban expressway in Beijing with 2-minutes interval, and the periodicity similarity is analyzed separately under 7 states. Road section detectors are fixed as shown in Fig. 1. On the basis of above-mentioned criteria, the similarity coefficient and variation coefficient are calculated in Table 1.

**Table 1** SC & VC in Different States

ID	d-d	w-d	m-d	one w-d	two w-d	three w-d	four w-d
Similarity Coefficient							
1	0.974	0.889	0.647	0.858	0.740	0.698	0.740
2	0.970	0.949	0.928	0.981	0.979	0.974	0.951
3	0.961	0.953	0.946	0.983	0.979	0.975	0.958
4	0.982	0.931	0.859	0.926	0.898	0.862	0.849
Variation Coefficient							
1	0.022	0.114	0.497	0.066	0.412	0.594	0.506
2	0.065	0.050	0.132	0.011	0.098	0.080	0.141
3	0.005	0.040	0.085	0.007	0.006	0.010	0.116
4	0.010	0.044	0.059	0.077	0.058	0.048	0.048

According to the calculation results, it's learned that short-term traffic flow periodicity similarity is stronger, and weakens as time increases. If the similarity coefficients of seven states all meet criteria of periodicity similarity, periodicity similarity of traffic flow in consecutive two weeks is the strongest.

### 3 Short-Term Traffic Flow Forecasting

The methodology of ENN is proposed by Elman in 1990. As the delay operator, an acceptor layer is added in the hiding layer of feedback network to realize purpose of memory [9].

Generally, ENN is divided into 4 layers: input layer, intermediate layer (hidden layer), continual layer and output layer. The input layer unit has only the signal deliver function, and the output layer unit has linear weighting function. Hidden layer unit delivering function can adopt linear or non-linear function, continual layer also is called the context layer or the state layer, and it is used to memory the output value before hidden layer unit and it can be regarded as one step to postpone operator. Based on ENN, nonlinear states spatial of short-term traffic flow forecasting are expressed:

$$y(k) = g(w^3 x(k)) \quad (3)$$

$$x(k) = f(w^1 x_c(k) + w^2 (q(k-1))) \quad (4)$$

$$x_c(k) = x(k-1) \quad (5)$$

Where  $y$ ,  $x$ ,  $q$  and  $x_c$  are denoted output node vector of  $m$  dimension,  $n$  dimension node unit vector of middle layer, input vector of  $r$  dimension and feedback state vector of  $n$  dimension.  $w_3$ ,  $w_2$  and  $w_1$  are denoted connecting weights from middle layer to output layer, from input layer to middle layer, from acceptor layer to middle layer respectively.  $g(\bullet)$  is transfer function of output neural unit, usually defined as  $S$  function. ENN learning target function is described as follows:

$$E(w) = \sum_{k=1}^n [y_k(w) - \tilde{y}_k(w)]^2 \quad (6)$$

Where  $\tilde{y}_k(w)$  is the output vector. Previous researches have testified that traffic flow time series represent high chaos characteristic and short-term predictability when its length equals to 5, thereby input vectors are established as five-dimension. The relationship between inputs and outputs can be expressed:

$$y = q_i(t) = F(q_i(t-1) + q_i(t-2) + q_i(t-3) + q_i(t-4) + q_i(t-5)) \quad (7)$$

Where  $F(\bullet)$  is input-output fitting function,  $q_i(t)$  is normalized traffic flow data. This paper uses the daily, weekly, monthly traffic flow data sets before testing

sample as training sample to train ENN. Moreover, as above mentioned, weekly traffic flow has the highest level of periodicity similarity and the strongest regularity, traffic flow time series of same working day in one month is selected as training data to study the capability of ENN based on periodicity similarity of traffic flow. According to the chaos characteristics of traffic flow, input nodes number is fixed for 5 and output nodes number is 1. The quantity of hidden nodes is determined by trial-and-error method. Initially, the range of hidden nodes number was set for 10~21 by experience, after that ENN was tested by training sets. Furthermore, forecasting results are evaluated and finally obtained the optimal model. It found that network with 16 hidden nodes performs best, so it is adopted to forecast short-term traffic flow based on periodicity similarity of traffic flow.

### 4 Forecasting Results

Two conventional evaluation indices of short-term traffic flow forecasting are: mean absolute percentage error (MAPE) and mean absolute deviation (MAD). The two indices reflect forecasting efficiency from different aspects, and forecasting efficiency is better when the value of them is smaller. Besides, another index are introduces to evaluate forecasting effect: accepted error percentage (AEP).

Comparison of forecasting efficiency indices of four sections under seven states are shown in Fig. 2, it can be concluded that MAPE of most sections decreases as training sample increases. In case training data size is small, forecasting effect will result doubled perfect by utilizing training modes states of the 4<sup>th</sup> and 5<sup>th</sup>.

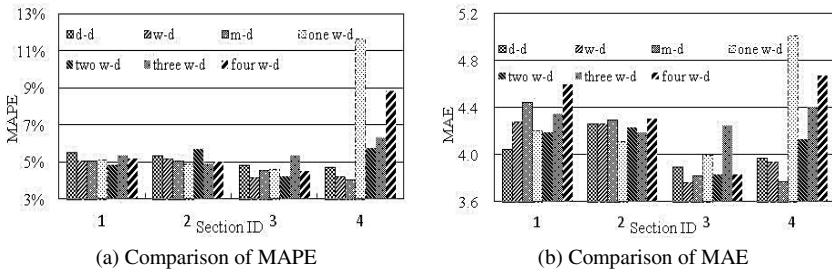
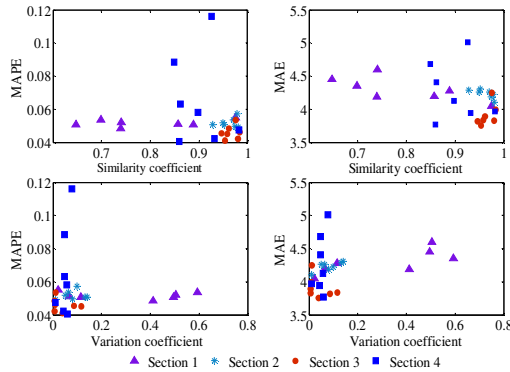


Fig. 2 Comparison of forecasting efficiency

The relationship between forecasting efficiency and single criteria of periodicity similarity of traffic flow is analyzed in Fig. 3. The value of forecasting efficiency index indicates a decreasing trend as the similarity coefficient increases, and decreases as variation coefficient reduces, that is to say, forecasting effect will get better if selected traffic flow data sets has strong periodicity similarity.



**Fig. 3** Sensitivity analyses between forecasting efficiency and periodicity similarity

## 5 Conclusions

The ENN based on periodicity similarity of traffic flow is designed in this paper. It can not only improve convergence rate, but also enhance the real-time property and forecasting accuracy. By analyzing the relationship between forecasting results of ENN and criteria of periodicity similarity of traffic flow, three conclusions are summarized as follows:

(1) Under normal circumstances, the forecasting results of ENN are much better when taking monthly periodicity similarity into account, instead of considering weekly periodicity similarity only, but both of them are superior to consideration of daily periodicity similarity.

(2) With regard to small samples, the forecasting efficiency is the best if input data are time series which are composed of the traffic flow of one and two weeks ago same time-period, the error reduced 5.81% in maximum.

(3) Forecasting efficiency and the periodicity similarity are positive correlated, which means the samples with a high level of periodicity similarity help to enhance forecasting efficiency. Forecasting effect and the periodicity similarity are not obvious causality, which means the samples with a high level of periodicity similarity can't enhance forecasting effect obviously.

**Acknowledgments.** This paper is supported by the National Basic Research Program of China (973 Program, 2006CB705505), Doctoral Thesis Fund, Research and Development Foundation for Chinese Young Professionals, sponsored by General Motors Company, 2009 and Excellent Doctoral Student Scholarship for Science and Technology Innovation, Beijing Jiaotong University, 2010(No: 141082522).

## References

- [1] Dong, C.J., Shao, C.F., Li, X.: Short-Term Traffic Flow Forecasting of Road Network Based on Spatial-Temporal Characteristics of Traffic Flow. In: 2009 WRI World Congress on Computer Science and Information Engineering Proceeding, pp. 645–650 (2009)

- [2] Ishak, S., Al-Deek, H.: Performance Evaluation of Short-Term Temporal-Series Traffic Prediction Model. *Journal of Transportation Engineering* 128(6), 490–498 (2002)
- [3] Stathopoulos, A., Karlaftis, M.G.: A multivariate state spatial approach for urban traffic flow modeling and prediction. *Transportation Research Part C* 11, 121–135 (2003)
- [4] Chen, H., Grant-Muller, S.: Use of Sequential for Short-term Traffic Flow Forecasting. *Transportation Research Part C* 11, 319–336 (2001)
- [5] Ou, X., Qiu, G., Zhang, Y., Li, Z.: Analysis of Similarity for Urban Traffic Volumes. *Central South Expressway Engineering* 28(2), 4–7 (2003)
- [6] Zheng, W., Lee, D.-H., Asce, M., Shi, Q.: Short-term Freeway Traffic Flow Prediction: Bayesian Combined Neural Network Approach. *Journal of Transportation Engineering* 132, 114–121 (2006)



# An Eigenvector-Based Kernel Clustering Approach to Detecting Communities in Complex Networks

Lidong Fu and Lin Gao

**Abstract.** To detect communities in complex networks, we generalize the modularity density( $D$ ) to weighted variants and show how optimizing the weighted function( $WD$ ) can be formulated as a spectral clustering problem, as well as a weighted kernel  $k$ -means clustering problem. We also prove equivalence of the both clustering approaches based on  $WD$  in mathematics. Using the equivalence, we propose a new eigenvector-based kernel clustering algorithms to detecting communities in complex networks, called two-layer approach. Experimental results indicate that it have better performance comparing with either direct kernel  $k$ -means algorithm or direct spectral clustering algorithm in term of quality.

## 1 Introduction

The study of complex network has become a fast growing subject in many disciplines, including social science, biology, and technological networks. One of the key properties in complex networks is the presence of communities. Communities are groups of nodes with a higher level of interconnection among themselves than with the rest of the graph[1]. These community structures may hinder important information on the functioning of the system, and can be relevant to understand the growth mechanisms of such networks. As a result, the problem of detecting and characterizing the community structure in a network from various fields has become one of the outstanding issues [2,3].

Recently, a new approach was developed by Li et al. [4] to overcome limitations of previous measures for measuring community structure. They proposed a

---

Lidong Fu

School of Computer, Xi'an University of Science and Technology, Xi'an, P.R. China 710054

e-mail: [fulidong2005@163.com](mailto:fulidong2005@163.com)

Lidong Fu · Lin Gao

School of Computer Science and Technology, Xidian University, Xi'an, P.R. China 710071

e-mail: [lgao@mail.xidian.edu.cn](mailto:lgao@mail.xidian.edu.cn)

quantitative measure, called "modularity density" (known as  $D$ ), to assess the quality of community structures, and formulated community discovery as an optimization problem. Since optimizing  $D$  is a NP-hard problem, kernel  $k$ -means has been employed to optimize the objective function. However, the kernel  $k$ -means algorithm is prone to problems of poor local minima and sensitive to initial starting partitions, which can be seen as a drawback.

In this paper, we first generalize modularity density function to a weighted form ( $WD$ ). Then we show how optimizing the  $WD$  function can be formulated as a spectral clustering problem, as well as a weighted kernel  $k$ -means clustering problem. We also prove equivalence of the both clustering approaches based on  $WD$  in mathematics. By the means of the equivalence, we use characteristics of spectral clustering, which can globally optimize  $WD$  in eigenvector space and outperform other traditional clustering algorithms, to initialize the weighted kernel  $k$ -means algorithm. The eigenvector-based kernel clustering approach, called two-layer algorithm.

## 2 Spectral Optimization of Weighted Modularity Density

We can generalize the modularity density to weighted variants. This will prove useful for building a general connection to weighted kernel  $k$ -means. We introduce a weight  $w_i$  for each vertex of the network, and for each community  $V_c$ , we define  $w(V_c) = \sum_{i \in V_c} w_i$ . We generalize the modularity density in Ref[4] to be

$$WD(\{V_c\}_{c=1}^p) = \sum_{c=1}^p \frac{\text{links}(V_c, V_c) - \text{links}(V_c, \bar{V}_c)}{w(V_c)} \quad (1)$$

Clearly, if all weights are equal to one, weighted modularity density is standard modularity density. Therefore, modularity density is a special case of weighted modularity density.

Let us introduce an indicator vector  $t_c(i) = 1$ , if community  $V_c$  contains vertex  $i$ , and  $t_c(i) = 0$  otherwise, for  $1 \leq c \leq p$ . Define a diagonal degree matrix  $B$  with  $B_{ii} = \sum_{j=1}^n A_{ij}$ . Then, the weighted modularity density objective may be written as

$$WD(\{V_c\}_{c=1}^p) = \sum_{c=1}^p \frac{t_c^T A t_c - t_c^T (B - A) t_c}{t_c^T W t_c} \quad (2)$$

where  $W$  is weighted matrix of nodes. The equalities hold since  $t_c^T W t_c$  gives us the size of weighted community  $V_c$ , whereas  $t_c^T A t_c$  equals the sum of the links' weight inside community  $V_c$ ,  $t_c^T (B - A) t_c$  equals the sum of links' weight of community connecting other communities. Furthermore, let us define  $z_c = t_c / (t_c^T W t_c)^{1/2}$ , then weighted  $D$  can be expressed as:

$$WD(\{V_c\}_{c=1}^p) = \sum_{c=1}^p z_c^T (2A - B) z_c \quad (3)$$

Let us define a corresponding  $n \times p$  vertices' assignment matrix  $X$  and  $X = W^{1/2}Z$ , where the  $c$ th column of  $Z$  equals  $z_c$ . Then we can further reduce  $WD$  as follows:

$$WD(\{V_c\}_{c=1}^p) = \text{trace}(X^T W^{-1/2}(2A - B)W^{-1/2}X) \quad (4)$$

where  $\text{trace}$  denotes the trace of a matrix. Clearly,  $X$  is an orthonormal matrix i.e.  $X^T X$  equals  $I_p$ . Thus the problem of maximizing  $WD$  can then be expressed as:

$$\begin{aligned} \max_X \{ & \text{trace}(X^T W^{-1/2}(2A - B)W^{-1/2}X) \} \\ & \text{s.t. } X^T X = I_p \end{aligned} \quad (5)$$

We know to maximize vertices' assignment matrix  $X$  in Eq.(5) is NP-complete. To address this we can attempt to derive a approximation by relaxing the discreteness constraints that the  $X_{ic} \in \{0, \frac{w_i^{1/2}}{(\sum_{i \in V_c} w_i)^{1/2}}\}$ , so that instead the  $X_{ic} \in R^1$ . By a standard result in linear algebra [6], we may relax the trace maximization in Eq.(5) such that  $X$  is an arbitrary orthonormal matrix.

### 3 Equivalence of Objectives

It has been shown that the weighted kernel  $k$ -means objective function can also be written as a trace optimization problem [5]. The minimization of the weighted  $k$ -means objective function is expressed as kernel matrix trace maximization:

$$\max_Y \{ \text{trace}(Y^T W^{1/2} K W^{1/2} Y) \} \quad (6)$$

Where  $Y$  is an orthonormal matrix ( $Y^T Y = I_p$ ) with  $Y_{ic} \in \{0, \frac{w_i^{1/2}}{(\sum_{i \in V_c} w_i)^{1/2}}\}$  if node  $i$  belongs to community  $V_c$ , 0 otherwise. As above, Eq.(6) can be achieved by taking the top  $p$  eigenvectors of  $W^{1/2} K W^{1/2}$ . It is easy to see that the matrix  $Y$  in this section is identical to  $X$  in section 2. Setting the kernel matrix  $K$  to  $W^{-1}(2A - B)W^{-1}$ , the trace maximization for weighted kernel  $k$ -means in Eq.(6) is seen to equal  $\text{trace}(X^T W^{-1/2}(2A - B)W^{-1/2}X^{-1/2})$ , which is exactly the trace maximization for the weighted modularity density in Eq.(5).

A requirement for the kernel matrix  $K = W^{-1}(2A - B)W^{-1}$  is positive semidefinite, so that weighted kernel  $k$ -means can be converged. We may resolve this problem by performing diagonal shifting for kernelizing  $K$ . This is done by adding  $\sigma W^{-1}$  to  $K$ . We get

$$\text{trace}(X^T W^{1/2}(\sigma W^{-1} + K)W^{1/2}X) = \sigma p + \text{trace}(X^T W^{-1/2}(2A - B)W^{-1/2}X) \quad (7)$$

Note that  $\sigma p$  is only a constant in Eq.(7). So the optimal solution of the objective function is the same for both the shifted and unshifted matrices. Thus, shifting the diagonal allows us to construct a positive semi-definite kernel matrix  $K$ , which guarantees monotonic convergence of the kernel k-means algorithm to a local optimum. Since adding  $\sigma W^{-1}$  to  $K$  increases the eigenvalues of kernel matrix  $K$  by  $\sigma$ .

## 4 Two-Layer Approach

Spectral methods typically perform well because they compute the globally optimal solution of a relaxation to the original clustering objective. In contrast, weighted kernel  $k$ -means is fast but is prone to problems of poor local minima and sensitivity to initial starting partitions. Our equivalence between spectral methods and kernel  $k$ -means based on weighted modularity density allows us to solve a relaxed problem using the eigenvectors of the matrix.

Thus, we may first compute the top  $p$  eigenvectors of the matrix  $W^{1/2}KW^{1/2}$  i.e.,  $W^{-1/2}(2A - B)W^{-1/2}$ . This maps the original nodes of networks to a lower dimensional space. Once post-processing is performed and a discrete clustering solution has been attained, we can treat the resulting partitioning as a good initialization to weighted kernel  $k$ -means on the full vertices of the network. We first run spectral clustering to get an initial partitioning and then refine the partitioning by running weighted kernel  $k$ -means on the partitioning. We call this the two-layer approach. The algorithm can be specifically described as follows:

**Input:**  $A$  : input adjacency matrix,  $p$  : number of clusters,  $w$  : weights for each point,  $t_{max}$  : optional maximum number of iterations

**Output:**  $\{V_c\}_{c=1}^p$  : final partitioning of the vertices

1. Generate a vector  $w$  of vertex weights to construct matrix  $W$
2. Compute the transition matrix  $M = W^{-1/2}(2A - B)W^{-1/2}$
3. Compute the first  $p$  eigenvectors  $u_1, u_2, \dots, u_p$  of  $M$ .
4. Let  $U \in R^{n \times p}$  be the matrix containing the vectors  $u_1, u_2, \dots, u_p$  as columns.
5. For  $i = 1, 2, \dots, n$ , let  $f_i \in R^p$  be the vector corresponding to the  $i$ -th row of  $U$ .
6. Cluster the vertices  $(f_i)_{i=1, \dots, n}$  in  $R^p$  with the k-means algorithm into initial clusters  $\{V_c^{(0)}\}_{c=1}^p$ .
7. Compute kernel matrix  $K = \sigma W^{-1} + W^{-1}(2A - B)W^{-1}$ .
8. Return  $\{V_c\}_{c=1}^p = \text{Weighted - Kernel - Kmeans}(K, p, t_{max}, w, \{V_c^{(0)}\}_{c=1}^p)$  as shown in Ref. [5].

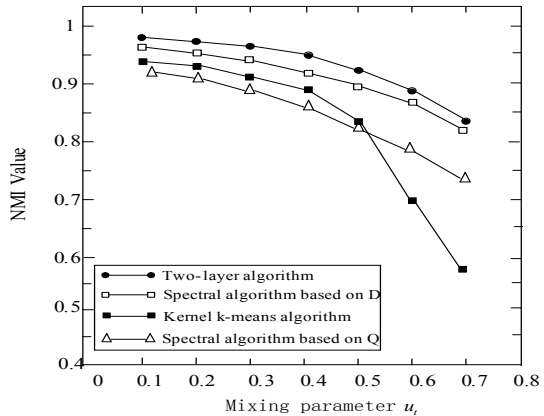
## 5 Experimental Results

### 5.1 Testing the Methods on LFR Benchmark

We have evaluated these algorithms including spectral method based on  $Q$  [7] on the new benchmark proposed by Lancichinetti et al (LFR)[8]. We generated 100

instances for each of LFR benchmark graphs. Using the algorithm presented in Ref [8], we generate each graph whose node degree was taken from a power law distribution with exponent 2 and community size from a power law distribution with exponent 2. Each graph has 3000 vertices, average degree  $\langle k \rangle = 15$ , maximum degree 50, maximum for the community sizes 50 and minimum for the community sizes 5. To evaluate clusters, we use Normalized Mutual Information (NMI) index as a measure of similarity between two partitions  $A$  and  $B$ [9].

**Fig. 1** Test of various algorithm on the LFR heterogeneous benchmarks. The NMI values are displayed as a function of the mixing parameter,  $u_t$ . Each point corresponds to an average over 100 graph realizations.



From the Figure 1, we can see the two-layer algorithm can achieve its peak above 0.98, spectral algorithm based on  $D$  roughly 0.96 and kernel k-means roughly 0.93 at  $u_t = 0.1$  respectively. In particular, spectral algorithm based on  $Q$  begins to fail even when communities are only loosely connected to each other. This is due to the fact that modularity optimization has an intrinsic resolution limit that makes small communities hard to detect. When  $u_t > 0.5$ , which means communities are no longer defined in the strong sense, i.e., such that each node has more neighbors in its own community than in the others, kernel k-means algorithm has worse performance than spectral algorithm based on  $Q$ . The reason is that kernel k-means algorithm is prone to problems of poor local minima and has more difficult random choice of centre of community structure in a network in each iteration step with community structure becoming more ambiguity. Conversely, when  $u_t < 0.5$ , bisection spectral algorithm based on  $Q$  has worse performance than kernel k-means algorithm. The reason is maximizing modularity by recursively bisecting the network unduly constrains their results, leading to a bias in the size of the communities they find and limiting their effectiveness [10]. We can shown the two-layer algorithm has better performance than others methods.

## 6 Conclusions

In this paper, we reverse engineer  $D$  function into a spectral framework and weighted kernel  $k$ -means objective function for finding community structures respectively. Following, a new algorithm two-layer method was developed to optimize the objective function  $D$ . The algorithm allows us to use the output of spectral clustering to initialize weighted kernel  $k$ -means, which can often produce better community results than weighted kernel  $k$ -means of using pure random initialization and spectral method.

**Acknowledgment.** This work was supported by the Key National NSF (Grant No.60933009), the Programs Foundation of Xi'an University of Science and Technology of China (Grant No.2010029).

## References

1. Fortunato, S.: Community detection in graphs. *Physics Reports* 486, 75–174 (2010)
2. Newman, M.E.J.: Detecting community structure in networks. *Eur. Phys. J. B* 38(2), 321–330 (2004)
3. Kashtan, N., Itzkovitz, S., Milo, R., et al.: Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* 20, 1746–1758 (2004)
4. Li, Z.P., Zhang, S.H., Wang, R.S., et al.: Quantitative function for community detection. *Phys. Rev. E* 77(3), 036109 (2008)
5. Kulis, B., Basu, S., Dhillon, I., et al.: Semi-supervised graph clustering: a kernel approach. *Math. Learn.* 74, 1–22 (2009)
6. Golub, G., Van, L.C.: *Matrix computations*. Johns Hopkins Univ. (1989)
7. Newman, M.E.J.: Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* 103(23), 8577–8582 (2006)
8. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithm. *Phys. Rev. E* 78, 046110 (2008)
9. Danon, L., Diaz-Guilera, A., Duch, J., et al.: Comparing community structure identification. *J. Stat. Mech. Theory Exp.*, PO9008 (2005)
10. Sun, Y., Danila, B., Josic, K., Bassler, K.E.: Improved community structure detection using a modified fine-tuning strategy. *Europhysics Letters* 86(2), 28004 (2009)

# Distributed Gaussian Mixture Learning Based on Variational Approximations

Behrouz Safarinejadian

**Abstract.** In this paper, the problem of density estimation and clustering in sensor networks is considered. It is assumed that measurements of the sensors can be statistically modeled by a common Gaussian mixture model. We develop a distributed variational Bayesian algorithm (DVBA) to estimate the parameters of this model. This algorithm produces an estimate of the density of the sensor data without requiring the data to be transmitted to and processed at a central location. Alternatively, DVBA can be viewed as a distributed processing approach for clustering the sensor data into components corresponding to predominant environmental features sensed by the network. To verify performance of DVBA, we perform several simulations of sensor networks.

## 1 Introduction

Distributed Data Mining (DDM) has recently emerged as an extremely important area of research. The objective, here, is to perform data mining tasks (such as association rule mining, clustering, classification) on a distributed database, that is, a database distributed across several sites (nodes) connected by a network. For example, an algorithm for distributed association rule mining in peer-to-peer systems has been presented in [1]. K-means clustering has been extended to the distributed scenario in [2]. Techniques for performing non-parametric density estimation over homogeneously distributed data have been studied and experimentally evaluated in [3].

A distributed EM (Expectation Maximization) algorithm has been developed in [4], [5] for density estimation in sensor networks assuming that the measurements are statistically modeled by a mixture of Gaussians. A distributed EM algorithm has also been developed in [6] for density estimation in peer-to-peer networks. An important problem of the EM algorithm is that singularity may happen in the estimated parameters. Especially, if the model order is not properly selected and the assumed order is greater than the real order of the observed data, singularity will be inevitable.

Recently, variational methods have gained popularity in the machine learning literature and have also been used to estimate the parameters of finite mixture models. The variational Bayesian method aims to construct a tight lower bound on the data marginal likelihood and then seeks to optimize this bound using an iterative scheme [7]. An important advantage of the variational approach is that, despite of the EM algorithm, we do not have the singularity problem.

In this paper, it is assumed that measurements of the sensors are statistically modeled by a common mixture of Gaussians. A distributed variational Bayesian algorithm (DVBA) is then developed for estimating the Gaussian components which are common to the sensor network. DVBA can also be used as a general distributed data mining algorithm for density estimation and clustering of the data distributed over the nodes of a network.

The rest of the paper is organized as follows. In Section 2, the basic problem statement is given, and observations and data models are defined. Section 3 develops a distributed variational Bayesian algorithm to estimate the parameters of the mixture model. The results of simulations are presented in Section 4. Finally, section 5 concludes the paper.

## 2 Problem Statement

Assume that distribution of measurements is represented by a mixture of Gaussian components:

$$f(\mathbf{y}_m; \boldsymbol{\pi}_m, \boldsymbol{\varphi}) = \sum_{j=1}^J \pi_{m,j} \mathcal{N}(\mathbf{y}_m; \boldsymbol{\mu}_j, \mathbf{T}_j) \quad (1)$$

where  $J \in \mathbb{Z}_{\geq 1}$  is the number of mixture components,  $j \in \mathbb{Z}_{\geq 1}$  represents the component index,  $\{\pi_{m,j}\}_{j=1}^J$  are the mixture probabilities at node  $m$ ,  $\boldsymbol{\mu}_j$  is the mean and  $\mathbf{T}_j$  the precision (inverse covariance) matrix.  $\boldsymbol{\varphi}_j = \{\boldsymbol{\mu}_j, \mathbf{T}_j\}$  is the set of parameters defining the  $j$ th component.

## 3 A Distributed Variational Bayesian Algorithm

Here, we assume that the observed data is distributed over nodes of a network. Suppose that distribution of the observations is represented by the finite mixture of components shown in (1). Here, conjugate priors are assigned to the parameters  $\boldsymbol{\pi}$  and  $\boldsymbol{\varphi}$ .

$$f_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}; \alpha_1^0, \dots, \alpha_J^0)$$

$$f_{\boldsymbol{\mu}, \mathbf{T}}(\boldsymbol{\mu} \setminus \mathbf{T}) = \prod_{j=1}^J \mathcal{N}(\boldsymbol{\mu}_j; m_j^0, (\beta_j^0 \mathbf{T}_j)^{-1})$$



$$f_T(\mathbf{T}) = \prod_{j=1}^J \mathcal{W}(T_j; \mathbf{v}_j^0, \Sigma_j^0)$$

We define a vector of sufficient statistics as:

$$\mathbf{s}^t = \{q_j^t, \mathbf{a}_j^t, \mathbf{b}_j^t\}$$

where

$$\begin{aligned} q_j^t &\triangleq \sum_{m=1}^M \sum_{i=1}^{N_m} q_{m,i,j}^t \\ \mathbf{a}_j^t &\triangleq \sum_{m=1}^M \sum_{i=1}^{N_m} q_{m,i,j}^t \mathbf{y}_{m,i} \\ \mathbf{b}_j^t &\triangleq \sum_{m=1}^M \sum_{i=1}^{N_m} q_{m,i,j}^t \mathbf{y}_{m,i} \mathbf{y}_{m,i}^T \end{aligned}$$

Assume that  $q_{m,i,j}^t$  is the local quantity computed at node  $m$  and iteration  $t$  using

$$q_{m,i,j}^t = \frac{\phi_{m,i,j}^t}{\sum_{j=1}^J \phi_{m,i,j}^t} \quad (2)$$

$$\phi_{m,i,j}^t = \tilde{\pi}_j^{t-1} (\tilde{T}_j^{t-1})^{1/2} e^{-\frac{1}{2}(\mathbf{y}_{m,i} - \mathbf{m}_j^{t-1})^T E[\mathbf{T}_j^{t-1}](\mathbf{y}_{m,i} - \mathbf{m}_j^{t-1}) - \frac{d}{2\beta_j^{t-1}}}$$

The proposed DVBA performs as follows. Initialize  $\{\alpha_{m,j}^0\}$ ,  $\{m_{m,j}^0\}$ ,  $\{\beta_{m,j}^0\}$ ,  $\{v_{m,j}^0\}$  and  $\{\Sigma_{m,j}^0\}$  at some chosen value. Assume that the algorithm proceeds in a cyclic manner. The following processing and communication are carried out at each node in sequence. At iteration  $t+1$  node  $m$  receives  $q_j^t$ ,  $\mathbf{a}_j^t$  and  $\mathbf{b}_j^t$  from the immediate previous node. The node then computes the hyperparameters using:

$$\alpha_j^t = \alpha_j^0 + q_j^t \quad (3)$$

$$\beta_j^t = \beta_j^0 + q_j^t \quad (4)$$

$$\mathbf{m}_j^t = \frac{\beta_j^0 \mathbf{m}_j^0 + \mathbf{a}_j^t}{\beta_j^t} \quad (5)$$

$$\Sigma_j^t = \Sigma_j^0 + \mathbf{b}_j^t + \beta_j^0 \mathbf{m}_j^0 \mathbf{m}_j^{0T} - \beta_j^t \mathbf{m}_j^t \mathbf{m}_j^{tT} \quad (6)$$

$$\mathbf{v}_j^t = \mathbf{v}_j^0 + \mathbf{q}_j^t \quad (7)$$

The mean value of the means  $\mu_j$  and the precisions  $T_j$  at this node are obtained through

$$E[\boldsymbol{\mu}_j^t] = \mathbf{m}_j^t$$

$$E[\mathbf{T}_j^t] = \mathbf{v}_j^t (\boldsymbol{\Sigma}_j^t)^{-1}$$

Then  $q_{m,i,j}^{t+1}$  is computed using (2) and the local values of the sufficient statistics vector is then calculated as

$$q_{m,j}^{t+1} = \sum_{i=1}^{N_m} q_{m,i,j}^{t+1}$$

$$\mathbf{a}_{m,j}^{t+1} = \sum_{i=1}^{N_m} q_{m,i,j}^{t+1} \mathbf{y}_{m,i}$$

$$\mathbf{b}_{m,j}^{t+1} = \sum_{i=1}^{N_m} q_{m,i,j}^{t+1} \mathbf{y}_{m,i} \mathbf{y}_{m,i}^T$$

Finally, node  $m$  updates its mixing probabilities and sufficient statistic vectors  $q_j^t$ ,  $\mathbf{a}_j^t$  and  $\mathbf{b}_j^t$  according to

$$\pi_{m,j}^{t+1} = \frac{q_{m,j}^t}{N_m} \quad (8)$$

$$q_j^{t+1} = q_j^t + q_{m,j}^{t+1} - q_{m,j}^t$$

$$\mathbf{a}_j^{t+1} = \mathbf{a}_j^t + \mathbf{a}_{m,j}^{t+1} - \mathbf{a}_{m,j}^t$$

$$\mathbf{b}_j^{t+1} = \mathbf{b}_j^t + \mathbf{b}_{m,j}^{t+1} - \mathbf{b}_{m,j}^t$$

The updated values  $\{q_j^{t+1}, \mathbf{a}_j^{t+1}, \mathbf{b}_j^{t+1}\}$  are then transmitted to the next node and the above process is repeated.

## 4 Simulation Results

Here, we use a network of 50 nodes, each containing 100 data points, to evaluate performance of DVBA in environmental modeling. It is assumed that each node in the network senses an environment that can be described as a mixture of some elementary conditions. The measurements are thus statistically modeled with a mixture of Gaussians; each Gaussian component corresponding to one of the elementary conditions. Second elements of the estimated mean vectors in 50 nodes are shown in Fig. 1. It can be seen that the estimated mean values in all nodes are

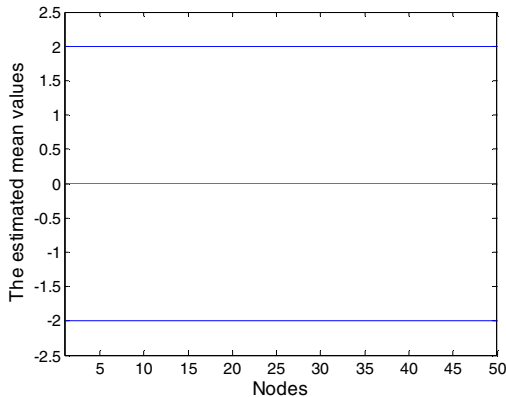
very close to their true values (the true values are 2, 0 and -2). The true and estimated parameters of the components using DVBA are shown in tables 1 and 2, respectively. As seen, good estimates of the true values have been obtained. The values offered in these tables are the mean value of the estimated parameters at all nodes of the network. Standard deviations of the estimated parameters at various nodes are in order of  $10^{-5}$ . The small values of the standard deviations imply that the estimated values obtained at each node are almost the same.

**Table 1** True mean and covariance matrices for the 2D data set.

Component	Mean vector	Covariance matrix
1	$[0, -2]$	$\begin{bmatrix} 2 & 0 \\ 0 & 0.2 \end{bmatrix}$
2	$[0, 0]$	$\begin{bmatrix} 2 & 0 \\ 0 & 0.2 \end{bmatrix}$
3	$[0, 2]$	$\begin{bmatrix} 2 & 0 \\ 0 & 0.2 \end{bmatrix}$

**Table 2** Fitted mean and covariance matrices using the DVBA.

Component	Mean vector	Covariance matrix
1	$[0.0004, -1.9890]$	$\begin{bmatrix} 2.0912 & 0.0112 \\ 0.0112 & 0.2077 \end{bmatrix}$
2	$[0.0297, -0.0169]$	$\begin{bmatrix} 2.0582 & -0.0292 \\ -0.0292 & 0.1836 \end{bmatrix}$
3	$[0.0553, 1.9944]$	$\begin{bmatrix} 1.9480 & -0.0201 \\ -0.0201 & 0.1970 \end{bmatrix}$



**Fig. 1** Three estimated mean values using DVBA in the network with 50 nodes.

## References

1. Wolff, R., Schuster, A.: Association rule mining in peer-to-peer systems. *IEEE Transactions on Systems, Man and Cybernetics, Part B* (34), 2426–2438 (2004)
2. Dutta, S., Gianella, C., Kargupta, H.: K-means clustering over peer-to-peer networks. In: 8th International Workshop on High Performance and Distributed Mining, SIAM International Conference on Data Mining (2005)
3. Giannella, C., Dutta, H., Mukherjee, S., Kargupta, H.: Efficient kernel density estimation over distributed data. In: 9th International Workshop on High Performance and Distributed Mining, SIAM International Conference on Data Mining (2006)
4. Safarinejadian, B., Menhaj, M.B., Karrari, M.: Distributed data clustering using expectation maximization algorithm. *Journal of Applied Sciences* 9, 854–864 (2009)
5. Safarinejadian, B., Menhaj, M.B., Karrari, M.: A distributed EM algorithm to estimate the parameters of a finite mixture of components. *Knowledge and Information Systems* 23, 267–292 (2010)
6. Safarinejadian, B., Menhaj, M.B., Karrari, M.: Distributed Unsupervised Gaussian Mixture Learning for Density Estimation in Sensor Networks. *IEEE Transactions on Instrumentation and Measurement* 59, 2250–2260 (2010)
7. Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes. In: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (1999)

# The Function and Relationship of Verbal and Nonverbal Cues in IM Interpersonal Communication<sup>\*</sup>

Lu Xiaoyan and Yao Jinyun

**Abstract.** Chatting by instant messaging (IM) is a very popular way of computer-mediated communication. QQ is the most popular instant messaging in China. This research attempts to examine the function and relationship of verbal cues (topics) and nonverbal cues (emoticons) in IM interpersonal communication by using quantitative research method. Results show that one kind of emoticon using and three kinds of topics discussed can predict the behavior of QQ interpersonal communication. We term this kind of emoticon ‘expect-responsive’ emoticon.

## 1 Introduction

Instant messaging (IM) has become one of the most popular online applications. Tencent QQ is the most popular free instant messaging in Mainland China. Its concurrent users reached 100 million in March 2010, a majority of whom are college and university students. Previous studies assumed that computer-mediated communication (CMC) lack nonverbal communication cues. However, Shao-Kang Lo’s research (2008) found that emoticons as a communication tool performed nonverbal communication functions. Emoticons allow IM users to correctly understand the level and direction of emotion, attitude, and attention expression. He termed emoticons as ‘quasi-nonverbal’ cues. This research attempts to find the relationship between nonverbal (emoticons) and verbal (topics) cues in QQ interpersonal communication, and examine the functions of these cues.

---

Lu Xiaoyan

College of Media and International Culture, Zhejiang University, Hangzhou, China

Yao Jinyun

Zhejiang Vocational College of Commerce, Hangzhou, China

<sup>\*</sup> This paper was supported by the Fundamental Research Funds for the Central Universities and the Fundamental Research Funds for the Central Universities.

## 2 Method

A sample of 40 students from four colleges was gathered. Each participant was requested to report ten default emoticons and ten topics they frequently used in QQ interpersonal communication. Based on the frequency of usage of emoticons and topics, we collected 30 default emoticons and 22 topics used most often on QQ. Then, we conducted a web-based survey as a pilot test to examine students' habits of emoticons usages and topics discussed in QQ. The Participants were 45 students from one college. Finally, a 74-item questionnaire with a five Likert response scale was designed. A total of 184 students from three colleges were sampled for the purpose of this study. Cronbach's  $\alpha$  of the whole questionnaire is 0.922.

This questionnaire was designed to measure the following: frequency of emoticon usage, topic discussed and behavior of interpersonal communication. Survey items were rated on five-point Likert scales ranging from (1) seldom use to (5) often use. Furthermore, the stem about the behavior of interpersonal communication is expressed as, 'I use QQ to' and it was followed by the items: 'chat with relatives, classmates and friends ,' 'chat in QQ groups ,' 'email,' 'update blog and read other friends' blogs.'
























## 3 Results

### 3.1 Usage of Emoticons

Analysis included principal axis factoring analysis with varimax rotation to extract possible factors of emoticon usage. Analysis of the correlation matrix, KMO (0.880), and Bartlett's test of sphericity ( $P < .000$ ) for the 30 items related to emoticons usage suggested that the correlation matrix was factorable. Items with eigenvalues of 1.0 or higher and item loadings of 0.40 were retained. The factor analysis of the emoticons usage in QQ yielded five interpretable factors. They can explain 55.26% of the total item variance. Responses to the retained items which value of factor loading of more than 0.50 were summed and averaged to form the scales representing each factor.

- The first factor, accounted for 18.58% of the variance after rotation ( $M=2.83$ ,  $SD=1.00$ ), and its ten containing emoticons. Seven items' loadings are more than 0.5, and they can express the emotion of grief, sadness, intimacy and torment. Most emotions are used to express strong feelings.
- The second factor containing four emoticons, accounted for 11.05% of the variance after rotation ( $M=3.59$ ,  $SD=1.18$ ). Three items' loadings are more than 0.5, and they can express the feeling of 'my god' and happy. These emoticons are used most often in the five factors.
- The third factor containing four emoticons, accounted for 9.46% of the variance after rotation ( $M=2.30$ ,  $SD=0.98$ ). Three items' loadings are more than 0.5, and they mean astonished and 'I want to strive.'
- The fourth factor containing two emoticons, accounted for 8.30% of the variance after rotation ( $M=2.05$ ,  $SD=1.14$ ). They can express feeling of anger and they are used least often in the five factors.

**Table 1** Results of factor analysis in emoticons usage

Factor	Emoticon	Factor loading
<b>F1: grief, sadness, intimacy and torment</b>		.733
		.649
		.647
		.641
		.589
		.531
		.506
		.497
		.433
<b>F2: my god and happy</b>		.423
		.826
		.572
		.532
		.483
<b>F3: astonished and strive</b>		.702
		.619
		.502
		.461
<b>F4: anger</b>		.928
		.879
<b>F5: play a joke</b>		.712
		.567
		.436

Note: item loadings of 0.50 were retained.

- The last factors containing three emoticons, accounted for 7.88% of the variance after rotation ( $M=2.78$ ,  $SD=1.26$ ). Two items' loadings are more than 0.5, and they mean 'playing a joke.'
- All five factors are correlated significantly with each other ( $p < 0.01$ ). The strongest correlations were between Factor 1 (grief, sadness, intimacy and torment) and Factor 5 (play a joke) ( $r = .582$ ). And the weakest correlations were between Factor 2 (my god and happy) and Factor 4 (angry) ( $r = .254$ ).

### 3.2 Topics Discussed

We use principal axis factoring analysis with varimax rotation to extract possible factors of topics discussed. Analysis of the correlation matrix, KMO (0.861), and Bartlett's test of sphericity ( $P < .000$ ) for the 22 items related to topics discussed in

QQ suggested that the correlation matrix was factorable. Items with eigenvalues of 1.0 or higher and item loadings of 0.40 were retained. The factor analysis of topic talking in QQ yielded six interpretable factors. They can explain 50.88% of the total item variance. Responses to the retained items with value factor loading of more than 0.50 were summed and averaged to form the scales representing each factor.

- The first factor, accounted for 14.71% of the variance after rotation ( $M=3.19$ ,  $SD=0.90$ ), and it contained six topics. They are ‘own recent life’, ‘recalling past events’, ‘feeling, hobbies’, ‘student event’, ‘own event online’. So we name this factor ‘private businesses’.
- The second factor containing three topics, accounted for 10.69% of the variance after rotation ( $M=2.33$ ,  $SD=0.87$ ). They are ‘physical education and sports’, ‘digital products’, ‘future and dream’. We name this factor ‘active living and dream’.
- The third factor containing two topics, accounted for 7.99% of the variance after rotation ( $M=1.96$ ,  $SD=0.84$ ). They are ‘stars and gossip’, ‘web celebrity’. We name this factor ‘Stars and celebrities’. These topics are talked about least often in the 6 factors.

**Table 2** Factor analysis of topics discussed

Factor	Topic	
F1: private businesses	own recent life	.712
	recalling past events	.690
	feeling	.678
	hobbies	.620
	student event	.598
	own event online	.598
F 2:active living and dream	physical education and sports	.636
	digital products	.530
	future and dream	.526
F 3 :stars and celebrities	stars and gossip	.731
	web celebrity	.602
F 4 :leisure Life	amusement and travel	.709
	affection and love	.561
	vogue and shopping	.507
	movies and music	.468
F 5: food and drink	food and drink	.619
	employment	.484
	Facial and slim	.480
F 6: information from friends and society	recent life of friends andclassmates	.543
	current news	.502

Notes: item loadings of 0.50 were retained.



- The fourth factor containing four topics, accounted for 7.43% of the variance after rotation ( $M=2.76$ ,  $SD=0.88$ ). They are ‘amusement and travel’, ‘affection and love’, ‘vogue and shopping’, ‘movies and music’. We name this factor ‘Leisure Life’.
- The fifth factor containing three topics, accounted for 6.30% of the variance after rotation ( $M=2.33$ ,  $SD=1.10$ ). Only one item loading is more than 0.5, so we name this factor ‘food and drink’.
- The last factors containing two topics, accounted for 3.76% of the variance after rotation ( $M=3.23$ ,  $SD=0.89$ ). They are ‘recent life of friends and classmates’, ‘current news’. We name this factor ‘information from friends and society’. These topics are discussed most frequently of the six factors
- All six factors are all correlated significantly with each other ( $p < 0.01$ ). The strongest correlations were between topic two and topic five ( $r = .508$ ). And the weakest correlations were between topic three and topic six ( $r = .204$ ).

### 3.3 Predictors of QQ Interpersonal Communication

The variables of five factors in emoticons usage and 6 factors in topics discussed were all entered on the first step of the regression analysis in order to predict frequency of QQ interpersonal communication (QQ IC). Factor one of emoticons usage is the only significant predictor, and three factors of topics discussed are also significant predictors. The regression equation is significant ( $P < 0.001$ ).

**Table 3** Regression analysis of emoticons and topics

Dependent Variable /independent variable	$\beta$	R(%)	F
frequency of QQ IC		26.8	17.68
Factor 1 of emoticons	.235***		
Factor 1 of topics	.270***		
Factor 6 of topics	.161*		
Factor 5 of topics	-.111*		

Notes : \* :  $P < 0.05$  ; \*\*\* :  $P < 0.001$

### 3.4 Demographic Differences

There is no difference in age variable in college students. However, we found several differences in gender variable.

- Girls ( $M=3.03$ ,  $SD = 1.06$ ) use emoticons of sadness, grief and intimacy more often than boys ( $M=2.54$ ,  $SD = 0.82$ ) ( $P < 0.001$ ).
- Girls ( $M=3.35$ ,  $SD = 0.93$ ) talk more about topics of private business than boys ( $M=2.96$ ,  $SD = 0.80$ ) ( $P < 0.01$ ).
- Girls ( $M=2.08$ ,  $SD = 0.87$ ) talk about more topics of stars and celebrities than boys ( $M=1.79$ ,  $SD = 0.77$ ) ( $P < 0.05$ ).

## 4 Conclusion and Discussion

Results show that the interaction of F1 of emoticons usage (grief, sadness, intimacy and torment) with another factor of topics discussed (private businesses) is significantly correlated with the behavior of QQ interpersonal communication. Moreover, one factor of topics discussed (information from friends and society) can also predict the behavior of QQ interpersonal communication, and it has no interaction with factors of emoticons usage. In fact, the F1 of emoticons are different from the other four factors of emoticons. The F1 of emoticons can express strong feelings, while others express the general feelings. Accordingly we can divide all the emoticons into two kinds: the F1 of emoticons can be termed 'expect-responsive' emoticons, and others can be termed responsive emoticons. People's 'expect-responsive' emoticons usage may mean expecting receiving responsive information or getting help, and the responsive emoticon usage may mean responding information to others after receiving information. When people use 'expect-responsive' emoticons in QQ chatting, most of them are talking about topics of private things, such as their own recent lives, recalling their past events, their feelings and hobbies, their own event online and so on. We can also assume that instant messaging as a computer-mediated communication tool has the function of transforming feelings of people, and the dynamic emoticons play an important role in IM interpersonal communication.

## References

1. DeVito, J.A.: *The Interpersonal Communication Book*. Taipei Yang-Chih Book Co., Ltd. (2000)
2. Carter, K.A.: Type Me How You Feel: Quasi-Nonverbal Cues in Computer-mediated Communication. *ETC*, 29–39 (spring 2003)
3. Lo, S.-K.: The Nonverbal Communication Functions of Emoticons in Computer-Mediated Communication. *Cyber psychology & Behavior* 11(5), 595–597 (2008)

# Using Version Control System to Construct Ownership Architecture Documentations

Po-Han Huang, Dowming Yeh, and Wen-Tin Lee

**Abstract.** Ownership architecture was usually constructed by investigating the comments at the top of source files. That is, to associate developer names with source files is to examine the comments manually. If such documentation can be produced automatically, it will be more immediate to indicate the status of the project. This research focus on the logs in the version control system. The data within version control logs is in a regular form and information can be retrieved quickly. The importance of developers can also be estimated by the number of own files and frequency of making a change. In order to understand the system architecture, the directory structure of source code can be used to identify function components of the system essentially. The source files in a directory implement the same function component, and the owners of these source files can be considered a team. Using the documents, researcher can know the ownership architecture and more information about the status of the project.

## 1 Introduction

Instead of buying commercial software, more and more people try to use open source software to meet their needs in recent years. Open source software is characterized by its open source code and free distribution. Many volunteers participate in open source projects to contribute their efforts. A successful open source project with well-managed development team and community can always produce high quality software.

---

Po-Han Huang

Graduate Institute of Information Education, National Kaohsiung Normal University,  
Kaohsiung, Taiwan

e-mail: [hbh.no1@msa.hinet.net](mailto:hbh.no1@msa.hinet.net)

Dowming Yeh · Wen-Tin Lee

Department of Software Engineering, National Kaohsiung Normal University,  
Kaohsiung, Taiwan

e-mail: [{dmyeh, wtlee}@nkn.edu.tw](mailto:{dmyeh, wtlee}@nkn.edu.tw)

Software projects should have documentations that describe the architecture of the system. Documentations can help software development and maintenance. The ownership architecture was defined in a previous research [1]. The Ownership architecture that shows the relationship between developers and source files is useful to understand a software system. Besides, ownership architecture documentations have other use:

1. Identification of Experts: If a developer has any questions about some part of the system, the most effective way is to consult another experienced developer face to face. Developers can rapidly find the appropriate expert via the ownership architecture documentation.
2. Finding Non-Functional Dependencies: To understand a system, function call and data access is often used to find dependencies of the source code. But in some cases, this approach cannot find relationships that are not visible in the source code. If a developer worked on several parts of a system, we may reasonably hypothesize that these parts have non-functional dependencies. By the ownership architecture, the ownership of every source code and subsystem can be easily found.
3. Quality Estimates: An experienced developer usually can write code with good quality. By reading the list of the developers in the ownership architecture documentation, we can estimate the quality of source code.
4. Adjusting number of team members: Wrong number of developers may reduce development efficiency. Too many developers may not be able to adequately partition the work, and too few developers may cause delay. Ownership architecture can be used to help adjusting the number of team members.

However, many software projects do not have such documentations or the content is not up-to-date. Therefore the purpose of this research is to reconstruct system documents based on existed project development data and open source projects are main research objects.

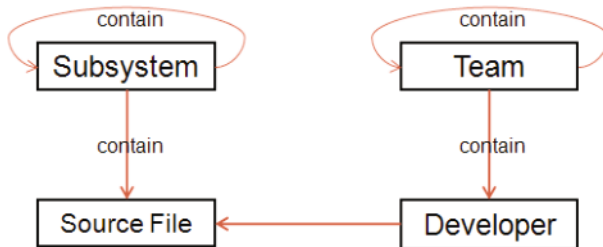
## 2 Ownership Architecture

The ownership architecture was defined in the previous research [1]. The elements of ownership architecture are source files, subsystems, developers, and teams (see Fig. 1). It uses subsystems to group source files, teams to group developers, and shows the ownership between developers and source files. The main purpose of the study is to construct ownership architectures to help understanding software systems.

Several studies have adopted the concept of ownership to help understanding systems. For example, the ownership of source files was used to understand how the developers drove the evolution of the system, including the number of developers and the behaviors of developers [3]. There are also other applications like domain expert identification, understanding the organization's development or team structure [2]. For most software projects, ownership architecture documentations do not exist. Therefore the ownership architecture has to be reconstructed by existed data.

Much useful information that reveals status of projects has been reserved in the process of software development, including comments in source code, logs in version control systems, and project documentations. The data is public and easy to get for open source projects. The characteristics and constraints of each data are as follows:

1. Source code: Most of source files have a comment at the top of the file. This comment may contain a copyright notice, the developer's name, and change information of the source file. The owner of each source file can be found by retrieving the developer's name. But this method has some constraints. First, not all source files have comments or the developer's name is omitted. In this case, to find the owner is almost impossible. Secondly, the format of comments is not fixed. In the previous research [1], ownership architecture was constructed by investigating the comments at the top of source files. However, the method to associate developer names with source files is to examine the comments manually. To retrieve the information in comments automatically is difficult and advanced techniques are needed.
2. Project documentations: The lists of developers usually can be found in project documentations or project websites. Some detailed lists may also describe achievements of each developer. By the documentation, we can know how many developers attend a project and the work allocation. But the relationship between developers and source files is not demonstrated, and the number of developers may increase as the software develops. If the document update is not frequent, some developers may not appear in the list. Although the list of developers cannot be used to construct the ownership architecture, it is helpful to identify developers.
3. Version control systems: Version control systems are used generally to manage source code. After developers doing a change to the source file, the version control system will record the modification. Each change log contains the developer's name, the date of change made, and comments. We can find at least one developer of each source file by analyzing the change logs. Because change logs have a fixed format, to program a tool which can extract information automatically is practicable. There are several studies that have utilized the information in version control systems. Version control logs are investigated to explain the rational of dependencies and software architecture [4]. Other applications are like measures of expertise [6] and social network analysis [5].



**Fig. 1** Ownership Architecture

### 3 Design and Implementation

#### 3.1 Data Collection

In this research, the Password Safe project [7] in SourceForge [8] was chosen as the research object. SourceForge is an open source software development web site and provides free hosting to open source software development projects with a centralized resource for managing projects, issues, communications, and code.

The Password Safe project has many public areas, like bug tracking system, forums, and software repositories. Because ViewVC was used, we can use web browsers to browse SVN repository of the project. ViewVC [9] is a browser interface for CVS and SVN repositories. It generates HTML to present navigable directory, revision, and change log listings. The developers of Password Safe project use SVN to manage the source code. Detailed information like source code and change logs can be found by clicking hyperlink. The research goal is to know the ownership between developers and source files. Change logs in the software repository are necessary. In order to collect data efficiently, we use the offline browser to download html files that contain change logs for later analyzing.

#### 3.2 Data Analysis

Each change logs contains revision number, date and developer, file length, and comment. Because every change log has the same format, we can design a tool to retrieve the information automatically. To find the pattern, we use a text editor to open the change log and analyze its format in the form of text. After analysis, the pattern is shown in Table 1.

**Table 1** Pattern of a change log

Information	Html code
Revision number	<strong>number</strong>
Date	Modified \n \n <em>date</em>
Developer	by <em>developer</em>
File length	 File length: length(s)
Diff to	<a>previous number</a>

Table 2 shows how to use JAVA regular expression to retrieve the name of the developer in a change log. First, a regular expression, specified as a string, must be compiled into an instance of Pattern class. The resulting pattern can then be used to create a Matcher object that can match arbitrary character sequences against the regular expression. In this case, the string between by "<em>" and "</em>" will be found. Finally, the Matcher object matches the entire input text and returns matched strings.

**Table 2** Pattern Matching in Java

Function	Java code
Patter definition	Pattern p = Pattern.compile("(?<=by<em>).*?(?=</em>)");
Pattern matching	Matcher m = p.matcher(text);
Matched strings	while(m.find()) String name = m.group();

Because developers may modify a source file many times, the retrieved names are stored in an array and duplicates are eliminated. The frequency which a developer made a change can indicate the contribution to the source file. Each name that appears in change logs is counted to help estimating importance of developers. In this phase, the analysis tool will analyze input data and generate a document. The document lists all source files and corresponding developers to show the ownership.

### 3.3 Software Architecture

Besides the relationship between developers and source files, the ownership architecture also shows the software architecture and how developers are grouped into teams. The software architecture describes function components (subsystems) of the software and indicates which source files implement them. In this research, we use directory structure of source code to identify the function components essentially.

The tool analyzes the directory structure of source code to construct the software architecture. Names of each directory and source files in directories will be retrieved. Each directory is considered a subsystem, and source files in the directory are related to the subsystem.

### 3.4 Information Integration

The software architecture can be used to group developers. The source files in a directory implement the same function component, and the owners of these source files can be considered a team. In this phase, developers in directory level will be retrieved to know the team's composition. The number of source files which a developer owns will be also counted, and it can help estimating the developer's importance in the team. Finally, results of the three steps will be integrated into ownership architecture documentations.

## 4 Conclusions

The purpose of this research is to construct ownership architecture documentations. We chose a software project as the research object and analyzed the version control system. Change logs in version control system contain developer information and the format is fixed. To design a tool which can extract information automatically is

practicable, and the tool can be applied to software projects that SVN or CVS is used to manage source code.

We implemented the tool, and the result indicated the possibility of generating documentations by a program. By the ownership architecture documentations, developers can know project status and find system experts easily.

**Acknowledgements.** This work was supported by the National Science Council, Republic of China, under the grant number NSC- 96-2221-E-017-006-.

## References

1. Bowman, I.T., Holt, R.C.: Reconstructing ownership architectures to help understand software systems. In: International Conference on Program Comprehension, p. 28 (1999)
2. Ganesan, D., Muthig, D., Knodel, J., Yoshimura, K.: Discovering organizational aspects from the source code history log during the product line planning phase—a case study. In: Proceedings of the 13th Working Conference on Reverse Engineering, pp. 211–220. IEEE Computer Society, Washington, DC, USA (2006)
3. Girba, T., Kuhn, A., Seeberger, M., Ducasse, S.: How developers drive software evolution. In: Proceedings of the Eighth International Workshop on Principles of Software Evolution, pp. 113–122. IEEE Computer Society, Washington, DC, USA (2005)
4. Hassan, A.E., Holt, R.C.: Using development history sticky notes to understand software architecture. In: Proceedings of the 12th IEEE International Workshop on Program Comprehension, pp. 183–193. IEEE Computer Society, Washington, DC, USA (2004)
5. Luis, G.B., Robles, G.: Applying Social Network Analysis to the Information in CVS Repositories. In: Proceedings of the Mining Software Repositories Workshop. 26th International Conference on Software Engineering (2004)
6. Mockus, A., Herbsleb, J.D.: Expertise browser: a quantitative approach to identifying expertise. In: Proceedings of the 24th International Conference on Software Engineering, ICSE 2002, pp. 503–512. ACM, New York (2002)
7. PasswordSafe, <http://sourceforge.net/projects/passwordsafe/>
8. SOURCEFORGE, <http://sourceforge.net>
9. ViewVC, <http://viewvc.tigris.org>



# DF-ReaL2Boost: A Hybrid Decision Forest with Real L2Boosted Decision Stumps

Zaman Md. Faisal\*, Sumi S. Monira, and Hideo Hirose

**Abstract.** In this hybrid decision forest each individual base decision tree classifiers are integrated with an additional classifier model, the *boosted* decision stump. In this boosting, observation weights for subsequent iterations are updated according to the binomial log-likelihood (L2) loss function. This boosted decision stump trained on the extra samples different than the base tree classifiers (which are defined as out-of-bag samples). This extra sample along with the subsample on which the base tree classifiers are trained approximates the original training set, so in this way we are utilizing the full training set to construct a hybrid decision forest with larger feature space. We have applied this hybrid decision forest in a real world applications: prediction of short term extreme rainfall. To check its performance we have also compared the results with relevant prediction methods of the two applications. Overall results suggest that the new hybrid decision forest is capable of yielding commendable predictive performance.

**Keywords:** decision forest, real adaboost, logistic loss, credit classification, rainfall forecas.

## 1 Introduction

Ensemble learning is one of the main research directions in recent years, due to their potential to improve the generalization performance of the predictors. It has attracted scientists from several fields, such as statistical data mining [1], machine learning [2] and pattern recognition [3] to seriously explore

---

Zaman Md. Faisal · Sumi S. Monira · Hideo Hirose

Kyushu Institute of Technlogy

680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

e-mail:  [{zaman,sumi}@ume98.ces.kyutech.ac.jp](mailto:{zaman,sumi}@ume98.ces.kyutech.ac.jp) ,  [hirose@ces.kyutech.ac.jp](mailto:hirose@ces.kyutech.ac.jp)

\* Corresponding author.

the theory and the use of ensemble methodology. Several studies demonstrate that the practice of combining several base classifier models into one aggregated classifier leads to significant gains in classification performance over its constituent members [4], [5].

A *bagging type* [4] hybrid ensemble method was proposed by Hothorn and Lausen, defined as *Double Bagging* [6] to add the outcomes of arbitrary classifiers to the original feature set for bagging of classification trees. Double bagging is a combination of linear discriminant analysis and classification trees. Motivated by the, “main ideas” of Double Bagging, we proposed in this paper a novel ensemble classifier generation method integrating small subsampling aggregation and real logitBoosting. In this decision forest class probabilities of real logitBoost with decision stump as base classifier is utilized to enlarge the feature space of the component base decision tree classifier of the decision forest. Real adaboost is efficient in discarding the course information of predicted labels by transforming the class posteriori probability of the outcomes into real valued scale. But still this method lacks robustness, so we have incorporated the binomial log likelihood loss function instead of the exponential loss function of the original adaboost algorithm. So this decision forest is comprised of a decision forest with component decision trees trained from real logitBoost module.

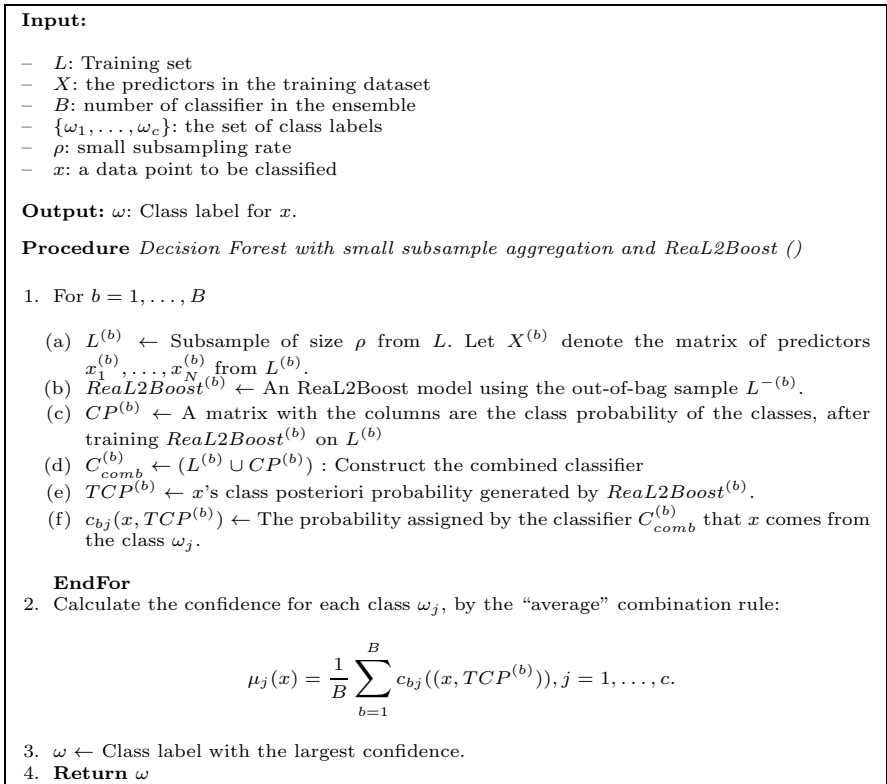
The rest of the paper organized as: in Section 2 we have discussed briefly about the constitutional steps of the new decision forest, emphasizing on the real logitBoosting. In Section 3 we have described the two real world problems handled in this paper with the description of the experiment and discussion of the results of both the problems. This is followed by the Conclusion of the paper.

## 2 DF-ReaL2Boost: A Hybrid Decision Forest with Real L2Boosted Decision Stumps

In this section, we briefly discuss about the construction of the new decision forest. When a decision tree is adopted as the base learning algorithm, only splits that are parallel to the feature axes are taken into account even though the decision tree is non-parametric and can be quickly trained. Considering that other general splits such as linear ones may produce more accurate trees, a “Double Bagging” method was proposed by Hothorn and Lausen [6] to construct ensemble classifiers. In double bagging framework the out-of-bag sample is used to train an additional classifier model to integrate the outputs with the base learning model.

In [7] it is shown that the bias and variance of the regression bagging ensemble can be reduced by smaller subsamples. Based on this theory we have used small subsampling ratio to create subbagging ensemble for classification task. Also as error rate of an ensemble is inversely related to the sample size

and on the other hand the variance of an ensemble increase with the decreasing of the subsample size, these two competing effect will not increase the error rate with decreasing sample size as usually expected in prediction. Bagging type ensemble methods reduce the prediction variance by a smoothing operation but without much effect on the bias of the base model. Though the constitutional steps this new decision forest is similar to bagging but this is trained on an enlarged feature space. This entails an enhanced representational power for each of the base decision tree, which will lower the bias of the decision forest when combined. Moreover the Real2Boost module will low biased estimates of class probabilities, which will in a sense increase the efficiency of the additional features for each base tree classifier and hence will increase the prediction accuracy of the decision forest all together.



**Fig. 1** Generic Framework of DF-Real2Boost

In our framework each out-of-bag samples are utilized to construct an Real2Boost and then this Real2Boost is applied back to the subsample to extract the CPP, then all the CPPs of the Real2Boost are stored in the matrix  $CPP$  and are used as the additional features with the original feature

$\mathcal{X}$  as  $[\mathcal{X} \text{ CPP}]$  which is an  $r \times p(c + 1)$ , where  $p$  is number of features,  $c$  is number of classes in  $D_i$ . The detailed steps of the proposed method for constructing an decision forest ensemble is described in Fig. [1](#)

## 2.1 Real LogitBoosting (Real L2Boost)

Friedman et. al [\[8\]](#) proposed an alternative boosting algorithm called Real AdaBoost. This algorithm outputs a real-valued prediction rather than class labels at each stage of boosting. And the class probability estimate is converted using the half-log ratio to a real valued scale. This value is then used to represent an observations contribution to the final overall model. But the observation weights for subsequent iterations are updated according to the exponential loss function of AdaBoost. This exponential loss function is not robust against the noisy outcomes. Another boosting algorithm developed by Friedman et. al [\[8\]](#), called the LogitBoosting (logitBoost from now on) fits an additive logistic regression model by stagewise optimization of the binomial log-likelihood is more robust in noisy problems where the misclassification risk is substantial. In this paper we have incorporated the real adaboost with binomial log-likelihood loss function, and hence the method, *Real LogitBoosting*. In addition to this we have used the logitBoost framework with decision stumps as the base classifier.

## 3 Problem Setup and Discussion of the Experimental Results

### 3.1 Short Term Extreme Rainfall Forecast Problem

In this section we describe the experimental setup of short term extreme rainfall forecast problem. Our main goal is to check the performance of the new decision forest in short term forecasting, so we have taken a rainfall dataset of a large area comprising several weather stations. We have also compared the new method with other renowned relevant data-mining ensemble techniques : a) LogitBoosting (LB) [\[9\]](#), b) Random Forest (RF) [\[5\]](#), and with a popular single classifier model named, c) Least Square Support Vector Machine (LSSVM) [\[10\]](#).

**Dataset description.** In this paper we have used a dataset containing meteorological records averaged over several weather stations of Bihar region in India. The data was collected from Indian Statistical Institute (Calcutta), which contains rainfall update from 1990 to 2004. The output parameter is a numerical variable which gives us the amount of rainfall (in mm) in a day, but we do the next transformation to have a categorical output based on the accumulated rainfall in a day; the categorization is done as follows:

Rainfall amount Category	
0 mm–1 mm	1
2 mm–15 mm	2
30 mm >	3

The first category can also be defined as ‘no rain’ category and the last category can be defined as ‘extreme rainfall’ category. In this paper we are interested to perform the binary forecast of ‘extreme rainfall’ against the ‘normal rain’ event. This is done by defining category 1 and 2 together as ‘normal rain’ event and category 3 is defined as ‘extreme rainfall’ event. We define this forecast problem as ‘Extreme rainfall forecast problem’. As the probability of such high amount of rainfall is rear, this forecast problem is about forecasting a rear event.

**Description of experimental setup.** In this problem our aim is to forecast whether or not it will not extreme the next day; so we have reconstructed our data to single lag. It is also very important to note that as most of the methods in this paper are based on resampling technique so we should be careful in constructing the lagged dataset for the methods. The value of all the input variables are set to current day and the next day rainfall is forecasted with all these variables. In this paper we have conducted our experiment in two phases, a) Training phase, b) Out of sample (Test) phase. For this purpose we have partitioned our data in two equal parts, training set and test set. The training period is from 1990 to 2000 and the test phase is from 2001 to 2004.

Signal detection theory was first applied to the verification of meteorological forecasts in the pioneering studies of Mason [11] and provides a universal framework for evaluating the joint probability distribution of forecasts and observations. The computed metrics are fraction correct (FC), hit rate (H), false rate (F) and false alarm ratio (FAR), odds ratio (OR) [12]. We have also computed some skill scores to check the skill of the methods in out of sample forecast; the scores we computed are, Bias, Pierce Skill Score (PSS), Extreme Dependency Score (EDS) [20], Odds Ratio Skill Score (ODSS) [13], Symmetric EDS (SEDS) [14] and Area under the ROC Curve (AUC).

In the out of sample or test phase, the extreme rainfall forecast problem can be viewed as a rare event forecasting. Some usual metrics (FC, H, F, FAR) can be misleading because these measures are heavily influenced by the frequency of the climatological events, so inference on extreme events based on these measures will be rather biased. Stephenson in [12], for verification of rare event forecasting (also extreme event), proposed to use Bias, OR (Odds Ratio), PSS (Pierce Skill Score) and ORSS (Odds ratio skill score). Recently Stephenson et.al [13] proposed a measure EDS (Extreme Dependency Score) and Hogan in [14] proposed SEDS (Extreme Dependency Score), these two measures are till now quite robust for verification of the rare event forecast. In addition to these we have also computed the AUC (Area under the ROC curve) of models for both the problems. As we know higher AUC is desirable

for binary prediction problem and this measure is now used more frequently than accuracy in binary prediction problems.

**Discussion of the results.** In Table [1](#) we have presented the values of the metrics of the prediction algorithms for the extreme rainfall forecasting. The best values of each metric for each of the methods are marked bold (here best is relative to the behavior of the metrics). It is worth notable that the new decision forest has best skill scores in terms of the forecast verification measures (PSS, EDS) and better values than most other methods for ORSS and SEDS. Also the AUC of the new ensemble is far better than other methods.

**Table 1** Metric values of the methods in test phase

Metrics	DF-Real2Boost	LB	RF	LSSVM
FC	0.8980	<b>0.9162</b>	0.8897	0.8932
H	0.5000	0.5271	<b>0.5312</b>	0.3543
F	0.0083	<b>0.0056</b>	0.0282	0.0468
FAR	0.5361	0.4167	<b>0.2277</b>	0.5287
Bias	1.540	0.7429	0.7768	0.7411
OR	11.011	<b>16.182</b>	11.763	10.704
ORSS	0.8334	<b>0.8836</b>	0.8433	0.8291
PSS	<b>0.4167</b>	0.2777	0.3193	0.2935
EDS	<b>0.6332</b>	0.3891	0.3882	0.3568
SEDS	0.4198	0.388	<b>0.4651</b>	0.4459
AUC	<b>0.8160</b>	0.5389	0.6596	0.6468

## 4 Conclusion

In this paper a new hybrid decision forest is proposed. The decision forest is incorporated with small subsample aggregation and real logitBoosted decision stumps. The decision forest is applied in a real world problems: predicting extreme rainfall and its performance is compared with some very well known machine learning algorithms available now. The results suggest that the new decision forest has capability to efficiently perform the classification task.

## References

1. Hastie, Trevor, Tibshirani, Robert and Friedman, J.: The elements of statistical learning: data mining, inference and prediction. 2 edn. Springer (2009)
2. Dietterich, T.G.: Ensemble Methods in Machine Learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems, Springer-Verlag (2000) 1—15
3. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. 1 edn. Wiley-Interscience (2004)

4. Breiman, L.: Bagging Predictors. *Machine Learning* **24**(2) (1996) 123–140
5. Breiman, L.: Random Forests. *Machine Learning* **45**(1) (2001) 5–32
6. Hothorn, T., Lausen, B.: Double-bagging: combining classifiers by bootstrap aggregation. *Pattern Recognition* **36**(6) (June 2003) 1303–1309
7. Friedman, J., Hall, P.: On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference* **137**(3) (March 2007) 669–683
8. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28**(2) (2000) 337–407
9. Dettling, M., Buhlmann, P.: Boosting for tumor classification with gene expression data (June 2003)
10. J.A.K. Suykens, Vandewalle, J.: Least Squares Support Vector Machine Classifiers. *Neural Processing Letters* (1999) 293–300
11. Mason, I.: A model for assessment of weather forecasts. *Australian Meteorological Magazine* **30** (1982) 291–303
12. Stephenson, D.: Use of the “Odds Ratio” for Diagnosing Forecast Skill. *Weather Forecasting* **15**(2) (2000) 221–232
13. Stephenson, D.B., Casati, B., Ferro, C.a.T., Wilson, C.a.: The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications* **15**(1) (March 2008) 41–50
14. Hogan, R.J., O’Connor, E.J., Illingworth, A.J.: Verification of cloud-fraction forecasts. *Quarterly Journal of the Royal Meteorological Society* **135**(643) (July 2009) 1494–1511

# Magazine Image Retrieval with Camera-Phone

Cheng Yang, Jie Yang, and Deying Feng

**Abstract.** In this paper, we describe an approach for image-based retrieval with a camera-phone. We begin by getting the features in salient region of images. Then we quantify each image to a vector using the clustering-based bag-of-words model and sparse matrix algorithm. Finally, the invert document algorithm is used for speeding up the real-time query. The result of experiment shows the high efficiency in precision, query time and memory.

**Keywords:** image retrieval, features, salient region, bag-of-words, invert document.

## 1 Introduction

Content-based image retrieval (CBIR) has been studied for years, many different approaches have been developed and the local feature-based method has also been proved effective during recent years. James Philbin has used the SIFT descriptors and bag-of-words model in large-scale object retrieval system [1]. Weijun Wang has used the method of combination of information from the local co-occurrence of salient regions with the sparse vector representation [2]. We therewith, propose an approach for magazine pages image retrieval with a camera-phone.

We scan a lot of magazine covers and advertisement pages as the standard images. The query images are gotten by a camera-phone, which result in differences in scale, illumination and rotation from standard images. The phone user sends the query image to local server. After computing the similarity of query image and standard images, the server returns the most likely scanned images. The whole system includes five modules: feature extraction, salient region extraction, cluster, quantization and query.

Compare to other content-based images retrieval system, our system have some distinctive characters and challenges. First, the query images gotten from camera-phone have lower resolution and definition, even have some noise. Second, the change of illumination between query and standard images is often big but the rotation is often little. Third, the images of magazine often contain a large number



of little characters which will interfere with the retrieval. Lastly, the requirement of real-time capability is high, and it requires a processing time of less than 5 seconds. The last two problems are the main tasks to solve.

The presentation of the system is organized as follows: in section 2 we present local features and salient region extraction; in section 3, cluster and quantization; in section 4, the invert document algorithm in query; in section 5, the result of our experiment, and in section 6, conclusions and further work.

## 2 Local Feature and Salient Region

Local features, as an image feature extraction technology, have enjoyed growing popularity in recent years. Local features have been proved to be very successful in applications such as image matching, image retrieval, image recognition, texture recognition, video data mining, image mosaic, recognition of object categories, and so on. They describe the local information of the images. Compared with the global features, the local features are more distinctive, invariable and robust. Thus, the local features are more adaptive when the image has blur background, partial occlusion or illumination changes.

Many different local features techniques have been developed. Johnson and Hebert [3] have proposed an expressive representation for 3D object recognition in the context of range data. Zabih and Woodfill [4] have developed an approach robust to illumination changes. To resolve the problem of scale change, Lowe [5] has proposed a scale invariant feature transform (SIFT), which combines a scale invariant region detector and a descriptor based on the gradient distribution in the detected regions. The descriptor is represented by a 3D histogram of gradient locations and orientations. The quantization of gradient locations and orientations makes the descriptor robust to small distortion. SIFT descriptor is invariant to changes of scale, translation, rotation and illumination. So we decide to use SIFT descriptor in our approach.

However, the SIFT detector detects too many key points whereas the magazine pages have a large number of little characters which is harmful to the retrieval. Hence before the feature extraction, we extract the salient region of images [6].

## 3 Cluster and Quantization

### 3.1 Visual Vocabulary

In order to compute the similarity between images, we should code each image to a vector with same dimensions. To solve this problem, visual vocabulary is necessary. The concept of visual vocabulary is rooted in text retrieval and it is created from local features. We cluster the features of all standard images to  $K$  centers, and given the cluster centers set is  $\{C_1, C_2, \dots, C_K\}$ . Then each center and features which belong to the center can be regard as a visual word, and the belonging relationship between each feature and centers can be regard as a visual

vocabulary. A given visual word represents a set of similar feature, and the images can be decomposed to a set of visual words. Namely, we can code the images to vectors with dimensions of  $K$ .

### 3.2 Approximate K-Means(AKM)

Then, we discuss the algorithm of cluster. As the increment of the corpus of images and the number of centers, the K-means cluster become very slowly. So we use the AKM (Approximate K-means) algorithm. Instead of using accurate nearest neighbor search in K-means cluster, the AKM use an approximate nearest neighbor search with randomized k-d tree [1]. In the randomized algorithm the splitting dimension is chosen at random from among a set of the dimensions with highest variance. In the same vein, the split value, rather than the median value along that dimension as the value to split on, is randomly chosen using a point close to the median.

A new data point finds the approximately nearest cluster center as follows. Each tree is descended to a leaf and the distances to the separating boundaries are pushed into a single priority queue for all trees [7]. Then, we iteratively choose the most probable branch from all trees and keep adding invisible nodes into the queue. There is a fixed maximum number of searches and it stops once the fixed number of tree paths has been explored. The speed of AKM is at least two times faster than K-means. Moreover as the increment of corpus of images the advantage is more obvious (see Table 1).

**Table 1** Comparison of K-means and AKM

Algorithm of cluster	Cluster centers	Cost time
K-means	10,000	72h
AKM	5,000	12h
AKM	10,000	23h

In the following paragraphs, we discuss several parameters of the cluster. In this paper, we evaluate the cluster using the precision of retrieval. The definition of precision and the reason of using precision will be present at section 5. To begin with, we define a parameter  $\rho$  which is the ratio of the number of cluster centers and the number of total features to be clustering. Then we consider the number of cluster centers  $K$ . The  $K$  is a very important parameter which can impose direct effect upon the precision of retrieval. It is, therefore, reasonable that the greater the  $K$  is the more precise the retrieval is. In our case, when the  $\rho$  is less than approximately 0.3, the criterion is right. When the  $\rho$  is greater than it, however, it is invalid.

It is important to note the adjustment of a significant cluster parameter *Iter* which is the number of iteration of cluster. It seems that the perception of the cluster result grows in proportions with that of the iteration. Yet just as the retrieval precision of result of experiment shows, apart from improving the result

of cluster, the increment of iteration also brings the slow speed and inefficiency. Especially when the  $\rho$  becomes sufficiently great, e.g. 0.1, in the condition of unchanging the precision of retrieval, the iteration even can be set to 1. It reduces the time consumption of cluster enormously (see Table 2).

**Table 2** Cost time of AKM with different iteration

Iteration	Cost time	Iteration
40	4h	40
20	1.3h	20
1	183s	1

### 3.3 Stop List and tf-idf

Before quantization, we use a stop list to select useful visual words in vocabulary. A stop list is the visual words (centers) occurring in most images representing little information. Suppose the  $N_i$  is the number of images containing center  $C_i$ , namely there are  $N_i$  images have at least one feature belonging to center  $C_i$ . We remove the words which satisfy the following inequation:

$$N_i > k * N \quad (1)$$

Where  $N$  is the number of images and  $k$  is a experiential parameter. In our case we set the  $k$  equal to 0.4, and usually leave 80% words.

The quantization uses the method which is known as ‘term frequency–inverse document frequency’ (tf-idf). Suppose there is a vocabulary of  $T$  words after using stop list, then each image can be quantized to a vector  $V_d = [v_1, v_2, \dots, v_T]$ . To eliminate the difference of numbers of features among images, we normalize the  $V_d$ .

## 4 Query Module

The treatments ahead are all off-line, but the query module must be real-time. The camera-phone user shoots the magazine page which he feels interesting, and then sends the image to the local server. After computing, the server returns the result of retrieval and sends the result back to the user. It requires the query module a fast speed of computation and communicates with the phone user in time. As the latter is related to the technology of wireless communication, we only discuss the computation in local server in this paper.

When the server get query image from user’s phone, it first extracts the SIFT features of query image, then find the approximate nearest neighbor of each feature in cluster centers using 8 randomized k-d trees, so that the query image can be quantized to a vector  $V_q$  which has same dimensions with  $V_d$ . The  $V_d$  is stored as sparse vector too. At last, we compute the similarities between query image and standard images using the dot product of  $V_q$  and  $V_d$ . The standard images which have the top biggest similarities will return as the result of retroviral. If some parts

of the returned image are as same as that of query image (though may have changes in scale, illumination and rotation), we call the returned image “positive”.

To speed up the query, we use the invert document algorithm. If most corresponding components of two vectors are not both nonzero, the dot product of them is unlikely to be very big. So it does the  $V_q$  and  $V_d$ , so it is inefficiency to compute the similarity one by one. The invert document algorithm finds those  $V_d$  which have most nonzero corresponding components with  $V_q$ . It needs some off-time treatment before query. Given the size of visual vocabulary is  $T$ , we establish  $T$  stacks  $\{S_1, S_2, \dots, S_T\}$ . If the image  $d$  has the visual word  $i$ , we push the number  $d$  to the back of stack  $S_i$ . After getting the  $V_d$ , we establish a set of accumulators with size of  $N$ , where  $N$  is the number of standard images. The initial value of each accumulator is zero. If the component  $i$  of  $V_d$  is nonzero, then look up the stack  $S_i$ , and the accumulators corresponded the images which have been pushed in the stack add one. At last, we sort the  $N$  accumulators as descending order, and only get the  $M$  images corresponding the top  $M$  accumulators to dot product with  $V_d$ . In our case, it is efficient to get a satisfied retrieval result that set the  $M$  to 5.

## 5 Result of Experiment

The eventual corpus of images has 18,755 standard images. After salient region extraction, the number of total SIFT features is 1,083,816, and the  $\rho$  is 0.3 namely the number of cluster center is 325,144. We use the AKM which the number of iteration is 5 to cluster and the number of test query images is 474.

As the main interesting of our system users is weather the returned images is positive, we use the precision to evaluate the retrieval which is defined as follows:

$$precision = \frac{correct \ retrieval \ number}{test \ query \ images \ number} \quad (2)$$

We call retrieval “correct” when the number of positive images among returned images reaches the required.

If only the first returned image is required to be positive, the precision of retrieval can reach 95%. Table 3 shows the performance of retrieval in the condition of different required number of retrieved positive images. As the number of standard images with the same image part is small, we won't gather the precision when the required number of retrieved positive images exceeds 3.

**Table 3** Precision with different retrieved positive images

Required number of retrieved positive images	Precision
1	95%
2	71%
3	42%

## 6 Conclusion and Further Work

Compare to the other image retrieval systems, our system have follow characters: the query images are a part of standard images, the resolution and definition of images get from camera-phone are lower, the query images often have noise, and the rotation between query and standard images are not very big.

To solve these problems, we first down-sampled the standard images with a factor of 2 and query images with 3 and smooth these images using Gaussian filter. As the SIFT descriptor is invariant to changes of scale, this approach doesn't lose the important key features of images but filter some noise of them. The last result of experiment shows that it doesn't reduce the precision of retrieval and on the contrary it speeds up the retrieval. And the other key steps of our approach are salient region extraction, AKM cluster, using stop list and sparse matrix presenting.

Using the technique above, we can enlarge the corpus of images easily to 100,000 or even bigger. In that case, we should sample local features randomly from the features of each image. Especially, in the step of cluster we can only choose a part of the images as the training samples. Then the all images are quantized to vectors using the visual vocabulary created from the training samples. Otherwise, it can hardly store a bulk of local features in memory and cluster them. In addition, it's better to set the number of iteration of cluster small, preferable is to reduce it to 1.

## References

1. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1-8, pp. 1545–1552 (June 2007)
2. Wang, W., Luo, Y., Tang, G.: Object retrieval using configurations of salient region. In: CIVR 2008, pp. 67–74 (July 2008)
3. Johnson, A., Heber, M.: Object recognition by matching oriented points. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA, pp. 684–689 (1997)
4. Zabih, R., Woodfill, J.: Non-Parametric Local Transforms for Computing Visual Correspondance. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 801, pp. 151–158. Springer, Heidelberg (1994)
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
6. Hou, X., Zhang, L.: Saliency Detection: A Spectral Residual Approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, vol. 1-8, pp. 2280–2287 (2007)
7. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *Journal of the ACM* 45(6), 891–923 (1998)

# Statistical Clustering and Times Series Analysis for Bridge Monitoring Data

Man Nguyen, Tan Tran, and Doan Phan

## 1 Introduction

The process of implementing a damage detection strategy for bridges is referred to as *Bridge Health Monitoring* (BHM). The BHM process involves the observation of a system over time using periodically sampled dynamic response measurements from an array of sensors, the extraction of damage-sensitive features from these measurements, and the statistical analysis of these features to determine the current state of the system's health [12]. Therefore, the achieved data from attached sensors would be very huge in dimensions, would make researchers confused in further examinations on data bridge. There have been many approaches to solve the BHM sensors reduction problem, range from univariate analysis between couples of variables [13] to carefully selecting measurement points based on specific bridge knowledge [7]. However, they are either inapplicable for interrelated nature data sets, or using too much mechanical knowledge in its process.

Suppose we obtained a huge data set  $D$  after continuously on-line measuring a structure or system  $S$  by using many sensors distributed in a certain way on the structure, we need a scheme for evaluating reliability/health of  $S$ , exploiting the spatial and temporal characteristics of  $D$ , to answer two key questions:

1. Which sensors could provide most information about the structure status/ health?
2. From the most informative sensors determined by the first answer, how much certain we could conclude that they potentially indicate damaged places of the observed structure?

Having a mathematical answer of the first question helps us to optimize the resource for investigation the structure's lifetime or meaningful usage; and to some extent, evaluating the structural performance helps engineers and managers to respond to unexpected accidents, and make the right decision at the right time.

---

Man Nguyen · Tan Tran · Doan Phan

Ho Chi Minh City University of Technology, VNU-HCM, Viet Nam

e-mail: [tan@cse.hcmut.edu.vn](mailto:tan@cse.hcmut.edu.vn)

The structure of this paper is as follows. Section 2 mentions our proposed method for reducing the dimensions of bridge database. An application of the proposed method and its results obtained from Sai Gon bridge is presented in Sect. 3. In Section 4 we report the use of probabilistic approach for analysing the same data sets, section 5 shows the experimental validation to the second method of time series modelling. Finally we summarize the paper by reviewing what have been done and pointing out some open concerns.

## 2 Statistical Reduction for Bridge Monitoring Data

### 2.1 Principal Component Analysis (PCA)

We follow [3] [5] for PCA techniques. Let us consider a sample  $\mathbf{x} = (x_1, x_2, \dots, x_p)^\top$  is a  $p$  random variables vector. In essence, PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations (the PCs) of the original variables with the largest variance. For many data sets, the first several PCs explain most of the variance, so that the rest can be disregarded with minimal loss of information.

Now assuming a standardized data with the empirical covariance matrix  $\Sigma_{p \times p} = \frac{1}{n}XX^\top$ , where  $X$  is a  $(n \times p)$  matrix consists of  $n$  observations on  $p$  variables in  $\mathbf{x}$ , we can utilize the connection between PCA and the *singular value decomposition* (SVD) of the mean-centred data matrix  $X$  which takes the form:

$$X = USV^\top, \quad (1)$$

where  $U^\top U = \mathcal{I}_p$ ,  $VV^\top = V^\top V = \mathcal{I}_p$  and  $S$  is diagonal with diagonal elements  $s_1, s_2, \dots, s_p$ . Here,  $s_1 \geq s_2 \geq \dots \geq s_p$  are the non-negative square-roots of the eigenvalues of  $X^\top X$ , the columns of  $U$  are the  $p$  orthogonal eigenvectors of  $XX^\top$  and the rows of  $V^\top$  are the orthogonal eigenvectors of  $X^\top X$ . *PC scores* are given by  $US$ .

A approach using *cross-validation*, that has been developed for problems involving either model choice or assessment of the performance of a predictor (or both) as economically as possible, will be used in this paper. Associated with a given value  $k$  is the predictor  $\hat{X}^{(k)}$ ; an estimate of  $X$  which arises from fitting only the first  $k$  PCs. Thus the prediction model is given by

$$X = \hat{X}^{(k)} + E^{(k)}, \quad (2)$$

where  $E^{(k)}$  is the  $(n \times p)$  matrix of error scores and  $k = 0, 1, 2, \dots$ . Each row of  $E^{(k)}$  has a multivariate normal distribution under the usual distributional assumptions. The errors in any row of  $E^{(k)}$  are statistically independent of the errors in any other row since the rows of a data matrix generally represent randomly sampled subjects.

To compute the discrepancy between actual and predicted values, we use

$$\text{PRESS}(k) = \frac{1}{n \times p} \text{trace} \left\{ \left( E^{(k)} \right)^\top \left( E^{(k)} \right) \right\} \quad (3)$$

and some suitable function of these PRESS values is considered in order to choose the optimum value of  $k$ . The notation PRESS, stands for PREDICTION SUM of SQUARES, and this is taken in a similar sense as in linear regression. These PRESS( $k$ ) values are a measure of how well the model in (2) predicts the data for each  $k$ .

As noted before, we know that write  $X = USV^T$ . Denote the updated values of  $U$ ,  $V$ , and  $S$  by  $\hat{U}$ ,  $\hat{V}$  and  $\hat{S}$  when the  $j$ th column of  $X$  is deleted, and by  $\bar{U}$ ,  $\bar{V}$ , and  $\bar{S}$  when the  $i$ th row of  $X$  is deleted. Corresponding notation is attached to the elements of all these matrices in the obvious way. We predict  $x_{ij}$  by

$$\hat{x}_{ij}^{(k)} = \sum_{t=1}^k (\hat{u}_{it} \sqrt{\hat{s}_t}) (\sqrt{\bar{s}_t} \bar{v}_{tj}). \quad (4)$$

To choose the optimum value of  $k$ , we finally consider a suitable function of PRESS( $k$ ). Analogy with regression analysis suggests some function of the difference between successive PRESS values. One such possibility is the statistic

$$W(k) = \frac{\text{PRESS}(k-1) - \text{PRESS}(k)}{D_k} \div \frac{\text{PRESS}(k)}{D_r}. \quad (5)$$

where  $D_k$  is the number of degrees of freedom required to fit the  $k$ th component and  $D_r$  is the number of degrees of freedom remaining after fitting  $k$ th component. Consideration of the number of parameters to be estimated, together with all the constraints on the eigenvectors at each stage, shows that  $D_k = n + p - 2k$ . Also, since there are  $np - p$  degrees of freedom at the outset (each column of  $X$  being mean-centred),  $D_r$  can be found easily at each stage.  $W$  represents the increase in predictive information supplied by the  $k$ th component, divided by the average information in each of the remaining components. We therefore suggest that the optimum value for  $k$  is the last value of  $k$  at which  $W$  is greater than a chosen unity.

## 2.2 Selecting the Variables

The technique used in this paper is motivated by the long standing and well established technique of *Canonical Correlation Analysis* [4]. We used measures of multivariate association based on canonical correlations as criteria for selecting variables in PCA. The idea is to maximize the similarity or overlap between the spaces spanned by the sets of two PCs, one arises from the full set data while the other arises from the subset data.

Let  $Y$  be the  $(n \times k)$  transformed data matrix of PC scores yielding the best  $k$ -dimensional approximation to the original data determined in the previous section. Let  $\tilde{X}$  denote the  $(n \times q)$  reduced original data matrix which retains only  $q$  selected variables, and  $\tilde{Y}$  be the corresponding  $(n \times k)$  matrix of PC scores. It would seem



reasonable to set  $q$ , the number of variables to retain, to  $k$ . Now let  $\tilde{Z} = (Y|\tilde{Y})$ . The corresponding  $(2k \times 2k)$  partitioned correlation matrix between PCs is given by

$$R = \begin{pmatrix} R_{YY} & R_{Y\tilde{Y}} \\ R_{\tilde{Y}Y} & R_{\tilde{Y}\tilde{Y}} \end{pmatrix}, \quad (6)$$

where,  $R_{Y\tilde{Y}}$  is the  $(k \times k)$  matrix of correlations between the PCs of the set  $Y^\top$  and  $\tilde{Y}^\top$ . Therefore, the *squared canonical correlations* between the two sets of PCs are given by the  $k$  eigenvalues of the matrix  $R_{Y\tilde{Y}}R_{\tilde{Y}\tilde{Y}}^\top$ , arranged in descending order.

For a reasonable index of the total association between the sets (of variables), we use an index, recommended in [9], as follows:

$$\hat{\gamma} = \frac{1}{k} \left( \sum_{j=1}^k \hat{\rho}_j^2 \right), \quad (7)$$

where  $\hat{\rho}_1, \hat{\rho}_2, \dots, \hat{\rho}_k$  are the canonical correlations between the sets of PCs in  $Y$  and  $\tilde{Y}$  arranged in descending order.

### 3 Justification of the First Method on Realistic Datasets

**The data set.** The combined methods are validated by Sai Gon bridge data. Sai Gon bridge is a major entrance to Ho Chi Minh city. 59 sensors were plugged on the bridge transferring vibration signals to a central computer. The sensors map is shown in Fig. 1. Essentially the data is a matrix  $M$  whose rows spread a finite interval of the time evolution, with unit scale  $1/64$  of second, and the columns stand for the attached sensors. Each entry  $M[i, j]$  holds a electronic signal, transformed from the mechanical vibration signal at the time instant  $i$  and recorded by sensor  $j$ , caused by vehicle movements, by wind, by water streams under the bridge, and by noise (undefined factors).

**Justification of the method.** To examine the dimensionality reduction on this database, we firstly choose a demonstrated  $(500 \times 59)$  dataset, namely A1. Each row in A1 is electrical signal in volt measured on 59 sensors, which are numbered from 0 to 58, timely increasing ordered with 1 second interval, from 2:00 PM, 25 August 2009. We chose the lower bound value of  $W$  to be 1 and using backward elimination to achieved  $k$ -sized subset. The results suggest that we should retain  $k = 10$  sensors as in the first row of Table 1.

In addition, we exploit six more datasets based on some special time features from the same original database to check the consistency in the results and performance of the procedure. Carefully examine Table 1, we can see that although the kept variables are not exactly the same, most of them, however, appears in almost every or every results. We can point out most likely uncorrelated, important sensors, appeared six to seven times, such as sensors number **1, 5, 21, 41, 48, 50, 55** and **56**. If we consider our experiment as a repeated process over many time points in the whole datasets, we would want to “union” every sensors appeared in the results to

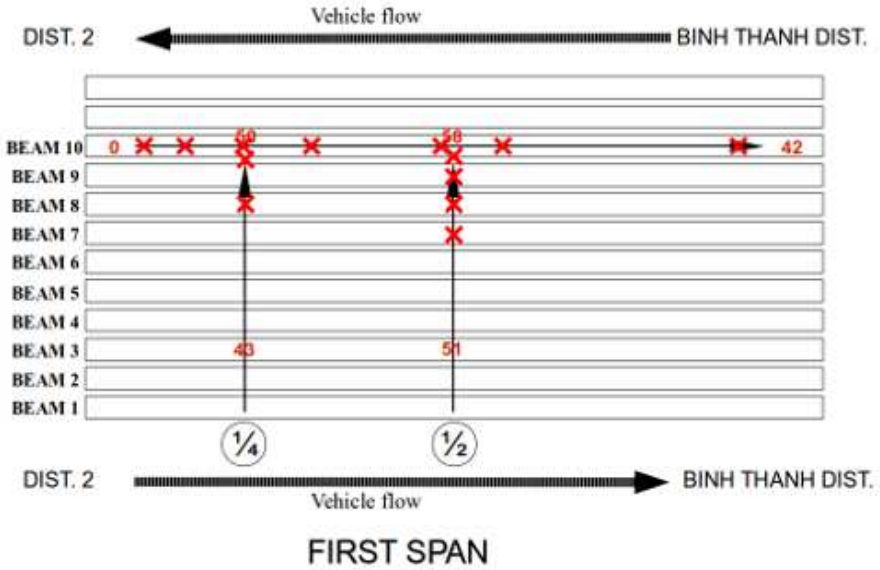


Fig. 1 Sensors map and notable sensors positions (x)

have the total number of measured variables be only thirteen variables, much less than 59 original sensors.

Table 1 Experimental datasets and results

Dataset name	Measurement start time	Rows no.	Measure Time (sec)		Retained $k$
			$t_1$	$t_2$	
A1	25-08-09 02:23 PM	500	43	22	10 1, 5, 15, 21, 27, 41, 48, 50, 55, 56
A2	25-08-09 08:30 PM	500	51	13	9 1, 10, 21, 41, 48, 50, 54, 56, 58
A3	26-08-09 12:00 AM	500	48	20	9 1, 5, 21, 27, 41, 48, 50, 55, 56
A4	26-08-09 08:00 AM	500	49	20	9 1, 5, 15, 27, 41, 48, 50, 55, 56
A5	27-08-09 08:00 AM	500	49	30	9 1, 5, 15, 21, 41, 48, 50, 55, 56
A6	27-08-09 02:40 PM	1000	105	25	10 1, 5, 15, 21, 27, 41, 48, 50, 55, 56
A7	26-08-09 08:11 AM	1500	172	32	9 1, 5, 15, 27, 41, 48, 50, 55, 56

Moreover, when we once again look at the Fig. 1 to get an spatial view of the results and check where thirteen sensors are, we can be visually confident that they are distributed at particular points on every part of the measured spans, which are, on two tails, in the middle and two sides of the bridge. Briefly speaking, they appear at important points on the bridge and represent many unique features of the bridge.

## 4 Time Series Modelling for Damage Prognosis

After the dimensionality reduction step discussed in previous parts, that allowed us to select meaningful predictor sensor variables, we combine the approaches suggested in works [8, 10, 11, 14] to find solutions for the following questions:

1. *Existence*: Is there damage in the system?
2. *Location*: Where are damages in the system?

We will employ time series analysis, data clustering and hypothesis testing to make decision about the state of a bridge.

### 4.1 Data Clustering

The data that measured at the first time can be used as *reference database*, i.e. the data were measured at the undamaged state of the bridge. Then, the new data are recorded at current state of bridge, used as *unknown (new) database*. The data clustering procedure is a process to select the previously recorded signal from the reference database which is recorded under environmental and operational conditions closest to that of new obtained signal.

After standardizing the databases, all time series signals  $x(t)$  and  $y(t)$  in reference and new database, respectively, are fitted with AR model of order  $r$  such that:

$$x(t) = \sum_{j=1}^r \theta_x(j)x(t-j) + e_x(t) \quad (8)$$

$$y(t) = \sum_{j=1}^r \theta_y(j)y(t-j) + e_y(t) \quad (9)$$

where  $\theta_x(j), j = 1 \dots r$  are the AR coefficients,  $e_x(t)$  is white noise input with variance of  $\sigma_x^2$  and  $\{\theta_x(j), \sigma_x^2\}$  can be regarded as signal feature. The same meaning apply for  $y$ . Two feature spaces,  $\Omega^R$  and  $\Omega^N$  respectively corresponding to the reference database and new database can be obtained. The data clustering procedure is then implemented by searching in space  $\Omega^R$  a point  $\{\theta_x(j), \sigma_x^2\}$  that is similar to the target point  $\{\theta_y(j), \sigma_y^2\}$  in  $\Omega^N$ .

Specifically, for a certain target point  $\{\theta_y(j), \sigma_y^2\}$  in  $\Omega^N$  we compute a subspace of feature points in  $\Omega^R$  satisfying the constraint

$$|\sigma_x^2 - \sigma_y^2| \leq \varepsilon_1, \quad (10)$$

After this step, a subspace  $\Omega_1^R \subset \Omega^R$  is obtained. We make another refinement by giving another constraint on  $\Omega_1^R$

$$\frac{\sum_{j=1}^r \theta_x(j)\theta_y(j)}{\sqrt{\sum_{j=1}^r \theta_x^2(j)}\sqrt{\sum_{j=1}^r \theta_y^2(j)}} \geq \varepsilon_2 \quad (11)$$

to obtain a smaller subspace  $\Omega_2^R$  of  $\Omega^R$ . Finally, the Euclidean distance between the target point and each feature point in subspace  $\Omega_2^R$  is calculated.

After this data clustering procedure, the time series signals which are measured in similar environmental and operational conditions with each time series signal in new database are chosen to make the signal pairs which are used in further phases.

## 4.2 Damage Extraction

Our practical calculator indicated that that AR model is not well predicted the time series signal. Hence, we use ARX model to model time series signals. Based on the assumption that the error between the measurement and the prediction in AR model mainly caused by unknown external input, the ARX model is used to represent the input/output relationship between  $e_x(t)$  and  $x(t)$ :

$$x(t) = \sum_{i=1}^p \alpha_i x(t-i) + \sum_{j=1}^q \beta_j e_x(t-j) + z_x(t) \quad (12)$$

where  $z_x(t)$  is the residual error after fitting the ARX( $p, q$ ) model to  $e_x(t)$  and  $x(t)$ . After that, this ARX( $p, q$ ) model is used to reproduce the input/output relationship between  $e_y(t)$  and  $y(t)$ :

$$z_y(t) = y(t) - \sum_{i=1}^p \alpha_i y(t-i) + \sum_{j=1}^q \beta_j e_y(t-j), \quad (13)$$

here  $\alpha_i$  and  $\beta_j$  are the coefficients in Eq (12). If the ARX model of reference signal  $x(t)$  and  $e_x(t)$  were not good for representing the new signal  $y(t)$  and  $e_y(t)$ , there would be a significant change in the standard deviation of residual error  $z_y(t)$  compared to that of  $z_x(t)$ . Consequently, the standard deviation of residual error can be defined as the *damage-sensitive feature*.

## 4.3 Sequential Probability Ratio Test (SPRT)

When the data are collected sequentially, the SPRT procedure will be an appropriate technique to analyse data. The SPRT frequently results in a saving of 50% in the number of observations over the most efficient test procedure based on the fixed

number of observations. As in classical hypothesis test, the SPRT starts with a pair of hypotheses, say  $H_0$  and  $H_1$  for the null hypothesis and alternative hypothesis respectively. The cumulative sum of the log-likelihood ratio is

$$T_n = T_{n-1} + \ln \frac{f(z_n; \sigma_1)}{f(z_n; \sigma_0)}$$

where  $f(z_n; \sigma_1)$  is the probability density function of  $z_n$  at  $\sigma = \sigma_1$ . The stopping rule is a simple threshold scheme:

- i)  $a < T_n < b$ : continue monitoring (critical inequality)
- ii)  $T_n \leq b$ : Accept  $H_0$
- iii)  $T_n \geq a$ : Reject  $H_0$  (Accept  $H_1$ )

where  $a$  and  $b$  ( $0 < a < b < \infty$ ) depend on the desired type I and type II errors,  $\alpha$  and  $\beta$ . They may be chosen as follows:

$$a \cong \log \frac{\beta}{1 - \alpha}, \quad b \cong \log \frac{1 - \beta}{\alpha}.$$

In the damage detection, the standard deviation of residual errors is considered as parameter of the hypothesis testing:

$$H_0 : \sigma \leq \sigma_0 \quad H_1 : \sigma \geq \sigma_1 \quad \text{where } 0 < \sigma_0 < \sigma_1.$$

In this case,  $H_0$  is the hypothesis that the location (where the sensor is plugged) is considered possibly undamaged. Otherwise, it is concluded to be potentially damaged. There are many ways to choose two user specified standard deviation values. One can choose these values from the training database obtained from structure or initialize these values by experiments and then adjust whenever there are more data.

If modified observations  $\{t_i\} (i = 1, 2, \dots)$  are defined by

$$t_i = \ln \frac{f(z_i; \sigma_1)}{f(z_i; \sigma_0)}$$

then, the cumulative sum of the log-likelihood ratio  $T_n$  is calculated by:

$$T_n = \ln \frac{f(z_1, z_2, \dots, z_n; \sigma_1)}{f(z_1, z_2, \dots, z_n; \sigma_0)}.$$

If  $T_n > a$  then  $f(z_1, \dots, z_n; \sigma_1) > e^a * f(z_1, \dots, z_n; \sigma_0)$ . The value of joint probability density function with standard deviation  $\sigma_1$  of  $(z_1, z_2, \dots, z_n)$  is greater than the value of joint probability density function with standard deviation  $\sigma_0$  ( $e^a$  times), then we accept  $H_1$ . This argument can be applied for the two remaining stopping rules.

## 5 Experimental Validation Using Observed Data

To validate the second proposed technique, a combination of time series analysis and SPRT, we used the reduced Sai Gon bridge data set as in the above section. To be precise, because some sensors have not provided stable data, we only exploit data from the best reliable 6 sensors from the 8 important sensors already obtained in Section 3.

The *reference database*  $\mathbf{R}$  contains data recorded in 2 continuous hours on 20 September 2008. The *unknown state database*  $\mathbf{U}$  contains data measured in 20 continuous minutes on these 6 sensors on 25 August 2009. We conducted two experimental validations, in experiment 1 each data sample (matrix) in  $\mathbf{U}$  has size  $1024 \times 6$ , consisting of 1024 signals measured at these 6 sensors, and  $4096 \times 6$ -data sample in experiment 2. The order  $r$  of AR model in data normalization step is set to 25. The value of  $\varepsilon_1$  at step 1 of data clustering is set to 0.1 and the value of  $\varepsilon_2$  at step 2 is set to 0.45. The order  $p$  and  $q$  of ARX model are chosen  $p = 20$  and  $q = 5$ . Ljung [6] suggested keeping the sum of  $p$  and  $q$  in the ARX model smaller than  $r$  ( $p + q \leq r$ ).

Although the orders  $p$  and  $q$  of ARX model are chosen arbitrarily, similar results can be obtained for different combinations of  $p$  and  $q$  as long as the sum of  $p$  and  $q$  is kept smaller than  $r$ . In this experiment, the order of AR model is chosen equal 25, the orders of ARX model are chosen to satisfy the condition as Ljung [6] suggested. The upper bounds of Type I and Type II errors are set to 0.001. The corresponding two bounds are  $a = 6.91$  and  $b = -6.91$ . Whenever the Z statistic goes below the lower bound  $b = -6.91$ , the hypothesis  $H_0$  is accepted.

On the other hand, when the Z statistic becomes larger than the upper bound  $a = 6.91$ , the null hypothesis  $H_0$  is rejected, and the alternative hypothesis  $H_1$  is accepted. The standard deviation of residual errors of 6 sensors after applying AR models are shown in Table 2. In this table the standard deviation of residual errors is larger than 30% of the standard deviation of original data.

**Table 2** The standard deviation of residual errors of AR and ARX model

Sensor	1	2	3	4	5	6
AR	0.69	0.65	0.54	0.63	0.68	0.61
ARX	0.013	0.05	0.008	3.6e-15	0.048	0.016

Therefore, we use ARX model instead of AR to fit data in damage extraction step. Table 2 also tells us that ARX model decreases the standard deviation of residual errors to approximately 0. This shows that ARX model is more suitable to fit bridge data. After damage extraction step, the damage feature is passed to SPRT procedure to get the conclusion.

The result of SPRT procedure at  $\sigma_0 = 0.5$  and  $\sigma_1 = 0.65$  in experiment 1 is shown in Table 3. Each row of this table represents a data sample of size  $1024 \times 6$ , recorded continuously in 17.06 sec (60 signals/sec). Hence Table 3 contains results of 20 data samples in the set  $\mathbf{U}$ , recorded in 20 disjoint time intervals, each 17.06 sec on 25 August 2009. Table 3 shows that number 1 appears 4 and 3 times on column 2 and 5

**Table 3** The result of SPRT procedure in experiment 1 (left) and 2 (right) (1: **damaged**, -1: **suspected**, 0: **undamaged**)

Sensor	1	2	3	4	5	6	1	2	3	4	5	6
0	0	<b>1</b>	0	0	0	<b>-1</b>	0	<b>1</b>	0	0	0	0
0	0	0	0	0	0	0	0	<b>1</b>	0	0	0	0
0	0	0	0	0	0	<b>-1</b>	0	<b>1</b>	0	0	0	0
0	<b>1</b>	0	0	0	<b>-1</b>	<b>-1</b>	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	<b>1</b>	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	<b>-1</b>	0
0	0	0	0	0	0	0	0	0	0	0	<b>1</b>	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	<b>-1</b>	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	<b>1</b>	0
0	0	0	0	0	<b>1</b>	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	<b>1</b>	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	<b>1</b>	0	0	<b>1</b>	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	<b>1</b>	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	<b>1</b>	0	<b>1</b>	0	0	0	<b>1</b>

respectively. The locations where **sensors 2 and 5** are plugged on must be considered more carefully to get conclusion about their status. Column **1, 3, 4** are all number 0. The locations where these sensors put on are likely undamaged. The other locations need many more observations to get better result.

More convincing conclusion can be validated in the half right of Table **3**, that number 1 still appears more often- 5 and 4 times- on **columns 2 and 5** respectively, with other data samples taken in experiment 2. All statistical models used in our experiments and all computations have been coded and validated on the **R** system, an open source but powerful statistical computing software.

## 6 Conclusion

Nevertheless, to have a more concise conclusion, both theoretically and practically, the research team would collaborate with bridge engineers to compare the result with mechanical approaches. Many tough problems remains, as the followings.

1. *Constant selecting.* How could we select constants ( $\varepsilon_1, \varepsilon_2, \sigma_0, \sigma_1, d_{min} \dots$ ) for a certain bridge to get highly reliable prediction? A feasibly empirical way may be applying the proposed solution to other bridges to get more experience on choosing these values.

2. *Efficient data-mining*. When a damage is discovered at certain positions on a bridge, besides repairing or upgrading, what more efficient data-mining methods can be realized and designed to automatically set off extreme frequencies or deviations that caused by vehicle movements?
3. Finally, although we tried to use a better (mathematical and statistical) approach to choose the number  $k$  of PCs, the research on this direction still needs more attention.

**Acknowledgements.** Our gratitude first devotes to Nhi Kieu Ngo, the Director of LAM, University of Technology at HCMC for allowing us to use LAM's highly valuable realistic databases to validate our approach. The work has been partially funded by the WB grant – based TRIG 2 project of VNU- HCM.

## References

1. Farrar, C.R., Keith, W.: An introduction to structural health monitoring. In: Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, vol. 365(1851), pp. 303–315 (2007)
2. Eastment, H.T., Krzanowski, W.J.: Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. *Technometrics* 24(1), 73–77 (1982)
3. Hrdle, W., Simar, L.: Applied multivariate statistical analysis, 2nd edn. Springer (2007)
4. Hotelling, H.: Relations Between Two Sets of Variates. *Biometrika* 28(3-4), 321–377 (1936)
5. Jolliffe, I.T.: Principal component analysis, 2nd edn. Springer (2002)
6. Ljung, L.: System identification: theory for the user. Prentice Hall, Englewood Cliffs (1987)
7. Papadimitriou, C.: Optimal sensor placement methodology for parametric identification of structural systems. *Journal of Sound and Vibration* 278(4-5), 923–947 (2004)
8. Rytter, A.: Vibration based inspection of civil engineering structures. Ph.D. Dissert., Department of Building Technology & Structural Engineering. Aalborg University, Denmark (1993)
9. Sithole, M.M., Ganeshanandam, S.: Variable selection in principal component analysis to preserve the underlying multivariate data structure. In: ASC XII 12th Australian Stats Conference, Monash University, Melbourne (1994)
10. Hoon, S., Allen, D.W., Worden, K., Farrar, C.R.: Statistical damage classification using sequential probability ratio test. *Structural Health Monitoring*, 57–74 (2003)
11. Hoon, S., Worden, K., Farrar, C.R.: Statistical Damage Classification under Changing Environmental and Operational Conditions. *Journal of Intelligent Materials Systems and Structures* (2007)
12. Hoon, S., Farrar, C.R., Hemez, F.M., Shunk, D.D., Stinemat, D.W., Nadler, B.R., Czarnecki, J.J.: A Review of Structural Health Monitoring Literature: 1996-2001. *Structural Health Monitoring*. Los Alamos National Laboratory Report (2004)
13. Lingyun, Y., Schopf, J.M., Dumitrescu, C.L., Foster, I.: Statistical Data Reduction for Efficient Application Performance Monitoring. CCGRID (2006)
14. Zhang, Q.W.: Statistical damage identification for bridges using ambient vibration data, pp. 476–485. Elsevier (2006)



# Towards Smart Advisor's Framework Based on Multi Agent Systems and Data Mining Methods

Madjid Khalilian

**Abstract.** Data mining is a method for extraction hidden pattern from huge data sets. It can help us to use this information for predicting future and offer solution or suggest an option. On the other hand many applications need to know customer's behavior and suggest best option to them. These applications are distributed and can be implemented by multi agent system to use advantages of autonomous intelligent agents. In this paper we present an automated approach for a recommendation system using autonomous intelligent agent with data mining techniques. A case study has been implemented and examined under mentioned situations.

## 1 Introduction

Many applications need to have a smart advisor component to recommend best option for using and making decision by customers. Tourist, travel, product and reservation services (hotel, flight, car, video ...) can better work, if they offer option to customer. These applications usually have distributed environment (it means we have many offices). Distributed environments have own difficulties like communication, coordination and being up to date. on the other hand information has become the most valuable commodity in many applications. We are overwhelmed with an influx of data that is stored in distributed data repositories. Analyzing these data and extracting meaningful pattern demands the system having powerful analytical tool. The manual analysis of this data set is slow, expensive and highly subjective. Indeed, such manual data analysis is becoming impractical as data volumes grow exponentially. Some difficulties are list below:

- Efficiency is a very strict requirements in the case of development of a knowledge extraction system

---

Madjid Khalilian  
Islamic Azad University, KARAJ Branch, Iran  
e-mail: khalilian@kiaiu.ac.ir

- Sharing information is one of the most important requirements for knowledge processing system
- Extendibility in applications with data mining methods is emergent requirements
- Changing environments caused having dynamic system.

Nevertheless, these limitations can be avoided, since the agent paradigm distinguishes clearly agent theory, which provides concepts related to the agent field and agents architectures that offer specific solution. In addition the complexity of tasks that we are capable of automating and delegating to computers has grown steadily that leads us to intelligence definition. Delegation and intelligence imply the need to constructing system with properties like independency and efficiency. Multi agent systems (MAS) offer a number of advantages with respect to computer supported cooperative working, resource sharing, distributed computation, robustness, sharing of expertise.

Our main objective is proposing a framework to combine MAS and Data mining methods for achieving higher efficiency and avoiding mentioned problems against both MAS and DM. delimitation for our project is focusing on recommendation system and using classification for offering based on previous information.

In section 2 we review related works, section 3, 4 describe MAS and DM concepts, in section 5 we demonstrate our framework, section 6 is about a case study and finally we have conclusion and future work.

## 2 Related Works

Analyzing huge amount of data and extracting valuable pattern in many applications is interesting for researchers Ref[1] proposed a methodology for distributed data community and heterogeneous data repositories. Data mining methods has been used for individual activities. They try to enterprise data mining with MAS to achieve automation but they focus on specific application and organizing multiple agents to achieve user centric goal. Ref[2, 3] respect to combining MAS and DM. Ref[4] has a compression between single agent and multiple agents for text classification in terms of some criteria. Ref[5] described an extendable multi agent data mining system. They have concentrated on agent based on data classification. They proposed a framework for mining data based on MAS. It has been also leveraged JADE for implementation. For these purpose 8 classification algorithms have been used in context. Main problem for this research is limitation to classification algorithms. Diversity is not included in this research. Robustness is another weakness for this project. In Ref[6] demonstrated process mining in multi agent activity logs. They have shown that the potential of structured log analysis to gain more high level insight into behavior of multi robot system.

Ref[3] presented the developed and implemented distributed data mining technology architecture of the multi agents' software tool supporting this technology. It mentioned some key problems that are related to distribution of data sources. Ref [7]respect to problem from different point of view with other researchers. It

considered MAS for evaluating classification algorithms. Most researchers focus on finding novel algorithms for classification but it may be not suitable for different situations and data sets. It also proposed a solution for evaluating situation with distributed computing and multi agents' technology. It can help users to choose most suitable classification algorithms with less consuming time and cost. It can be studied for different methods in data mining.

Ref[8] used a framework for describing data mining process with multi agent system. It tried to use advantages of multi agents like intelligence to choose best algorithms and methods but there is no demonstration for their framework. Ref[9] proposed an immune multi agent system for network intrusion. For adapting dynamic changing network environment detector set can vary according to the changes of the network environment. It is a good solution for network security but using data mining can improve it. Ref[10] suggested a method for designing the attack scenario construction module of correlation function with in their framework(SATA). The main point of their approach is transferring the alerts data base into attack sequences to solve problems of multi agent and multi stage attack scenario construction. They have used mining frequent attack sequences as an important tool in their research. Ref[11] proposed a framework for data mining based MAS. Indeed, it is in opposite our goal. It means they want to use MAS for data mining purpose but we use data mining for MAS. They presented a new approach for data mining with considering component for mining as an agent. Conservation in time has been shown in their experiments on spatial data. Ref[12] developed a smart travel service advisor. They use semantic web and agent technology which help users in the decision making process before finalizing a service provider. Previous knowledge about customer's behavior has not been considered.

### 3 Multi Agent Systems (MAS)

An agent is a computer system that is capable of *independent* action on behalf of its user or owner (figuring out what needs to be done to satisfy design objectives, rather than constantly being told)[13]. A multi agent system is one that consists of a number of agents, which *interact* with one-another. In the most general case, agents will be acting on behalf of users with different goals and motivations and to successfully interact, they will require the ability to *cooperate*, *coordinate*, and *negotiate* with each other, much as people do.

#### 3.1 Overview of the System

Our proposed system has two main parts, agent and environment. Agent system also has some subsystems including:

- Sensor which see the environment for perception,
- Next function which is a responsible for changing the internal state regard to the new perceptions,

- Knowledge Base for classifying data based on their attributes and making the classification model,
- Action which is a set of suggestions based on data classes.

But, environment is a database of object' attributes, diagnosis and also given offer for them.

### 3.2 Scenario

First, we need to have a complete database of object' attributes, diagnosis and also given specific class for them. After that we need a classifier for classifying objects regard to the database as training dataset and make a classification model. In agent operation, it can see the object attributes through its sensors or by entering the object's attributes with user and then classifies it and finds the proper class and then suggest best option to a user as an agent action. Albeit, with any seeing situation and new perception of agent, next function changes the internal state of it and also helps the classifier to find the target class.

## 4 Data Mining

Classification is a form of data analysis that can be used to extract models describing important data classes[14]. It is a kind of data mining for recognizing hidden pattern. Classification is included two main steps:

- A model is built describing a predetermined set of data classes or concepts
- The model is used for classification.

Bayesian classifier are statistical classifiers can predict class membership probabilities based on Bayes theorem. Most advantages are easy to implement and good results obtained in most of the cases on the other hand most important disadvantage is class conditional independence assumption, therefore loss of accuracy. Besides practically, dependencies exist among variables e.g. hospitals: patients, Profile, age and family history etc. Symptoms: fever, cough etc., Disease: lung cancer, diabetes etc. Dependencies among these cannot be modeled by Naïve Bayesian Classifier.

## 5 Smart Advisor Framework

Our proposed framework is included four components:

- Registration: in this component information from customers and other objects is being registered.
- Reservation: this component up dates Data Base and base on recommendation option reserves target objects.

- Offer: base on data mining methods offer optimum option to customer.
- Data Base: all data is stored and used by other components.

In fact offer agent has main role in our framework for this purpose we can use Naïve Bayesian classifier. It assumes conditional independence of attributes with respect to the class. The Bayes rule is suitable for this purpose:

$$P\langle C_k | V \rangle = P(C_k) \cdot \prod_{i=1}^{\alpha} \frac{P\langle C_k | v_i \rangle}{P(C_k)} \quad (1)$$

The task of the learning algorithm is to use learning examples to estimate unconditional probabilities  $P(C_k)$ ,  $k=1 \dots m_0$ , and conditional class probabilities  $P\langle C_k | v_i \rangle$ ,  $k=1 \dots M_0$ , given the value  $v_i$  of attributes  $A_i, i=1 \dots a$ , for estimation of prior (unconditional) class probabilities, Laplas's law of succession is used:

$$P(C_k) = \frac{n(C_k) + 1}{n + m_0} \quad (2)$$

where  $n(C_k)$  is the number of learning examples from the class  $C_k$ ,  $n$  is the total number of learning examples, and  $m_0$  the number of classes (values of the  $A_0$  attribute). For estimation of class probabilities we act the same way.

In the naïve Bayesian classifier, all probability estimations are rather reliable and thus the danger of over fitting is relatively small. Another point in favor of the naïve Bayesian classifier is that, even when the conditional independence assumption is not entirely met. The class probability estimations usually differ enough that the independence assumption error does not change their order[15].

## 6 Case Study: Video Club System

In a video club, customers can find and rent their desired movies and video titles. In this regard, we want to design and implement the system that can help them to find and checkout video titles, record chosen movie titles for each customer and offer them new movies based on their interests, provide them a return date, allow them to reserve unavailable video titles, notification of arrival of reserved video titles by using Multi Agent system approach. Our system consisted of three agents, Checkout Agent, Offer agent, Reservation Agent, that each of them is responsible for a given scenario. This system has been implemented by JADE and run under ECLIPS. It is clear that we have autonomous agent so they can run completely individual and caused speed up, extendibility, flexibility, and robustness.

## 7 Conclusion and Future Work

Combining advantages of MAS and DM can help us in both sides of these technologies. When we use MAS for DM we will avoid limitations in DM processing

e.g. speed up, robustness, scalability... On the other hand with DM methods we will be able to have the intelligent agents. Many directions exist to improve and extend proposed framework. Different applications can be used and examined framework. Other methods in DM can be applied in proposed framework.

## References

- [1] Zaidi, S.Z.H., Abidi, S.S.R., Hashmi, Z.I.: Engineering of Multi-Agent Systems to Effectuate Distributed Data Mining Activities
- [2] Kargupta, H., Hamzaoglu, I., Stafford, B.: Scalable, distributed data mining using an agent based architecture, pp. 211–214 (1997)
- [3] Gorodetsky, V., Karsaeyv, O., Samoilov, V.: Multi-agent technology for distributed data mining and classification, pp. 438–441 (2003)
- [4] Peng, S., Mukhopadhyay, S., Raje, R., Palakal, M., Mostafa, J.: A comparison between single-agent and multi-agent classification of documents, pp. 935–944 (2001)
- [5] Albashiri, K.A., Coenen, F., Leng, P.: EMADS: An extendible multi-agent data miner. Knowledge-Based Systems (2009)
- [6] Rozinat, S.Z.A., Veloso, M., van der Aalst, W.M.P., McMillen, C.: Analyzing Multi-agent Activity Logs Using Process Mining Techniques
- [7] Zhuang, F., He, Q., Shi, Z.: Multi-Agent based automatic evaluation system for classification algorithms, pp. 264–269 (2008)
- [8] Rajan, J., Saravanan, V.: A Framework of an Automated Data Mining System Using Autonomous Intelligent Agents, pp. 700–704 (2008)
- [9] Wang, D.G., Li, T., Liu, S.J., Liang, G., Zhao, K.: An Immune Multi-agent System for Network Intrusion Detection, pp. 436–445 (2008)
- [10] Huang, S., Li, Z., Wang, L.: Mining Attack Correlation Scenarios Based on Multi-agent System. In: Smith, M.J., Salvendy, G. (eds.) HCII 2007. LNCS, vol. 4557, pp. 632–641. Springer, Heidelberg (2007)
- [11] Baazaoui Zghal, S.F.H., Ben Ghezala, H.: A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data. In: Proceedings of World Academy of Science, Engineering and Technology, vol. 5, pp. 22–26 (April 2005)
- [12] Tariq, M., Kumar, V., Khoja, S., Chowdhry, B., Khan, M.K.: Smart travel service advisor using Semantic Web and Agent Technology. Computers and Simulation in Modern Science 1, 126–131 (2008)
- [13] Wooldridge, M.: An Introduction to MultiAgent Systems. John Wiley & Sons (2002)
- [14] Han, J., Kamber, M.: Data mining concepts and techniques, 2nd edn. Morgan kaufmann (2006)
- [15] Kononenko, I., Kukar, M.: Machine learning and data mining. West sussex, Horwood (2007)

# Automatic LSA-Based Retrieval of Synonyms (for Search Space Extension)

Kamil Ekštein and Lubomír Krčmář

**Abstract.** This paper describes a research, experiments, and theoretical considerations leading towards automatic computational thesaurus construction based upon identification of synonyms in large sets of texts for the needs of question-answering (QA) systems. The method benefits from and is founded on Latent Semantic Analysis (LSA) technique. LSA serves as a hypothesis generator which produces hypotheses about the words that might be synonyms. Subsequently, the generated hypotheses are proven right or wrong by means of examination of morphologic bindings between the two words and of the overall syntactic structure of the context in which they appear, namely the subject-object relation. The retrieved synonyms are used to extend the search space where a QA system mines the answers.

## 1 Introduction

In the course of the last few years, extensive attention has been aimed to research in fields of web search, question answering (QA) and generally information retrieval from large collections of distributed textual data, simply speaking from the web. The motivation of such a research is obvious: The web nowadays presents nearly ultimate source of global knowledge. Millions of people of various levels of education and skills contribute to its highly linked and structured content. Thanks to it, the most of pieces of information available from the web are usually highly redundant and thus it is easy to assess their validity using simple statistics-based techniques.

---

Kamil Ekštein

Laboratory of Intelligent Communication Systems, Dept. of Computer Science and Engineering, University of West Bohemia, Plzeň, Czech Republic  
e-mail: [kekstein@kiv.zcu.cz](mailto:kekstein@kiv.zcu.cz)

Lubomír Krčmář

Text-Mining Research Group, Dept. of Computer Science and Engineering, University of West Bohemia, Plzeň, Czech Republic  
e-mail: [lkrcmar@kiv.zcu.cz](mailto:lkrcmar@kiv.zcu.cz)

As it is, vast majority—and maybe all—of the publicly available information is already (or will be in few next years) stored in the web. *The fundamental question is thus how to find the proper piece of information.*

Currently, the widely used method is to use a web search engine (like e.g. Google, Yahoo, Bing, etc.) and feed it with keywords that are expected to be found in the requested piece of information. However, the choice of the keywords is crucial and dramatically influences the result of such a search. Thus, such a way of searching is undoubtedly suboptimal in the case of both human-performed search and machine search via the search engine API<sup>1</sup>. If the keywords are estimated improperly, the requested information—although existing and available from the web—might not be found.

Due to the above mentioned shortcoming, a lot of effort has been spent on design and development of the so-called Question Answering (QA) systems. These systems accept a natural language question on certain topic instead of keywords. The question is syntactically and semantically processed and as a result of the processing, a set of keywords is generated. These generated keywords are entered into a “traditional” web search engine which provides a large set of documents containing the keywords. Simpler QA systems end up at this point and return the obtained documents ordered by some appropriate metric like e.g. PageRank. The more sophisticated QA systems go further on and try to analyze the content of the obtained documents in order to derive the actual answer to the originally entered question.

The most notable instance of the sophisticated QA system is the Wolfram|Alpha engine. The most powerful web search engines like e.g. Google and Bing also offer the possibility to enter the search phrase in a natural language, however, the result is not processed to derive the answer. Instead, the ranked documents are returned—thus classifying the mentioned search engines into the group of simpler systems.

Our research team at the Laboratory of Intelligent Communication Systems (LICS, <http://liks.fav.zcu.cz>) developed two experimental QA systems: (i) LASE (LICS Advanced Search Engine), publicly available for testing and evaluation at <http://liks.fav.zcu.cz:8180/LASE2/>, and (ii) more experimental QAS (Question-answering System). These systems are maximally semantically oriented. They apply statistical semantic analysis onto the input question to generate the set of keywords for traditional web search engines and consequently process the returned set of documents by means of template-based semantic analysis to retrieve the actual answer [1].

Considering the above depicted situation, it is obvious that any refinement of the keywords entered into the web search engine improves the precision and reliability of the provided answers because the QA system has a larger search space. The answers are synthesized from the data contained in the documents provided by a web search engine. The more documents there are available for the analysis and processing, the higher probability that the right answer is found.

A straightforward way to increase the probability that the correct piece of data is found, is to increase the overall number of the documents provided by the web

---

<sup>1</sup> Application Programming Interface. Mostly RPC (Remote Procedure Call) in the case of web search engines.



search engine. It can be done by entering more than one keyword with the same or similar meaning, i.e. *entering the keyword and its synonyms*<sup>2</sup>.

**Example:** The question “When the president of the U.S.A. was born?” is semantically processed into the following set of keywords  $S_1 = \{\text{president, U. S. A., date of birth}\}$ . Google returns approx. 1,960,000 documents within 0.21 sec (the exact search phrase was ["president"] ["U.S.A."] ["date of birth"]). The fourth from the top contains the necessary information in a suitable form that can be used to synthesize the answer.

When an extended set  $S_2 = \{\text{president, Barack Obama, U.S.A, USA, date of birth, birthdate, birthday}\}$  with synonyms is entered, Google returns 43,900,000 documents within 0.13 sec, i.e.  $22.4\times$  more documents in half the time.

The example shows the benefit of using the synonyms in the search phrases generated by the semantic analysis modules of QA systems. The goal of the research behind this paper is thus to be able to generate the extended keyword sets using an automatically built thesaurus and extend the search space this way.

## 2 Related Work in Synonyms Retrieval

There exists a number of algorithms for synonyms recognition based on the definition of the task given by Turney [6]. However, these algorithms are designed and optimized for best performance on the TOEFL<sup>3</sup> synonyms test-like tests, as they are widely used to assess the systems. The TOEFL synonyms test is a multiple-choice answer test where the task is to mark one word out of four provided which is the synonym to a given word. The best-rated answer is the selection of a word that is the closest (by means of an expert opinion) to the given word.

Generally, there are two major approaches to the problem: The paradigm based upon the attributional similarity as shown in [4] and others, and the paradigm based upon a definition of ad hoc similarity measures [6]. Moraliyski and Dias [5] provide a brief summary of the performance of the most notable works on synonyms recognition—see table 1. In other words the mentioned systems take the provided candidate answers and try to prove the hypothesis about the synonymy to the given word right or wrong by various means, including the web mining.

However, the task in our situation is not to find the best matching synonym out of a provided set but to generate as many synonyms as possible in order to extend the search and obtain more documents related to the entered question. This is a considerable difference to the systems available for comparison.

---

<sup>2</sup> The designation “synonym” is probably not entirely correct from the linguistic point of view in the scope of this paper, however. It should better be “alternative keyword” or alike.

<sup>3</sup> Test of English as Foreign Language. Standard test of English language proficiency for foreigners that want to work or study in an English-speaking environment.

**Table 1** Accuracy of synonyms recognition systems on TOEFL test (after [5]).

Work	Best result	Work	Best result
Landauer and Dumais, 1997	64.40%	Terra and Clarke, 2003	81.25%
Sahlgren, 2001	72.00%	Elhert, 2003	82.00%
Turney, 2001	73.75%	Freitag et al., 2005	84.20%
Jarmasz and Szpakowicz, 2003	78.75%	Turney et al., 2003	97.50%

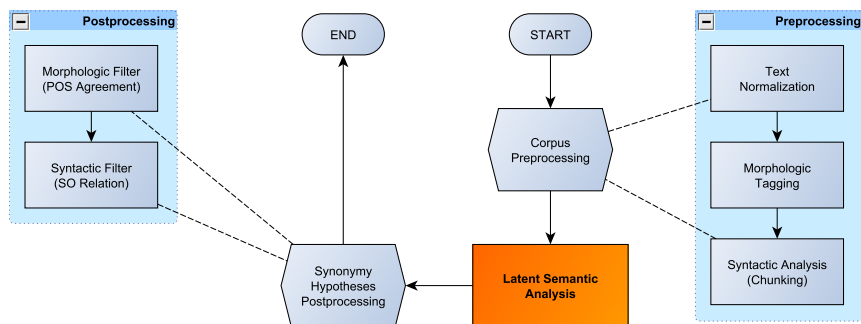
### 3 Thesaurus Building—The Method

The projected method of automatic thesaurus building uses Latent Semantic Analysis (LSA) as a foundation. Landauer and Dumais in [4] showed that “the accuracy of LSA is statistically indistinguishable from that of a population of non-native English speakers on the same questions” [5] (meant the questions from TOEFL synonyms test). In the other words, LSA itself should be capable of picking up a correct synonym to the input word in reasonably complicated cases. It, in our opinion, sufficiently justifies the use of LSA as a basic part of the projected synonyms retrieval system.

The course of the proposed algorithm is shown in Fig. 1 and comprises of the following steps: Before the LSA is applied the documents used to retrieve the synonyms (corpus) are preprocessed. The preprocessing consists of text normalization and morphological analysis.

During the text normalization phase, any non-textual information is removed from the document including particularly HTML tags, formatting sequences, etc. Also the coding of the text is normalized to UTF-8 (as the UTF-8 is used as an internal coding in Java, the programming language the LSA library is written in).

Consequently, the corpus documents are morphologically tagged so that each word is marked with a 15-position tag (details can be found in [2]) containing the

**Fig. 1** The flowchart of the developed algorithm.

part-of-speech (POS) indication and other morphological categories. The morphological tagging is performed by Hajič’s Feature-based tagger [2].

The preprocessed documents are passed into the LSA module. It is deployed in a standard way and returns a reduced term-term matrix, i.e. each element represents a correlation value between two terms. Where the value surpasses a predefined threshold<sup>4</sup>, the two terms are included into the set of synonyms hypotheses. It results in a list of word pairs preceded with a correlation value in the form shown in tab. 2 (the example comprises of 5 randomly chosen entries from the resulting file).

**Table 2** Example of the LSA module output—in Czech (left) and translated to English (right).

Score	Pairs (in Czech)	Pairs (translated)
0.809796	jurisdikce ↔ nástupnictví	jurisdiction ↔ succession
0.836097	nádor ↔ rakovina	tumour ↔ cancer
0.806359	seminář ↔ zámek	seminar ↔ castle
0.881768	kancléř ↔ Kohl	chancellor ↔ Kohl
0.833124	plyn ↔ narkóza	gas ↔ anaesthesia
...	...	...

The LSA code is taken from the S-Space Package developed and maintained by the Natural Language Processing group at UCLA [7].

Due to the properties of LSA, the provided pairs of words demonstrate certain semantic relation, however, the relation is a superset of synonymy. In most of the cases the relation is either more loose or vague. Nonetheless, the synonymy relation is naturally included as the expressions of the IS-A relation<sup>5</sup> are very common in most of the texts.

It is thus necessary to filter out the pairs that prove more loose or vague relation than synonymy. The last stage of the method is the filtering: At first, all hypotheses are rejected where the two words are not of the same POS type. It reflects the property of natural languages where the synonyms are of the same POS type in vast majority of cases. This step drastically reduces the hypothesis set potency.

Even if the hypothesis set is now significantly smaller, it still contains pairs that are not synonyms. To further refine the selection, a simple syntactic test is applied: The two words must have a subject-object (SO) relation, either direct or indirect (transitive), i.e.

$$W_1 \xleftrightarrow{\text{SO}} W_2,$$

<sup>4</sup> The threshold is set heuristically in order to balance the quality and the number of hypotheses.

<sup>5</sup> The IS-A relation between two words means that one of the words is a more generic denomination of the other, like in e.g. “dog **is an** animal”.

or in the transitive form

$$W_1 \overset{\text{SO}}{\longleftrightarrow} W_3 \wedge W_2 \overset{\text{SO}}{\longleftrightarrow} W_3, W_1 \overset{\text{SO}}{\longleftrightarrow} W_4 \wedge W_2 \overset{\text{SO}}{\longleftrightarrow} W_4, \dots$$

From the above shown transitive form it can be concluded that  $W_3$  and  $W_4$  are (likely) synonyms. The transitive form can be without generality detriment extended to arbitrary length.

The detection of synonyms from the transitive SO relation is programmatically procured in a very simple way: Each generated hypothesis  $W_a \overset{\text{SO}}{\longleftrightarrow} W_b$  is used to constitute a linked list (at the beginning with only the head  $W_a$  and the second member  $W_b$ ). When a next pair is found where either  $W_a$  or  $W_b$  is contained, the second member of the pair is chained to the existent linked list where the first member already resides. The clauses used to verify the SO relation are reliably identified by a syntactic chunker (partial parser).

## 4 Case Study: Synonyms Retrieval from PDT 2.0-Based Corpus

Laboratory of Intelligent Communication Systems develops primarily NLP algorithms and techniques for the Czech language<sup>6</sup>. The synonyms retrieval task was also targeted to the Czech language, however, the algorithms and techniques are generally language-independent.

Verification and testing of the above described synonyms retrieval algorithm was carried out on the Prague Dependency Treebank 2.0<sup>7</sup> [3]. PDT 2.0 is a top Czech computational linguistics project and contains a large number of representative annotated texts suitable to serve as a corpus for the synonyms retrieval task.

We used 800 documents from PDT 2.0 and fed them to the developed system. After the above described preprocessing, the data underwent the LSA. The resulting pairs of words—synonym hypotheses—were morphologically filtered to agree in POS tag. The selected POS was noun and personal name. As a result, 470 pairs of nouns and personal names were produced.

A native Czech speaker, linguist expert, marked 18 of them as correctly found synonyms. Thus, LSA followed by POS agreement filtering reaches rather poor accuracy of 3.83%. After deploying the direct SO relation filtering onto the previously obtained 470 pairs, the resulting file contains only 15 pairs. 14 of them are approved by the expert to be synonyms. Such a result can be observed (depending on the point of view) as 77.7% accuracy (because 14 of 18 synonym pairs contained in the data were identified) or 93.3% (as 14 pairs out of 15 produced are correct synonyms). For the web search optimization task, for which the algorithm was originally developed, the more optimistic result can be considered.

<sup>6</sup> This is also due to the fact that the laboratory also co-operates with one of the largest Czech web search providers.

<sup>7</sup> Detailed information on the PDT 2.0 project are available online and can be found at <http://ufal.mff.cuni.cz/pdt2.0/>.

When deploying the transitive version of the SO relation filtering, the algorithm produces 5 more synonym pairs. 3 out of these 5 pairs are correct. In total, 20 candidate pairs are provided and 17 are correct—the overall accuracy is thus 85%.

## 5 Conclusion

Thorough examination of the data computed by the LSA algorithm reveals that the extent of the corpus is not completely sufficient. The actual number of correct synonym pairs (those found in the data by an expert) is quite low (18 out of 470 hypotheses). Therefore, the resulting performance must be apprehended rather as preliminary values. However, the proposed algorithm proved its ability to retrieve synonyms from a corpus data and to create a thesaurus automatically. Using the available PDT 2.0-based corpus the resulting thesaurus is very small—20 entries (17 correct). Thus, it cannot be used for the task the algorithm was originally developed. However, if more corpus data were available, the resulting performance is promising.

**Acknowledgements.** The research was supported by the grant of Ministry of Education, Youth and Sports of the Czech Republic No. 2C06009, COT-SEWing, and UWB grant SGS-2010-028 Advanced Computer and Information Systems. The access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is highly appreciated.

## References

1. Konopík, M., Rohlík, O.: Question Answering for Not Quite Semantic Web. In: Proc. of 13th International Conference on Text, Speech and Dialogue TSD 2010, Brno, Czech Republic. Springer (2010)
2. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech), Prague, Czech Republic. Charles Univeristy Press, Karolinum (2004)
3. Hajič, J., Böhmová, A., Hajičová, E., Vidová Hladká, B.: The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: Abeillé, A. (ed.) Treebanks: Building and Using Parsed Corpora, pp. 103–127. Kluwer, Amsterdam (2000)
4. Landauer, T.K., Dumais, S.T.: A solution to Platós problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2), 211–240 (1997)
5. Moraliyski, R., Dias, G.: Combination of Global and Local Attributional Similarities for Synonym Detection (2007), <http://www.di.ubi.pt/~ddg/publications/Pliska2007.pdf>
6. Turney, P.D.: Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In: Flach, P.A., De Raedt, L. (eds.) ECML 2001. LNCS (LNAI), vol. 2167, pp. 491–502. Springer, Heidelberg (2001)
7. Jurgens, D., Stevens, K.: The S-Space Package: An Open Source Package for Word Space Models. System Papers of the Association of Computational Linguistics. University of California Los Angeles, Los Angeles (2010)

# Research on Method of Wavelet Function Selection to Vibration Signal Filtering

Xiangzhong Meng and Jianghong Wang

**Abstract.** The wavelet multi-resolution analysis is very suitable for the analysis and processing of vibration signal. During the vibration signal filtering, the different wavelet function had immediate effect on filtering; as the different vanishing moment, the support length and filter length are different which caused to the different filtering efficiency. The wavelet function selection method is put forward included the waveform of wavelet function, the vanishing moment of wavelet function and the curvilinear smoothness of filtered signal. Emulated with the vibration signal of rotor experiment platform as an example, selected Db6 as filtering wavelet function, filtering processed the vibration signal, and the effect of signal filter was very obviously.

## 1 Introduction

The Mallat signal decomposition and reconstruction algorithm based on wavelet multi-resolution analysis (WMA) is very suitable for the signal analysis and processing. The wavelet function is different, the curvilinear smoothness and vanishing moment are different. The different vanishing moment, the support length and filter length are different which caused to the different filtering efficiency. The wavelet function selection method is put forward including the filtered signal curve smoothness, the wavelet function waveform, and the wavelet function vanishing moment. Targeting to match the signal characteristics, Aleksandra set up the rule of selection wavelet basis so that wavelet functions to adapt to the trend of the signal. Ding ailing gave the construction algorithm of optimal match the wavelet basis function used the structured wavelet filter bank and the new waveform similarity rule.

---

Xiangzhong Meng

School of Optoelectronic Engineering, Xi'an Technological University, Xi'an, China  
e-mail: mx.xiangzhong@163.com

## 2 Signal Decomposition and Reconstruction Based on Wavelet Multi-resolution

In 1986, Mallat and Meyer proposed the concept of multi-resolution, leading wavelet theory to produce a breakthrough. Mallat put forward the signal decomposition and reconstruction algorithm at the same time.

As to WMA, it is to the square integrable function  $f(t) \in L^2(\mathbb{R})$  as the limit of gradual approach, which each level is the results to smooth the function  $f(t)$  by the low-pass smoothing function  $\varphi(t)$ , and the function  $\varphi(t)$  make step by step expansion, that is, sequential analysis the function  $f(t)$  used different resolution.

Given the WMA  $\{V_m\}_{m \in \mathbb{Z}}$  of  $L^2(\mathbb{R})$ ,  $\varphi(t)$  is scale function, defined the sequence  $\{g_k\}_{k \in \mathbb{Z}} \in l^2(\mathbb{Z})$ , and  $g_k = (-1)^k \bar{h}_{1-k}$ ,  $h_{(j-2k)} = \langle \varphi_{j+1,n}(t), \varphi_{j,k}(t) \rangle$ ,  $g_{(j-2k)} = \langle \varphi_{j+1,n}(t), \Psi_{j,k}(t) \rangle$ ,  $\frac{1}{\sqrt{2}} \Psi(\frac{t}{2}) = \sum_k g_k \varphi(t-k)$ , and signal decomposition and reconstruction algorithms is as follows.

Mallat signal decomposition algorithm,

$$x_k^{(j)} = \sum_n \bar{h}_{(j-2k)} x_n^{(j+1)}, d_k^{(j)} = \sum_n \bar{g}_{(j-2k)} x_n^{(j+1)} \quad (1)$$

In which  $x_k^{(j)}$  is signal decomposed low frequency,  $d_k^{(j)}$  is signal decomposed high frequency,  $\bar{h}_{(j-2k)}$  is scale factor of WMA,  $\bar{g}_{(j-2k)}$  is wavelet coefficients of WMA,  $n=0,1,2,\dots,k-1$  is the number of input sequence,  $j$  is the scale of wavelet signal decomposition (there,  $x_k^{(0)}$  is the original signal when  $j=0$ ).

Mallat signal decomposition diagram is shown in Fig. 1(a). The signal is decomposed into different frequency bands by a band pass filter. The decomposition scale factor is the low-pass filter coefficients, and the wavelet coefficients is just high-pass filter coefficients,  $x_k^{(j)}$  and  $d_k^{(j)}$  are the approximation and error signal of  $x_n^{(j+1)}$  respectively. Filtered the input signal by  $h_k$  or  $g_k$ , sampled the filter output signal, and then the resolution of the signal obtained is only half of the input signal.

Mallat signal reconstruction algorithm

$$x_k^{(j+1)} = \sum_n h_{(k-2n)} x_n^{(j)} + \sum_n g_{(k-2n)} d_n^{(j)} \quad (2)$$

Mallat reconstruction diagram is shown in Fig. 1(b). Interpolated the approximation signal and error signal, filtered the input signal by  $h_{(k-2n)}$  or  $g_{(k-2n)}$ , and then, added the filter output signal. The resolution obtained is reduced.

The original signal is decomposed into low and high frequency signal, then, the low-frequency signal is decomposed into low and high frequency signal again. During the decomposition, the signal data is half and the sampling interval is double of the former layer decomposition signal. At last, the multi-scale wavelet decomposition is completed. The various frequencies that wavelet decomposed is equivalent to the original signal time-domain analysis through the filter and reduce

of sampling points, and the resolution is inevitable declined. Then, the wavelet reconstruction is done, and the resolution of the signal is reached to the original signal. The linear phase filter is adopted during the reconstruction, so that, the signal reconstruction error is reduced and reserved the advantage of the good coherence between the orthogonal wavelet bands, and there is no redundancy information between the wavelet coefficients of each scale. So, the original signal can be accurately reconstruct by using FIR.

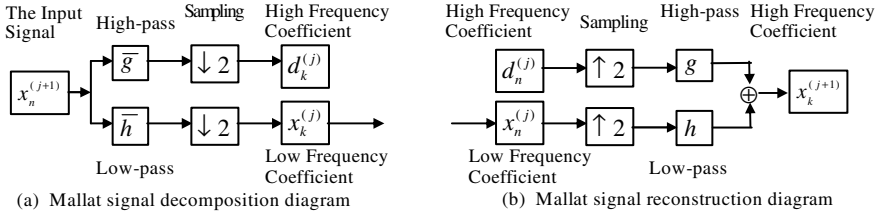


Fig. 1 Mallat signal decomposition and reconstruction diagram.

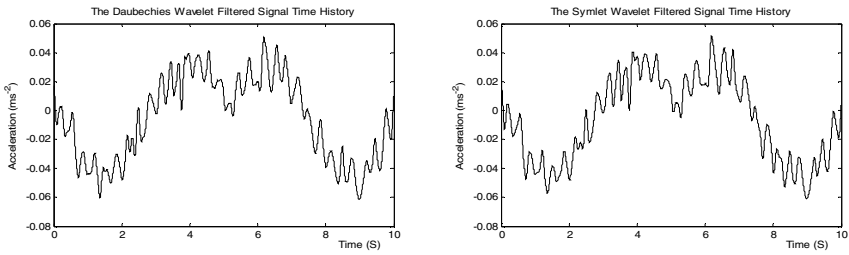
### 3 The Method of Wavelet Function Selection to Vibration Signal Filtering

It is very important to consider both the signal curve smoothness and the filtering efficiency. The wavelet function is different, the curvilinear smoothness, and vanishing moment and filtering effect are totally different. The wavelet family is inappropriate, the filtered signal curve is not smooth and more peaks, and caused to reduce the real-timeliness of active vibration control system. On the other hand, the vanishing moment is different, the support length and filter length are different, and obviously, the filtering efficiency is different. As thus, the selection of wavelet function is decided by the wavelet family and the vanishing moment during vibration signal filtering.

During the signal filtering process, the filter waveform should be as much as possible consistent with the signal waveform. As the vibration signal is superposed by a series of simple harmonic waves, the wavelet function waveform should be as much as possible similar to the harmonic waveform. The wavelet function waveform of Db3 is almost exactly same as Sym3, but there is very large difference between the waveform of Db6 and Sym6. It is obviously, the waveform of Db6 is more similar to the harmonic.

The Daubechies wavelet family and Symlet wavelet family are more suitable signal filtering according to the preliminary experiments and the related references. Selected the wavelet function of Daubechies wavelet family and Symlet wavelet family, the vanishing moments, the decomposition scale and the threshold method are same, filtered the vibration signal of rotor experiment platform as an example. The wavelet filtered signal curves are shown in Fig. 2.





**Fig. 2** The wavelet filtered signal curves

It is shown in Fig. 2 that the Daubechies wavelet filtered signal curve is smoother than the Symlet wavelet with same vanishing moment. In active vibration control system, Selected the Daubechies wavelet family, the output signal and the control effect are steadier.

During the vibration signal decomposition and reconstruction, the vanishing moment of the same wavelet family is different, the entropy values are different. The smaller entropy value means the signal decomposition and reconstruction efficiency, that is the filtering efficiency is higher, and more to meet the real-timeliness requirement of active vibration control system. The vibration signal decomposition and reconstruction entropy values of Daubechies and Symlet wavelet are shown in Tab. 1.

**Table 1** Entropy Values of Daubechies and Symlet Wavelet

Vanishing Moment	Wavelet Function	Entropy Values
4	Daubechies	7.5992
	Symlet	7.5953
5	Daubechies	7.5732
	Symlet	7.5856
6	Daubechies	7.5425
	Symlet	7.6030
7	Daubechies	7.5833
	Symlet	7.6621
8	Daubechies	7.5936
	Symlet	7.7234

It is shown in Tab. 1 that, on the whole, the vibration signal decomposition and reconstruction entropy values of Daubechies wavelet are smaller than the Symlet wavelet with the same vanishing moment, and the entropy value of Db6 wavelet is the smallest.

In summary, during the vibration signal filtering process, selected the Daubechies wavelet family which the vanishing moment is 6, that is, the wavelet function Db6 as filter. The wavelet function Db6 is orthogonal and compactly supported, has extreme value phase and the maximum order of vanishing moments in the given support space. So, the wavelet function Db6 is very suitable for the real-time filtering of vibration signal.

### 4 The Vibration Signal Filtering Based on Wavelet Multi-resolution Analysis

The vibration signal filtering based on the wavelet multi-resolution analysis as follows: firstly, wavelet multi-scale decomposed the vibration signals: Daubechies wavelet function is adopted, and a 5-scale multi-resolution analyzed; secondly, threshold quantification processed the high-frequency signal approach to mandatory filtering. And finally, reconstructed the vibration signal using the low and high frequency coefficients according to the wavelet multi-resolution analysis, and the reconstructed signal is the filtered signal.

The wavelet transform and inverse transformation of the wavelet function Db6 can be calculated with limited discrete convolution, which is very suitable for embedded systems programming calculations online.

In order to verify the accuracy of wavelet decomposition and reconstruction, during the signal reconstruction, all of the low and high frequency component are preserved. The vibration signal and the wavelet reconstructed signal are shown in Fig. 3.

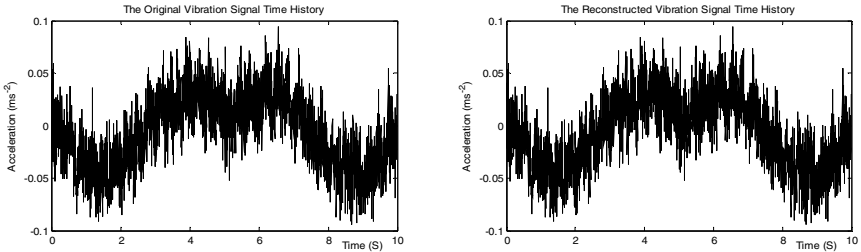


Fig. 3 The vibration signal and the wavelet reconstructed signal

Quantified the high-frequency signal coefficients by threshold, reconstructed the vibration signal, the reconstructed signal is the wavelet filtered signals. The original vibration signal and the wavelet Db6 filtered signal are shown in Fig. 4.

During the active vibration control process, the low frequency components of wavelet decomposition can be used directly as the input signal, which is very effective to the effectiveness and efficiency.

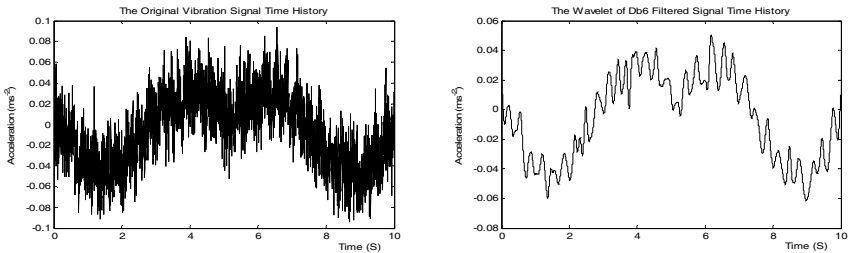


Fig. 4 The original vibration signal and the wavelet filtered signal

## 5 Closing

Selected an appropriate wavelet function, achieved a higher signal to noise ratio, reduced waveform distortion, enhanced the calculation efficiency, and easy to the vibration active control real-time processed. The selection method of wavelet function is researched including the filtered signal curve smoothness, the wavelet function waveform, and the wavelet function vanishing moment. The wavelet transform and inverse transformation of the wavelet function Db6 can be calculated with limited discrete convolution, which is very suitable for embedded systems programming calculations online. Selected Db6 wavelet function, emulated with the vibration signal of rotor experiment platform as an example, the vibration signal filtering realized successfully, and the effect of signal filter is very obviously.

## References

- [1] Cohen, A., Daubechies, I., Feauveau, J.: Biorthogonal bases of compactly supported wavelets. *Commun. Pure Appl.* 45(3), 485–560 (1992)
- [2] Ding, A., Shi, G., Zhang, N., Jiao, L.: Signal Compression and Design of Wavelet Based on Waveform Matching. *Journal of Electronics & Information Technology* 29(4), 804–807 (2007)
- [3] Xia, D., Zhou, B., Wang, S.: Application of wavelet's real-time filter in silicon micro-machined gyroscope. *Journal of Chinese Inertial Technology* 15(1), 92–95 (2007)
- [4] Kugarajah, T., Zhang, Q.: Multidimensional Wavelet Frames. *IEEE Transactions On Neural Network* 6(6), 1552–1556 (1995)
- [5] Guan, S., Shi, X.: Fabric Defect Detection Based on Wavelet Lifting Scheme. *Computer Engineering and Applications* 44(25), 219–221 (2008)
- [6] Feng, G., Li, J., Liu, P.: Noise and Vibration Source Identification in Light Bus Based on Wavelet. *Journal of Vibration, Measurement & Diagnosis* 26(3), 196–199 (2006)

# Time Series Subsequence Matching Based on Middle Points and Clipping

Nguyen Thanh Son and Duong Tuan Anh

## 1 Introduction

The similarity search on time series data is classified into two classes: whole sequence matching and subsequence matching. In whole matching, it is supposed that the time series to be compared have the same length. In the subsequence matching, the result of this problem is a consecutive subsequence within the longer time series that best matches the query sequence. Many solving methods have been proposed for time series similarity search problem. The most promising one is the method that carries out similarity search in two steps: reducing on the dimensionality of time series data, then indexing the reduced data with a multidimensional index structure.

Several techniques have been proposed in the literature for time series dimensionality reduction, such as the Discrete Fourier Transform (DFT) ([4]), Piecewise Aggregate Approximation (PAA) [6], and the methods that base on important points such as Extreme Points [3] and Perceptually Important Points (PIP) ([2]). In addition, Ratanamahatana et al, 2005 [8] proposed an intuitive and effective method that transforms time series into sequence of bits. This bit level time series representation, called *clipping* method, has several major advantages over existing methods. First, the clipped distance is less than or equal to the Euclidean distance. Secondly, the bit level representation can be efficiently stored and manipulated. Thirdly, the bit level representation allows us to use many available algorithms which are applicable to binary data only. However, this clipped representation has two major shortcomings. First, it does not allow user to choose the reduction ratio since the data itself dictates the reduction ratio and the

---

Nguyen Thanh Son

Faculty of Information Technology

Ho Chi Minh City University of Technical Education, Vietnam

Duong Tuan Anh

Faculty of Computer Science and Engineering,

Ho Chi Minh City University of Technology, Vietnam

user has no choice to make. Secondly, it is not equipped with an index mechanism to support similarity queries.

In this paper, we introduce MP\_C, a new method for time series dimensionality reduction and a subsequence matching approach which uses this new bit-level time series representation. The MP\_C method is performed by dividing time series into segments, some points in each segment being extracted and then these points are transformed into a sequence of bits. In our method, we choose the points in each segment by dividing a segment into sub-segments and the middle points of these sub-segments are selected. Our new method not only preserves the virtues of the clipping method but also overcomes its two weak points. We can prove that MP\_C satisfies the lower bounding condition and make MP\_C indexable by showing that a time series subsequence compressed by MP\_C can be indexed with the support of Skyline index. Our experiments show that our MP\_C method is better than PAA in terms of tightness of lower bound and pruning power, and in subsequence matching, our MP\_C method with Skyline index can perform faster than PAA based on traditional R\*-tree.

## 2 Background

### 2.1 Index Structures for Time Series

The popular multidimensional index structures are R-tree and its variants ([1], [5]). In a multidimensional index structure (e.g., R-tree or R\*-tree), each node is associated with a minimum bounding rectangle (MBR). A MBR at a node is the minimum bounding box of the MBRs of its child nodes. A potential weakness in the method using MBR is that MBRs in index nodes can overlap. Overlapping rectangles could have negative effect on the search performance.

Skyline Index, another elegant paradigm for indexing time series data which uses another kind of minimum bounding regions, is proposed by Li et al., 2004 [7]. Skyline Index adopts new Skyline Bounding Regions (SBR) to approximate and represent a group of time series data according to their collective shape. An SBR is defined in the same *time-value* space where time series data are defined. Therefore, SBRs can capture the sequential nature of time series. SBRs allow us to define a distance function that tightly lower-bounds the distance between a query and a group of time series data. SBRs are free of internal overlaps. Hence using the same amount of space in an index node, SBR defined a better bounding region. For k-nearest-neighbor (KNN) queries, Skyline index approach can be coupled with some well-known dimensionality reduction technique and can improve its performance remarkably ([7]).

### 2.2 Clipping Method

The clipping method is a method for time series compression that transforms all points of a time series into a sequence of bits, according to the value of each point is above or below the average value.

Let a time series  $C = \{c_1, \dots, c_n\}$ . The point  $c_i$  is transformed into binary representation by the following formula:

$$c_i = \begin{cases} 1 & \text{if } c_i > \mu \\ 0 & \text{otherwise} \end{cases}$$

where 1 represents above the time series average  $\mu$  and 0 represents below.

### 3 MP\_C Representation

#### MP\_C – Approximation and Clipping

MP\_C method is based on the dividing the sequence into segments. Some points in each segment are extracted. Next, the chosen points are transformed into a sequence of bits representing whether each value is above or below the average of the corresponding segment. The mean of each segment and the bit sequence are recorded as segment features.

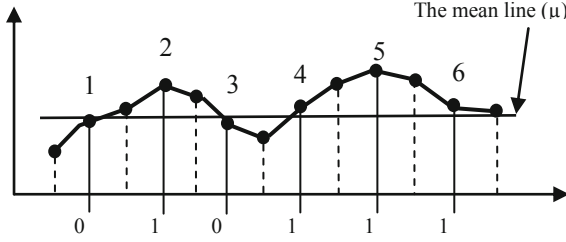
We can choose some points in each segment with different ways in time order. For example, in order to choose  $k$  points we can extract the first or the last  $k$  points of each segment and so on. In our study we use the following simple algorithm: (1) Dividing each segment into sub-segments, and (2) Choosing the middle points of the sub-segments. This brings out a clipped representation of middle points. Hence MP\_C has all the advantages of the previous bit level representation proposed by Ratanamahatana et al. [8] while it still allows the user to have a choice of compression ratio through determining the number of middle points chosen.

Given a time series subsequence  $C$  and a query  $Q$ , without loss of generality, we assume  $C$  and  $Q$  are  $n$  units long.  $C$  is divided into segments. We use the above algorithm to choose  $l$  middle points in each segment of  $C$ . Next, these middle points are transformed into a sequence of bits, where 1 represents above the segment average and 0 represents below, i.e., if  $\mu$  is the mean of segment  $C$ , then

$$c_i = \begin{cases} 1 & \text{if } c_i > \mu \\ 0 & \text{otherwise} \end{cases}$$

Figure 1 shows the intuition behind this technique, with  $l = 6$ . In this case, the sequence of bits 010111 and the  $\mu$  value are recorded.

In order to match the query  $Q$  with a time series  $C$  in the database, first  $Q$  is transformed into the same feature space as  $C$ . But the middle points of  $Q$  are not transformed into a sequence of bits since the clipped representation allows raw data to be directly compared to the clipped representation. Then we shift the segment mean lines in  $Q$  to meet those in  $C$  so that we can compare the similarity in shape between  $Q$  and  $C$ .



**Fig. 1** An illustration of MP\_C method

### Similarity Measure Defined for MP\_C

In order to guarantee no false dismissals we must produce a distance measure in the reduced space,  $D_{MP\_C}$  which is less than or equal to the distance measure in the original space.

**Definition 1 (MP\_C Similarity Measure)** Given a query  $Q$  and a subsequence  $C$  (of length  $n$ ) in raw data. Both  $C$  and  $Q$  are divided into  $N$  segments ( $N \ll n$ ). Suppose each segment has the length of  $w$ . Let  $C'$  be an MP\_C representation of  $C$ . The distance measure between  $Q$  and  $C'$  in MP\_C space,  $D_{MP\_C}(Q, C')$ , is computed as follows.

$$D_{MP\_C}(Q, C') = \sqrt{D_1(Q, C') + D_2(Q, C')} \quad (1)$$

$D_1(Q, C')$  and  $D_2(Q, C')$  are defined as

$$D_1(Q, C') = \sum_{i=1}^N w(q\mu_i - c\mu_i)^2 \quad (2)$$

$$D_2(Q, C') = \sum_{j=1}^N \sum_{i=1}^l (d(q_i, bc_i))^2 \quad (3)$$

where

$q\mu_i$  is the mean value of the  $i^{\text{th}}$  segment in  $Q$ ,  $c\mu_i$  is the mean value of the  $i^{\text{th}}$  segment in  $C$ ,  $bc_i$  is binary representation of  $c_i$ .

$d(q_i, bc_i)$  is computed by the following formula:

$$d(q_i, bc_i) = \begin{cases} q_i' & \text{if } (q_i' > 0 \text{ and } bc_i = 0) \text{ or} \\ & (q_i' \leq 0 \text{ and } bc_i = 1) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$q_i'$  is defined as  $q_i' = q_i - q\mu_k$ , where  $q_i$  belongs to the  $k^{\text{th}}$  segment in  $Q$ .

**Lemma 1.**  $D_{MP\_C}(Q, C') \leq D(Q, C)$  where  $D(Q, C)$  is the Euclidean distance between  $Q$  and  $C$ .

The proof of Lemma 1 can be seen in our previous work [9].

## 4 A Skyline Index for MP\_C

In this section, we describe how we can adopt Skyline index for MP\_C time series compression method.

### MP\_C Bounding Region

In traditional multidimensional index structure such as R\*-tree, minimum bounding rectangles (MBRs) are used to group time series data which are mapped into points in a low dimensional feature-space. If a MBR is defined in the two-dimensional space in which a time series data exists, the overlap between MBRs will be large. So by using the ideas from Skyline index, we can represent more accurately the collective shape of a group of time series data with tighter bounding regions. To attain this aim, we use MP\_C bounding regions (MP\_C\_BRs) for bounding a group of time series data.

**Definition 2 (MP\_C Bounding Region).** Given a group  $C'$  consisting of  $k$  MP\_C sequences in a  $N$ -dimensional feature space. The MP\_C\_BR  $R$  of  $C'$ , is defined as follows

$$R = (C'_{max}, C'_{min})$$

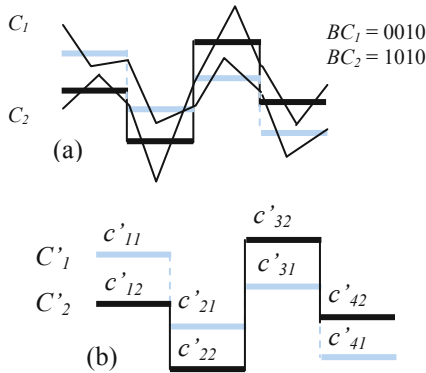
where

$$C'_{max} = \{c'_{1max}, c'_{2max}, \dots, c'_{Nmax}\}, C'_{min} = \{c'_{1min}, c'_{2min}, \dots, c'_{Nmin}\}$$

and, for  $1 \leq i \leq N$ ,

$$c'_{imax} = \max\{c'_{i1}, \dots, c'_{ik}\} \text{ and } c'_{imin} = \min\{c'_{i1}, \dots, c'_{ik}\}$$

where  $c'_{ij}$  is the  $i^{\text{th}}$  mean value of the  $j^{\text{th}}$  MP\_C sequence in  $C'$ .



**Fig. 2** An example of MP\_C\_BR. (a) Two time series  $C_1$ ,  $C_2$  and their approximate MP\_C representations in four dimensional space. (b) The MP\_C\_BR of two MP\_C sequences  $C'_1$  and  $C'_2$ .  $C'_{max} = \{c'_{11}, c'_{21}, c'_{32}, c'_{42}\}$  and  $C'_{min} = \{c'_{12}, c'_{22}, c'_{31}, c'_{41}\}$



Figure 2 illustrates an example of MP\_C\_BR. In this example,  $BC_i$  is a bit sequence of time series  $C_i$  and the number of middle points in each segment is one.

### Time Series Indexing Based on MP\_C\_BR

Now, we have to define the distance function  $D_{region}(Q, R)$  of the query  $Q$  from the MP\_C\_BR  $R$  associated with a node in the index structure such that it satisfies the group lower bound condition  $D_{region}(Q, R) \leq D(Q, C)$  for any time series  $C$  in the MP\_C\_BR  $R$ . The proof of this group lower-bound condition is given in our previous work [9].

We can index the MP\_C representation of time series data by first building a Skyline index which bases on a spatial index structure such as  $R^*$ -tree. Each leaf node in the  $R^*$ -tree contains a MP\_C sequence and a pointer refer to an original time series data in the database. The MP\_C\_BR associated with a non-leaf node is the smallest bounding region that spatially contains the MP\_C\_BRs associated with its immediate children.

We use a sliding window of size  $w$  to divide the time series  $C$  into subsequences of length  $w$  from  $C$ , extract  $k$  middle points from each subsequence and apply MP\_C transformation on each such subsequence. Then we store the features transformed from all such subsequences in Skyline index.

## 5 Experimental Evaluation

We compare our MP\_C using Skyline index to the popular method PAA based on  $R^*$ -tree. We also compare MP\_C to Clipping technique. Time series datasets for experiments come from various sources publicly available through the Internet and are organized into five datasets. The five datasets are EEG data, Economic data, Hydrology data, Production data and Wind data. The comparison among three methods MP\_C, PAA and Clipping is based on the tightness of lower bound, the pruning power and the implemented system. All the experimental results in terms of each of the three criteria show that MP\_C is better than PAA and Clipping.

## References

1. Beckman, N., Kriegel, H.P., Schneider, R., Seeger, B.: The  $R^*$ -tree: An Efficient and Robust Access Method for Points and Rectangles. In: Proc. of 1990 ACM-SIGMOD Conf., Atlantic City, NJ, pp. 322–331 (May 1990)
2. Chung, F.L., Fu, T.C., Luk, R., Ng, V.: Flexible Time Series Pattern Matching Based on Perceptually Important Points. In: Proc. of Int. Joint Conf. on Artificial Intelligence-Workshop on Learning from Temporal and Spatial Data, pp. 1–7 (2001)
3. Fink, E., Pratt, K.B.: Indexing of Compressed Time Series. In: Last, M., Kandel, A., Bunke, H. (eds.) Data mining in time series Databases, pp. 43–65. World Scientific, Singapore (2003)

4. Faloutsos, C., Ranganathan, M., Manolopoulos, Y.: Fast Subsequence Matching in Time-Series Databases. In: Proc. of ACM SIGMOD Int'l Conf. on Management of Data, Minneapolis, MN, May 24-27, pp. 419-429 (1994)
5. Guttman, A.: R-trees: a Dynamic Index Structure for Spatial Searching. In: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, June 18-21, pp. 47-57 (1984)
6. Keogh, E., Chakrabarti, K., Pazzani, M., Mehrotra, S.: Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. *J. Knowledge and Information Systems* 3(3), 263-286 (2000)
7. Li, Q., Lopez, I.F.V., Moon, B.: Skyline Index for Time Series Data. *IEEE Trans. on Knowledge and Data Engineering* 16(6) (2004)
8. Ratanamahatana, C., Keogh, E., Bagnall, A.J., Lonardi, S.: A Novel Bit Level Time Series Representation with Implication of Similarity Search and Clustering. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 771-777. Springer, Heidelberg (2005)
9. Son, N.T., Anh, D.T.: Similarity Search using MP\_C, a Combination of Middle points and Clipping. Technical Report, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (November 2010)

# Effects of Spatial Scale in Cellular Automata Model for Land Use Change

Guanwei Zhao

**Abstract.** Cellular automata (CA) is an efficient model to simulate land use and coverage change (LUCC) process. However, the spatial scale decisions of geographic cellular automata are often made arbitrarily. This article investigates the effect of changing cell size and neighborhood configuration on the result prediction accuracy of the CA-Markov model and the morphology of land use change simulation result. The research shows that spatial scale has great impact on the simulation results of CA-Markov model. Therefore, the selection of cell size must be careful. Neighborhood configuration also has impact on the simulation results of CA-Markov model.

## 1 Introduction

Land use and coverage change (LUCC), is the core of global change and sustainable development research, while urban area is one of the key research areas [1]. The process of LUCC is very complicated as there are many factors impact on the LUCC process, thus the role of various models which designed for understanding and predicting the land use dynamic is very important [2]. In recent years, Cellular Automata (CA) is a cool issue in spatial explicit modeling of urban LUCC. It is proved that cellular automata model can simulate and predict the dynamic of urban land use change more accurately compared with traditional model [3-4]. The last few years have seen a burst of publications in the field of land-use cellular automata (for short, LUCA) [5-13], with many applications demonstrating apparent correspondences between model and real world process. However, the majority of land-use cellular automata model researches now focus more on transition rules and model implementation than on spatial scale. As a result, the spatial scale decisions

---

Guanwei Zhao

School of Geographical Sciences, Guangzhou University, Guangzhou, China

e-mail: zhaogw@gzhu.edu.cn

of land-use cellular automata are often made arbitrarily or decided by data availability. Studies show that the spatial scale has significant impact on the result of land-use cellular automata [14], and the simulation result is close to the real situation only when the model in specific spatial scale range. For this reason, some scholars have begun to research the issue of spatial scale in LUCA [14-17].

The objective of this study is to investigate the effects of cell size and neighborhood configuration in LUCA. In a context in which models are increasingly built to study and eventually predict the dynamics of urban land use systems, it is crucial to acquire the best possible knowledge on the sensitivity of the model components.

## 2 Materials and Methodology

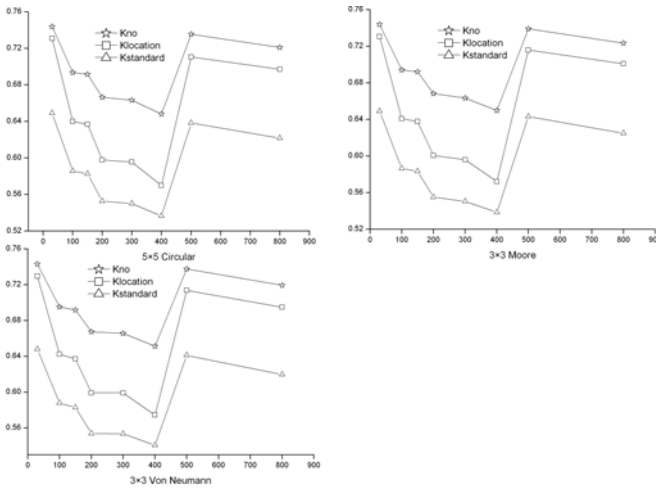
The study area chosen for this study is the Huadu district in Guangzhou city, china. The dataset used comprises three land-cover maps of the Huadu district, one for year 1995 and one for year 2000 and another for year 2005, derived from three Landsat-TM remote sensing images. The original spatial resolution of these images is 30 meters and the land-cover classes are farmland, orchard land, woodland, grassland, industry/ming land, public management and service land, transportation land, water and other land. The additional dataset includes administrative map of Huadu district in year 2005 and topographic map with the scale of 1: 50000 in year 2005.

To evaluate the effects of cell size and neighborhood configuration in LUCA, it is necessary to establish diverse cell size scenarios. Eight cell sizes were selected: 30m (original resolution), 100m, 150m, 200m, 300m, 400m, 500m, 600m, 700m, and 800m. These cell sizes cover the extent of the cell size spectrum commonly use in LUCA. The datasets at the seven nonoriginal resolutions were produced by nearest neighbor resample tool in ESRI Arcgis 9.3.

## 3 Results and Discussion

### 3.1 *Effects of Spatial Scale on Accuracy*

Figure 1 show that the kappa coefficient of simulation result decreased while the cell size increased from 30m to 400m. When the cell size increase to 500m, the kappa coefficient of simulation result increased unexpectedly and is higher than the value of 400m. Then, while the cell size increased to 800m, the kappa coefficient of simulation result decreased again, but its value still higher than the value of 400m. Actually, the accuracies of simulation results with the cell size larger than 400m were lower than the accuracy of 400m through comparison between simulation land cover map and actual land cover map. Thus it can be seen that kappa coefficient can not be simply used for evaluate the quality of simulation results. Therefore, we should pay more attention to the limit value of cell size while using CA-Markov model to simulate the land use change.



**Fig. 1** Kappa coefficients of simulation result with diverse cell sizes

From the neighborhood’s point of view, the kappa coefficient of simulation results generated by three neighborhood types has no significant difference. It shows that neighborhood type has no notable impact on the kappa coefficient of simulation results to some extent.

### 3.2 Effects of Spatial Scale on land Use Spatial Pattern

According to the results of landscape pattern analysis shows (see Table 1, Table 2 and Table 3), as the cell size increased, the patch number (NP), patch density (PD), landscape shape index (LSI), mean patch shape index (SHAPE\_MN) and mean patch fractal dimension index (FRAC\_MN) of simulation results decreased no matter what kind of neighborhood was chosen.

**Table 1** Landscape indices of simulation results (using 5×5 circular neighborhood)

Cell size(m)	NP	PD	LSI	SHAPE_MN	FRAC_MN
30	6581	3.019	44.610	1.629	1.073
100	5824	2.672	26.887	1.212	1.032
150	3600	1.648	20.671	1.191	1.026
200	2141	0.981	16.786	1.192	1.028
300	1389	0.640	13.293	1.170	1.023
400	573	0.263	9.291	1.199	1.026
500	361	0.166	8.241	1.276	1.030
800	211	0.098	7.833	1.277	1.026

**Table 2** Landscape indices of simulation results (using 3×3 Moore neighborhood)

Cell size(m)	NP	PD	LSI	SHAPE_MN	FRAC_MN
30	11803	5.415	45.226	1.372	1.051
100	6744	3.094	29.163	1.207	1.033
150	4132	1.892	22.668	1.190	1.028
200	2331	1.068	17.983	1.194	1.029
300	1466	0.676	14.154	1.178	1.025
400	657	0.302	10.348	1.205	1.027
500	509	0.234	10.856	1.236	1.026
800	256	0.119	8.295	1.238	1.025

**Table 3** Landscape indices of simulation results (using 3×3 Von Neumann neighborhood)

Cell size(m)	NP	PD	LSI	SHAPE_MN	FRAC_MN
30	13914	6.383	57.594	1.577	1.079
100	8127	3.729	33.746	1.213	1.034
150	4965	2.273	25.793	1.180	1.027
200	2774	1.270	20.513	1.194	1.029
300	1806	0.832	15.863	1.145	1.022
400	770	0.354	11.590	1.196	1.026
500	453	0.209	10.465	1.240	1.027
800	211	0.098	7.833	1.277	1.026

In terms of number of patch, when the cell size increased to 400m, the NP value of simulation result fell to below 1000. It can be found that the quality of simulation results is very poor through comparison between simulation map and actual map. NP value of simulation result using Von Neumann neighborhood is the largest one in three neighborhood types. From the patch density point of view, the value of PD declined while the cell size increased, suggesting that landscape fragment level of simulation result decreased. As the cell size increased to greater than 400m, value of LSI dropped to around 10. Landscape shape index of simulation results generated by Moore and Circular neighborhood was significant lower than the LSI of Von Neumann neighborhood. Mean patch size also showed a downtrend. As cell size increases, value of FRAC\_MN appeared slightly reduced, fluctuations between 1.07 and 1.02. Thus, for three neighborhood types, Von Neumann type can generate larger landscape indices than other two types.

## 4 Conclusions

This research explored the spatial scale effect in LUCA. This effect is very apparent in our results. LUCA simulation outcomes are both influenced by the choice made about cell size and neighborhood configuration. In our results, choosing a smaller cell size created a more accurate simulation result. When the cell size exceeds a threshold value, the simulation accuracy will decline rapidly. As the cell size exceeds a limit value, the simulation accuracy dropped continually while the kappa coefficient was still high. Therefore, the choice of cell size should be very careful when using CA model to simulate LUCC. The spatial scale analysis of LUCA model should be performed before model application, combined with consideration of the actual situation of study area.

Furthermore, the choice of a neighborhood configuration is less influential on simulation results. For the three neighborhood types, Von Neumann neighborhood can generate simulation results with higher number of patches and patch density than other two. But the kappa coefficient of simulation results using three neighborhood types has no significant difference. In addition, it can be seen that the number of patches and patch density of simulation results will decrease gradually as the neighborhood size increased through comparison between three neighborhood types.

The spatial scale sensitivity analysis does not remove the scale problem, but is the simplest way of limiting its effects. This is only the beginning. We can improve our works in various directions. For example, it is worth noting that the analysis undertaken does not include a common land use categories (urban and non-urban). Future development in this area should attempt to solve this problem.

**Acknowledgments.** The authors would like to thank the Guangzhou Land and Resources and the Housing Authority Huadu Branch for their support of this research. The authors also would like to thank reviewers for their comments that help improve the manuscript.

## References

1. Lambin, E.F., Turner, B.L., et al.: The causes of land-use and land-cover change: moving beyond the myths. *Global Environmental Change* (2001), doi:10.1016/S0959-3780(01)00007-3
2. Meyer, W.B., Turner, I.B.L.: *Change in land use and land cover: a global perspective*. Cambridge University Press, Cambridge (1994)
3. Batty, M.: Urban evolution on the desktop: simulation with the use of extended cellular automata. *Environment and Planning A* (1998), doi:10.1068/a301943
4. Batty, M., Xie, Y.C., Sun, Z.L.: Modelling urban dynamics through GIS-based cellular automata. *Computers, Environment and Urban Systems* (1999), doi:10.1016/S0198-9715(99)00015-0
5. Li, X., Yeh, A.G.: Modelling sustainable urban development by the integration of constrained cellular automata and GIS. *International Journal of Geographical Information Science* (2000), doi:10.1080/136588100240886

6. Li, X., Yeh, A.G.: Neural-network-based cellular automata for simulating multiple land use changes using GIS. *International Journal of Geographical Information Science* (2002), doi:10.1080/13658810210137004
7. Clarke, K.C., Hoppen, S., Gaydos, L.J.: A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. *Environment and Planning B: Planning and Design* (1997), doi:10.1068/b240247
8. Couclelis, H.: From cellular automata to urban models: new principles for model development and implementation. *Environment and Planning B: Planning and Design* (1997), doi:10.1068/b240165
9. Jantz, C.A., Goetz, S.J., Shelley, M.K.: Using the SLEUTH urban growth model to simulate the impacts of future policy scenarios on urban land use in the Baltimore-Washington metropolitan area. *Environment and Planning B: Planning and Design* (2003), doi:10.1068/b2983
10. Yang, X., Lo, C.P.: Modeling urban growth and landscape change in the Atlanta metropolitan area. *International Journal of Geographical Information Science* (2003), doi:10.1080/1365881031000086965
11. Torrens, P.M., O'Sullivan, D.: Cellular automata and urban simulation: where do we go from here? *Environment and Planning B: Planning and Design* (2001), doi:10.1068/b2802ed
12. Ward, D.P., Murray, A.T., Phinn, S.R.: A stochastically constrained cellular model of urban growth. *Computers, Environment and Urban Systems* (2000), doi:10.1016/S0198-9715(00)00008-9
13. Wu, F.L., Webster, C.T.: Simulation of land development through the integration of cellular automata and multicriteria evaluation. *Environment and Planning B: Planning and Design* (1998), doi:10.1068/b250103
14. Benenson, I.: Warning! The scale of land-use CA is changing! *Computers, Environment and Urban Systems* (2007), doi:10.1016/j.compenvurbsys.01.001
15. Kocabas, V., Dragicevic, S.: Assessing cellular automata model behaviour using a sensitivity analysis approach. *Computers, Environment and Urban Systems* (2006), doi:10.1016/j.compenvurbsys.2006.01.001
16. Jantz, C.A., Goetz, S.J.: Analysis of scale dependencies in an urban land-use-change model. *International Journal of Geographical Information Science*, doi:10.1080/13658810410001713425
17. Menard, A., Marceau, D.J.: Exploration of spatial scale sensitivity in geographic cellular automata. *Environment and Planning B: Planning and Design* (2005), doi:10.1068/b31163



# Prediction Model Based on PCA - DRKM-RBF

Wang Zhe, Sun Wen-wen, Zhou Tong, and Zhou Chun-guang

**Abstract.** This paper presents a new neural network predictive model named PCA - DRKM - RBF, which combines Principal Component Analysis (PCA), Dynamic Rough set K-means (DRKM) and Radial Basis Function (RBF) neural network. The data processed by the principal component analysis is the neural network's input, and the RBF neural network's hidden nodes are the centers using DRKM. The paper forecasts the cyclodextrin closure constant, and the results indicate that the model has obviously improved the accuracy of prediction.

**Keywords:** Principal Component Analysis, Rough set, K-means, RBF neural network, prediction.

## 1 Introduction

Prediction plays an increasingly important role in scientific research data processing, for example the drug research is one of the important fields of scientific, because its composition is complicated and the data quantity is much larger, so using computer to process data can effectively improve the efficiency and reduce the cost. As a result, prediction algorithm and the model also become the research focus, and it has achieved remarkable achievements such as: support vector machine prediction [1], the decision tree prediction [2], simple bayesian prediction [3], neural networks prediction [4], principal component regression prediction [5] and so on, in many research fields. These prediction methods can apply some factors' data in some research fields, however, the prediction results were affected because various related factors have been omitted. Therefore, how to construct prediction models using relationship among the data becomes the focus. So it has become one important research direction that how to use the validity of neural

---

Wang Zhe · Sun Wen-wen · Zhou Tong · Zhou Chun-guang  
Department of Computer Science and Technology, Jilin University, Changchun, China

network in data mining areas, and optimize the principal component analysis and rough set method to improve the prediction accuracy.

The paper constructs a new prediction model named PCA - DRKM - RBF based on RBF (radial basis function) neural network, principal component analysis and rough set clustering method. The model includes the following two aspects:

Firstly, it proposes a new algorithm based on the rough set k-means algorithm, which is named dynamic rough set k-means algorithm. It can solve the problem calculating dynamic determine class number named K.

Secondly, it proposes the idea which can improve neural network hidden layer structure by dynamic rough set clustering method. When the RBF neural network was used, the result can be taken as hidden nodes center after the training sample was manipulated by rough set clustering method, and using the advantage of rough set can solve fuzzy boundaries between the two sets. As a result, the method can activate corresponding clustering center of the hidden layer.

Based on the two points mentioned above, the PCA - DRKM - RBF prediction model proposed in this paper can improve the prediction's accuracy by its advantage that it can reduce the mass data dimension, and combine clustering and prediction together, and utilize the information comprehensively. This article demonstrates the effectiveness of the model through processing cyclodextrin closure constant.

## 2 Dynamic Rough Set Clustering Algorithm

### A *K-Means Algorithm Based on Rough Set*

K-means [6] algorithm is the most classical algorithm to clustering methods. Lingras [7] proposed K-means algorithm based on rough set, the algorithm can solve the problem caused by the uncertain boundary, and the method can combine rough sets theory and classic k-means algorithm.

Rough set theory [8] in two aggregates of a given set C is approximate: C's lower approximation aggregate is  $\underline{A}(C)$ : It's sure that objects belong to the set C; C's upper approximation aggregate is  $\overline{A}(C)$ : It's possible that objects belong to the set C.

Rough set's K-means clustering center is defined as follows:

$$RCenters(i) = e_{lower} \times \frac{1}{N_{lower}} \times \sum_{X \in \underline{C}_i} X + e_{upper} \times \frac{1}{N_{upper}} \times \sum_{X \in \overline{C}_i} X \quad (1.1)$$

In the formula,  $e_{lower}$  is the lower approximation centralization weight,  $e_{upper}$  is the upper approximation centralization weight,  $e_{upper} + e_{lower} = 1$ , and  $e_{lower} > e_{upper} \cdot N_{lower}$  and  $N_{upper}$  are the objects' number of lower and upper approximation sets separately.

## ***B Dynamic k-Means Algorithm Based on Rough Set***

One of the biggest problems of rough set k-means algorithm is difficult to determine  $K$ 's value. This paper proposes a method determining dynamic set's number of  $K$  based on the rough set  $K$  clustering. It includes the following two aspects:

a. When the lower approximation sets are empty but the uppers are not empty, it shows that there are no objects belonging to the set purely, but some objects may be included to the set. The set's clustering center has no value, and set number is not reasonable, then  $K$  minus 1. The lower approximation set disappears if they are empty.

b. When two classes' upper approximation sets are same, it indicates that the two classes' comparability is small, so these two classes are combined into one class.

Using the two aspects above, dynamic k-means based on rough set (DRKM) algorithm procedure presents as follows:

Specific means include the steps as follows:

- (1)Getting clustering centers by k-means algorithm;
- (2)Classifying each object to the upper or lower approximation sets;
- (3)Recalculating class center according to the formula (1.1);
- (4)Judging the following three conditions:
  - ① If a class' lower approximation is empty, then  $K = K - 1$ ; turning to②;
  - ② If there are two classes whose upper approximation sets are equal, then  $K = K - 1$ ; turning to①;
  - ③ If there are no conditions meeting the two conditions above, then turning to (5);
- (5)Repeating (2) (3) (4) until  $RCenters$  remains as one value.

In the algorithm, the step (4) embodies the dynamic change of  $K$ . In the procedure, it makes the classification more reasonable because  $K$  always should be the most suitable numerical.

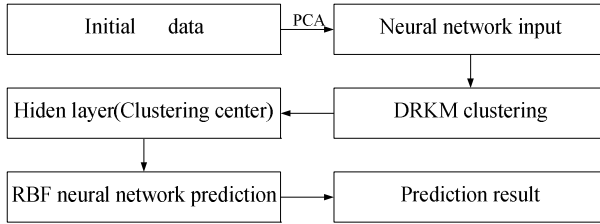
## **3 Prediction Model PCA-DRKM-RBF**

### ***A Neural Network RBF***

Neural network RBF (radial basis function) is a type of feedforward neural networks containing one hidden layer, input layer, and output layer (figure 1 shows). Hidden layer of transmission function is Gaussian Function [9] with radial. Input layer node presents data attributes generally, as a result it is necessary to take principal component analysis firstly for multi-dimension data. Principal component analysis [10] not only can extract main information, but also can simplify the neural network's input structure.

## B Characteristics of Prediction Model PCA-DRKM-RBF

The new prediction model PCA -DRKM - RBF's framework shows as Figure1:



**Fig. 1** Prediction model PCA-DRKM-RBF

When the new model PCA-DRKM-RBF is used, three procedures need as following:

(1) Making the result of the original data's principal component analysis as the RBF neural network's input;

(2) Grouping the samples into the training sample and prediction sample, and making dynamic rough set k-means clustering to the training sample, then getting clustering centers;

(3) Establishing RBF neural network model, using the first step (1) as input layer, taking second step (2) as hidden nodes center, then getting predicting result by the neural network (RBF).

## 4 Experiments and Analysis Results

Cyclodextrin is generation name of a series of annular oligosaccharides coming from amylose when it was worked by cyclodextrin glucosyltransferase which is produced by bacillus, the cyclodextrin closure constants can reflect cyclodextrin's character. Extracting more attribute variables from cyclodextrin and researching connection between cyclodextrin complex closure constants and attribute variables, can set up the model predicting cyclodextrin complex closure constants.

There are 97 samples of cyclodextrin data processed in this paper, and each sample contains 66 attributes and 1 cyclodextrin constant. It is cyclodextrin closure constant K reflecting cyclodextrin nature. The purpose of this experiment is predicting cyclodextrin closure constant K through researching cyclodextrin's 66 attributes, and cyclodextrin properties were also researched.

### A Principal Component Analysis

After principal component analysis, cyclodextrin original data's 66 attribute were reduced to 18 dimensions, the results are shown in Table 1.

**Table 1** Contribution rate of PCA

principal component	eigenvalue	contribution rate%	cumulating contribution rate%	principal component	eigenvalue	contribution rate%	cumulating contribution rate%
1	8.412	12.746	12.746	10	2.168	3.285	61.657
2	5.955	9.023	21.769	11	2.064	3.127	64.784
3	4.483	6.793	28.562	12	1.816	2.751	67.535
4	4.261	6.456	35.018	13	1.680	2.546	70.081
5	4.017	6.087	41.105	14	1.573	2.383	72.464
6	3.172	4.806	45.911	15	1.501	2.274	74.737
7	3.035	4.598	50.509	16	1.295	1.962	76.700
8	2.742	4.155	54.664	17	1.258	1.906	78.606
9	2.447	3.708	58.372	18	1.115	1.690	80.295

The principal component can reflect more than 80% of the information from original information data after reducing dimension number. At the same time, it can reduce data information and remove redundancy.

## ***B DRKM Clustering***

In the 97 data samples, there are 80 training samples and 17 testing samples.

After processed by DRKM clustering method, the class number is  $K = 6$ . So 80 samples were gathered into six classes, and each class contains the number as showing in Table 2:

**Table 2** Cluster of sample

clustering	Sample number of each cluster (total number:80; deletion number:0)					
	1	2	3	4	5	6
class						
sample number	8	13	20	11	10	18

## ***C The New Model and Predicting Results***

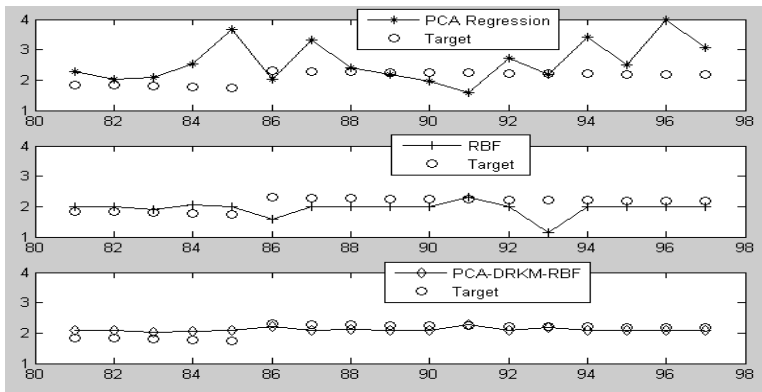
Neural network model RBF is established firstly, input layers is the data processed by principal component analysis, then input nodes number is 18. Hidden layer neurons number is  $K = 6$ , which is equal to samples' classification number, so hidden layer node number is 6, and the output layer is the cyclodextrin complex closure constant.

The result of principal component analysis is RBF neural network's input; the dynamic rough set clustering center is the center of hidden layer; and the cyclodextrin closure constant is output node. The neural network achieves stability when 80 training samples were trained for 75 times; then 17 testing samples were used for testing, and prediction mean square error is 0.0039% by the new model. Comparing with the other two methods, predicted results are shown in Table 3:

**Table 3** Mean square errors of different methods

Prediction methods	Mean square error
principal component regression forecast (PCA Regression)	0.7144
RBF neural networks	0.9151
PCA-DRKM-RBF	0.0039

From the figure 2, several results can be got. Firstly, the principal component regression curve distribution is scattered, and the distance is farther from target. Secondly, predictive value of RBF neural network is closer to the target, but its error is still greater; the predicted value of PCA-DRKM-RBF is the closest to the target, and the prediction is more effective.



**Fig. 2** Comparison of Prediction with different methods

## 5 Conclusions

The results of cyclodextrin experimental data show that the neural network prediction method based on PCA-DRKM-RBF can improve the accuracy of the predictions successfully. The experiment result shows that the model uses principal

component to simplify neural network input structure, and the innovation dynamic clustering algorithm is used to neural network hidden nodes center. In all, the method proposed in this paper has a better ability in predicting data.

**Acknowledgment.** This paper is supported by (1) the National Natural Science Foundation of China under Grant No. 60873146、60973092、60903097, (2) project of science and technology innovation platform of computing and software science (985engineering), (3) the Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China, (4)Jilin Science and Technology Development Priorities Project (contract number: 20090304).

## References

- [1] Chen, C., Wu, Z., Sun, S., et al.: Forecasting the ionospheric f0F2 parameter one hour ahead using a support vector machine technique. *Journal of Atmospheric and Solar-Terrestrial Physics* 72(18), 1341–1347 (2010)
- [2] Better, M., Glover, F., Samorani, M.: Classification by vertical and cutting multi-hyperplane decision tree induction. *Decision Support Systems* 48(3), 430–436 (2010)
- [3] Palacios-Alonso, M.A., Brizuela, C.A., Sucar, L.E.: Evolutionary learning of dynamic naive bayesian classifiers. *Journal of Automated Reasoning* 45(1), 21–37 (2010)
- [4] Rajput, N.S., Das, R.R., Mishra, V.N., Singh, K.P., Dwivedi, R.: A neural net implementation of SPCA pre-processor for gas/odor classification using the responses of thick film gas sensor array. *Sensors and Actuators* 148(2), 550–558 (2010)
- [5] Salem, M.Y., El-Zanfaly, E.S., El-Tarras, M.F., et al.: Simultaneous determination of domperidone and cinnarizine in a binary mixture using derivative spectrophotometry partial least squares and principle component regression calibration. *Analytical and Bioanalytical Chemistry* 375(2), 211–216 (2003)
- [6] Jain, A.K.: Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31(8), 651–666 (2010)
- [7] Lingras, P., West, C.: Interval set clustering of web users with rough K-means. *Journal of Intelligent Information Systems* 23(1), 5–16 (2004)
- [8] Shyng, J.-Y., Shieh, H.-M., Tzeng, G.-H.: An integration method combining Rough Set Theory with formal concept analysis for personal investment portfolios. *Knowledge-Based Systems* 23(6), 586–597 (2010)
- [9] Boyd, J.P., Lei, W.: Asymptotic coefficients for Gaussian radial basis function interpolants. *Applied Mathematics and Computation* 216(8), 2394–2407 (2010)
- [10] Lam, K.-C., Tao, R., Lam Mike, C.-K.: A material supplier selection model for property developers using Fuzzy Principal Component Analysis. *Automation in Construction* 19(5), 608–618 (2010)

# A Topic Detection and Tracking System with TF-Density

Shu-Wei Liu and Hsien-Tsung Chang

**Abstract.** In the past, news consumption took place predominantly via newspapers and were hard to track. Nowadays, the rapid growth of the Internet means that news are continually being shared and stored on a previously unimaginable scale. It is now possible to access several news stories on the same topic on a single web page. In this paper, we proposed a topic detection and tracking system with a new word measurement scheme named TF-Density. TF-Density is a new algorithm modified from the well-known TF-IWF and TF-IDF algorithms to provide a more precise and efficient method to recognize the important words in the text. Through our experiments, we demonstrated that our proposed topic detection and tracking system is capable of providing more precise and convenient result for the tracking of news by users.

## 1 Introduction

With the development of new technologies in recent years, people nowadays can engage in an increasingly wide range of activities on the Internet, e.g. making new acquaintances, shopping, e-mail and reading news online via dedicated news sources. In a FMCG report on the behavior of netizens, the foremost two most popular activities are “news reading” (65%) and “music appreciation” (65%). In fact, reading news online is so common among adults that it has become a necessary component of the daily routine of many.

However, there exists a glut of news sources on the Internet, such as [1,2]. Visiting just one news source no long suffices for many who desire different perspectives and are eager to partake in discussions on breaking or developing stories. As such, Internet users are spending ever more time to visit different news sources and

---

Shu-Wei Liu · Hsien-Tsung Chang

Department of CSIE, Chang Gung University, Taoyuan, Taiwan  
email: ftcloud@gmail.com, smallpig@widelab.org



perform internet searches on related topics, to the extent certain web services such as GoogleNews [3] have taken the initiative to gather over 350 news sources and classify them into different categories, e.g. politics, sports or art. Similar news will be clustered under the same news topic, e.g. Republicans Divided on Obama's Proposal to Extend Middle-Class Tax Cuts, with the topic containing the various news stories which Google has gathered from the different news sources. Such services have vastly reduced the time that users spend on searching for different news sources.

Topic Detection and Tracking (TDT) tasks are commonly utilized to structure news stories from newswires and broadcasts into topics [4]. When a story is received by the system, the system will gauge whether the story is breaking news or old news. If the stream has never before been seen, it may very well be breaking news. In this age of information explosion, people regularly resort to the search engines to determine "what is new" or "what is going on" in the world. However, with an explosion of information and documents on the internet, new tools are needed to better organize this information.

Just as in the corporate world, laypeople are also eager to be the first to know when a particular company releases certain information on its products, to save on purchases or to make a quick buck. In this paper, we focused on Chinese articles, whose search focus is different from that of English articles, and proposed a new "term weight algorithm" for structuring news. Our algorithm can help people more efficiently and conveniently locate the news they desire.

## 2 Related Works

Topic detection and tracking (TDT) techniques have been under development for several years [5], with various algorithms for calculating the terms for clustering the media stream. TFIDF [5, 6, 7] is a widely used algorithm in TDT, and its essence is that if a term appears several times in a story, it is considered important. If the term appears in a few sources, it is considered an important word in the story. TFIDF is specifically on the lookout for two features: DF and TF, but in some situations, the term will be overestimated or underestimated. Therefore, so someone proposed a new algorithm, TFIWF [8, 9, 10], which is different from TFIDF in that it uses a new feature WF (word frequency). The algorithm will be more efficient and accurate in assigning term weight for structuring the stream. In [11] the objective is to extract the useful terms and filter the noise terms. After term filtering, the process will become more accurate and efficient in TDT. And in [12], the focus is on cluster efficiency and cluster accuracy. For the cluster efficiency, an indexing tree has been proposed to speed up the cluster structure and to reweigh certain terms such as the glossary of people's name or certain key points in the article, to bestow greater weight than others and render the cluster more accurate.

## 3 System Architecture and Algorithms

Our system is running as following steps. The first step was data collection. We collected news stories in Chinese by using the RSS technology. RSS is a family of

web feed formats used to publish frequently updated work [13]. We subscribed to feeds from various Chinese news websites and stored these data on our server. Since the news data were in the HTML format, for the second step we analyzed the news data to extract useful data, including the titles and main bodies. In the third step, after extracting the titles and main body contents, we sent them to the CKIP (Chinese Knowledge Information Processing Group) system constructed by Academia Sinica [13] to split the article into terms. After the third step, we were able to determine the word composition of the article. In step four, we processed the filter module which would remove the words that contribute nothing but noises, as well as stop words (such as is, are, you, I). Finally a term list was created for each story. For the fifth step, we created a term information table including term information such as TF (term frequency), DF (word frequency) and WT (total number of terms) that are useful for weighting the article. Next, we applied our proposed TF-Density algorithm to weigh the term list in every story and create a vector list for every story. In the last step, we clustered the stories using cosine similarity.

The Incremental TF-IDF algorithm [5, 6, 7] has been widely used in information retrieval and text mining. The TF-IDF value consists of two components: the TF and IDF values. The TF (term frequency) refers to how many times the term  $w$  appears in the story  $d$  while the IDF (inverse document frequency) refers to the number of documents containing the term  $w$ .

The TF-IWF algorithm was proposed by [8], and it further incorporates the wf (word frequency) to more accurately assign weight to terms of importance in the article. The wf (word frequency) of the term  $w$  at time  $t$  is calculated as follows:

$$wf_t(w) = wf_{t-1}(w) + wf_{st}(w) \quad (1)$$

Where  $wf_t(w)$  refers to the number of times the term  $w$  appears at the time  $t$ ,  $wf_{st}(w)$  refers to the number of times the term  $w$  appears before the breaking story at the time  $t$ ,  $wf_{t-1}(w)$  refers to the number of times the term  $w$  appears before the time  $t$ , and  $w_t$  refers to the total number of times the term  $w$  appears before the time  $t$ .

$$weight_t(d, w) = \frac{tf(d, w) \log((w_t + 1) / (wf_t(w) + 0.5))}{\sqrt{\sum_{w' \in d} (tf(d, w') \log((w_t + 1) / (wf_t(w') + 0.5)))^2}} \quad (2)$$

We proposed a new TF-Density algorithm. The algorithm combines features from the TFIDF and TFIWF algorithms, because there are two importance features in these two algorithms, namely DF (Document Frequency) and WF (word frequency), that we wanted to retain while providing a more precise and efficient method to recognize the important words in the text. Our approach is that we will calculate the number of times each term appears in all the documents, and divide it by number of documents in which the term  $w$  appears. This way, we can obtain the density of the term  $w$ , i.e. the average number of times it appears in all the documents. Therefore, if a particular term  $w$  appears more times in the stories than the average density, it is likely that this particular term in this particular story may be more important than others. The formula is given as follows:

Where WF refers to the number of times the term  $w$  appears in all documents, and DF refers to the number of documents in which the term  $w$  appears.

Therefore, we can calculate  $Density_{term}$ , which refers to the average number of times the term  $w$  appears in one document.

$$Density_{term} = \frac{WF}{DF} \quad (3)$$

$$weight_t(d, w) = \frac{TF(d, w) / Density_{term}}{WF / Wt} \quad (4)$$

We use cosine similarity to calculate the similarity between two stories  $d$  and  $d'$  at the time  $t$  with the formula given as follows:

$$similarity = \cos(\theta) = \frac{A \times B}{|A| |B|} \quad (5)$$

$$similarity_t(d, d') = \sum_{w \in d \cap d'} weight_t(d, w) \times weight_t(d', w) \quad (6)$$

When a news story breaks at the time  $t$ , it will become a candidate for a new topic and be compared with previous topics, by means of their pair-wise similarities. If the similarity value is greater than the threshold  $d\theta$ , the topic will be considered as having previously appeared before the time  $t$ . If the value is less than the threshold, the candidate will be deemed a new topic. This way, we can ascertain whether the topic is old or new.

Our algorithm will execute as following. Step 1 every new incoming story will be put into a new cluster, which will contain just one story. Step 2 To calculate the pair-wise similarities for each Topic clusters, we choose the top 30% term vector for calculating with one another since, because in the Chinese news context after CKIP, there are many noise words but not stopwords such as Modal or Adjective. Therefore, in order to prevent the unnecessary words from affecting the performance, we apply this method to filter unnecessary words. Step 3 finally, we can know the largest similarity  $\theta$  with new Cluster and Cluster B in all clusters. If  $\theta$  is larger than the threshold  $d\theta$ . We conclude that A appears to be a new topic, so we combine A and B into a new cluster. If  $\theta$  is smaller than the threshold, then cluster B will be considered a new topic.

Each new topic clusters compared with previous topics to check whether the story is new. When a new cluster is compared with Topic cluster B, if the cluster B contains only one story, the calculation is straight or ward. However, if cluster B has more than one story, we will combine all the term vector for the stories in the cluster, and choose the top 30% terms vector according to the term weight for calculation. This is due to the fact that news topic has the Times feature, so that when an event began at the time  $t$ , the event's topic may refer to other subjects such as the names of people or places. Therefore, we combine all the term vectors of the stories in the cluster and to Increase the story can assigned in correct cluster.

## 4 Experiments and Discussions

We collected data from Chinese news web sites such as Apple Library, UNN, etc. and famous portals such as Yahoo!, etc. The data included news stories, discussions and reviews. We used Dataset of different sizes to experiment with our algorithm and for comparison with one another.

DataSet1: We collected 232 sports related stories with 100 topics from several Chinese news web sites on 2010-08-07, with the topics pre-clustered by users. The number of stories within a topic is between 1 and 15. DataSet2: We collected 523 stories from 2010-08-7 to 2010-08-9, containing 135 topics pre-clustered by users. The number of stories within a topic is between 1 and 40.

In our experiment, we compared our algorithm with TFIWF. However, the WF feature of TFIWF is updated dynamically, and we needed to calculate the initial WF. Therefore, we used data from almost 30 days to train the algorithm to implement TFIWF. We used the Recall, Precision and F-measure to analyze the story has been assigned to the correct Topic cluster. The metric is the traditional evaluation metric, which is widely used in information retrieval and clusters [14]. Recall and Precision usually contradict each other for information retrieval. Therefore, if we cluster all the stories in one cluster, Recall will become 100%. If we cluster every story as a topic, the Precision will become very high. As such, we used the F-measure to average the two value objectives to gauge the performance.

**Table 1** Performance of the three algorithms.

	DataSet1	DataSet1	DataSet1	DataSet2	DataSet2	DataSet2
	Recall	Precision	F-Measure	Recall	Precision	F-Measure
TF-Density	95.2%	95.2%	93.6%	88.3%	91.4%	87.2%
TF-IDF	95.2%	91.7%	91.3%	89.9%	85.7%	84.4%
TFIWF	95.2%	91.6%	88.7%	87.0%	87.1%	83.6%

Our proposed term weighting model, TF-Density, combines the DF (Document Frequency) and WF (word frequency) features of IDF and IWF respectively. Therefore, our experiments compared its performance with those of the aforementioned two algorithms based on Dataset1 and Dataset2.

Table 1 shows the average Recall, Precision and F-measure for the three algorithms. One can see that the density for our proposed weighting model TF-Density in F-measure was better than those of IDF and IWF. As for Recall and Precision, the Precision of our algorithm was better than that of IDF. However, as we mentioned in the Evaluation Metric section, Recall and Precision usually contradict each other. Therefore, while we were able to tune both Recall and Precision values better than for the other algorithms, the performance in F-measure was not the best for our algorithm.

## 5 Conclusions

In this paper, we have created detection and tracking system for news topics. In this system, we proposed a new words measurement scheme named TF-Density, derived by modifying the well-known TF-IWF and TF-IDF algorithms to provide

a more precise and efficient method for recognizing important words in the text. Our experiment collected data from various Chinese news web sites and the results indicated that our proposed TF-Density algorithm performed better than TFIDF and TFIWF. The results showed our proposed topic detection and tracking system can provide a more precise and convenient result for users to track news.

## References

1. <http://udn.com/NEWS/main.html>
2. <http://news.google.com.tw/>
3. <http://news.google.com.tw/>
4. <http://www.nist.gov/speech/tests/tdt/index.htm>
5. Yang, Y., Pierce, T., Carbonell, J.: A Study of Retrospective and On-line Event Detection. In: 21th ACM SIGIR Conference, Melbourne, Australia. ACM Press (1998)
6. Brants, T., Chen, F., Farahat, A.: A System for New Event Detection. In: SIGIR 2003, Toronto, Canada (2003)
7. Zheng, D., Li, F.: Hot topic detection on BBS using aging theory. In: Liu, W., Luo, X., Wang, F.L., Lei, J. (eds.) WISM 2009. LNCS, vol. 5854, pp. 129–138. Springer, Heidelberg (2009)
8. Wang, C., Zhang, M., Ma, S., Ru, L.: Automatic Online News Issue Construction in Web Environment WWW 2008, Beijing, China (2008)
9. Wang, C., Zhang, M., Ma, S., Ru, L.: Automatic Online News Topic Ranking Using Media Focus and User Attention Based on Aging Theory. In: CIKM 2008, NapaValley, California, USA (2008)
10. Wang, C., Zhang, M., Ma, S., Ru, L.: An Automatic Online News Topic Key - phrase Extraction System. In: 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (2008)
11. Lee, S., Kim, H.: News Keyword Extraction for Topic Tracking. In: Fourth International Conference on Networked Computing and Advanced Information Management
12. Kuo, Z., Zi, L.J., Gang, W.: New Event Detection Based on Indexing-tree and Named Entity. In: SIGIR 2007 (2007)
13. <http://ckipsvr.iis.sinica.edu.tw/>
14. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill (1983)

# Community Identification of Financial Market Based on Affinity Propagation

Lei Hong, Shi-Min Cai, Zhong-Qian Fu, and Pei-Ling Zhou

**Abstract.** Community identification in complex financial system is an important task in the exploratory analysis of stock time series. In this paper, a recently proposed message-passing-based algorithm called affinity propagation is introduced to identify stock groups. First, the similarities computed between all pairs of stocks of portfolio by considering the synchronous time evolution of their logarithm return are mapped into spatial distances. Then, the spatial distances are used to cluster the stocks into different communities of financial network via choosing appropriate preference according to affinity propagation. The results suggest that the approach is demonstrably effective in identifying multiple stock groups without any extra knowledge of stocks, and provide a meaningful economic taxonomy.

**Keywords:** Community, affinity propagation, time series, financial market.

## 1 Introduction

Financial markets have been well studied as evolving complex systems, in which the companies interact with each other by competition and cooperation. The performance of a company is compactly drawn by its stock price. Thus, the interaction among companies in financial markets can be uncovered by the study of correlations between all pairs of stock price, which has attracted much interest of experts in real fields. In the early years, Mantegna and his coworkers first investigated a correlation-based taxonomy of stocks by the method of hierarchical

---

Lei Hong · Shi-Min Cai · Zhong-Qian Fu · Pei-Ling Zhou  
Department of Electronic Science and Technology,  
University of Science and Technology of China  
Hefei Anhui, 230026, PR China  
e-mail: leihong@mail.ustc.edu.cn, csm1981@mail.ustc.edu.cn

tree [1, 2]. Recently, the hierarchical organization and scale-free dynamics of financial network was discussed via complex network theory and minimum spanning tree technique [3-6]. These properties suggest that the stocks are tendency to cluster into groups (i.e. communities in financial networks), and urge us to further use clustering algorithms to identify community structure in financial market on account of the application to the portfolio optimization in real market.

In the field of complex network, many algorithms have been proposed to detect the community structure of network [7-11]. However, the disaster of most of these available algorithms is their high computational complexity, which is the main reason that prevents the use of such algorithms for large size of financial systems. Hence, we pay close attention to data cluster analysis on the purpose of straightly identifying communities of financial markets. So far, many existing methods of data cluster, such as K-means, K-centers and spectral clustering, need some prior-knowledge of datasets. Here we introduce a message-passing-based algorithm called Affinity Propagation (AP), recently proposed by Frey and Dueck [12], which could achieves a considerable improvement over aforementioned data cluster methods (i.e. K-means, K-centers and spectral clustering, see also [12]). AP is approximated following the idea of belief-propagation, yet it is implemented for  $N$  data points with only  $O(N^2)$  messages, far less than the  $O(N^3)$  messages that have to be determined self-consistently. Therefore, the algorithm is feasible even in the very large datasets.

In this paper, we consider each stock as a special data point, and propose the similarity function between data points based on spatial distance mapping from the correlations of returns of stock price. Based on AP, we observe that the stocks can self-organizedly cluster into multiple groups.

## 2 Affinity Propagation Algorithm

Clustering is a form of unsupervised machine learning in which observations are partitioned into groups that within-cluster similarity is larger than that of between-cluster. AP is one of most efficient clustering algorithms, which seeks to identify each cluster by one of its elements, so called *exemplar*, based on the certain proper definition of similarity function between data points. In AP, each exemplar is required to refer to itself as a self-exemplar, i.e. there is only one central node and all other nodes are directly connected it. According to this constraint, AP must be at maximizing the overall similarity of all data points to their exemplars. Specifically, AP emphatically uses data points themselves to characterize clusters, and simultaneously considers all data points as candidate exemplars.

In AP, for a given set  $D = \{x_i\}$  ( $i = 1, \dots, N$ ), this process is realized via an efficient message-passing-based formulation which can be derived as an instance of the max-sum algorithm for factor graphs [13]. There are two different kinds of messages, responsibility  $r(i, c)$  and availability  $a(i, c)$  ( $i, c$  label data points  $x_i$  and  $x_c$ ), which are exchanged between observations, and each needs consider

a different kind of competition. Messages can be combined at any stage to determine which observations are exemplars, and for every other one, which exemplar it belongs to. Herein,  $r(i, c)$  denotes the suitability that the candidate exemplar  $c$  is to serve as a true exemplar of point  $i$ , compared with other potential ones, while  $a(i, c)$  reflects the possibility that point  $i$  chooses point  $c$  as its exemplar. To keep our description as self-contained as possible, we briefly review the procedure of algorithm.

AP takes as input of a collection of real valued similarities  $s(i, c)$  between data point  $i$  and  $c$ , which are measured by the negative squared error ( i.e. Euclidean distance ). It is noted that the similarities can be generally computed in alternative ways [12]. The algorithm wants to find a mapping,  $c: \{1, \dots, N\} \mapsto \{u_1, \dots, u_N\}$ , which classifies each data point  $i$  to its exemplar  $u_i$ , as well as  $u_i$  itself must be a data point. Whenever a data point is selected as an exemplar by another data point, it has to be its own self-exemplar.

Initially, availability  $a(i, c)$  is set as  $a(i, c) = 0$  ( $i \neq c$ ), and  $a(i, i)$  is given by the value of preference  $p$ . Then, the messages, responsibilities and availabilities, are updated sequentially using the following equations

$$\begin{aligned} r(i, c) &\leftarrow s(i, c) - \max_{c': c' \neq c} \{a(i, c') + s(i, c')\}; \\ a(i, c) &\leftarrow \min \left\{ 0, r(c, c) + \sum_{i': i' \in \{i, c\}} \max \{0, r(i', c)\} \right\}, i \neq c; \\ a(c, c) &\leftarrow \sum_{i': i' \neq c} \max \{0, r(i', c)\}, i = c. \end{aligned} \quad (1)$$

When updating messages are subject to the above rules, the computations are easily implemented. In addition, the numeric oscillations resulting from some circumstance may effects on the convergence of message-passing procedure. Thus, it introduces a damping factor  $\lambda \in [0, 1]$  to avoid this case. Each message is set to a weighted combination of its value from the previous iteration and its updated value, weighted by  $\lambda$  and  $1 - \lambda$ , respectively. At last, we can get the exemplar decision by maximizing over the sum of responsibility and availability. The updating iteration of algorithm is terminated until the exemplar decision remains unchanged for a fixed time steps. In our experiments, we set  $\lambda = 0.5$ , and the converging conditions of datasets are that exemplar decisions keep stable for 50 time steps or the overall maximum number of iterations to 1000.

It is valuable to be mentioned that the preference  $p$  affects the number of clusters, of which high values discover more number of clusters than lower ones. The number of clusters self-organizedly emerges from data, which suggests that  $p$  is closer in spirit to regularization strength than a preset number of clusters.



### 3 Empirical Results

In order to get the community structure of financial market, our starting point is to quantify the degree of similarity between the synchronous time series evolution of a pair of stock price. Herein, we survey the high frequency time series of trading stock price from the Trades and Quotes (TAQ) databases for one pool of  $N=147$  US stocks selected from S&P500. Their duration is the whole year 1996 with 250 trading day. In each day, the stock trades recorded with sample frequency per 30 minutes last 390 minutes (9:30-16:00). The degree of similarity is quantified by correlation coefficient of logarithmic returns of stock price

$$\rho_{i,j} = \frac{1}{T} \sum_{t=1}^T \frac{(R_i(t) - \langle R_i \rangle)(R_j(t) - \langle R_j \rangle)}{\sigma_i \sigma_j} \quad (2)$$

where  $T$  is the total number of stock time series,  $i$  and  $j$  are the numerical labels of stocks,  $\delta$  is the standard deviation corresponding to return  $R$ , and  $\langle \rangle$  denotes the statistical average in whole time scale. The return  $R_i$  is defined as

$$R_i = \ln P_i(t+1) - \ln P_i(t), \quad (3)$$

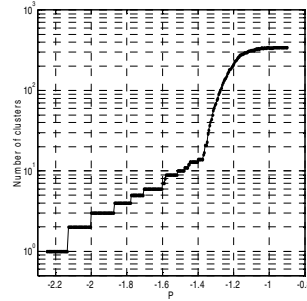
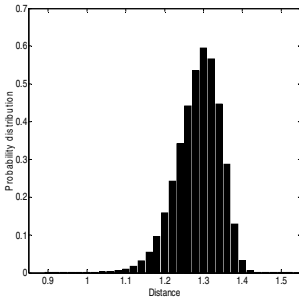
where  $P_i(t)$  describes the price of stock  $i$  at timescale  $t$ . We obtain the  $N(N-1)/2$  correlation coefficients computed between all pairs of stocks of portfolio by considering the synchronous time evolution of their logarithm return.

The correlation coefficients can't be straightly used to detect the community structure of financial market because they do not fulfill the three axioms that define a metric [1]. An alternative approach is to use the distance between two stocks to determine the spatial relationship. The measure of similarity of data points to be clustered hence depends on the mapping function. We map the similarity into spatial distance via a function of the correlation coefficients

$$d_{i,j} = \sqrt{2(1 - \rho_{i,j})}, \quad (4)$$

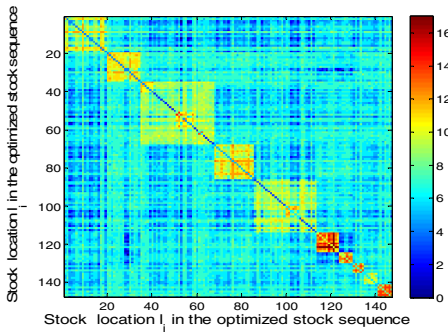
where  $d_{i,j}$  is equivalent to the Euclidean distance between two vectors. The probability distribution of elements of spatial distance matrix is shown in Fig.1. Therefore, the input of a collection of real valued similarities in AP can be set as the negative value of spatial distance,  $s(i,c) = -d(i,c)$ .

According to AP, we firstly investigate the number of clusters for 147 stocks as a function of  $p$ , shown in Fig.2. The empirical result shows that the number of clusters increases with the growth of  $p$ , yet the lower  $p$  brings the all data points are aggregated into one cluster while the larger  $p$  makes the all data points consider themselves as exemplars leading to a amount of clusters. Therefore, the



**Fig. 1** The probability distribution of elements of spatial distance matrix. **Fig. 2** The number of clusters as a function of preference  $p$

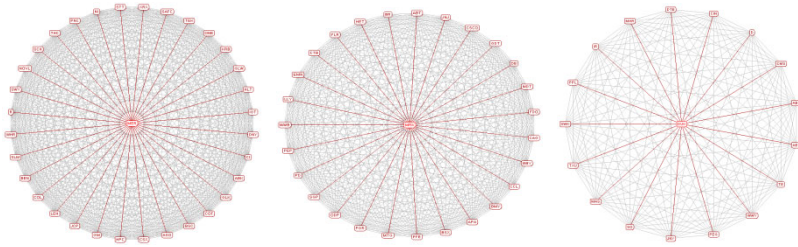
proper choice of  $p$  is important to identify the underlying communities (i.e. the number of clusters). In [12], Frey and Dueck suggest that  $p$  can be set as the median value of all similarities of data points. We follow the same idea that the median value of negative spatial distance of stocks is computed as the value of  $p$ . For this choice, the 147 stocks are clustered into 10 groups visualized by the mapping spatial distance matrix elements  $d_{l_i, l_j}$  with the optimized stock sequence (see Fig.3). Figure 3 clearly shows the multiple independent blocks of mapping spatial distance matrix in the diagonal, which ranges from finance, utility, healthcare, energy, consumer cyclical, software technology, semiconductor, services.



**Fig. 3** (Color online) The visualizaition of the mapping spatial distance matrix with the optimized stock sequence.

It is noted that the elements of mapping spatial distance matrix are transformed via the linear relation of color mapping. In addition, because the S&P500 index has traditionally been market-vlaue weighted, the price fluctuations of stocks with larger market capitalizaitons, have a greater effect on the index. It means that

these stocks can pass much more messages to the other stocks with relatively less market capitalizations, hence they can easily become the exemplars of groups. In fact, the 10 exemplars, e.g. Merrill-Lynch (MER), Merck (MRK), Duke Energy (DUK), Intel (INTC), Bank of American (BAC), Texas Instruments (TXN), etc., are consistency with a real observation of financial market. To comprehensively understand the community structure of financial market, in Fig. 4, we present the visualizing networks of the former three largest groups including the financial, healthcare, and utility, of which the exemplars are Merrill-Lynch, Merck, and Duke Energy, respectively.



**Fig. 4** (Color online) The visualizing network of the former three largest groups. (a) Finance, (b) Healthcare, (c) Utility.

## 4 Conclusions

In conclusion, the community identification of financial market associating with a meaningful economic taxonomy is important for the portfolio optimization. AP has been proved that it is powerful tool for unsupervised clustering, which behaves efficient and fast convergence to the final clustering even for large datasets. Hence, we apply AP to identify the communities of financial market via computing the similarity between all pairs of stocks of portfolio by considering the synchronous time evolution of their logarithm return. The present study shows that the approach is demonstrably effective in identifying multiple stock groups without any extra knowledge of stocks. We also find that identified groups of stocks can correspond to the specific industrial sectors, which suggests that time series of stock price are carrying valuable economic information.

**Acknowledgment.** This work is supported by the National Natural Science Foundation of China under Grant Nos. 60874090, 60874111, 60974079, 61004102. S-MC appreciates the financial support of K.C. Wong Education Foundation, China Postdoctoral Science Foundation, and China Postdoctoral Science Foundation, and the Fundamental Research Funds for the Central Universities.

## References

1. Manganella, R.N.: Hierarchical structure in financial markets. *Eur. Phys. J. B* 11, 193–197 (1999)
2. Bonanno, G., Vandewalle, N., Mantegna, R.N.: Taxonomy of stock market indices. *Phys. Rev. E* 62, R7615 (2000)
3. Bonanno, G., Caldarelli, G., Lillo, F., Mantegna, R.N.: Topology of correlation-based minimal spanning trees in real and model markets. *Phys. Rev. E* 68, 046130 (2003)
4. Onnela, J.P., Chakraborti, A., Kaski, K., Kertész, J., Kanto, A.: Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E* 68, 056110 (2003)
5. Onnela, J.P., Kaski, K., Kertész, J.: Clustering and information in correlation based financial networks. *Eur. Phys. J. B* 38, 353–362 (2004)
6. Cai, S.M., Zhou, Y.B., Zhou, T., Zhou, P.L.: Hierarchical organization and disassortative mixing of correlation-based weighted financial networks. *Int. J. Mod. Phys. B* 21, 433–441 (2010)
7. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* 99, 7821–7826 (2001)
8. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. *Phys. Rev. E* 69, 026113 (2004)
9. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 435, 814–818 (2005)
10. Rosvall, M., Bergstrom, C.T.: An information-theoretic framework for resolving community structure in complex networks. *Proc. Natl. Acad. Sci.* 104, 7327–7331 (2007)
11. Fortunato, S.: Community detection in graphs. *Phys. Rep.* 486, 75–174 (2010)
12. Frey, B.J., Dueck, D.: Clustering by Passing Messages Between Data Points. *Science* 315, 972–976 (2007)
13. Kschischang, F.R., Frey, B.J., Loeliger, H.A.: Factor graph and the sum-product algorithm. *IEEE Trans. Inf. Theory* 47, 498–519 (2001)

# Research on Optimization of Target Oil Wells for CO<sub>2</sub> Huff and Puff in Yushulin Oil Well

Erlong Yang, Huaidong He, and Lei Wang

**Abstract.** CO<sub>2</sub> huff and puff is one of the effective methods to develop the low permeability oil fields. But the optimization for target wells before conducting CO<sub>2</sub> huff and puff is the key step, since it is directly related to the success or failure of CO<sub>2</sub> huff and puff measures. The influence of various factors such as remaining oil saturation, water cut, reservoir thickness, permeability, porosity and flowing bottom hole pressure (FBHP) on the effect of CO<sub>2</sub> huff and puff is analyzed firstly in this paper so as to establish an evaluation index. And then analytic hierarchy process (AHP) is used to determine a reasonable weight of single factor. And finally, two oil wells, Shu 11Y68-57 and Shu 11Y67-611, are optimized for CO<sub>2</sub> huff and puff according to the fuzzy comprehensive evaluation method. The two wells optimized have a better effect of improving oil, which shows that the CO<sub>2</sub> huff and puff technology is feasible and has a certain generalized value.

## 1 Introduction

Aiming at low permeability, insufficient natural energy, fast production decline rate, high water injection pressure, poor injectability, we implement CO<sub>2</sub> huff and puff to increase oil production [1,2]. A great number of both domestic and foreign research results show that CO<sub>2</sub> huff and puff can substantially enhance oil recovery, especially on the condition that there are rich CO<sub>2</sub> resources in Daqing, and conducting CO<sub>2</sub> huff and puff has an expansively developing prospects. Since the optimization

---

Erlong Yang

Key Laboratory of Enhanced Oil Recovery (Ministry of Education),

Northeast Petroleum University, Daqing, Heilongjiang, China

e-mail: [erlongyang.dqpi@gmail.com](mailto:erlongyang.dqpi@gmail.com)

Huaidong He

Oil recovery Plant No.3, Daqing Oil Field Corp. Ltd., Daqing, Heilongjiang, China

Lei Wang

Oil recovery Plant No.1, Daqing Oil Field Corp. Ltd., Daqing, Heilongjiang, China

**Table 1** Scale table.

Scale	Meaning
1	Two factors have the same importance.
3	One factor is slightly more important than the other one.
5	One factor is obviously more important than the other one.
7	One factor is strongly more important than the other one.
9	One factor is extremely more important than the other one.

for target oil wells is fuzzy, this paper uses fuzzy comprehensive evaluation and analytic hierarchy process to establish the evaluation system for selecting wells and provide theoretical guidance for the on-site implementation of CO<sub>2</sub> stimulation.

## 2 Single Factor Analysis of Influence on CO<sub>2</sub> Huff and Puff

According to the CO<sub>2</sub> stimulation mechanism, combining the reservoir, crude oil and reservoir characteristics, the dynamic effects of reservoir parameters on CO<sub>2</sub> huff and puff process are summarized and the evaluation indexes of CO<sub>2</sub> huff and puff with single well are established to provide evaluation criteria of screening CO<sub>2</sub> huff and puff target Wells [3].

There are several influential factors on CO<sub>2</sub> huff and puff: Remaining oil saturation, Oil-well water cut, Reservoir thickness, Permeability, Porosity, and Bottom hole flowing pressure(BHFP). According to the above mainly influential factors on CO<sub>2</sub> huff and puff, calculate influential factor weights, use fuzzy comprehensive evaluation method to determine CO<sub>2</sub> huff and puff suitability for oil wells, and make it be the basis for screening single-well CO<sub>2</sub> huff and puff.

## 3 Optimization Method of Target Oil Wells

### 3.1 Analytic Hierarchy Process (AHP)

(1) Build pairwise comparison judgment matrix

According to certain rules, compare the two factors  $u_i$  and  $u_j$ , decide which one is more important, and then give a certain value to the importance, generally using 1 ~ 9 ratio scale, as shown in Table 1. We can insert compromise values 2, 4, 6, 8 into the adjacent two levels. For  $n$  factors, obtain pairwise comparison judgment matrix  $A$  [4].

(2) Calculation method of weight vector

First, calculate the product of all the matrix line elements:  $M_i = \prod_{j=1}^n a_{ij}$ . Then, calculate the  $n$ th root of  $M_i$  and draw the new vector  $B$ :  $B_i = \sqrt[n]{M_i}$ . Normalize each  $B_i$  and get each weight vector:  $W_i = B_i / \sum_{k=1}^n B_k$ . Finally obtain the weight vector:

**Table 2**  $R \cdot I$  and matrix order relation table.  $MO$  denotes the matrix order.

$MO$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$R \cdot I$	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.524	1.56	1.58	1.59

$W = (W_1, W_2, \dots, W_n)^T$ . After achieving the weight vector, calculate the maximum eigenvalue according to the following approximation method:  $\lambda_{\max} = \sum_{i=1}^n \frac{(AW)_i}{nW_i}$ .

(3) consistency check

The index of consistency check is consistency ratio  $C \cdot R$ , defined as:  $C \cdot R = (C \cdot I)/(R \cdot I)$ . In the formula,  $R \cdot I$  is mean random consistency index that is related to the order of matrix and calculated according to the listed values in Table 2.  $C \cdot I$  is consistency index:  $C \cdot I = (\lambda_{\max} - n)/(n - 1)$ . The inspection standard is  $C \cdot I < 0.1$ . So we can take it that the judgment matrix is acceptable and the calculated weight is reasonable.

### 3.2 Determination Method of Membership Degree

If higher value of factors is better for CO<sub>2</sub> huff and puff, then use half-ascent-trapezoid distribution to evaluate membership degree:  $R_{kj} = \frac{r_{kj} - (r_{kj})_{\min}}{(r_{kj})_{\max} - (r_{kj})_{\min}}$ . If lower value of factors is better for CO<sub>2</sub> huff and puff, then use half-descent-trapezoid distribution to evaluate membership degree:  $R_{kj} = \frac{(r_{kj})_{\max} - r_{kj}}{(r_{kj})_{\max} - (r_{kj})_{\min}}$ . In the formula,  $R_{kj}$  denotes the membership degree of  $j$ th well and  $k$ th factor,  $r_{kj}$  the value of  $j$ th well and  $k$ th factor,  $(r_{kj})_{\min}$  the minimum value of  $k$ th factor and  $n$ th well, and  $(r_{kj})_{\max}$  denotes the maximum value of  $k$ th factor and  $n$ th well.

### 3.3 Theory of Fuzzy Comprehensive Evaluation

Fuzzy comprehensive evaluation method is a kind of synthetic evaluation that apply fuzzy transformation principle and maximum membership degree principle and consider various factors relevant to the evaluation object [5,6]. Main steps of fuzzy comprehensive evaluation are as follows: (1) Determine the evaluation object; (2) Determine the comment set  $V = (v_1, v_2, \dots, v_n)$ ; (3) Determine the factor set  $U = (u_1, u_2, \dots, u_n)$ ; (4) Determine  $r_i$  according to factors and then build matrix  $R = |r_{ij}|_{\min}$ ; (5) Determine the weighting set  $W = (w_1, w_2, \dots, w_n)$ ; (6) Select appropriate calculation model to do fuzzy transformation, obtain  $Y = W \otimes R$ , and then translate it into conclusion of required forms.

### 3.4 Optimum Selection of Well Location

(1) Determine the factor set Select 6 oil wells for well selection analysis in Shu 16 Block of Yushulin Oilfield. Determine 6 evaluation factors that are remaining

**Table 3** Evaluation factor index of oil wells. *Wn* denotes the well number, *Wc* the water cut, *Pr* the porosity, *Per* the permeability, *Os* the oil saturation, and *Rt* the reservoir thickness.

<i>Wn</i>	<i>Wc</i>	<i>Pr</i>	<i>Per</i>	<i>Os</i>	<i>Rt</i>	FBHP
Shu 11Y66-62	0.0602	0.1057	0.8014	0.4610	11.20	0.91
Shu 11Y67-59	0.0303	0.1180	0.9730	0.4895	11.80	1.14
Shu 11Y67-611	0.0321	0.1115	0.7595	0.4737	13.40	5.70
Shu 11Y68-57	0.0260	0.1037	0.7865	0.4538	10.40	0.51
Shu 11Y68-58	0.0361	0.1407	1.0135	0.5587	11.20	1.29
Shu 11Y68-60	0.4449	0.1519	0.8103	0.5489	11.80	2.32

oil saturation, water cut of oil well, reservoir thickness, permeability, porosity and FBHP respectively, as shown in Table 3.

Give judgment matrix *A* according to scale table (Table 1):

$$A = \begin{matrix} & h & K & S_o & f_w & \varphi & P_{wf} \\ \begin{matrix} h \\ K \\ S_o \\ f_w \\ \varphi \\ P_{wf} \end{matrix} & \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1/2 & 1 & 2 & 3 & 4 & 5 \\ 1/3 & 1/2 & 1 & 2 & 3 & 4 \\ 1/4 & 1/3 & 1/2 & 1 & 2 & 3 \\ 1/5 & 1/4 & 1/3 & 1/2 & 1 & 2 \\ 1/6 & 1/5 & 1/4 & 1/3 & 1/2 & 1 \end{bmatrix} \end{matrix}$$

(2) Calculate weight vector:

$$W = (0.381, 0.252, 0.16, 0.101, 0.064, 0.042).$$

After obtaining weight vectors, calculate  $\lambda_{max} = 6.12$  according to formula (5).

(3) Check consistency

According to table 2,  $R \cdot I = 1.26$  when the judgment matrix is six-rank. Calculate  $C \cdot I = 0.024$ ,  $C \cdot R = 0.019$  according to formula (6) and (7). The inspection standard is  $C \cdot R = 0.019 < 0.1$ . So we can take it that the judgment matrix is acceptable and the calculated weight is reasonable.

(4) Calculate membership degree array

If higher value of remaining oil saturation, reservoir thickness and FBHP is better for CO<sub>2</sub> huff and puff, then use half-ascent-trapezoid distribution to evaluate membership degree. If lower value of water cut, permeability and porosity is better for CO<sub>2</sub> huff and puff, then use half-descent-trapezoid distribution to evaluate membership degree. The membership degree array of 6 oil wells to 6 factors is shown in Table 4.



**Table 4** Membership degree array of oil wells. *Wn* denotes the well number, *Ef* the effective thickness, *Per* the Permeability, *Ws* the water saturation, *Wc* the water cut, and *Pr* the porosity.

<i>Wn</i>	<i>Ef</i>	<i>Per</i>	<i>Ws</i>	<i>Wc</i>	<i>Pr</i>	FBHP
Shu 11Y66-62	0.9184	0.9593	0.8352	0.0689	0.2667	0.0771
Shu 11Y67-59	0.9895	0.7039	0.1593	0.3406	0.4667	0.1214
Shu 11Y67-611	0.9852	0.8392	1	0.1895	1	1
Shu 11Y68-57	1	1	0.894	0	0	0
Shu 11Y68-58	0.9757	0.2323	0	1	0.2667	0.1503
Shu 11Y68-60	0	0	0.8	0.9065	0.4667	0.3487

**Table 5** Evaluation results for oil wells.

Well number	Evaluation result
Shu 11Y68-60	0.2644
Shu 11Y68-58	0.5542
Shu 11Y67-59	0.6487
Shu 11Y66-62	0.752
Shu 11Y68-57	0.7754
Shu 11Y67-611	0.8721

(5) Determine the comment set

Classify evaluation results as 4 levels, excellent: 0.75~1, good: 0.5~0.75, moderate: 0.25~0.5, poor: 0~0.25.

(6) Fuzzy comprehensive evaluation results

According to the principle of fuzzy comprehensive evaluation, make fuzzy transformation to weighting set and membership degree array. Final evaluation results for each well are shown in Table 5.

## 4 Field Application Effect

According to the optimization result, select Shu 11Y68-57 and Shu 11Y67-611 for CO<sub>2</sub> huff and puff field tests.

(1) Shu 11Y68-57. Gas injection volume 120m<sup>3</sup>, liquid CO<sub>2</sub>, gas injection rate 60m<sup>3</sup>/d, soaking time 15d, minimum FBHP 5MPa. Contrasted with the production effects of water flooding in the same period, the cumulative oil increment by CO<sub>2</sub> huff and puff was 727m<sup>3</sup> within a year and a half, oil exchange ratio was 0.0110m<sup>3</sup>·m<sup>-3</sup> and the degree of reserve recovery was 9.805%.

(2) Shu 11Y67-611. Gas injection volume  $120\text{m}^3$ , liquid  $\text{CO}_2$ , gas injection rate  $60\text{m}^3/\text{d}$ , soaking time 15d, minimum FBHP 5MPa. Contrasted with the production effects of water flooding in the same period, the cumulative oil increment by  $\text{CO}_2$  huff and puff was  $1161\text{m}^3$  within a year and a half, oil exchange ratio was  $0.0176\text{m}^3\cdot\text{m}^{-3}$  and the degree of reserve recovery was 8.69%.

The field test shows that  $\text{CO}_2$  huff and puff effect of Shu 11Y67-611 is better than Shu 11Y68-57. This is consistent with theoretically selected results and verifies the rationality of selecting well.

## 5 Conclusion

1. Remaining oil saturation, water cut of oil well, reservoir thickness, permeability, porosity and FBHP are determined as the evaluation system.
2. Two optimum wells, Shu 11Y68-57 and Shu 11Y67-611, has been selected that are suitable to implement  $\text{CO}_2$  huff and puff. And the field test has proved its good effects of stimulation.

**Acknowledgements.** This work is supported in part by National Science Foundation of China under Grant No.50634020.

## References

1. Qian, W., Cheng, S., Ge, Y.: Practice and recognition on the parameters of carbon dioxide huff-and-puff in low permeability reservoirs and its application. *Fault-block Oil and Gas Field. Pragmatics* 11(5), 40–42 (2004)
2. Chen, M., Jiang, H., Wu, Y.: Research on fuzzy comprehensive evaluation method of  $\text{CO}_2$  puff and huff well site optimization. *Oil-Gasfield Surface Engineering. Pragmatics* 27(10), 1–2 (2008)
3. He, Y., Mei, S., Yang, Z.: Numerical simulation analysis of the influence factors on  $\text{CO}_2$  stimulation in palogue oilfield. *Special Oil and Gas Reservoirs. Pragmatics* 13(1), 64–67 (2006)
4. Chang, J., Jiang, T.: Research on the weight of coefficient through analytic hierarchy process. *Journal of Wuhan University of Technology (Information and Management Engineering). Pragmatics* 29(1), 153–155 (2007)
5. Liang, X., Sun, L., Sun, L.: The research of fuzzy inter- grated evaluation method in wells and layers choice for  $\text{CO}_2$  puff and huff. *Natural Gas Geoscience. Pragmatics* 16(5), 658–660 (2005)
6. Fu, G., Ma, L., Qu, X.: Quantitative evaluation for the potentiality of remaining oil by multilevel fuzzy judgement method. *Journal of Earth Science and Environmental. Pragmatics* 26(2), 38–40 (2004)

# A Secret Embedding Scheme by Means of Re-indexing VQ Codebook upon Image Processing

Cheng-Ta Huang, Wei-Jen Wang, Shiuh-Jeng Wang, and Jonathan C.L. Liu

**Abstract.** A VQ-based, reversible information hiding method embeds secret data into a VQ-encoded image to achieve data protection. It produces either a VQ-encoded image or a compressed codestream as the output. The embedded secret data and the original VQ-encoded image can be extracted from the encoded output when needed. This paper proposes a novel information hiding method for VQ-encoded images. The proposed method allows the users to decide the embedding capacity of secret data in an image. The proposed method utilizes the fact that two neighboring blocks in a VQ-encoded image are usually similar. First, it re-orders the codebook according to the similarity to the codeword of the first block in an image. Second, it sequentially picks up two consequent blocks, calculates the difference of their new indices that are mapped from the re-ordered codebook, and encodes the secret bits along with the difference value as the compressed output for each block. We conducted several experiments, and used the results to compare the performance of the proposed method with the performance of a recent similar method. The experimental results showed that the proposed method is very efficient in terms of bit rate, in particular when an image has many similar neighboring blocks in it.

**Keywords:** Vector Quantization, Information Hiding, Image Processing.

---

Cheng-Ta Huang · Wei-Jen Wang

Department of Computer Science and Information Engineering, National Central University, Taoyuan 320, Taiwan

Shiuh-Jeng Wang

Department of Computer Information Science and Engineering, University of Florida Gainesville, FL, USA

Shiuh-Jeng Wang

Department of Information Management, Central Police University, Taoyuan 333, Taiwan  
e-mail: sjwang@mail.cpu.edu.tw

## 1 Introduction

The Internet has become more popular and more important in people's daily life nowadays. Through the Internet, people can share photos with their friends, transmit documents to their co-workers, and upload their videos to a video-sharing service such as YouTube [1]. However, the digital multimedia has some well-known security weaknesses such as copyright and authorship. These weaknesses can be greatly enhanced by information hiding [2-9], which is a technique for embedding secret data, such as copyright and trademark, into the cover multimedia. Information hiding can be used for better protection of secret communication by embedding secret data into the cover multimedia to avoid censorship from unauthorized people and organizations. Some information hiding techniques also compress the cover multimedia with the embedded secret data [5], [10-12], and then produce a compressed codestream as the output. They help reduce the network transmission time while delivering the cover multimedia and the embedded secret data to the receiver.

Vector quantization (VQ), one of the image compression techniques, has been widely studied in the context of image information hiding. VQ-based methods for information hiding can be classified into two non-overlapping groups of methods according their output. A method in the first group produces a VQ-based image (stego-image) in which the secret data are embedded; a method in the second group produces a compressed codestream consisting of the cover image and the embedded secret data. This study focuses on the related problems of the second group. A novel VQ-based embedding method is proposed to optimize the bit rate of the compressed codestream, given a fixed number of secret bits per block for embedding.

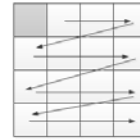
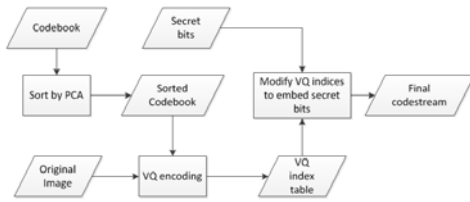
The rest of this paper is organized as follows. Section 2 introduces Wang et al.'s embedding scheme. Section 3 presents the proposed method. Section 4 shows the experimental results, and Section 5 concludes this paper.

## 2 Background

This section introduces Wang et al.'s embedding scheme [5], which compresses the cover image and the secret data into a codestream as the output.

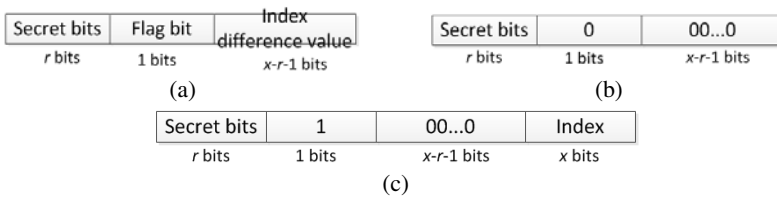
### 2.1 Lossless Image Information Hiding by Wang et al.'s Scheme

In 2010, Wang et al. proposed an encoding method for both image compression and lossless information hiding [5]. The method sorts the codebook by PCA first. Then, the embedding phase and the extraction phase use the sorted codebook to process the input data. Figure 1 shows the concept of the embedding phase while Figure 2 demonstrates the embedding order of the blocks. In the embedding phase, the difference value between the current embedding index and the previous index is computed, and then the difference value is used to decide how the encoding is proceeded, as shown in Figure 3.



**Fig. 1** The embedding flowchart of Wang et al.'s scheme.

**Fig. 2** The embedding order of Wang et al.'s scheme (2010).



**Fig. 3** The encoding rules proposed by Wang et al. (2010), where  $t = 2^{x-r-1} - 1$ ,  $x = \log_2$  (the codebook size), and  $r$  = the size of the secret bits.

### 3 The Proposed Method

This research proposes a VQ-based information hiding scheme based on the concept of index difference. We devise a new strategy, namely codebook re-ordering, to shorten the range of the index-difference values of the neighboring blocks. This strategy reduces the space overhead while embedding the secret data. The method uses a coefficient variable  $s$  to decide the number of secret bits to be embedded per block. Given a VQ-encoded image, the total embedding capacity of the proposed method is equal to the number of image blocks multiplies a user-defined integer coefficient  $s$ . In this section, we will introduce how codebook re-ordering works.

#### 3.1 Codebook Re-ordering

We observed that, in most cases of the VQ-encoded images, the difference value of the indices of any two neighboring blocks is not close. Moreover, the difference value of the indices of two neighboring blocks is usually large even though their corresponding codewords are very similar. In order to compact the distribution of the difference values, we sort the codebook by a similarity function, which calculates the similarity of the pointed codeword of a given index and the pointed codeword of the first index. Figure 4 shows the concept of codebook re-ordering.

The index values in a VQ index table usually have a wider distribution. When the indices in the codebook are sorted according to their similarity to the first index, each index value in the corresponding VQ index table (*Index table*' in Figure 4) should have more similar index values in its neighborhood. The re-ordered codebook is then used in the data-embedding phase and in the extraction phase. This strategy shifting the distribution of the difference values to a smaller range, and transforms the difference values into smaller values as well.

### 3.2 Data Embedding

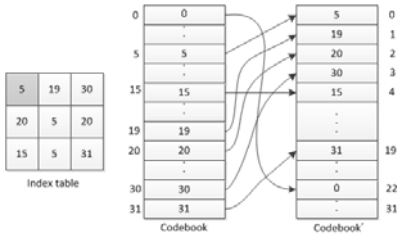
In this sub-section, we will show how secret data are embedded. The algorithm follows the concept of difference-value encoding, and uses the technique of codebook re-ordering to improve the bit rate of the output codestream (*CS*). The algorithm uses a user-defined coefficient to decide the number of secret bits to be embedded per block. For convenience of description, a function called *sizeof* is used in the proposed embedding method shown in the following to calculate the number of elements in a codebook.

#### Algorithm DataEmbedding:

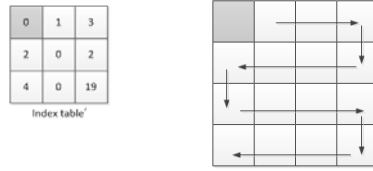
**Input:** cover image (*CI*), codebook (*CB*), secret bits (*S*), user-defined coefficient (*s*)

**Output:** codestream (*CS*)

1. Partition *CI* into 4×4 blocks and set *CS* to NULL. Find the most similar codeword  $cb_0$  of the first block in *CB*, and then append the index of codeword  $cb_0$  to *CS*.
2. Sort *CB* according to how similar a codeword and  $cb_0$  are. Thus, a new re-ordered codebook *RCB* is created for encoding.
3. Use *RCB* to encode all blocks, and produce a new VQ-encoded image *CI'*.
4. Follow the order shown in Figure 5 to calculate the difference values from the second block for image *CI'*. A difference value *d* is defined as the result of the index value of the current block subtracting the index value of the previous block.
5. If the absolute difference value  $|d|$  is equal to zero, go to Step 6. If  $|d|$  is larger than or equal to  $2^{(\log_2(\text{sizeof}(CB)) - s - 2)}$ , go to Step 7. Otherwise, go to Step 8.
6. Append (*s*-bit *S* || 10) to *CS*.
7. Append (*s*-bit *S* || 11 || the current index) to *CS*.
8. If the absolute difference value  $|d|$  is large then zero, append (*s*-bit *S* || 00 ||  $d$ ) to *CS*. Otherwise, the difference value is small then zero, append (*s*-bit *S* || 01 ||  $|d|$ ) to *CS*.
9. Output *CS* and terminate the algorithm unless all blocks are processed. Otherwise, go to Step 4.



**Fig. 4** The concept of codebook re-ordering.



**Fig. 5** The embedding phase for the proposed method.

### 3.3 Data Extraction and Reversion

To extract the secret data and the cover image from an encoded codestream, we have to construct the re-ordered codebook that was used in the embedding phase. This first step is to extract the first index of the original VQ-encoded image, which is stored at the beginning of the encoded codestream. The next step is to use the index to find its corresponding codeword, and then sort the codebook to get the re-ordered codebook. Consequently, the one-to-one mapping relations between the original codebook and the re-ordered codebook are calculated. With the mapping relations, the original VQ index in a block along with the embedded secret bits can be easily reversed and extracted from the corresponding code pattern in the codestream.

## 4 Experimental Results

We have conducted two experiments to compare the proposed method with Wang et al.'s scheme in terms of bit rate. In the experiments, six 512×512 gray-level images were used as the cover images (a) Lena, (b) Peppers, (c) Baboon, (d) Boat, (e) Goldhill, and (f) F16. Two codebooks, one consisting of 256 codewords and the other consisting of 512 codewords, were generated by the LGB algorithm. The secret data were generated randomly in bits (0 and 1). In the experiments, both methods embed  $s$  secret bits per block, where  $s$  is equal to either 1, 2, and 3. Table 1 shows the results of the first experiment in terms of bit rate. The experiment used different images with a codebook size of 512 to evaluate the bit rate values of the proposed method and Wang et al.'s scheme. The results show that, compared with Wang et al.'s scheme, the proposed method only uses about 90% of the bit rate to deliver the same secret data and image given  $s$  is 1. While  $s$  is 1 or 2, the proposed method still outperforms Wang et al.'s scheme for all the test images. Table 2 shows the results of the second experiment, which uses different images with a codebook size of 256 to evaluate both the proposed method and Wang et al.'s scheme. When each block only embeds one secret bit ( $s=1$ ), the average bit rate of the proposed method is about as small as 85% of the bit rate of Wang et al.'s scheme. While  $s$  is equal to 2, the proposed method still significantly outperforms Wang et al.'s scheme by about 9% less bit rate. While  $s$  is equal to 3, the proposed method is on average better but the difference is not significant.

**Table 1** Comparisons on the proposed method and Wang et al.'s scheme using different images with a codebook size of 512

<i>CB=512</i>	<i>s = 1</i>		<i>s = 2</i>		<i>s = 3</i>	
	<i>Wang et al.</i>	<i>Proposed</i>	<i>Wang et al.</i>	<i>Proposed</i>	<i>Wang et al.</i>	<i>Proposed</i>
Lena	0.611	0.564	0.670	0.624	0.741	0.715
Pepper	0.601	0.541	0.645	0.613	0.715	0.709
Baboon	0.612	0.606	0.727	0.686	0.857	0.782
Boat	0.600	0.484	0.641	0.549	0.696	0.642
GoldHill	0.590	0.531	0.641	0.607	0.726	0.699
F16	0.599	0.509	0.642	0.563	0.689	0.651

**Table 2** Comparisons on the proposed method and Wang et al.'s scheme using different images with a codebook size of 256.

<i>CB=256</i>	<i>s = 1</i>		<i>s = 2</i>		<i>s = 3</i>	
	<i>Wang et al.</i>	<i>Proposed</i>	<i>Wang et al.</i>	<i>Proposed</i>	<i>Wang et al.</i>	<i>Proposed</i>
Lena	0.549	0.484	0.598	0.549	0.662	0.648
Pepper	0.544	0.463	0.580	0.546	0.640	0.639
Baboon	0.568	0.547	0.668	0.626	0.722	0.720
Boat	0.538	0.414	0.569	0.499	0.613	0.572
GoldHill	0.529	0.465	0.541	0.537	0.569	0.633
F16	0.537	0.409	0.566	0.486	0.606	0.593

## 5 Conclusions

This work proposed a reversible information hiding method for VQ-compressed images. It has high compression rate and high embedding capacity while encoding the secret data and the image. It can extract the original image and the embedded secret data easily. Since a block in an image usually has similar neighboring blocks, the proposed method reconstructs a new codebook by rearranging the original codebook based on the concept of similarity. The new codebook is used to transform the original VQ-encoded image into a new VQ-encoded image. Then, the proposed method uses the concept of index difference to transform the new VQ-encoded image into codestream. We conducted some experiments to compare the proposed method and the latest similar work proposed by Wang et al. in 2010. The experimental results showed that the proposed method has better bit rate than Wang et al.'s scheme, in particular when a block and its neighbors in an image are similar.

**Acknowledgements.** This research was partially supported by the National Science Council of the Republic of China under the Grant NSC 99-2218-E-008-012, NSC 98-2221-E-015-001-MY3, and NSC 99-2918-I-015-001.



## References

1. Youtube, <http://www.youtube.com/>
2. Shen, J.J., Ren, J.M.: A robust associative watermarking technique based on vector quantization. *Digital Signal Processing* 20(5), 1408–1423 (2010)
3. Chen, T.H., Wu, C.S.: Compression-unimpaired batch-image encryption combining vector quantization and index compression. *Information Sciences* 180(9), 1690–1701 (2010)
4. Chang, C.C., Pai, P.Y., Yeh, C.M., Chan, Y.K.: A high payload frequency-based reversible image hiding method. *Information Sciences* 180(11), 2286–2298 (2010)
5. Wang, Z.H., Chang, C.C., Chen, K.N., Li, M.C.: An encoding method for both image compression and data lossless information hiding. *Journal of Systems and Software* 83(11), 2073–2082 (2010)
6. Luo, W., Huang, F., Huang, J.: Edge adaptive image steganography based on LSB matching revisited. *IEEE Transactions on Information Forensics and Security* 5(2), 201–214 (2010)
7. Shie, S.C., Lin, S.D., Jiang, J.H.: Visually imperceptible image hiding scheme based on vector quantization. *Information Processing & Management* 46(5), 495–501 (2010)
8. Chow, S.S.M., Yap, W.S.: Partial decryption attacks in security-mediated certificateless encryption. *IET Information Security* 3(4), 148–151 (2009)
9. Lin, I.C., Lin, Y.B., Wang, C.M.: Hiding data in spatial domain images with distortion tolerance. *Computer Standards & Interfaces* 31(2), 458–464 (2009)
10. Chen, W.J., Huang, W.T.: VQ indexes compression and information hiding using hybrid lossless index coding. *Digital Signal Processing* 19(3), 433–443 (2009)
11. Lu, Z.M., Wang, J.X., Liu, B.B.: An improved lossless data hiding scheme based on image VQ-index residual value coding. *Journal of Systems and Software* 82(6), 1016–1024 (2009)
12. Wang, J.X., Lu, Z.M.: A path optional lossless data hiding scheme based on VQ joint neighboring coding. *Information Sciences* 179(19), 3332–3348 (2009)

# Knowledge Management System: Combination of Experts' Knowledge and Automatic Improvement

Hiroshi Sugimura and Kazunori Matsumoto

**Abstract.** This paper proposes a knowledge management system that acquires knowledge from time series data by using users background knowledge. To obtain if-then rules as knowledge, we apply decision tree learning. However usual methods of decision tree learning targets discrete value data, thus a new approach is needed for dealing with this type of data. Experts forecast future events by using their knowledge. They, in typical cases, focus on a set of useful patterns and then apply knowledge relevant to them. We apply this idea into the framework of decision tree learning. We prepare a set of patterns, which is called clues, and then express time series data in terms of the clues. Thus the clues are attributes by which features of data are described. In addition, to make a better prediction with the learning process, we develop a mechanism that improves the quality of the clues. The essential idea of the mechanism is based on a genetic algorithm. The clue is evaluated by using entropy of information theory, and is improved by GA operators. We can obtain new knowledge from improved clues and the extracted decision tree. This paper details the system and results of the experiment.

## 1 Introduction

This paper focuses on time series data which are sequences of numerical values. This type of data is occurred widely in many practical areas such as finance, medical, science, and so on. Thus, analysis or datamining methods of them is very important [1, 2]. This study incorporates pattern discovery into the decision tree learning. The decision tree learning has an advantage in that it easily transforms knowledge to if-then rules, and a lot of successful cases have been reported [3, 4]. However, the node of typical decision tree learning cannot deal with numerical attributes, especially

---

Hiroshi Sugimura · Kazunori Matsumoto

Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi-shi,

Kanagawa 243-0292, Japan

e-mail: [hiroshi.sugimura@gmail.com](mailto:hiroshi.sugimura@gmail.com), [matumoto@ic.kanagawa-it.ac.jp](mailto:matumoto@ic.kanagawa-it.ac.jp)

time series data. We thus need to preprocess data in this case. A simple method for this purpose is to rearrange data to its average. However, this disregards the shape of behavior of sequence, thus two sets of data may be incorrectly deemed to be similar even if they have greatly different shapes. Time series has many more features that must be taken into account of it.

In [5], they develop a method that acquires knowledge by using decision tree learning. This method considers data sets that consist of multiple time series attributes. On the other hand, we aim to discover the knowledge that of from one-long-period. In [6], a method to predict future behavior by using clustering is proposed. However, several studies point out that brute force exploration often identifies meaningless simple behavior like a cosine curve [7, 8].

A useful and non-trivial pattern is hard to identify automatically. Thus, we use human experience, which is expressed in clues, as hints of useful and non-trivial patterns. They become features on attributes in the learning. In another characteristic of time series data, we need a method to manage ambiguity in data. We need to identify sets of data that have similar shapes, different lengths, different values at some points, etc. For this propose, we use a technology developed in image processing.

In addition, we describe a method that improves clues by using the genetic algorithm, GA for short. GA is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics [9]. This heuristic is generally used to generate useful solutions for optimization and search problems. In [10], they propose a method to encode and decode a decision tree to and from a chromosome where genetic operators such as mutation and crossover can be applied are presented. Our approach indirectly increments the quality of a decision tree by improving of clues. This paper also outlines the entire system and the methodology.

## 2 Time Series Data Mining by Using Clues

Much empirical knowledge is in the financial areas, especially stock price analysis and prediction. In this area, typical knowledge is used to predict future price changes on the basis of examinations of past price changes. Most knowledge in this area is described in terms of charts and patterns. A pattern is a subsequence of time series data and identifies a distinctive behavior of the sequence. In many cases, patterns are effectively used to represent knowledge of data. The existing knowledge of patterns plays an important role in our discovery process. Discovery of a pattern is a difficult problem so that we adopt a stepwise approach that gradually acquires a pattern starting with a user-specified clue.

Our method has two advantages. Firstly, the system can enable to use the user's knowledge. When the user knows already effective patterns in similar situations, such knowledge way is efficiently reused in case of a new situation. Secondly, the system can discover a set of superior clues. If the user does not know any applicable patterns, the system randomly generates the initial clues, and discovers useful knowledge by improving these clues. Improved clues are finally discovered as new knowledge.

## 2.1 Outline of the System

Fig. 1 shows an outline of the proposed system. The system consists of five steps. First, a user inputs and selects target data: clues, raw databases, and real-time data. The target data is then cleaned. Noise and missing data are processed in a predefined manner. Initial clues are also given here. Second, the system carries out a preprocessing operation for common processes. The collected data come in various types. We thus need to normalize them. Third, the knowledge is discovered by the machine learning with clues. Fourth, discovered knowledge is evaluated. In accordance with this evaluation, the system chooses “stop” or “improvement”. If it chooses “stop”, the system outputs the result that is a set of discovered knowledge and used clues. When it chooses “improvement”, the system carries out step five. Fifth, the clues are automatically improved on the basis of the evaluation, and return to the third step. This process is repeated until a termination condition has been reached in the fourth step. Next, we describe the data mining method based on clues and also the method for improving clues.

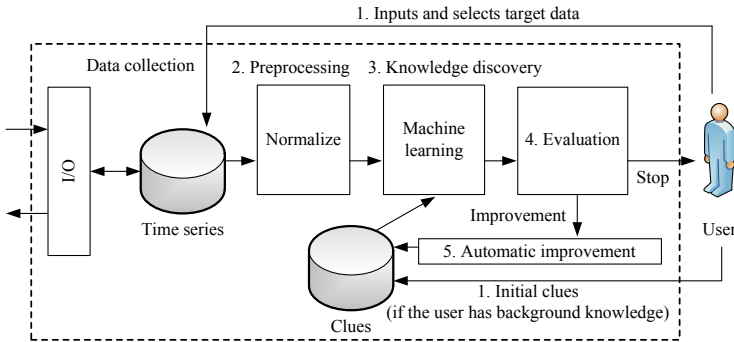


Fig. 1 An outline of the system

## 2.2 Training data

Fig. 2 shows a simple example of the method that makes decision tree from training data. Clues are given to a data mining processor and used as 'focus points' in the mining algorithm. We thus use these as attributes. Fig. 2(a) illustrates an example of training data. Clues  $P_0$  to  $P_2$  is the given set of clues in this example. For all training data, we compute the degree of matching to each clue, these values are attribute values. Each instance in the training data is associated with a class label, which shows future behavior. A table consists of tuples including the value and classes is training data. After applying a learning procedure, we obtain knowledge as shown in Fig. 2(b).

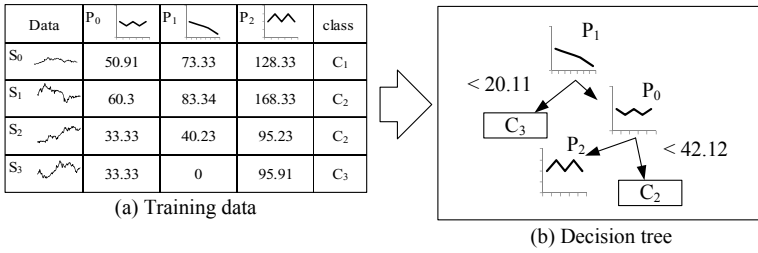


Fig. 2 Example of training data and result

Ambiguity of knowledge is a crucial issue in dealing with clues. Clues define abstract behaviors of time series data. Thus, actual data seldom match the clues. First, this clue may correspond to short or long period behavior. Thus we can say there is horizontal ambiguity. Similarly, there is vertical ambiguity that handles differences of values.

### 2.3 Pattern Matching

Euclidean distance or its simple extensions are brittle distance measure. They may fail to produce an intuitively correct measurement of similarity between two sequences because it is sensitive to small distortions in the time axis. Fig. 3(a) shows correspondence point pairs by using Euclidean distance. The two sequences have approximately the same component behavior, but this behavior does not line up in time axis. Fig. 3(b) shows the nonlinear alignment that allows a more sophisticated distance measurement to be calculated.

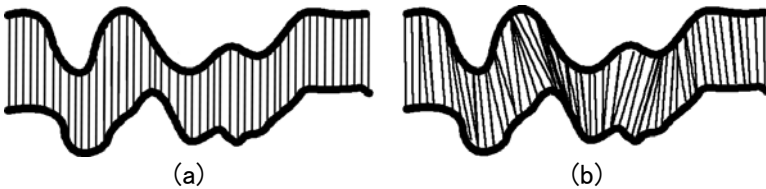


Fig. 3 Euclidean distance and DTW distance

For these ambiguities, we apply the dynamic time warping [11], DTW for short. DTW computes dissimilarity of two sequences by using predefined cost parameters. For two sequences that have different lengths  $i$ , and  $j$ , the dissimilarity degree  $g(i, j)$  is defined in the following equation, where  $i - 1$  and  $j - 1$  is the sequences with the last values removed.

$$g(i, j) = \min \{ g(i, j - 1) + q, g(i - 1, j) + r, g(i - 1, j - 1) + s \} \tag{1}$$

In the equation (1),  $q$  and  $r$  represent the cost of shortening and expanding the sequences along the time axis, and  $s$  is the distance cost of values. A more similar pair has a smaller value, which becomes zero when they are exactly the same.

### 3 Improvement of Clues

As we described above, a set of knowledge works as clues of knowledge discovery. We naturally expect better clues discover better knowledge. The quality of knowledge depends on the quality of clues. The system improves the quality of decision tree by improving clues automatically. For the automatically improvement, the local search (hill climbing, for example) and the reinforcement learning (Q-learning, for example) are applicable. However, it may obtain a local optimal solution. Our solution is to use the genetic algorithm.

The representation of a clue is an array of numerical values. Each clue becomes a gene by normalization. A set of current genes becomes current generation of a family. Fig. 4 shows representation of a gene that corresponds to a clue.

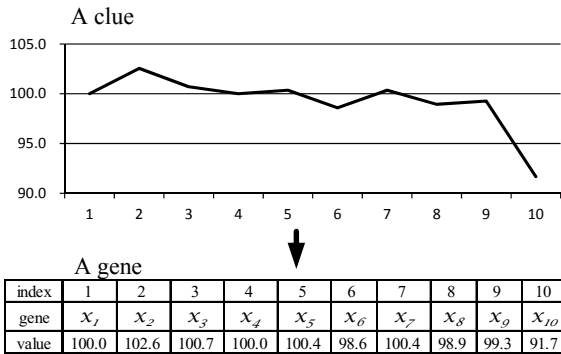


Fig. 4 A clue and a gene

The GA obtains a better next generation of family by combining two genes. The GA makes next generation of family by using GA operators which are selection and reproduction step. Selection operator chooses several genes in accordance with fitness. To compute the fitness of genes, the system utilizes the information gain ratio criterion. The most appropriate gene for classifying all instances of training data is selected by using this criterion. As we describe in Fig. 2(a), all instances are classified by future behavior, and similarities are computed among their sequences and all genes.

Let the total number of positive instances be  $a$ , total number of negative instances is  $b$ , then gain ratio ( $G(X_i)$ ) by using attribute  $X_i(1 \leq i \leq q)$  is defined by following equations:

$$G(X_i) = \frac{I(a,b) - E(X_i)}{S(X_i)} \quad (2)$$

$$I(a,b) = -\frac{a}{a+b} \log_2 \frac{a}{a+b} - \frac{b}{a+b} \log_2 \frac{b}{a+b} \quad (3)$$

$$E(X_i) = \sum_{j=1}^v \frac{a_j^i + b_j^i}{a+b} I(a_j^i, b_j^i) \quad (4)$$

$$S(X_i) = -\sum_{j=1}^v \frac{a_j^i + b_j^i}{a+b} \log_2 \frac{a_j^i + b_j^i}{a+b} \quad (5)$$

Selection step rates the fitness of each gene and preferentially selects the best gene. We use roulette wheel selection [9] to do this. A proportion of the wheel is typically allocated to each of the possible selections on the basis of their fitness value. In accordance with fitness proportionate selection, some weaker genes may survive the selection step. As although a gene may be weak, it may contain some components that could prove useful following the reproduction step. We thus consider as an advantage. The roulette wheel selection algorithm consists of the following steps.

1. Compute the total number of fitness of all family members as  $n$ ,
2. Generate a random number between 0 and  $n$ , and
3. Select the family members whose fitness added to the fitness of the preceding family members is greater than or equal to  $n$ .

The reproduction step is to create a next generation family of genes from those selected through genetic operators that are crossovers and mutated. To create each new gene, a pair of parent genes is selected for breeding from the pool selected previously. By creating a child gene using these methods, a new gene is created that has many of its parents' characteristics. The process continues until a new family of genes of suitable size is generated. The mutation step adds to the current value of a gene with a randomly generated valid value that is in the predefined range.

## 4 Experiment

To confirm the effectiveness of our proposed system, experiments in this paper use both artificial and real data. The artificial data are generated by synthesizing predefined and random patterns. The real data are two kinds: stock price and climate data. By using these kinds of data, our problem is to predict a future behavior based on immediately before 20 days observations.

Time series data are extracted by using the slide window method. According to the problem here, we set the slide window length to 20. As for a future behavior to be predicted, we take 5 data into consideration. In accordance with behavior of the period, time series data is classified. We discretize into three classes: *up*, *down*, and *stay*. If at least a value in this period, 5 days, exceeds the initial value of this period more than  $x\%$ , then we regard this behavior as *up*. On the other hand, at least a value decreases more than  $x\%$ , then we say that *down*. From this definition, a behavior

could be classified simultaneously *up* and *down*, but we simply put priority over *up*. A behavior which belongs to neither is called *stay*.

This rate  $x$  is determined by a prior experiment to reduce the class distribution bias of the dataset. This is because an extremely unbalanced dataset will increase the predictability using the bias. We use the C4.5 algorithm [12] within the WEKA [13], which is developed at the University of Waikato in New Zealand.

In the experimental for the improvement of clues, GA increases the family of clues one hundred times. The system shows the clues and the decision tree that has the highest accuracy. When accuracy is the same, we prefer smaller trees. To evaluate the prediction accuracy, we use the 10-fold cross validation. The variance is calculated to confirm the stability of accuracy.

### 4.1 Artificial Data

To make the artificial data sets, we prepare a set of typical patterns. This is called a knowledge set. We also prepare a set of random patterns. These patterns, typical and random, are merged into a sequence, and then its future behavior is set to one of *up*, *down* or *stay*. Fig. 5 summarizes this operation. In this figure, from  $C_0$  to  $C_2$  corresponds to *up*, *down*, *stay*, respectively. Priorities are used to decide future behaviors.

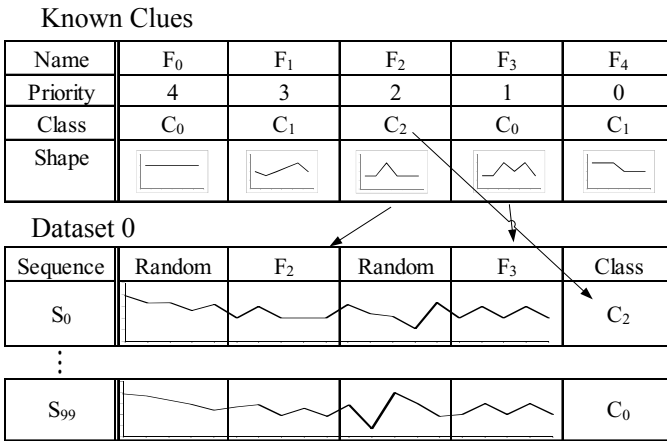


Fig. 5 Generating artificial data

To compare the techniques, we carry out three kinds of experiments which are direct approach, using a knowledge set, and using a random set. We experiment on 20 artificial time series data sets. The value of  $x$  is decided to 0.05. In the case of the system using clues, we measure the accuracy with various populations. Table 1 shows the average accuracy, the variance of accuracy, and the average size of a decision tree.



**Table 1** Experimental results of artificial data

		Accuracy	Variance	Size
Direct		69.16	38.7	16.50
Random clues	5 genes	65.63	5.8	48.83
	10 genes	71.02	9.9	44.39
	20 genes	72.34	14.7	34.44
GA	5 genes	68.11	13.1	32.52
	10 genes	72.02	8.8	35.64
	20 genes	73.31	14.9	48.01
Known clues	5 genes	71.80	8.4	46.48
	10 genes	75.16	15.0	35.35
	20 genes	77.99	16.2	22.40
GA	5 genes	79.87	17.0	16.60
Known	10 genes	82.29	11.2	20.00
Clues	20 genes	84.27	4.4	22.68

## 4.2 Stock Price Data

The stock price data is consisted of the opening price, the closing price, the highest price, and the lowest price of the trading days. We compare the accuracy and the stability of three methods, which are a direct approach, the trading rule mining method explained in the section 1, and our approach. The direct approach and our method use the closing prices. Our approach generates ten random clues, they become initial genes.

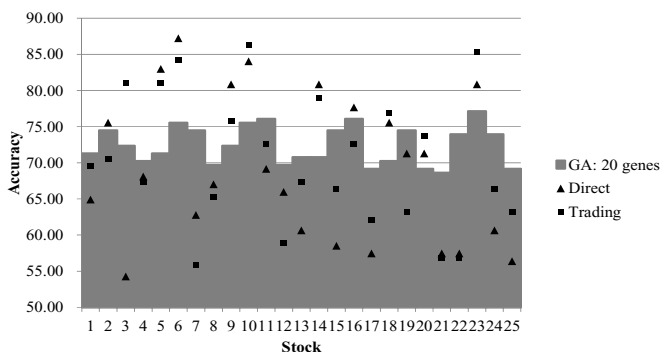
The trading rule mining method is explained in the section 1. In addition to the four types of price values in the above, we use trading volume, and 13 trend index values, which are Moving Average, Bollinger Band, Envelope, HLband MACD, DMI, volume ratio, RSI, Momentum, Ichimoku1, Ichimoku2, Ichimoku3, and Ichimoku4.

We obtained time series data consisting of the above mentioned attributes about twenty five companies. The period, from we have collected the time series stock data, is from 5th January 2006 to 29st December 2006.

Table 2 shows experimental results of stock price data. The size of the tree and the accuracy are the average of twenty five companies. Fig. 6 shows the graph of the accuracy obtained by each approach. In this graph, the horizontal axis shows the number of corporations, and the vertical axis shows the accuracy. The gray bars show the results of the proposed method with 20 generations GA. As a result, the accuracy is in fact improved. Further, the proposed method is superior in that it reaches high stability in average with low variances.

**Table 2** Experimental results of stock price data

Method	Size	Accuracy	Variance
Direct approach	22.52	69.15	97.69
Trading rule mining	14.92	70.32	80.84
5genes	20.68	60.47	66.29
GA: 5genes	28.96	67.05	20.06
10genes	34.44	67.09	10.49
GA: 10genes	37.45	70.82	9.97
20genes	33.48	67.00	14.27
GA: 20genes	36.35	72.29	6.76



**Fig. 6** Accuracy of each approach

### 4.3 Climate Data

We use daily data of atmosphere. The data set totals for 10 stations in Japan are obtained from the Global Historical Climate Network (GHCN) [14] for 50 years from 1st June 1959 to 31th May 2009. Table 3 shows results of the atmosphere.

**Table 3** Results of the atmosphere

Method	atmosphere			
	Size	Accuracy	Variance	Time(m)
Direct	5.2	99.79	0.01	0.15
5 genes	0.6	99.77	0.00	2.15
GA: 5 genes	2.4	99.78	0.00	7.50
10 genes	2.6	99.79	0.00	2.35
GA: 10 genes	5.2	99.81	0.00	14.50
20 genes	2.6	99.79	0.00	2.83
GA: 20 genes	5.0	99.83	0.00	28.30

## 5 Discussion

In Fig. 6 the prediction accuracy of our approach is higher than other methods, and its stability is also good. On the other hand, in Table 3 the direct approach obtains high prediction accuracy as with the proposed method. We think that we should apply the proposed method to data which have complicated behavior, such as stock price, rather than simple data, such as atmosphere. Because our approach takes longer than others for the processing time.

Table 3 shows that our approach takes longer for processing than the direct approach. The reason that is the decision tree is made many times in the process of GA. In the experiment in this paper, the GA process is simply terminated after producing 100th generations. However, it is thought that the processing time can be reduced by improving the termination condition. For example, we set the termination condition that is an accuracy threshold of a high value, such as 80%. It seems that we should apply our method to predict the long period rather than real time.

Next, we compare the degree of improvement. In Table 1 when five known clues are given, the accuracy improves by 8.07% (from 71.8% to 79.9%). When five random clues are given, the accuracy improves by 2.48% (from 65.63% to 68.11%). Table 2 shows the improvement in accuracy, but, Tables 3 do not. These results, demonstrate that the GA-based mechanism enhances the prediction accuracy.

## 6 Conclusion

This paper described a knowledge management system that acquires knowledge from time series data by using experts' background knowledge. To obtain if-then rules as knowledge, we apply decision tree learning. However usual methods of decision tree learning targets discrete value data, thus a new approach is needed for dealing with this type of data. Experts forecast future events by using their knowledge. They, in typical cases, focus on a set of useful patterns and then apply knowledge relevant to them. The main point of the system is a use clues of which are suggestive patterns for analysis. We regard clues as features of data and then express time series data in terms of the clues. The clue is evaluated by using information theoretic entropy, and is improved by the genetic algorithm. We can obtain new knowledge from improved clues and the extracted decision tree. The experimental results demonstrated the effectiveness of the system by using both the artificial and real data. As a result, the method is effective in the view points of accuracy and stability.

## References

1. Dong, G., Pei, J.: Sequence Data Mining. Springer (2007)
2. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Pub. (2005)

3. Pješivac-Grbović, J., Bosilca, G., Fagg, G.E., Angskun, T., Dongarra, J.: Decision trees and MPI collective algorithm selection problem. In: Kermarrec, A.-M., Bougé, L., Priol, T. (eds.) Euro-Par 2007. LNCS, vol. 4641, pp. 107–117. Springer, Heidelberg (2007)
4. Kubota, K., Nakase, A., Sakai, H., Oyanagi, S.: Parallelization of decision tree algorithm and its performance evaluation. *IPSJ SIG. Notes* 99(66), 161–166 (1999)
5. Yamada, Y., Suzuki, E., Yokoi, H., Takabayashi, K.: Decision-tree induction from time-series data based on a standard-example split test. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML), pp. 840–847 (2003)
6. Abe, H., Hirabayashi, S., Ohsaki, M., Yamaguchi, T.: Evaluating a trading rule mining method based on temporal pattern extraction. In: The Third International Workshop on Mining Complex Data (MCD 2007) In Conjunction with ECML/PKDD 2007, pp. 49–58 (2007)
7. Keogh, E., Lin, J.: Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and Information Systems* 8(2), 154–177 (2005)
8. Warren Liao, T.: Clustering of time series data—a survey. *Pattern Recognition* 38(11), 1857–1874 (2005)
9. Holland, J.H.: *Adaptation in Natural and Artificial Systems*. University of Michigan Press (1975)
10. Cha, S.-H., Tappert, C.: A genetic algorithm for constructing compact binary decision trees. *Journal of Pattern Recognition Research (JPRR)* 4(1), 1–13 (2009)
11. Berndt, D.J., Clifford, J.: Using Dynamic Time Warping to Find Patterns in Time Series. In: Proceedings of KDD 1994: AAAI Workshop on Knowledge Discovery in Databases, Seattle, Washington, pp. 359–370 (July 1994)
12. Ross Quinlan, J.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers (1993)
13. The University of Waikato, Machine learning project at the university of waikato in new zealand, <http://www.cs.waikato.ac.nz/ml/>
14. NESDIS-National Environmental Satellite, Data, and Information Service, Nndc climate data online, <http://www.nesdis.noaa.gov/>

# Evaluating Recommender System Using Multiagent-Based Simulator

## Case Study of Collaborative Filtering Simulation

Ryosuke Saga, Kouki Okamoto, Hiroshi Tsuji, and Kazunori Matsumoto

**Abstract.** This paper describes a agent-based simulation system to evaluate recommender systems. Recommender systems have attracted attention to present items found by preference of users. Many algorithms for recommender system are developed but the comparisons between their algorithms are difficult because of limited data set and the difficulty of constructing simulator environment. In order to resolve them, we develop agent-based recommender system simulator. This simulator constructs the simulator environment based on the network model, and lets recommender agent recommend items to agents, evaluates the items, and summarizes(outputs) the recommendation results. In the experiment on 100 agents, we can confirm the usability of our simulator because of recapturing the feature of collaborative filtering by this simulator.

## 1 Introduction

Recommender systems have been used for several applications and systems such as news sites, information sharing system, e-commerce and soon [1, 2, 3]. The systems offer benefits to consumer and item providers. These recommender systems help consumers in particular acquire new as well as preferable items and users can expect effective acquisition of the information. Therefore, an appropriate algorithm is needed for the recommender system.

Several researchers have proposed and developed many algorithms since collaborative filtering, one of the most successful technologies for recommender

---

Ryosuke Saga · Hiroshi Tsuji

Osaka Prefecture University, 1-1 Gakuen-cho, Naka-ku, Sakai, Osaka, Japan

e-mail: {saga, tsuji}@cs.osakafu-u.ac.jp

Kouki Okamoto · Kazunori Matsumoto

Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Kanagawa, Japan

e-mail: matumoto@ic.kanagawa-it.ac.jp

systems, was introduced and attracted many attentions [4, 5, 6, 7]. However, developing and applying the algorithm of collaborative filtering are difficult. One reason is based on the difficulty of evaluating these algorithms. For example, the algorithms were developed for any purposes and have validated specified and limited datasets in many cases. Therefore, the generality of the algorithms is not clear. Another reason for the difficulty is that the validation of the algorithm needs massive dataset. From the reasons, we developed a multi-agent-like simulator for evaluating the collaborative filtering and it is described in this paper.

## 2 Motivation and Requirement

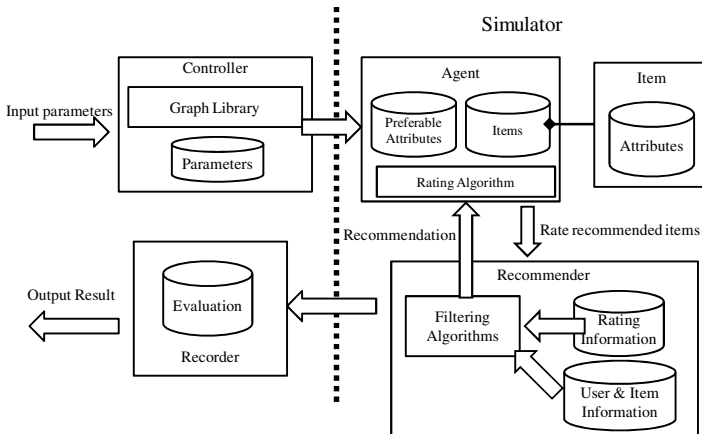
Recommender systems aim to recommend preferable items to users from user profile [1, 2, 3]. The profile is constructed by analyzing the content, user's voting and rating, and access logs, etc. Two types of filtering algorithms are used for dynamic recommendation: content-based filtering and collaborative filtering [6]. Especially, collaborative filtering is the most successful algorithms, and its profile is based on relationships among users or items [7]. It has an advantage wherein collaborative filtering is applicable for any items because the algorithm does not need to analyze the content itself. Also, the hybrid algorithm combining collaborative filtering and content-based filtering [8] has also been developed.

Generally, the recommender system algorithms work better as the dataset that includes the user's rating and item information becomes more massive. However, the algorithms do not work for a small dataset because the dataset is insufficient for calculating similarity and predicting items (called cold-start problems).

Therefore, developing the algorithms has various problems associated with it. The first problem is a limited dataset. To evaluate the algorithms, we need to use various environments by collecting various datasets. However, collecting the various datasets is difficult because we generally do not make use of recommender systems and do not have the data source. Therefore, many researchers have utilized limited dataset such as MovieLens Dataset and EachMovie Dataset. The evaluation measurement may change according to the goals of the algorithm. The algorithm is developed for specified goals, and in order to evaluate the algorithm, the proper evaluation measurements and data set are needed [3]. Also, each method has strong/weak points, and if we apply the algorithms for other goal, we cannot easily judge whether any of the algorithms are suitable. In order to identify the suitable algorithm, to compare the algorithms is useful; however, an experiment with a limited dataset and with different goals is difficult.

Therefore, we build the simulator to enable the comparison of the algorithms. The requirements/goals of simulator are defined as the followings.

1. The simulator can build the evaluation environment for the recommender system.



**Fig. 1** Simulator architecture

2. The simulator can compare the filtering algorithms of collaborative filtering and content-based filtering.
3. The simulator can output the results of evaluations to compare the filtering algorithms.

### 3 Instructions to Authors

This simulator consists of agents, items, Recommender, Controller, and Recorder as shown in Fig. 1. In this simulator, a simulator user gives the number of agents, items, thresholds, and algorithms as parameters. Agent acts as the user of recommender systems, and the algorithm of collaborative filtering is modeled into Recommender. Recommender has information on agents and items' ratings for each user. Controller receives parameters from the simulator user and handles the simulator such as initialization, progression, and suspension. Controller accepts not only the number of agents and items, but also thresholds for the simulator environment, preference, and agent status.

The simulation steps are as follows:

1. The simulator user inputs parameters.
2. Building simulator environment: Controller initializes the status of the agents and items and configures the simulator based on the parameters.
3. Recommendation: Recommender calculates the user similarity and recommends items to agents.
4. Rating items: The agents vote on the rating of recommended items and update the status.

5. Output result: Recorder compiles the result of recommendation and evaluates the recommendation using several measurements such as MAE, recall, precision, novelty, diversity, and discovery [9, 10].
6. Controller updates agent status and preference.

The simulator regards step 3 to step 6 as one turn, and it goes through the steps and iterates the turns.

## 4 Architecture of Simulator

### 4.1 Agents, Items, and Ratings: Definition of Attributes

The key point of modeling recommender systems is the preferences of agents for items. This paper assumes that the agents and items have many attributes. Also, we assume that the agents have specified preferable genres and domain, and in order to express the assumption, the simulator lets agents have a positive real number for all attributes for their degree of preference. The simulator also defines the attributes whose degree of preference is over the *preference threshold* as preferable attributes. In contrast, an item has the 0/1 flags for each attribute.

### 4.2 Rating Items

In this simulator, the agents and Recommender evaluate each item. For the agents, the rating  $r_{ij}$  of an item  $j$  for an agent  $i$  is generally evaluated by the equation,

$$r_{ij} = A(U_i, I_j, u_i) \quad (1)$$

Here  $U_i$  is the preferable attribute set of agent  $i$ . Also,  $I_j$  is the attribute set of Item  $j$  whose value is 1.  $u_i$  is the information about agent. Also, we call the function a rating algorithm. This function  $A()$  is updated according to the agent preference model. For an example of rating algorithm, we can configure it as follows

$$r_{ij} = \frac{\sum_u \frac{u}{\max(U_i)}}{|U_i \cap I_j|} \quad (2)$$

where  $|E|$  indicates the number of the elements of set  $E$ ,  $u$  is the elements of attributes of  $U_i$  corresponding to the index of  $U_i \cap I$  and  $\max(U_i)$  indicates the maximum of  $U_i$ , which is given by users as one of parameters.

On the other hand, Recommender, which is the implementation of the filtering algorithm, predicts the ratings of the items and recommends high-rated items to the agents. For example, the predicted rating of an item for the agent is generally shown in the following equation on collaborative filtering:

$$r_{ij} = R(\text{sim}(i, u), R) \quad (3)$$



where  $r_{ij}$  is the rating of item  $j$  for user  $i$ ,  $sim(u,v)$  indicates the similarity between user/item  $u$  and user/item  $v$  among a set of user  $U$  who evaluate the item  $j$ , and  $R$  shows the rating information. We can update the  $R()$  as well as  $A()$ . For example, GroupLens[4], which is a representative system using collaborative filtering, has the following equation.

$$r_{ij} = \bar{r}_i + \frac{\sum_{u \in U} (sim(u,i)(r_{uj} - \bar{r}_u))}{\sum_{u \in U} |sim(u,i)|} \quad (4)$$

In this equation,  $\bar{r}_i$  is the average rating when a user  $i$  has already voted.

### 4.3 Status of Agents

One of the problems that require attention is the *tiresome status*. This simulator allows agents to change the status in order to express the degree to which an agent is tiresome. In this simulator, agents have two statuses, *normal* and *tiresome*, and the trigger of changing status occurs in recommendation. The measurement for tiresome is calculated concretely by the evaluation equation, and when the measurement is less than the *tiresome threshold*, the agent changes the status to *tiresome*, otherwise it remains *normal*. The simulator implements the following equation, which is an example of the evaluation equation;

$$t = \sum_{i \in I} \frac{n_i r_{aj}}{rank_i} \quad (5)$$

Here,  $I$  is the set of recommended items from Recommender,  $r_{ij}$  is the rating of item  $j$  by agent  $i$  which is used for rating algorithm,  $rank_i$  is the rank of the item  $i$ , and  $n_i$  is 0 if  $i$  has already been recommended; otherwise it is 1.

### 4.4 Building Simulator Environment

The configuration of the simulator environment is one of the most important steps in the simulation process because an inappropriate environment leads to inappropriate and wasteful result.

Generally, communities tend to follow complex networks and, according to several references, they find that the networks tend to be small-world networks. Here, a small world is a phenomenon in a real world network, and the model has several features such as stability and compression of network. The structure was first formulated by Watts and Strogatz [10]. The structure appears in several networks, and the trend also appears in recommender system. Therefore, we create the environment based on the small-world network.

If we restrict the community to the recommender system, the network of the recommender system is scale-free network like it is in [11, 12, 13]. Therefore, we regard an agent as a node and initialize agents and items according to generating a scale-free network. The algorithm in detail is as follows, given  $n$  agents and  $m$  items as simulation parameters from the user,

1. Create the  $k$ -core clique in order to generate the scale-free network, and allocate the  $C_0$  common items among  $k$  nodes.
2. Add an agent ( $a_i (i=1, 2, \dots, n)$ ) to the network according to the algorithm of BA(Barabasi-Albert) model which is a scale-free network, and allocate the  $C_j$  common items to the clique including  $a_i$ . Note that  $C_j=m$ .
3. Iterate step 2 by allocating  $a_n$ .
4. Allocate the attributes of items to agents who have items in common.
5. Let Recommender calculate the similarity between the agents for recommendation.

## 5 Experiment

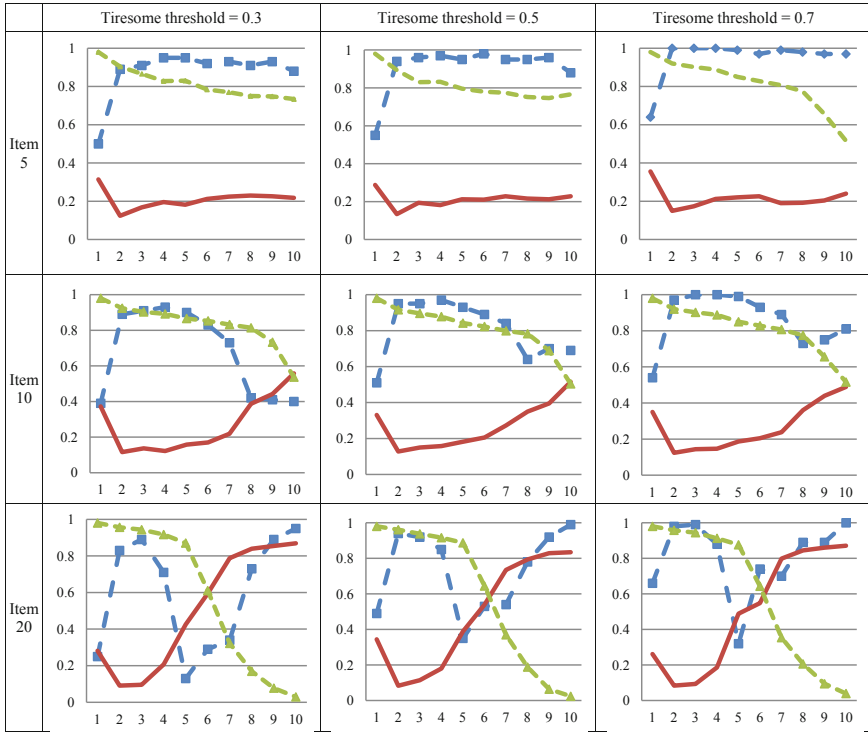
This section describes an experiment to confirm the usability of our simulator. We can confirm the usability when the simulator can show the feature that the user of recommender system becomes tired of the recommendation as precision of recommendation improves. To achieve this goal, we checked the result of precision, the percentage of tiresome agents, and discovery output by this simulator. Also, we changed several parameters and observe the effect of two of them, the tiresome threshold and the number of recommendation items, on the recommendation results.

For recommendation, we used the GroupLens algorithm and observed the phenomenon for 10 iterations. 100 agents joined and Recommender recommended 5, 10, and 20 items to each agent. Also, the tiresome threshold was changed from 0.3 and 0.7.

Figure 2 shows the result figures of the experiments. In each graph of Figure 2, the vertical axis is the measured value and the horizontal axis is the simulation turn. From left to right, the graphs are ordered by the tiresome threshold and from top to bottom, the graphs are ordered by the number of recommended items.

At a glance, we can see that the result pattern changed. When the number of recommended item was 5, the simulator showed many tiresome agents, low precision and high discovery. We found that precision did not improved and agents become tiresome because of a few recommended items. In such cases, it can be concluded that the recommendation itself ended in failure.

When the number of recommended items was 10, the simulator showed that the number of tiresome agents decreased just as precision tended to increase with the 7 or more turns. Then, we can guess that agent can receive the preferable and unknown items.



**Fig. 2** Experiment results of recommender system simulator Solid line: precision, Line marked by triangle: discovery, Line marked by square: tiresome

With 20 items, the simulator showed that the phenomenon in 10 items occurred at nearly 5 more turn: after that, the tiresome agents and precision tended to increase, and discovery tended to decrease. This feature shows the phenomenon of the collaborative filtering, so that we can confirm the usability.

## 6 Conclusion

This paper has proposed a simulator to allow a user to evaluate the algorithms for recommender systems. In order to evaluate the algorithms, the simulator builds an environment of a virtual recommender system based on a complex network model from parameters; Recommender makes recommendations to agents, and the simulator evaluates and outputs the results though Recorder. From experiments, we confirm the usability of our simulator by gaining a resurgence of the phenomenon of collaborative filtering.

## References

1. Resnick, P., Varian, H.R.: Recommender systems. *Commun. ACM*. 40, 56–58 (1997)
2. Adomavicius, G., Tuzhilin, A.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 734–749 (2005)
3. Saga, R., Tsuji, H.: Sales Records Based Recommender System for TPO-Goods. *IEEJ Transactions on Electronics, Information and Systems* 126, 661–666 (2006)
4. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of netnews. In: *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pp. 175–186. ACM, Chapel Hill (1994)
5. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based Collaborative Filtering Recommendation Algorithms. In: *Proc. 10th International Conference on the World Wide Web*, pp. 285–295 (2001)
6. Pazzani, M.J.: A Framework for Collaborative, Content-Based and Demographic Filtering. *Artificial Intelligence Review* 13, 393–408 (1999)
7. Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7, 76–80 (2003)
8. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining Content-Based and Collaborative Filters in an Online Newspaper. In: *Proceedings of ACM SIGIR Workshop on Recommender Systems* (1999)
9. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. *ACM Transaction on Information Systems* 22, 5–53 (2004)
10. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* 393, 440–442 (1998)
11. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. *Reviews of Modern Physics* 74, 47–97 (2002)
12. Amaral, L.A.N., Ottino, J.M.: Complex networks. *The European Physical Journal B - Condensed Matter* 38, 147–162 (2004)
13. Cano, P., Celma, O., Koppenberger, M., Buldú, J.M.: The Topology of Music Recommendation Networks. *Physics*, 0512266 (2005)
14. Breese, J., Heckerman, D., Kadie, C.: Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI 1998)*, pp. 43–52 (1998)

# On the Constraint Satisfaction Method for University Personal Course Scheduling

Yukio Hori, Takashi Nakayama, and Yoshiro Imai

## 1 Introduction

University Students have to decide their own course schedule by themselves. In order to make a course schedule, it is necessary to satisfy the student's interest and to meet course credit restrictions. On the other hands, the university's curriculum changes dynamically to catch up with social needs, advanced science and technology. In the many university, there are on-line course support system that allows students view all subjects information is available[4]. However, it is not easy for students to generate manually a course schedule from a large number of combination of classes, due to various constraints and/or criteria, especially for the freshman in the university.

This paper develops an automated system for generating the student's course schedule. A lot of students course schedule are manually generated based on syllabus information, such as class title, summary and categories scripted from the teachers. However, at a large number of classes, it is very exhaustive to manually generate the course schedule, due to various constraint condition or criteria. Thus, this paper formalizes this problem as a constraint satisfaction/optimization problem, develops an automated tool, and evaluates the effectiveness of our system. The system enables the students to make a suitable course schedule for the learning of the field they want.

## 2 Related Works

Nozawa et. al. had been proposed a syllabus analysis system[2]. In designing of an original curriculum by a higher education institution, or in external evaluation

---

Yukio Hori

Information Technology Center,

Kagawa University, 2-1 Saiwaicho, Takamatsu, Kagawa, Japan 760-0016

e-mail: [horiyuki@itc.kagawa-u.ac.jp](mailto:horiyuki@itc.kagawa-u.ac.jp)

Takashi Nakayama

Faculty of Science, Kanagawa University, 2946 Tsuchiya, Hiratsuka,

Kanagawa, Japan 259-1293

of an institution's curriculum, comprehending the curriculum contents of many institutions in the same field is necessary. However, this is very hard to do even for education experts in that field. This system uses clustering view from each class syllabus description. The main aim of this system is not for students, but for teachers to make specialized curriculum. This system treats syllabus data which constitute some curriculum and are expressed in a common format, calculates the similarity between the syllabus based on the occurrence frequency of technical terms, clusters the syllabus, and thus helps to find distinguishing features of the curriculum by visualizing and comparing the assignments of the syllabus to the clusters along various classification axes. Visualizing of syllabus data is useful to make student's study strategy, but it is not always the case the students find suitable course plan because of their lack of desire. We have tried to make a model of the student's interests.

Ohiwa et. al. has developed idea generation tool with card-handling[6]. Card handling is one of the most useful methods for information representation and idea generation. A generated card can be picked and moved by a mouse. Cards may be grouped by enclosing them with a curve. Relationships of cards and groups can also be marked by special lines. This system can be used for the course planning. However, there are a lot of restrictions about course credit in the curriculum. It is difficult to operate card-handling due to various constrain condition or criteria. Thus, We have formalized this problem as a scheduling problem[1].

The spreading-activation-based model has been proposed in previous studies investigating the syntactic priming (SP) effects in the document understanding [13] [14] [16]. Kojima has proposed a framework of document understanding using text information and context[17]. These past studies focused human's memory activation to find topics words. We applied this spreading-activation-based model to well-concerned course planning via all courses in curriculum.

The course planning is a multi-objective combinatorial optimization problem. In past study, some application system has been reported in a nurse work time scheduling[9] and a trainman allocation problem[10] [11]. We have formalized students course planning as a constrain satisfaction/optimization problem. This approach is particularly effective where there are many constraints. An objective function is first defined to measure the degree of active value by a course plan. We set out to optimize a problem defined by this objective function and the block of constraints for one student. Using this approach, we optimized every student's course plan in a semester. This suggested the proposed system could be used for discovering useful course schedule. This shows the proposed system supports the student to easily make their own course schedule and to easily to know new interesting course.

### 3 Generating Course Schedules

This section describes about how to generate time schedules using spreading activation on a course network. We define the feature space represented as user's interest and learning strategy, formalize as a constraint satisfaction/optimization problem to generating time schedules.

Feature vectors in the field of text are based on a set of nouns extracted from syllabus text; examples include Web mining [19]. In this case, the differences among the user’s interest and learning strategy and courses are well represented by term correspondences. We use the heuristic search technique to solve a constraint satisfaction/optimization problem. We show the system architecture of generating time schedules in Figure 1.

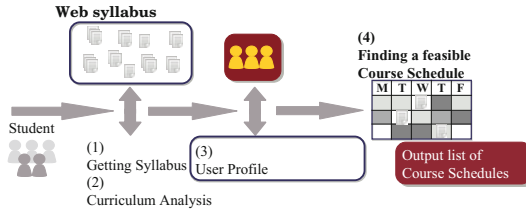


Fig. 1 The configuration of Active Syllabus

1. Getting syllabus text data

We consider the syllabus data discovery problem to find sets of Web page, each of which share some template, among many pages collected in the university web site by a Web crawler. A found set is a potential input for information extraction and wrapper generation algorithms. These results are restored into database and performing edit and browsing on our system.

2. Curriculum Analyzing

We represent feature vector based on a set of nouns extracted from syllabus text. The differences among the courses are represented by the similarity between the syllabus based on the occurrence frequency of technical terms. Next, we use the Repeated bisection clustering method to analyze the curriculum’s preference.

3. User Profile

We represent student’s interest and learning strategy as a set of nouns extracted from syllabus text. We has been developed 3 types of making user’s profile method.

4. Finding a feasible course schedule

The student’s course schedule is built via to solve a constraint satisfaction/optimization problem from the user profile.

In the next section, each step details and the system behavior are described.

3.1 Getting Syllabus Text Data

Obtaining syllabus information on the syllabus web site is described here. In general, there are some fundamental information and contents of the subject in the description of the syllabus. Fundamental information contains subject name, target faculty/grade, credit information (core/required/elective), teacher’s name, semester, lecture day and number of credits. The contents of the subject contains summary,

goal, teaching method, course content. Our system extracts all of these as a subject data and define  $c_i$  vector. The feature vector of each subject was constructed as follows. The weight of term  $w_{ij}$  was defined using tf-idf.

$$c_i = (w_{i1}, w_{i2}, \dots, w_{in})$$

$$w_{i,t_j} = \text{tf}(t_j, c_i) \cdot \text{idf}(t_j)$$

$$\text{idf}(t_j) = \log(N/\text{df}(t_j))$$

where  $\text{tf}(i, j)$  is the occurrence of term  $t_j$  in a subject  $j$ ,  $N$  is the number of total subjects, and  $\text{df}(t_j)$  is the number of subject term  $t_j$  appeared once or more. The followings were excluded as stop words.

1. Words of one letter hiragana and katakana in Japanese.
2. Low frequency terms: The terms under 70% of frequency were excluded.

### 3.2 Curriculum Analyzing

To facilitate comprehension of the curriculum's features, our system is utilizing document-clustering of syllabus data. The repeated bisection clustering is used in order to get student's intended course plan.

$$\begin{aligned} L_{ij} &= \frac{c_i c_j}{\|c_i\| \|c_j\|} \\ &= \frac{\sum_{k=1}^N w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^N w_{ik}^2} \sqrt{\sum_{k=1}^N w_{jk}^2}} \end{aligned} \quad (1)$$

Our system makes syllabus clusters using similarity between the syllabus based on the occurrence frequency of technical terms. The clusters help to find distinguishing features of the curriculum. Our system uses ward-method to make clusters. Moreover, because the size of the cluster is different in the subject and each curriculum, the faculty staff has adjusted the clustering results.

### 3.3 User Profile

There are many combinations of the course schedule in the term, and the student has an original intended plan. Our system makes the student profile as follow.

$$U = \{u_1, u_2, \dots, u_m\}_{u_i \in S}$$

Here,  $u_i$  is a word described in the syllabus,  $S$  is all syllabus set. However, it is not good way to manually choose various words via syllabus to make its profile. Our





Fig. 2 Screenshot of making a user profile

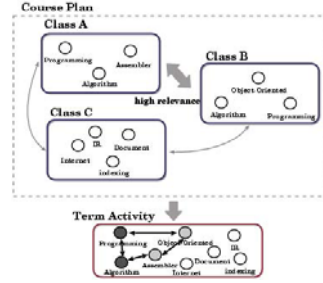


Fig. 3 The spreading activation model for a subject network

system has 3 types of strategy to make a student profile. The student can use these strategies by combining.

- Manual select  
The students select words from syllabus by manual operations.
- Cluster select  
By clustering the syllabus, classified fields of subject is generated. The student selects a cluster and it includes various keywords.
- Browsing behavior  
Students repeat browsing which follows links from one page to another, and examine the contents of the pages they got by link selection. Students expect to acquire useful information by following syllabus. In a word, a student’s syllabus browsing behavior could be utilized as an intended query. Our system extracts, keywords from the student’s syllabus browsing history.

The users are available to use combine these 3 profile setting methods. Figure 2 shows a user profile setting menu. Top menu is presented on upper left of Figure 2. These 3 profile setting methods are effective for arranging to select keyword into easily understanding their interest.

### 3.4 Course Scheduling

As discussed previously, making a course schedule is formalized as a scheduling problem. We use the courses contain the keywords in the user profile for an initial solutions of the scheduling problem, solve the scheduling problem using iterative scheme with best-first search. To solve a scheduling problem, There are five conditions as the follows.

- An output must consider about 1 semester’s schedule.
- An output contain classes from the class in each semester and grade.
- An output must consider the schedule within the restriction and number of credits of each semester.

- An output must include the required classes.
- An output must consider the student's interest.  
Out system sets the objective function and chooses the one with the maximum value for the objective function.
- An output must consider a student's personal schedule.

We formalized this as a constraint satisfaction/optimization problem.

Output:

$$\text{Course Schedule } C = \{c_1, \dots, c_o\}$$

Input:

$$\text{User Profile } U = \{u_1, \dots, u_m\} u_i \in S$$

Maximize:  $T + d$

$$T = \sum_i a_i \cdot O, \quad a_i = \sum_j a_j L_{ij} \quad (i, j = 1, 2, \dots, o) \quad (2)$$

Subject to

$$\min < \text{credit}(C) < \max$$

$$M\{m|m \text{ is a required course}\} \subset C$$

$$P\{p|p \text{ is a prohibited course pattern}\} \not\subset C$$

$$D\{d|d \text{ is a weight value which represents student's requests}\}$$

Here,  $\text{credit}(C)$  is number of credit in the course schedule  $C$ ,  $m$  is a required subject in the semester,  $p$  is a prohibited course pattern,  $d$  is a weight value which represents student's requests.

The objective function  $T$  in (2) is sum of  $c_i$  in the course schedule  $C$  with  $L_{ij}$  represents similarity between each classes. This function represents the network with the vector and similarities of classes in the course schedule.  $a_i$  is active value of class node  $c_i$ . Therefore,  $T$  means active value of the course network from the view point of student's interest.

The spreading activation based model has been proposed in previous studies investigating the syntactic priming (SP) effects in head-initial languages (e.g., English). The cohesiveness is calculated by spreading activation on a semantic network constructed systematically from word network. The spreading activation based model is defined as follows.

$$A(N) = C(N) + ((1 - \gamma)I + \alpha R(N))A(N - 1) \quad (3)$$

where  $\gamma$  is attenuation,  $\alpha$  is a propagation parameter,  $A(N)$  is a vector of active value when the firing count is  $N$ ,  $C(N)$  is a vector of active value from syllabus text,  $I$  is identity matrix and  $R(N)$  is a propagation matrix and its element  $r_{ij}$  represents a similarity metric between node  $i$  and  $j$  such as cosine similarity. In equation (2),  $r_{ij}$  means  $L_{ij}$ .  $a_i = \sum_j a_j \cdot L_{ij}$  is an element of activities in equation (2), a vector of active value presents  $A = (a_1, a_2, \dots, a_o)^t$  if propagation count  $N$  is left out. This means  $C(N) = 0$ ,  $\gamma = 1$ ,  $\alpha = 1$  in equation (3). Therefore equation (3) becomes as follows.

$$A(N) = L(N)A(N - 1)$$

Here, our system uses only one time propagation to the course network. Sum of each element  $A(N)$  is activation-based value in entire course network.  $O$  is a projection operator from syllabus feature vector to student interest vector, then activation-based value of entire course network represents  $T$  in equation (2). This projection means activation-based value via student’s interest on a course network. We use  $T$  and  $d$  represents student’s requests as the objective function.

The solution of the objective function is larger when a course schedule has similar courses. Therefore, as can be seen from Fig 2, we can represent the word network model which has an ability to optimize both connection weights and the activation functions. For the first trial of making a course schedule, a course schedule matched by a student profile is used. Our system chooses the one with the maximum value for the objective function. This becomes the new trial solution for the next iteration, and repeat this iterative process until a satisfactory and feasible schedule is obtained. This optimization is as follows.

1. Make an initial course schedule matched by a student profile.
2. Choose one subject from current course schedule, and replace with the most similarity other subjects.
3. If objective function’s value is updated, this trial solution is replaced to current one.
4. Repeat 2. until the whole of subjects is checked or replacing.

**Table 1** Experimental curriculum data

		Category	Required credits for graduate	
General Education Subjects		Theme subject 53	8 credits	
		General Education Seminar [elective] 52	2 credits	
		Common Subjects (70)	8 credits	
		Health · Sports [elective] 44	2 credits	
		General Education for upper-grade [elective] 6	4 credits	
	Foreign Language	Second Foreign Language 145	4 credits	subtotal 24 credits
		First Foreign Language 71	6 credits	
		Subtotal	30 credits	
Faculty of Engineering Classes	General Engineering Subject	Analytical Thinking Subject 9	8 credits	
		Communication Skill 4	6 credits	
	Specialized Subject	Fundamental Mathematics/Science Skill 9	10 credits	
		Fundamental Specialized Subject 25	30 credits	
		Advanced Specialized Subject 29	32 credits	
			Graduation Research	6 credits
	Elective courses			6 credits
			Subtotal	98 credits
		Total	128 credits	

### 3.5 Example of Output

Figure 4(upper-side) shows an example of course schedule by our system. Figure 4(down-side) shows an example of course schedule making by a real student. \* mark shows a required subject in the semester.

A course schedule in Figure 4(upper-side) is made by a profile includes words related computer science. This course schedule has more classes in the field of computer science and mathematics than student’s one.

1st grade time schedule by active syllabus(upper 1st semester/bottom 2nd semester)					
	Mon	Tue	Wed	Thu	Fri
1	Fundamental Mathematics	Introduction to Statistics	Computer System*		Modern History
2	General Biology C	Introduction to Economics			Introduction to Mechanical Eng.
3		Introduction to Mathematics	Artificial Intelligence		
4			Programming I *	Differential and Integral Calculus I	
5			Programming I *	Mathematical Science I *	
1	Fundamental Informatics	Introductory Career-design I	Probability and Statistics*		
2		Sociology	Mathematics Science II*		
3		Fundamental Life Science	Logical Circuit*		
4			Programming II*	Differential and Integral Calculus	
5			Programming II*		Basic Physics
1st grade time schedule by student (upper 1st semester/bottom 2nd semester)					
	Mon	Tue	Wed	Thu	Fri
1	Algebra	Basic Physics	Computer System *	German I	Biology
2	English I	Life and Human	Gender and Sexuality		German-Nordic Legend and Opera
3	German I				
4			Programming I*	Differential and Integral Calculus I	
5			Programming I*	Mathematical Science I *	English Com.
1		Life and Human and Statistics *	Probability	Introduction to Psychology C	Human Com.
2	English II	German II	Mathematical Science II*	Sociology D	
3			Logical Circuit *	German II	
4			Programming II*	Differential and Integral Calculus	Cross-cultural Communications
5			Programming II*		

Fig. 4 An example course schedule generated by Active Syllabus

## 4 Experiment

We evaluate the proposed methods using our university curriculum data. The evaluation test involved a total of 8 course schedules from different 8 students in grade from B4 to M2 and . In this evaluation test is a subjective experiment to choose an appropriate classes in the course schedule. Here, the course schedules by the students are their registered course schedules.

### 4.1 Outline of Test

We first describe about the curriculum for this experiment. Table 1 shows a credit requirements in our university. The number in () shows registerable number of classes in the division. In this curriculum consist of two basic elements, first one is

public lecture course to acquire broader knowledge and understanding, and the other is professional education to expertise acquisition in a particular area.

The number of courses, keywords, and the target grade are shown as Table 2. As there is a lot of variation in the 4th grade student's course schedule, it has been eliminated from this experiment.

**Table 2** Experimental data

Target curriculum:	Dept of Information Systems Engineering, Kagawa University
Number of courses:	513
Number of keywords:	7,819
Target grade:	1 ~ 3 1st/2nd semester

We defined precision is a ratio of suitable subjects in a course schedule to learn a field of curriculum.  $p/n$  is a precision of a course schedule, here  $n$  is the number of subjects and  $p$  is the number of suitable subjects in a course schedule. This precision is scored by subjects' judgement. Here we requested each subjects to judge about point of importance from specialized field, relativity and width in a course schedule.

The subjects scored the precision of a course schedule with seeing class title list in a same semester, and if there is a more suitable class in the list in a course schedule, then it's class is judged as not good class.

The subjects included 6 of B4 and 2 of M2, therefore the subjects have the knowledge about this curriculum.

Specifically, the subject name shown by the under line in figure 4 is not suitable subject from this subjective evaluation test. The students can choose the subjects from Mathematics, Physics and Computer Science in their course schedule. In this case, the precision is higher when a course schedule has much relative subjects.

## 4.2 Outline of Evaluation test

The evaluation test procedure is as follows. The test involved a total of 8 students ranging in grade from B4 to M1; 6 were B4 and 2 were M1. First, we got their course schedules from grade 1 to 3(human). We made the same number course schedules by Active Syllabus(as). This course schedules were made by history mode as shown in Fig 2 using subjects they registered. For the comparison, we conducted the same number course schedules chose subject at random(rand) and filtered the keywords in their profile(kw). The course schedule (kw) is the same as an initial solution of course scheduling problem of (as).

We evaluated the following precision in this test; How many appropriate subjects in the course schedule. Here, the subject doesn't know in which mode the course schedule was made.

By comparing each type of precision, we analyzed the subjective evaluation test.

### 4.3 Results

Table 3 shows a precision of this subjective evaluation result in each grade.

**Table 3** The results of the subjective evaluation

Grade:Term	B1:1	B1:2	B2:1	B2:2	B3:1	B3:2	Mean
rand	0.59	0.64	0.62	0.62	0.60	0.63	0.62
kw	0.75	0.74	0.70	0.74	0.76	0.73	0.74
as	0.83	0.80	0.80	0.83	0.85	0.80	0.82
human	0.83	0.83	0.81	0.81	0.85	0.82	0.83

As can be seen from Table 3, the precision is high in the order of (human), (as), (kw), (rand). The results of (human) and (as) are same level of precision. Considering the results of this subjective evaluation test, The experiment shows that Active Syllabus is almost as suitable as precision to select goodness class and is more suitable to make a course schedule efficiently. In comparison, the results of (kw) is decreased because random class is selected when there is no class that matches to the user profile. It is hard to specify all keywords represented student interest, Activation and Spreading Model is effective in in a university course scheduling problem.

We conducted analysis of variation in this 4 types of course schedules, and it showed differ significantly by precision ( $p < .01$ ). Multiple comparison between (human) and (as) didn't show differ significantly and the other types showed differ significantly ( $p < .05$ ). Therefore, Our proposal method can make the same level course schedules as (human).

### 4.4 Discussion

Considering the results of subjective evaluation test, we made an objective evaluation test. We calculated network active value  $T$  in 4 types of all course schedules.  $T$  was calculated by expression (2) Table 4 shows network active value  $T$ .

**Table 4** The results of the active value of the course schedule

	B1:1	B1:2	B2:1	B2:2	B3:1	B3:2	Mean
rand	18.44	8.26	22.32	21.08	10.15	10.09	15.06
kw	2.75	4.94	39.54	22.48	14.28	30.22	19.04
hm	5.87	9.19	44.83	57.30	25.13	11.17	25.58
as	42.44	18.74	62.52	33.66	30.22	23.84	35.24

As can be seen from Table 4, the  $T$  was high in the order of (as), (human), (kw) and (rand). (as) score was showed highest in average, but (hm) score was higher than (as) in the second semester of 2nd grade. This means our search method of optimization problem could not find a best solution.

We conducted analysis of variation in this objective evaluation test. Multiple comparison between (as) and the others showed differ significantly ( $p < .05$ ) and the other types didn't show differ significantly.

This objective evaluation means (as) has high relativity in the course schedule than others.

## 5 Conclusions and Future Work

In this paper, we have proposed a plan support system Active Syllabus aimed improving the efficiency of course schedules. This system made the course schedule using the spreading activation model in consideration of student's interest and the relation between subjects. The result of the experiment, precision was about 0.7. Our system should be doing more tests to confirm these preliminary results with in real use case.

Using activation spreading model into the syllabus description, Our system can provide not only existing classes matched by a student's interest, but also their relevant classes. This is differ from a lot of past research that merely provide show the class information. Promising direction for future works are the optimization of period to makes a student profile, the examination of syllabus descriptions and the similarity between classes. These problems should be considered for wide coverage.

## References

1. Hori, Y., Takimoto, M., Imai, Y.: A support system for course schedule design based on syllabus description. In: 8th International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 58–62 (2007)
2. Nozawa, T., Ida, M., et al.: Construction of Curriculum Analyzing System Based on Document Clustering of Syllabus Data(Education). Transactions of Information Processing Society of Japan 46(1), 289–300 (2005)
3. Miyazaki, K., Ida, M., Yoshikane, F., Nozawa, T., Kita, H.: Development of a Course Classification Support System for the Awarding of Degrees using Syllabus Data. IPSJ Journal 46(3), 782–791 (2005)
4. Yamada, S., Matsunaga, Y., Itoh, E., Hirokawa, S.: A Study of Design for Intelligent Web Syllabus Crawling Agent, The transactions of the Institute of Electronics, Information and Communication Engineers, D-I J86-D-I(8), 566–574 (2003)
5. Chubachi, Y., Ishii, Y., Ohiwa, H.: enTrance System Fundamental Concept and Value for a University Environment. In: Software Engineers Association. Software Symposium, pp. 181–184 (2001)
6. Ohiwa, H., Takeda, N., Kawai, K., Shimomi, A.: KJ editor: a card-handling tool for creative work support. Knowledge-Based Systems 10, 43–50 (1997)
7. Ohmi, Y., Kawai, K., Takeda, N., Ohiwa, H.: Pointing Operations in Cooperative Works using Card-handling Tool "KJ-Editor". Transactions of Information Processing Society of Japan 36(11), 2720–2727 (1995)
8. Ikegami, A., Niwa, A., Ookura, M.: Nurse Scheduling Problem in Japan. Operations Research as a Management Science Research 41(8), 436–442 (1996)

9. Aickelin, U., Dowsland, K.A.: An indirect genetic algorithm for a nurse-scheduling problem. *Comput. Oper. Res.* 31(5), 761–778 (2004)
10. Caprara, A., Fischetti, M., Toth, P.: A heuristic method for the set covering problem. *Operations Research* 47, 730–743 (1999)
11. Easton, K., Nemhauser, G., and Trick, M.: The traveling tournament problem description and benchmarks. In: *Proceedings of Seventh International Conference on Principles and Practice of Constraint Programming (CP 1999)*, pp. 580–584 (1999)
12. Horio, M., Suzuki, A.: Development of a General-purpose Scheduling Solver under the Framework of RCPSP/TAU. with Time Constraints. *Journal of Japan Industrial Management Association* 54(3), 203–213 (2003)
13. Anderson, J.R.: A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior* 22, 261–295 (1983)
14. Collins, A.M., Loftus, E.F.: A Spreading Activation Theory of Semantic Processing. *Psychological Review* 82, 407–428 (1975)
15. Lorch, R.F.: Priming and Searching Processes in Semantic Memory: A test of three models of spreading activation. *Journal of Verbal Learning and Verbal Behavior* 21, 468–492 (1982)
16. Waltz, D.L., Pollack, J.B.: Massively parallel parsing, A Strongly Interactive Model of Natural Language Interpretation. *Cognitive Science* 9, 51–74 (1985)
17. Kozima, H., Ito, A.: Context-sensitive word distance by adaptive scaling of a semantic space. In: Mitkov, R., Nicolov, N. (eds.) *Recent Advances in Natural Language Processing, Contemporary Issues in Linguistic Theory*, vol. 136, pp. 111–124. John Benjamins, Amsterdam (1997)
18. Huberman, B.A., Hogg, T.: Phase Transition in Artificial Intelligence Systems. *Artificial Intelligence* 33, 155–171 (1987)
19. Perkowski, M., Etzioni, O.: Towards adaptive Web site: Conceptual framework and case study. *Artificial Intelligence* 118, 245–275 (2000)



# Factors Affecting Productivity of Fractured Horizontal Wells

Li Tiejun, Guo Dali, Tang Zhihao, and Ke Xijun\*

**Abstract.** Fracturing, an important measurement used to develop low permeability reservoirs, has been used widely in hydrocarbon exploitation. Based on previous studies, this paper, applying mathematics and fluid mechanics, draws a productivity prediction model for fractured horizontal wells. According to the technical means of numerical simulation, the productivity of fractured horizontal well in an actual low permeability oil reservoir has been tested by using this model. The impact of fracture length, number of fracture and space distance on productivity are analyzed, which will provide guidance for optimization of fractured horizontal wells.

**Keywords** horizontal well, fracture, productivity prediction, numerical simulation, affecting factors.

## 1 Introduction

Although horizontal well is suitable for the development of special and hard-to-recovery reservoirs, for the low permeability reservoir, horizontal well can't achieve the desired effect. Therefore, fracturing of horizontal wells is often used to increase the productivity of horizontal wells. In the perspective of previous literatures [2, 3, 4, 5, 6, 7], the previous research on the fracturing of horizontal wells is usually based on mathematical model, which has certain assumptions and simplification, making it subjects to a certain degree of inaccuracy. For all these reasons, based on the literature [1], we re-derived and modified the productivity prediction model of fractured horizontal well. Also we take into account the actual situation that the productivity of each fracture is different, so that the new prediction model has more practical applications. Finally, we solve the model with the method of numerical simulation and analyze the impact of fracture length, number of fracture

---

Li Tiejun

School of Science in Southwest Petroleum University, Chengdu, China, 610500

e-mail: ltj@swpu.edu.cn

and space distance on productivity. This gives a reference basis to the optimization study of multi-section fractures in subsection fracturing of horizontal wells.

## 2 Simulation Method and Modeling

### A Pressure Model of a Random Point in the Stratum with Many Fractures

When the plane, which is paralleled with fractures, is perpendicular to the horizontal well axis, each fracture can be seen as that it is composed of countless number of points. The two wings of fracture are divided into  $n$  equal parts respectively, and each equal part can be seen as a point. The coordinate of the number  $j$  point in the left wing of the fracture is  $(x_j, y_j)$ , it can be represented by the center coordinates of number  $j$  piece,  $j$  ranges from 1 to  $n$ . According to the superposition principle of potential, the pressure of point  $(x, y)$  is:

$$(P_i - P(x, y, t))|_N = \sum_{i=1}^N \left( \sum_{j=1}^n \frac{q_{fij} \mu B}{4\pi kh} \left[ -Ei \left( -\frac{(x + \frac{1}{2}(\frac{2n-2j+1}{n}x_{fii}))^2 + (y - y_{fi})^2}{4\eta t} \right) \right] \right. \\ \left. + \sum_{j=1}^n \frac{q_{frij} \mu B}{4\pi kh} \left[ -Ei \left( -\frac{(x - \frac{1}{2}(\frac{2j-1}{n}x_{fri}))^2 + (y - y_{fi})^2}{4\eta t} \right) \right] \right) \quad (1)$$

Here  $P_i$  is initial reservoir pressure,  $q_{fij}$  is the production of number  $j$  section in the left wing of fracture  $i$ ,  $q_{frij}$  is the production of number  $j$  section in the right wing of fracture  $i$ ,  $x_{fii}$  is the length of the left wing of fracture  $i$ ,  $x_{fri}$  is the length of the right wing of fracture  $i$ ,  $y_{fi}$  is the coordinate of fracture  $i$  in the direction of  $y$ ,  $\mu$  is base oil viscosity,  $B$  is oil volume factor,  $k$  is in-place permeability,  $h$  is core intersection,  $\eta$  is pressure transmission coefficient,  $N$  is the number of fracture and  $t$  is the time.

### B Productivity Prediction Model of Fractured Horizontal Wells

Because of the artificial fracture, the stratum pressure re-distributes. Suppose the pressure at the end of the left wing of number  $j$  fracture is  $P_{fii}$  and at the other end is  $P_{fri}$ . Nevertheless, in the actual case, fractures are not necessarily symmetrical about the horizontal well shaft, so we take the average pressure of both ends of the fracture as the fracture end stress:

$$p(x_{fi}, y_{fi}, t) = \frac{P_{fii} + P_{fri}}{2} \quad (2)$$

The flow process of oil and gas from the fracture to the wellbore can be expressed as:

$$(x_{fi}, y_{fi}, t) - P_{wfi} = \frac{q_{fi} \mu B}{2\pi k_{fi} w_i} \left( \ln \sqrt{\frac{(x_{fli} + x_{fri})h}{\pi}} + s \right) \tag{3}$$

Here  $q_{fi}$  is the production of fracture  $i$ ,  $k_{fi}$  is the permeability of fracture  $i$ ,  $w_i$  is the width of fracture  $i$ ,  $s$  is skin coefficient,  $h$  is core intersection and  $r_w$  is wellbore radius.

Resistance generated when the fluid flowing in the horizontal well bore is much smaller than in the stratum, if we don't consider the pressure drop when the fluid flowing in the horizontal well bore, the pressure at the bottom of the fracture equals to sand face pressure, that is  $P_{wf1} = P_{wf2} = \dots = P_{wfn}$ , according to (1) (2) (3),

$$\begin{aligned} P_i - P_{wf} = & \frac{1}{2} \left( \sum_{k=1}^N \sum_{j=1}^n \frac{(\frac{1}{n} \frac{x_{fjk}}{x_{fjk} + x_{fjk}}) q_{jk} \mu B}{4\pi k h} \left[ -Ei \left( \frac{-(1-\frac{1}{2n})x_{fji} + \frac{1}{2}(\frac{2n-2j+1}{n})x_{fjk}^2 + (y_{fi} - y_{fj})^2}{4\eta t} \right) \right] \right. \\ & + \sum_{j=1}^n \frac{(\frac{1}{n} \frac{x_{fjk}}{x_{fjk} + x_{fjk}}) q_{jk} \mu B}{4\pi k h} \left[ -Ei \left( \frac{-(1-\frac{1}{2n})x_{fji} - \frac{1}{2}(\frac{2j-1}{n})x_{fjk}^2 + (y_{fi} - y_{fj})^2}{4\eta t} \right) \right] \\ & + \sum_{k=1}^N \sum_{j=1}^n \frac{(\frac{1}{n} \frac{x_{fjk}}{x_{fjk} + x_{fjk}}) q_{jk} \mu B}{4\pi k h} \left[ -Ei \left( \frac{((1-\frac{1}{2n})x_{fji} + \frac{1}{2}(\frac{2n-2j+1}{n})x_{fjk}^2 + (y_{fi} - y_{fj})^2)}{4\eta t} \right) \right] \\ & \left. + \sum_{j=1}^n \frac{(\frac{1}{n} \frac{x_{fjk}}{x_{fjk} + x_{fjk}}) q_{jk} \mu B}{4\pi k h} \left[ -Ei \left( \frac{((1-\frac{1}{2n})x_{fji} - \frac{1}{2}(\frac{2j-1}{n})x_{fjk}^2 + (y_{fi} - y_{fj})^2)}{4\eta t} \right) \right] \right) \\ & + \frac{q_{fi} \mu B}{2\pi k_{fi} w_i} \left( \ln \sqrt{\frac{(x_{fli} + x_{fri})h}{\pi}} + s \right) \end{aligned} \tag{4}$$

In accordance with the above derivation, we can get  $N$  linear equations with  $N$  unknown numbers and the equations can be solved. Using full pivoting Gaussian elimination we can obtain the productivity of each fracture, the productivity of the fractured horizontal wells is the total productivity of all the fractures:

$$Q = \sum_{i=1}^N q_{fi} \tag{5}$$

### 3 Model Solutions

In order to solve the above prediction model, we have to know the basic principles of this model. In fact, this productivity prediction consists of two elements: one is

the flowing model of fluid in the reservoir, which can be solved by analytic method; the other is the flowing model of fluid in the fracture, which can be solved by meshing method; also the more dense mesh the higher accuracy. The solving process is as follows:

- Step 1 Input the basic parameters: geometric size of reservoir, number fractures, fracture location, geometric size of fracture, time step, physical parameters of reservoir (such as permeability, viscosity, density, initial pressure, etc.);
- Step 2 According to the formula (4), combine the coefficient;
- Step 3 Calculate the coefficient matrix of equations;
- Step 4 Use full pivoting Gaussian elimination to obtain the productivity of each crack and if necessary the pressure at any point of the fracture in the reservoir can be calculated;
- Step 5 The productivity of the fractured horizontal wells is the total productivity of all the fractures;
- Step 6 According to the time step, calculate and record the average and cumulative production of the horizontal wells in a period time;
- Step 7 If the time is shorter than the production time, return to step 2, continue to the next step until the time attain the production time, and then output the average and cumulative production of the fracture and horizontal wells.

## 4 Example and Analysis

To illustrate the import of fracture parameters on the production, according to the data of an actual low permeability oil reservoir, we programmed to calculate the output under different conditions and do the sensitivity analysis. Table 1 shows the data of reservoir and fracture parameters.

**Table 1** The parameters of reservoir and fractures

parameter	data	parameter	data
core intersection ( <i>m</i> )	12	in-place permeability	
formation porosity	0.1	( $\mu\text{m}^2$ )	0.0075
coefficient of compressibility		skin factor	4
( $\text{MPa}^{-1}$ )	0.00035	wellbore radius ( <i>m</i> )	0.12
oil volume factor	1.084	base oil viscosity ( $\text{mPa}\cdot\text{s}$ )	4.8
initial reservoir pressure		bottom hole pressure ( $\text{MPa}$ )	3
( $\text{MPa}$ )	11.83	number of fracture	4
fracture width ( <i>mm</i> )	5.84	fracture permeability	
half of fracture length ( <i>m</i> )	75	( $\mu\text{m}^2$ )	30

### A The Impact of Fracture Length

With the other parameters are same as parameters in Table 1, we change the fracture length to  $10^m$ ,  $30^m$ ,  $50^m$ ,  $70^m$ , and  $90^m$ . Figure 1 shows the effect of different fracture length on the impact of production, with the increase of fracture length the production increased also. However, the rate of increase reduce, the increase in production are minimal when fracture length increased to a certain extent.

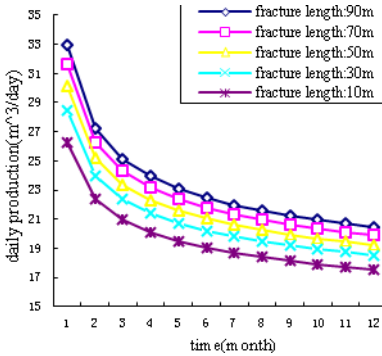


Fig. 1 Effect of Fracture Length on Productivity

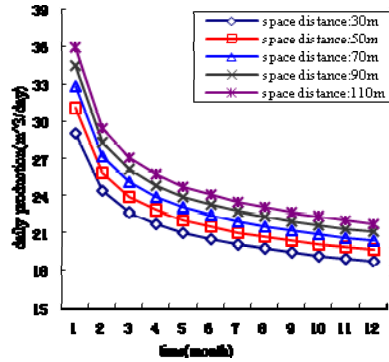


Fig. 2 Effect of Space Distance on Productivity

Therefore, from point of view of economic production and technical, in order to get the best value for money we have to choose the optimal value of the fracture length.

### B The Impact of Space Distance

Figure 2 shows the impact of different space distances on productivity with 4 fractures. From the figure we can see: When the number of fractures is constant, the greater the space distance, the higher the production. Although the rate of increase in production decreases as the distance increases, it is always on the rise. There is pressure interference between fractures in the production; at the beginning of production each fracture has its own control area and don't affect each other. With the passage of time, the control area is gradually expanding. There will be a pressure disturbance, resulting in lower productions, when the areas intersect with each other. It can be seen from the figure, with the increase of production time, the impact of space distance on the production reduction. When it reaches a certain production time, the distance gap of production of different crack is very small. In fracturing design, due to increased space distance will not increase too much capital investment; it is a good approach to increase the production by increasing the distance.

## 5 Conclusion

- (1) In consideration of many actual situations, the productivity prediction model is more reasonable and realistic.
- (2) The fracture length and number are the main parameters of production.
- (3) In terms of the fractured horizontal well, it is not the more fractures the better, generally, 4 or 5 fractures is enough.
- (4) The productivity of fractured horizontal wells increases with the increase of fracture length, number of fracture and space distance, but the growth rate decreases. For all kinds of reservoirs, there is a best combination of parameters.

**Acknowledgment.** This paper was supported by the Key Project of the Provincial Education Department of Si Chuan (10ZA150) and the Nature Science Foundation of Southwest Petroleum University (2010XJZ196).

## References

- [1] Lang, Z., Zhang, L.: Research on Productivity of Fractured Horizontal Wells. *Journal of the University of Petroleum* 18(2), 43–46 (1994)
- [2] Hu, J., Jia, Z., Wei, Z.: A New Method to Predict Performance of Fractured Horizontal Wells. In: SPE, 37051 (1996)
- [3] Guo, D., Liu, C., Zhao, J.: Dynamic Production Prediction and Parameter Identification for Gas Well With Vertical Fracture. *Applied Mathematics and Mechanics* 23(6), 563–568 (2002)
- [4] Wang, X., Duan, Y., Yan, X.: Optimization of Injection-production Pattern and Fracture System for Horizontal Wells, pp. 11–19. Petroleum Industry Press, Beijing (April 2008)
- [5] Zhang, X., Fang, H., Qiu, Y.: A Study on Factors Affecting the Performance of Hydraulically Fractured Horizontal Well in Low Permeability Reservoirs. *Acta Petrolei Sinica* 20(4), 51–55 (1999)
- [6] Li, G., Xiong, W., Song, J.: Influencing Factors to Productivity for High Pressure Water Jet Perforation. *Oil Drilling & Production Technology* 28(4), 19–22 (2006)
- [7] Ronaldo, C., Turgay, E.: Horizontal Well Design Optimization: a Study of the Parameters Affecting the Productivity and Flux Distribution of a Horizontal Well. In: SPE, 84197 (2003)

# Informal Lightweight Knowledge Extraction from Documents

Francesco Colace, Massimo De Santo, and Paolo Napoletano

**Abstract.** In this paper, we propose a method to automatically extract informal knowledge from a collection of documents. The method is mainly based on the definition of a kind of informal knowledge representation consisting of concepts (lexically indicated by words) and the links between them. We show that links can be inferred from documents through the use of the probabilistic topic model while the overall parameters optimisation procedure, based on a suitable score function, can be carried out through the Random Mutation Hill-Climbing algorithm. Experimental findings show that our method is effective and that, as side effects, the score function can be employed as a criterion to compute the homogeneity between documents, which can be considered as a prelude to a classification procedure.

## 1 Introduction

In literature, different approaches have been used to build ontologies: manual, semi-automatic and automatic methods [1]. Among semiautomatic and automatic methods we can distinguish those based on Machine Learning techniques from those based on pure Artificial Intelligence (AI) theories.

Nevertheless, the great majority of existing methods relies on the concept of ontology according to what is commonly acknowledged in computer science, that is an ontology is an *explicit specification of conceptualisations* [2], which is often represented as a directed graph whose nodes represent *concepts* and edges represent relations between *concepts*. Furthermore, the AI community considers that the backbone of an ontology graph is a taxonomy in which the relations are *is-a*, whereas the remaining structure of the graph supplies auxiliary information about the modelled domain and may include relations like *part-of*, *located-in*, *is-parent-of*, and

---

Francesco Colace · Massimo De Santo · Paolo Napoletano

Department of Information Engineering and Electrical Engineering, University of Salerno, Fisciano, 84084, Italy

e-mail: [fcolace, desanto, pnapoletano}@unisa.it](mailto:{fcolace, desanto, pnapoletano}@unisa.it)

many others. Although today there is no consensus on what concepts are [3], they are often lexically defined, and we refer to them by using natural language names. Therefore, when ontologies are visualised, their nodes are often shown with corresponding natural language concept names.

However, whatever the definition of ontology, we can affirm that knowledge organising systems (e.g. classification systems, thesauri, and ontologies) should be considered as systems basically organising *concepts* and their *semantic relations*. Starting from these considerations on the main aspects of knowledge definition, some questions arise:

1. *How formal should be the definition of the concepts and their relations in order to be universally shared and accepted?*
2. *How much human help do you need to build such a shared knowledge?*

The formality of an ontology can be understood in two ways: the degree of formality and expressivity of the language used to describe it; and how formal are the information sources used to build such a knowledge. Based on this consideration we can form a continuum of kinds of ontologies, starting from terms and web directories (also called *lightweight*), and continuing to rigorously formalised logical theories. However, most of the specifications do agree that an ontology should be defined in a formal language, which in practice usually means a logic-based language suitable for automatic reasoning, and should be fed with formal knowledge, that is the involvement of several experts. Nevertheless, the informality of an ontology is mainly related to the nature of the information which feeds the knowledge definition: has the information been, automatically or semi-automatically, extracted from some sources? If so, then this information contributes to form a kind of informal knowledge. As a consequence we can say that a kind of ontology is well determined by two aspects:

1. The formality, as well as the expressivity, of the language used to describe and represent it, which answers the first question;
2. The definition of the formality of the information sources, which answers the second question.

In our opinion, the most widely accepted definition of ontology, which considers formality in both these aspects, appears not to be suitable to describe the informal knowledge, for example, that we can automatically infer from text documents. On the contrary, the definition of a kind of *informal lightweight Ontology*, seems to provide a more flexible structure and applicability to an informal knowledge description of a context [4].

The main idea here is to introduce both the representation of a *Graph of Concepts* (GCs), which considers also the definition of a *Concept* and, a method to automatically learn such a representation from documents.



## 2 A Graph of Concepts

Let us define a *Graph of Concepts* as a triple  $\mathcal{G}_\mathcal{C} = \langle N, E, C \rangle$  where  $N$  is a finite set of nodes,  $E$  is a set of edges weighted by  $\psi_{ij}$  on  $N$ , such that  $\langle N, E \rangle$  is an a-directed graph, and  $C$  is a finite set of concepts, such that for any node  $n_i \in N$  there is one and only one concept  $c_i \in C$ . The weight  $\psi_{ij}$  can be considered as the degree of semantic correlation between two concepts  $c_i$  *is-related* $_{\psi_{ij}}$  *to*  $c_j$  and it can be considered as a probability:  $\psi_{ij} = P(c_i, c_j)$ . Since the relations between the nodes can only be of the type *is-related-to* then this representation can be considered as a lightweight conceptualisation that we call  $\mathcal{O}$ . The probability of  $\mathcal{O}$  given a parameter  $\tau$  can be written as the joint probability between all the concepts. By following the machine learning theory on the factorisation of undirected graph, we can consider such a joint probability as a product of potential functions where each of this function can be considered as the weight  $\psi_{ij}$ .

Each concept  $c_i$  can be defined as a rooted graph of words  $v_s$  and a set of links weighted by  $\rho_{is}$ . The weight  $\rho_{is}$  can measure how far a word is related to a concept, or in other words how much we need such a word to specify that concept. We can consider such a weight as a probability:  $\rho_{is} = P(c_i|v_s)$ . In order to compute weights defining the Graph of Concept we need to compute both  $\psi_{ij}$  and  $\rho_{is}$ .

### 2.1 Learning Concepts and Semantic Relations

A concept is lexically identified by a word of a vocabulary, and it can be represented by a set of connected words. Formally, a *word* is an item of a vocabulary indexed by  $\{1, \dots, V\}$ , each word using unit-basis vectors that have a single component equal to one and all other components equal to zero. Thus, the  $v$ th word in the vocabulary is represented by a  $V$ -vector  $w$  such that  $w^p = 1$  and  $w^q = 0$  for  $p \neq q$ . A document is a sequence of  $L$  words denoted by  $\mathbf{w} = (w_1, w_2, \dots, w_L)$ , where  $w_n$  is the  $n$ th word in the sequence. A corpus is a collection of  $M$  documents denoted by  $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ .

If we are considering an automatic extraction of knowledge from a document (or a set of documents), which in our case is represented by a graph of concepts, then it is obvious that we need to determine when a word denotes a concept and/or contributes to define a part of it. For this purpose, our method treats each word in the first instance as a possible concept and it calculates its degree of association with the remaining words, that is the computation of the probability of a concept.

It would be sufficient to calculate the probability of each concept to determine which of them are best specified by a set of words, in other terms which of them are most likely. In this way, we can determine which words of the corpus represent concepts and then calculate their statistical dependencies,  $\psi_{ij}$ , to finally obtain the graph of concepts. Note that since each concept is represented by a word, then the computation of  $\rho_{is} = P(c_i|v_s) = P(v_i|v_s)$  where the concept  $c_i$  is lexically identified by  $v_i$ , and  $\psi_{ij} = P(c_i, c_j) = P(v_i, v_j)$  where the concept  $c_i$  and  $c_j$  are lexically identified by  $v_i$  and  $v_j$  respectively. Those probabilities can be computed as a *word*

*association problem* that, as we show next, can be solved by using the probabilistic topic model introduced in [5].

### 3 Informal Knowledge Extraction Procedure

Given a set of documents, the whole procedure to automatically extract a GCs is composed of two stages: one for the identification of concepts from the vocabulary (*learning* of words-concepts relations) and another for the computation of relations between concepts (*graph learning*). Those stages involve the choice of parameters, which they can be manually or automatically set. In this section we show which parameters must be optimized, while in next section, we will explore how to develop a procedure to determine such parameters automatically by making use of a multi-objectives optimisation procedure based on the well known Random Hill-Climbing algorithm.

Let us consider one document from the corpus (the procedure still holds if we choose more than one document). It will contain  $V$  words/concepts and for each of them, by using the topic model, we can compute the probability distribution of concepts and we can select a specific number of concepts  $H$  out of a set of plausible ones. In this way we consider  $H$  as a variable, a parameter, which assumes a value within a set of plausible numbers, and the best value for  $H$  might be identified by performing a kind of optimization procedure.

Once that a number of concepts has been chosen, there are two steps left. Given each concept  $c_i$ , one step consists in determining a threshold  $\mu_i$  for selecting the number of concept-word relations  $\rho_{is}$ ,  $\forall i$ . The other step consists in computing the value  $\tau$  such that the most probable  $\psi_{ij}$ , representing concept-concept relations, are determined.

Summing up, we have to find a value for the parameter  $H$ , which establishes the number of concepts, a value for  $\tau$  and finally values for  $\mu_i$ ,  $\forall i \in [1, \dots, H]$ , thus we have  $H + 2$  parameters which modulate the shape of the ontology. If we let the parameters assuming different values, we can observe different ontology  $\mathcal{O}_t$  for each set of parameters,  $\Lambda_t = (H, \tau, \mu_1, \dots, \mu_H)_t$  extracted from the same set of documents, where  $t$  is representative of different parameter values.

A way of saying that an ontology, given the parameters, is the best possible for that set of documents is to demonstrate that it produces the maximum score attainable for each of the documents when the same ontology is used as a knowledge base for querying in a set containing just those documents which have fed the ontology builder. For this purpose let us suppose of using a corpus of  $M$  documents denoted by  $\mathcal{D} = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ , indexed following the *term frequency-inverse document frequency (tf-idf)* model, and let  $\mathcal{O}_t$  be the  $t$ -th possible ontology built from that repository with the set of parameters  $\Lambda_t$ . Let us suppose of using a common search engine, in this case Lucene, which has the ability to assign a rank or an order to documents  $\mathbf{w}$  that match a query  $\mathbf{q}$  following the *cosine similarity* model  $\mathcal{S}(\mathbf{q}, \mathbf{w})$ . Since our ontologies are represented as pairs of weighted and related words, we have

used the Lucene Boolean model and term boosting faculties to extend the original query with contributions from ontology.

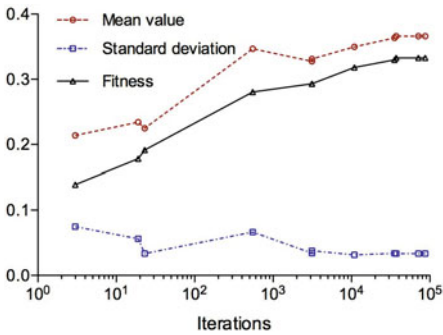
As a consequence, each ontology can be represented by the Lucene boolean syntax, which alternatively corresponds to two vectors, one representing all the pairs  $\mathbf{q}_t = \{q_1, \dots, q_U\}_t$ , and one representing each relation factor, that is the Lucene boost,  $\mathcal{B}_{\mathbf{q}_t} = \{\mathcal{B}_{q_1}, \dots, \mathcal{B}_{q_U}\}_t$ . By performing a Lucene search query that uses the ontology  $\mathcal{O}_t$  on the same repository  $\mathcal{D}$ , we obtain a score for each document  $\mathbf{w}_i$  and then we have  $\mathbf{S}_t = \{\mathcal{S}(\mathbf{q}_t, \mathbf{w}_1), \dots, \mathcal{S}(\mathbf{q}_t, \mathbf{w}_M)\}_t$ , where each of them depends on the set  $\Lambda_t$ . To compute the best value of  $\Lambda$  we can maximise the score value for each documents, which means that we are looking for the ontology which best describe each document. It should be noted that such an optimisation maximises at same time all  $M$  elements of  $\mathbf{S}_t$ . Alternatively, in order to reduce the number of the objectives to being optimised, we can contemporary maximise the mean value of the scores and minimise their standard deviation, which turns a multi-objectives problem into a two-objectives one. Additionally, we can reformulate the latter problem by means a linear combination of its objectives, thus obtaining a single objective function, i.e., *Fitness* ( $\mathcal{F}$ ), which depends on  $\Lambda_t$ ,  $\mathcal{F}(\Lambda_t) = E_m[\mathcal{S}(\mathbf{q}_t, \mathbf{w}_m)] - \sigma_m[\mathcal{S}(\mathbf{q}_t, \mathbf{w}_m)]$ , where  $E_m$  is the mean value of all element of  $\mathbf{S}_t$  and  $\sigma_m$  being the standard deviation.

The optimisation method we have chosen performs a search procedure through a zero-temperature Monte Carlo method, known in the Evolutionary Computation community as *Random Mutation Hill-Climbing* (RMHC), which extends the *Random Search* by generating new candidate solutions as variations of the best known solution. RMHC has evidenced his ability in many optimization problems under NP-hard, deceptive and neutral cost functions [6].

## 4 Experimental Results

The experiments have been conducted in different contexts and, for each context, a list composed of 10 web documents has been formed. In this paper, due to the limitation in space we show results obtained in the context of *Renaissance* (RA). Once a repository  $\mathcal{D}$  for each topic has been chosen, we have carried out the RHMC optimisation procedure. As regards the parameters to be optimised, we have considered the following ranges of variation:  $H \in [5, 15] \subseteq \mathbb{N}$  and  $\mu, \tau \in [0, 1] \subseteq \mathbb{R}$ .

For each context we have performed several runs of the RHMC algorithm with different pseudo-random generator initialisations. Moreover, for each run we have set a maximum number of evaluations equal to 100000. At the end of the optimisation, we have chosen the best run for each topic. In Fig. 1 the evolutions of the fitness, the mean value and standard deviation are shown for RA. In Fig. 2 the differences between the scores assigned to each document at the beginning and at the end of the run are shown (i.e., the best ones). Note that at the beginning the values across the documents are non-homogeneous, showing a low mean value and high standard deviation.



**Fig. 1** Mean value, standard deviation and fitness evolution for the topic RA

URLs	Rank for RA	
	$t = 3$	$t = 87127$
www.firenze-online.com	0,33	0,40
www.artistiinrete.it	0,28	0,40
www.arte-argomenti.org	0,25	0,34
www.arte.go.it	0,23	0,31
www.bilanciozero.net	0,22	0,41
www.visibilmente.it	0,22	0,38
digilander.libero.it	0,22	0,38
it.wikipedia.org	0,17	0,36
it.encarta.msn.com	0,16	0,34
www.salviani.it	0,06	0,33

**Fig. 2** Scores obtained for documents of RA at step  $t = 3$  and  $t = 87127$ .

## 5 Conclusions

The results obtained have confirmed that the proposed optimisation procedure is able to determine the optimal parameters at which the fitness function is maximised, thus demonstrating that our approach is effective.

The fact that the mean value and standard deviation of the final scores, assigned by the optimisation procedure to each document, are maximised and minimised at the same time, implies that such a method can be used as a semantic clustering method. As a consequence, the score function can be used as a criterion to compute the homogeneity between documents, which can be considered as a prelude to a classification procedure.

## References

1. Cimiano, P.: *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer (2006)
2. Gruber, T.R.: A translation approach to portable ontology specifications. *Knowl. Acquis.* 5, 199–220 (1993)
3. Hjørland, B.: Concept theory. *Journal of the American Society for Information Science and Technology* 60, 1519–1536 (2009)
4. Giunchiglia, F., Marchese, M., Zaihrayev, I.: Towards a theory of formal classification. In: *Proceedings of the AAAI 2005 International Workshop Contexts and Ontologies: Theory, Practice and Applications*. AAAI Press (2005)
5. Tenenbaum, J.B., Griffiths, T.L., Steyvers, M.: Topics in semantic representation. *Psychological Review* 114, 211–244 (2007)
6. Falco, I.D., Cioppa, A.D., Maisto, D., Scafuri, U., Tarantino, E.: Extremal optimization dynamics in neutral landscapes: The royal road case. *Artificial Evolution*, 1–12 (2009)

# A Classification Model Using Emerging Patterns Incorporating Item Taxonomy

Hiroyuki Morita and Yukinobu Hamuro

**Abstract.** By extracting frequent patterns efficiently, it is possible to enhance some existing algorithms. Using many candidate patterns causes the results of the classification model to be more powerful. Moreover, aggregating similar items within patterns increases the possibility of creating more powerful patterns. In our method, we define some taxonomies and extract more powerful frequent patterns to incorporate such taxonomies and items. Our aim is to improve Classification by Aggregating Emerging Patterns(CAEP) by using more promising patterns with taxonomy. Using certain computational experiments as a source of practical data, we show that our performance is better than the one that does not use taxonomy. By identifying the reason behind our performance, we show that our method can extract better candidate patterns incorporating taxonomy.

## 1 Introduction

Historical purchasing data have a lot of categorical attributes, such as individual characteristics and names of purchased items. Extracting frequent patterns from the data is a promising approach for CRMs. To employ this approach, we can implement market basket analysis and a classification analysis that extracts different features between loyal customers and other types of customer. Extracting patterns is very expensive from the viewpoint of computational cost. However, LCM [3] was a breakthrough, and we can now obtain many candidate patterns more easily. Such candidate patterns have the scope to improve existing classification models, such as Classification by Aggregating Emerging Patterns(CAEP) [2]. In CAEP, the process of extracting emerging patterns was very expensive because apriori method is

---

Hiroyuki Morita

Osaka Prefecture University, Sakai, Osaka, Japan

e-mail: [morita@eco.osakafu-u.ac.jp](mailto:morita@eco.osakafu-u.ac.jp)

Yukinobu Hamuro

Kwansei Gakuin University, Nishinomiya, Hyogo, Japan

e-mail: [hamuro@kwansei.ac.jp](mailto:hamuro@kwansei.ac.jp)

used, and for thick input data it was difficult to do so particularly. However by using LCM to extract emerging patterns, we can get them to smaller minimum support value and improve the performance. Additionally, an effective method of enhancing performance, is by using item taxonomy. The item taxonomy can aggregate some similar items and makes a pattern more powerful by incorporating the taxonomy. In particular, in case where the pattern aggregates more transactions in others, it is possible to improve classification more effectively.

In this paper, we propose a classification method with emerging patterns incorporating item taxonomy. Our paper studies the incorporation of item taxonomy into item patterns and the use of extracting patterns by LCM. Using computational experiments, we show a good performance and that the comparison of item patterns with both item and taxonomy patterns reveals certain improvements.

## 2 Related Works

The original CAEP was proposed by [1]. This model uses the emerging pattern proposed by [2] and decides class prediction by aggregating scores that are calculated by emerging patterns. Given two classes  $C_1$  and  $C_2$ , and a pattern  $\pi$ , support value for each class is denoted by  $SUP(\pi, C_k)$ . The growth rate is then defined as:

$$GR_{C_1}(\pi) = \frac{SUP(\pi, C_1)}{SUP(\pi, C_2)}. \quad (1)$$

The larger the growth rate of a pattern, the greater is its classification ability. Finally, the score for each transaction for each class is calculated as below, and the transaction is predicted to a class that has a larger score:

$$score(t, C_k) = \sum_{\pi \subseteq t, \pi \in E(C_k)} p(\pi, C_k) \times SUP(\pi, C_k), \quad (2)$$

$$p(\pi, C_k) = \frac{GR_{C_k}(\pi)}{GR_{C_k}(\pi) + 1}, \quad (3)$$

where  $t$  and  $E(C_k)$  denotes a transaction and a set of emerging pattern for class  $C_k$ , respectively. Then, the sum of normalized  $score(\pi, C_k)$  decides the class prediction of the transaction. It is a simple but interesting method. Another approach for using emerging patterns is proposed by [4]. They extract emerging patterns using apriori method to construct decision tree. [5] proposes decision tree based classification model by using a contrast set proposed by [6]. [5] extracts a contrast set by the LCM to save the extracting cost.

There was a limitation to the number of emerging patterns that can be extracted by using the existing method. Now, the LCM method is proposed and by replacing the extracting emerging patterns from the apriori method with the LCM we can acquire an adequate number of emerging patterns. Given the progress, more sophisticated improvements are desired. In short, it morphs many emerging patterns into fewer powerful emerging patterns to classify the classes. To solve this problem, we

propose the use of item taxonomy. Here, item taxonomy means a hierarchical structure to express a classification of items. For instance, given items  $\{a, b, c\}$  and a taxonomy  $\{X\}$  that is the taxonomy of  $\{a, b\}$ , pattern  $\{a, c\}$  and  $\{b, c\}$  are aggregated by a pattern  $\{X, c\}$  incorporating the taxonomy  $\{X\}$ . Therefore, we can aggregate counts of patterns  $\{a, c\}$  and  $\{b, c\}$  as a count of  $\{X, c\}$ . When there is a gap in aggregation among classes, emerging patterns are more powerful for classifying the classes.

To extract emerging patterns to incorporate item taxonomy, we have to expand item taxonomy into input data and to extract emerging patterns from the expanding input data, as shown in figure 1. In cases where patterns belong to the same taxon-

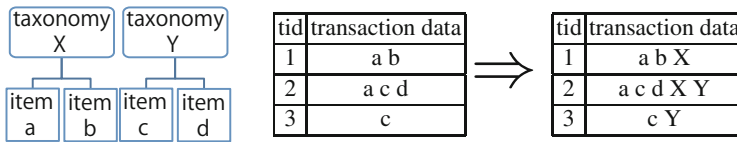


Fig. 1 Our transaction data

omy, it is redundant. Thus, we have to eliminate redundancy from such patterns. To do so, we use VSOP [7] after extracting patterns by LCM. VSOP is a software that compresses patterns and carries out calculations using a compressed data structure. By using VSOP, we can remove the redundant patterns for huge sets of patterns. We introduce our method in the next section.

### 3 Our Proposed Method

Our proposed method uses the basic CAEP flow without using emerging patterns incorporating item taxonomy. The main flow of the method is as below.

1. Original input data are transformed into expanding input data by using item taxonomy.
2. Extracting emerging patterns from the expanding data by using LCM.
3. Removing redundant emerging patterns by using VSOP.
4. The score for each class is calculated from emerging patterns that incorporate item taxonomy.
5. A classification model is constructed and class prediction is decided for the data.

Firstly, we create expanding input data from an original input data. Figure 1 shows an example of the expanding process, and figure 2 illustrates a definition of taxonomy. In the original input data, we check the definition of taxonomy and the corresponding item taxonomy is added to the transaction in the original data. After all item taxonomies are added, the expansion of input data is completed. As input data, such expanding data and class definition data such as in figure 2 are used and candidate emerging patterns are extracted. In this example, items in  $\{a, X\}$ ,  $\{c, Y\}$ , and  $\{a, d, Y\}$  include their taxonomy in their patterns. Including both an item and its

item	item taxonomy
a	X
b	X
c	Y
d	Y

tid	class
1	positive
2	negative
3	positive
4	positive
5	negative

candidate emerging patterns	
a	X
c	X
c	Y
a d	Y

**Fig. 2** Examples of item taxonomy, class definition, and candidate emerging patterns

taxonomy in a pattern is not unusual and the pattern does not need to be extracted. Thus, such redundant patterns are removed using VSOP, and only non redundant patterns like  $\{c, X\}$  remain as a set of emerging patterns. The remaining processes are similar to those of the original CAEP.

## 4 Numerical Experiments

We apply our method to certain input data and compare it with another method. Here, we illustrate the results for a practical POS data. The data is from a retail store in Japan, and have some individual attribute data, such as age and historical purchasing data for several kinds of items. From preliminary basic analysis, there is a large gap in the total sales of the store between customers who purchase an item only once and the other customers. Thus, we identify two customer classes: the former is class 1, and the latter is class 2. We apply our method to this classification problem. In the computation, we can use customer attributes, and only purchased historical data at the first instance is used as a common condition. We use taxonomies such

item	content taxonomy
item A	content $\alpha$
item B	content $\beta$
item C	content $\beta$

age	age taxonomy
18	teens
19	teens
20	twenties
:	:
:	:
29	twenties
:	:
:	:

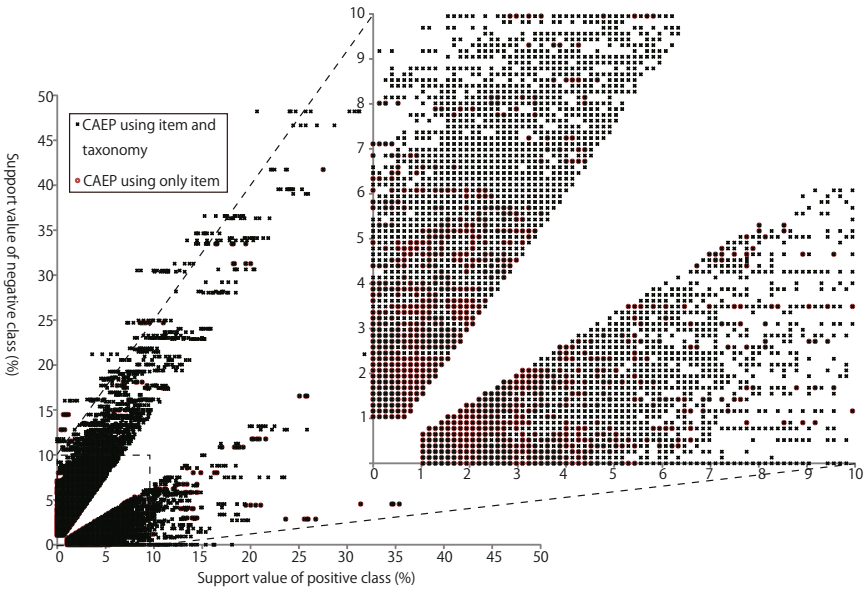
**Fig. 3** Examples of item taxonomy and class

as content taxonomy and age taxonomy. The content taxonomy which is defined by the store is provided from the beginning, whereas age taxonomy is provided by us additionally. For example, the relationship between item and taxonomy is as shown in figure 3 and it is defined by factors such as specific and attribute. While we can define an age taxonomy by deciding an interval of the age. In this experiment, we defined the age taxonomy by an interval of ten. In the other, zip code and some



**Table 1** Comparison of two methods

Data sets	CAEP with patterns of only item	CAEP with patterns of item taxonomy
1	76.22	80.30
2	76.68	79.91
3	75.06	78.04
4	76.36	79.26
5	76.49	80.36
Accuracy(best)	76.68	80.36
Accuracy(mean)	76.16	79.57
Computational time(mean)	86.30 sec.	227.92 sec.



**Fig. 4** Scatter graph of emerging patterns

demographical items are defined similarly. In the original data, deviation between the number of positive instances and negative instances is large, so we select the same instances by randomly changing random seeds. As a result, five data instances are generated. For the data sets, we apply our method, which is CAEP with patterns using taxonomy and CAEP with normal item patterns implementing ten cross validation; the gap between both is whether taxonomy is used or not. The accuracy for each data set is illustrated by figure 4. For every data set, our proposed method outperforms another one by several percentage points. However, in our method, we have to perform VSOP for every implementation, thus, the computational costs are higher compared to normal CAEP. Nevertheless, the computational time is acceptable, because it consumes approximately 20 seconds per implementation.

To identify the reason behind the result, figure 4 plot emerging patterns for each method. The symbols '×' and '○' denote emerging patterns for our proposed method and normal CAEP on both support value dimension, respectively. In the figure, upper right scatter plot is zoomed in a part of area of lower left figure. From these figures, we can see that emerging patterns in our method are more spread on the space than the another one, and see that in the bottom right and upper left domain where there are a few ○ symbols; our method can locate more emerging patterns. In short, our method enables us to find more powerful emerging patterns than another method, as a result, its overall improves accuracy better.

## 5 Conclusion and Future Works

In this paper, we propose an improvement of CAEP by using emerging patterns incorporating item taxonomy. The item taxonomy is very flexible and we can define them for factors such as age and zip code. From some computational experiments, we compare our method with normal CAEP using LCM, and show that our performance is better, and we can extract more powerful emerging patterns. Here, although we have defined item taxonomy, item taxonomy automatically to improve the performance is a more progressive method. In order to achieve this, it is an approach to create clusters that have similar emerging patterns. By incorporating such clustering method, we would like to improve our method better in the future.

## References

1. Dong, G., Zhang, X., Wong, L., Li, J.: CAEP: Classification by Aggregating Emerging Patterns. In: Arikawa, S., Nakata, I. (eds.) DS 1999. LNCS (LNAI), vol. 1721, pp. 30–42. Springer, Heidelberg (1999)
2. Dong, G., Li, J.: Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In: KDD 1999, pp. 43–52 (1999)
3. Uno, T., Asai, T., Uchida, Y., Arimura, H.: LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets. In: Proceedings of Workshop on Frequent itemset Mining Implementations, FIMI 2003 (2003)
4. Alhammady, H., Ramamohanarao, K.: Using Emerging Patterns to Construct Weighted Decision Trees. *IEEE Transactions on Knowledge and Data Engineering* 18(7), 865–876 (2006)
5. Morita, H., Nakahara, T., Hamuro, Y., Yamamoto, S.: Decision Tree-based Classifier Incorporating Contrast Pattern. In: Proceedings of the 13th IEEE International Symposium on Consumer Electronics, Kyoto, Japan (May 2009)
6. Bay, S.D., Pazzani, M.J.: Detecting Change in Categorical Data: Mining Contrast Sets. In: KDD 1999, pp. 302–306 (1999)
7. Minato, S.-i., Uno, T., Arimura, H.: LCM over ZBDDs: Fast Generation of Very Large-Scale Frequent Itemsets Using a Compact Graph-Based Representation. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 234–246. Springer, Heidelberg (2008)

# Synchronised Data Logging, Processing and Visualisation Framework for Heterogeneous Sensor Networks

Matthias Vodel, Rene Bergelt, Matthias Glockner, and Wolfram Hardt

**Abstract.** The proposed research work represents a generic concept for the synchronised logging, processing and visualisation of any kind of sensor data. The concept enables a chronological coordination and correlation of information from different, distributed sensor networks as well as from any other self-sufficient measurement systems. Based on the achieved relation between the several sensor sources, the information quality can be increased significantly. Accordingly, the aggregated, heterogeneous sensor information are convertible into multiple output formats. Dependent on the application specific requirements for the visualisation, we are able to consider additional meta-information from the test environment to optimise the data representation.

To evaluate of basic usability requirements and the efficiency of the proposed concept, an automotive sensor network represents a capable test system for the proposed framework. Within the demonstrator, the available on-board measurement systems were extended by high-precision sensor nodes, which establish an synchronised sensor network topology.

## 1 Introduction

Actual research projects in the field of wireless sensor networks operate on different, proprietary hardware platforms and contain multifaceted types of sensors. Currently, each measurement scenario consists of several application-specific and independent operating processes for the data-collection, -storage and -analysis.

It doesn't exist any uniform synchronisation techniques between the autonomous sensor systems. Accordingly, a detailed and target-oriented post-processing of the data sets within a shared knowledge base is not feasible. In consequence, we are not able to create unique relations between the different measurement information. Due to these missing relations, it is very hard to create a common primary index for the given, heterogeneous sensor platforms.

---

Matthias Vodel · Rene Bergelt · Matthias Glockner · Wolfram Hardt  
TU Chemnitz, Germany  
e-mail: [vodel, bere, mglo, hardt}@cs.tu-chemnitz.de](mailto:{vodel, bere, mglo, hardt}@cs.tu-chemnitz.de)

## 2 Related Work

During the last two decades, a couple of commercial tools for the measurement data recording and monitoring were developed. Unfortunately, most of them have functional or conceptual restrictions. Some of the vendors offer exclusive, hardware-specific analysis tools, which require special devices, predefined product series or vendors. Other sensor systems don't have any special software tools for extracting the measured data sets. There is no support for further post-processing steps.

In consequence, *LabView* [1], *jBEAM* [2] or *FlexPro* [3] offer multiple features to enhance the restricted vendor tools, which only provides a small set of general data recording and handling functions. These applications allow the interpretation of offline data from data bases or files as well as the live analysis of a given data source. Both *jBeam* and *LabView* operates platform-independent and all of these related tools support a lot of established data formats and communication interfaces. Especially *jBeam*, which integrates the *ASAM* standard [4], enables an easy and modular extension with user-defined components. Furthermore, *FlexPro* includes a lot of additional visual plugins and represents a complete visualisation framework for the given measurement data.

## 3 Concept

We are now looking for generic utilities and standards to route information from different sensor systems into a common data processing unit in a synchronised way, considering scenario-specific configuration schemes and sensor parameters. Hereby, synchronised time stamps for the heterogeneous sensor data sets are very important to allow correct correlations in the common knowledge base. Furthermore, such utilities allow us to define user-specific data analysis procedures during the measurement runtime and advanced data fusion techniques [5, 6, 7] to shrink the data volume directly within the sensor nodes.

To provide such features, we propose a capable and lightweight data management concept, which is able to bypass the already mentioned disadvantages. It combines advanced sensor network configuration features with resource-efficient operating parameters. Therefore, several basic demands have to be defined.

### 3.1 Requirements

The concept focuses on resource-limited systems and has to be feasible for a wide variety of application scenarios. Thus, the primary objective is a dynamic and flexible processing environment, which is adaptable to modifications in the configuration or in the analysis requirements. Furthermore, the data processing core has to be separable into two spatial, chronological and platform-specific operating modes. All components for the data measurement are working within the first mode. All

---

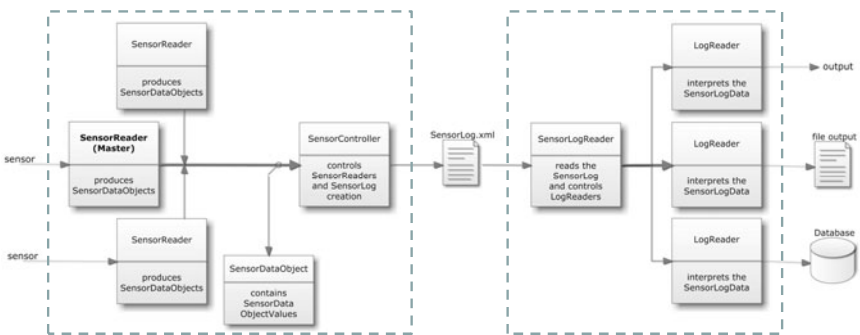
<sup>1</sup> ASAM - Association for Standardisation of Automation and Measuring Systems.

relevant modules for the data analysis as well as possible visualisation plugins operate independent within the second mode. Based on this requirement, we are able to map different data processing functions to predefined configuration scenarios. In contrast, other related software tools do not separate the data handling process into different phases in an efficient way.

The proposed concept deals with a standardised data transport and definable synchronisation parameters for collecting information from several distributed sensor components. Accordingly, changes in the data analysis process have no effects on the components of the data recording. This feature provides significant benefits, especially for complex sensor systems or inaccessible measurement environments.

### 3.2 Structure

As already mentioned, the structure of the proposed concept is divided into two operating modes. The first one encapsulates the data recording, synchronisation and correlation. A second mode processes the data analysis and generates a user-defined representation of the information sets. Due to the modular operating concept, the sensor net framework is completely independent from the given sensor configuration. Therefore, the environment uses an end-to-end communication design, called the *hourglass architecture*, which enables maximum interoperability between the several components. It means, that multiple sensor units and corresponding data processing units are connected by a dedicated *SensorController*. The sensor controller ensures a specific, universal mapping of the sensor information into a predefined format (see [Figure 1](#)). For the data output, the *SensorLogReader* component provides different modules for the information representation. The result is a very high diversity on both data input and data output components. In contrast, the binding middle part shows a strict uniformity.



**Fig. 1** Sensor net framework structure. The separation into two operation modes for the data collecting (left) and data analysis (right) is described.

This structure fulfils the central requirement of a separated data processing core for gathering and analysing the sensor measurements. Hereby, the common XML representation of the entire scenario is essential and allows us to transfer the information for any kind of application. All internal and external parameters of the environment as well as special meta-information regarding the measurement scheme are correlated together with the data sets. Accordingly, researchers are able to reconstruct the whole test scenario with synchronised data, time stamps and a detailed system configuration. The reuse factor, for instance in the field of automotive testing scenarios, increases substantially.

Furthermore, we have the ability to modify existing sensor configuration in a time efficient way. The structure allows us to add or remove single components without changing the data flows within the overall monitoring system. In principle, it is also possible to create a direct interconnection between data collecting and visualisation. This increases the runtime significantly and reduces the used resources. But without the common XML representation, is not possible to capture the entire measurement scenario in a reusable way.

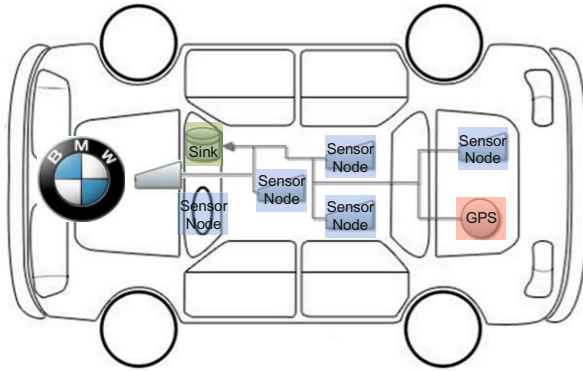
## 4 Application Scenarios

The proposed concept was developed to manage several sensor net scenarios at our computer engineering department. To clarify functional aspect of the implemented framework, we describe the data processing flow by a real-world automotive measurement system [8]. For this monitoring scenario, the existing sensor components of a given research vehicle were upgraded with high-definition sensor nodes. These nodes are placed at predefined positions to monitor the entire environment and provide independent measurement data about the current temperature, light intensity as well as the acceleration in two axes and the magnetic field strength. Thus, the established wireless sensor network provides meta-information about the measurement environment and external parameters. *Figure 2* illustrates the measurement scenario. In addition, we integrated a high-resolution GPS<sup>2</sup> sensor, which enables the correlation between absolute positioning information, speed, altitude and the available on-board vehicle data.

In our exemplary case, the system provides a high-resolution GPS unit. Besides positioning information, this sensor also provides an accurate time signal. In consequence, the given time stamps from the GPS sensor represent the global *synchronisation master*. Thereby, the framework is not restricted for using a time stamp as master index. Especially for the integration of multiple, autonomous sensor systems and a missing central scheduling entity, a user-defined choice for the synchronisation master provides important benefits.

---

<sup>2</sup> GPS - Global Positioning System.

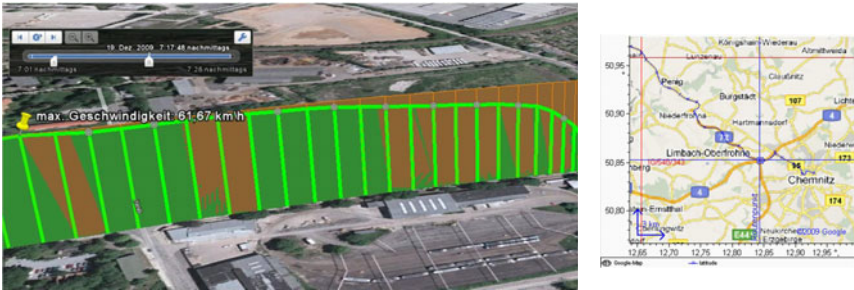


**Fig. 2** Measurement system - All data from the sensor nodes and the GPS module are transmitted to the data sink in the vehicle, represented by the proposed framework.

## 5 Data Analysis

For post-processing the collected data, two data analysis components were implemented. The first one is an export module, which prepares the sensor data sets for the storage in a given database system and accordingly transmits the chosen information. A second module is responsible for converting the sensor data with dedicated visualisation plugins, e.g. for Google Earth.

Within Google Earth, an additional 3D altitude track extends the visualised measurement curves (represented in *figure 3*).



**Fig. 3** Data visualisation in Google Earth.

## 6 Conclusion and Future Work

The proposed research work describes the implementation of a comprehensive data processing environment for heterogeneous sensor systems. The basic concept provides generic structures for many further research projects in the field of novel data

aggregation and data fusion techniques. For an easy data collecting and data analysis process, we are now able to synchronise and correlate the single data sets also on resource limited and embedded computer systems. The result is a common and extensive knowledge base, which integrates all information sources into complex data sets.

## References

1. National Instruments. LabView (2010), <http://www.ni.com/labview/>
2. AMS GmbH. jBEAM (2010), <http://www.jbeam.de/german/produkte/jbeam.html>
3. Weisang. FlexPro (2010), <http://www.weisang.com/>
4. ASAM Consortium. Association for Standardisation of Automation and Measuring Systems (2010), <http://www.asam.net/>
5. Gupta, V., Pandey, R.: Data Fusion and Topology Control in Wireless Sensor Networks. WSEAS Trans. Sig. Proc. 4(4), 150–172 (2008)
6. Qi, H., Iyengar, S.S., Chakrabarty, K.: Distributed Sensor Fusion - A Review of Recent Research. Journal of the Franklin Institute 338(1), 655–668 (2001)
7. Qi, H., Wang, X., Iyengar, S.S., Chakrabarty, K.: Multisensor Data Fusion in Distributed Sensor Networks Using Mobile Agents. In: Proceedings of International Conference on Information Fusion, pp. 11–16 (August 2001)
8. Vodel, M., Lippmann, M., Caspar, M., Hardt, W.: A Capable, High-Level Scheduling Concept for Application-Specific Wireless Sensor Networks. In: Proceedings of the World Engineering, Science and Technology Congress (ESTCON 2010) - 4th International Symposium on Information Technology (ITSIM), Kuala Lumpur, Malaysia, pp. 914–919. IEEE Computer Society (June 2010)



# Design of SENIOR: A Case Study Using NoGap

Per Karlström, Wenbiao Zhou, and Dake Liu

## 1 Introduction

The design and implementation of a new Application Specific Instruction-set Processor(ASIP) processor is usually the result of a substantial design effort, more details about the Application Specific Instruction-set Processor (ASIP) design process can be found in [10]. There are a number of different software tools that relaxes the design effort in one way or another. However all these tools forces the designer into a predefined architecture template. This limitation in design flexibility often makes designers of novel ASIP processors and programmable accelerators revert back to an Hardware Description Language (HDL), e.g. Verilog or VHDL. HDLs offers full design flexibility at the register transfer level, but the flexibility comes at the cost of increased design complexity. All details, e.g. register forwarding and/or pipeline control, has to be handled manually.

This paper will present a case study, where we have used Novel Generator of Accelerators and Processors (NoGap) to design an advanced Reduced Instruction Set Computing (RISC) processor with Digital Signal Processor (DSP) extensions, called SENIOR. The System Verilog code generated by NoGap, was successfully synthesized and targeted to both FPGA and Application Specific Integrated Circuit (ASIC) flows.

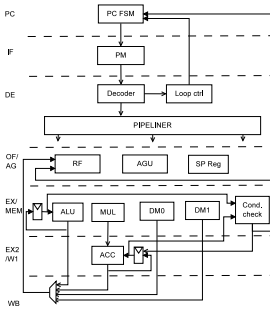
As the focus of this paper is not on NoGap the reader is referred to previous publications about NoGap for more detailed information [5, 7, 4, 6].

## 2 Related Work

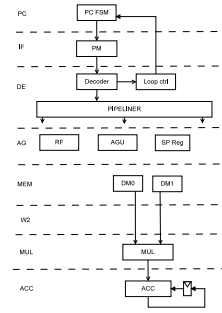
A number of tools such as LISA [12], EXPRESSION [2], nML [1], MIMOLA [9], ArchC [11], and ASIP Meister [3], tries to help ease the effort needed to design a

---

Per Karlström · Wenbiao Zhou · Dake Liu  
Linköping University, department of EE, Japan  
e-mail: [{{perk, zhou, dake}@isy.liu.se](mailto:{{perk, zhou, dake}@isy.liu.se)



**Fig. 1** Pipeline architecture for normal instruction in SENIOR



**Fig. 2** Pipeline architecture for convolution instruction in SENIOR

processor. They all have strengths and shortcomings. More details about how they compare to  $\mathcal{N}\mathcal{O}\mathcal{G}\mathcal{a}\mathcal{p}$  can be found in the related work section of [8].

### 3 SENIOR Architecture

The SENIOR processor is a single issue RISC processor with DSP extensions, based on the Harvard architecture. SENIOR is divided into a data path and sequence path. The data path consists of a single precision 16 bit Arithmetic Logic Unit (ALU), a double precision Multiply And Accumulate (MAC) unit, two data memories with dedicated address generated units, a condition checker, flag computation unit, a loop controller, and some data store units including a register file consisting of 32 16 bit registers, 18 16 bit special registers, and four 40 bit MAC register including eight guard bits. The sequence path contains a program counter, a Program Counter-Finite State Machine (PC-FSM) and an instruction memory.

The SENIOR processor has two kinds of pipelines, one normal pipeline and one longer pipeline. The different pipeline architectures are also illustrated in Fig. 1 and Fig. 2.

## 4 $\mathcal{N}\mathcal{O}\mathcal{G}\mathcal{a}\mathcal{p}$ Common Language ( $\mathcal{N}\mathcal{O}\mathcal{G}\mathcal{a}\mathcal{p}^{CL}$ ) for SENIOR

$\mathcal{N}\mathcal{O}\mathcal{G}\mathcal{a}\mathcal{p}^{CL}$  is a the default  $\mathcal{N}\mathcal{O}\mathcal{G}\mathcal{a}\mathcal{p}$  facet used to construct the Micro Architecture Structure Expression ( $\mathcal{M}\mathcal{a}\mathcal{s}\mathcal{e}$ ), Micro Architecture Generation Essentials ( $\mathcal{M}\mathcal{I}\mathcal{a}\mathcal{g}\mathcal{e}$ ) and Control Architecture STructure Language ( $\mathcal{C}\mathcal{a}\mathcal{s}\mathcal{t}\mathcal{l}\mathcal{e}$ ) descriptions, more details about  $\mathcal{N}\mathcal{O}\mathcal{G}\mathcal{a}\mathcal{p}^{CL}$  can be found in [8].

### 4.1 $\mathcal{M}\mathcal{I}\mathcal{a}\mathcal{g}\mathcal{e}$ in SENIOR

$\mathcal{M}\mathcal{I}\mathcal{a}\mathcal{g}\mathcal{e}$  Functional Units (FUs) are used to describe leaf module with functionality, like most of the components in Fig. 1. An example can be seen in Listing 1.

which is the Address Generation Unit (AGU) in SENIOR. The AGU is using dynamic clause selection to implement 10 different addressing modes for data memory access. Clauses 1 in NoCap represent the different functionality in a single FU.

Listing 1 M<sub>age</sub>

```
fu agu_0
{
  input [2:0] op_i; ...
  comb{
    switch(op_i) {
      0:%INC{ //post-increment
        addr_o=areg_i;
        areg_o=areg_i+step_i;
      }
      1:%DEC { ... }
      2:%OFS { ... }
      3:%MINC { ... }
      4:%BRV { ... }
      ... ; //other addr-mode
    }
  }
}
```

Listing 2 M<sub>ase</sub>

```
fu data_path {
  input [33:0] op_i;
  ... //More ports declaration

  fu::decoder_spec<instr_i,flush_i>()
    dec_unit;

  fu::agu_0(%INDR) agu_0;
  fu::agu_1(%INDR) agu_1;
  fu::addr_reg(%NOP) areg;
  ... //More FU declarations

  phase DE, OF, EX, EX2, WBL;
  phase ME, WA, MU, AC;
  stage ff() { cycle {ffo=ffi;} }
  pipeline long_pipe{ DE -> ff -> OF ->
    ff -> EX -> ff -> EX2 -> ff -> WBL;

  operation(long_pipe)
    dblld(dec_unit.dblld) {
      @DE;
      ...;
      @OF;
      rf;
      areg`%WRITEAB`;
      agu_0`%OFS,%INC,%DEC,%MINC,%BRV`;
      agu_0.imm12_i=0;
      areg.addr_a_i = dec_unit.arem_a;
      areg.addr_b_i = dec_unit.arem_b;
      agu_0.arem_i = areg.arem_a_o;
      areg.arem_a_i = agu_0.arem_o;
      ...;
    }
  operation .....//other operation
}
```

Listing 3 Decoder

```
fu:template<INSTRUCTION>
decoder_spec{
  input auto(("#") INSTRUCTION);
  output [3:0] rf_w;
  output [3:0] rf_x;
  output [3:0] areg_a;
  output [3:0] areg_b;
  output jump_o;

  immediate [3:0] imm_rxf_w;
  immediate [3:0] imm_rxf_x;
  immediate [3:0] imm_arem_a;
  immediate [3:0] imm_arem_b;
  ...;
  instruction dblld{
    source{
      areg_a = imm_arem_a;
      areg_b = imm_arem_b;
    }
    destination{
      rf_w = imm_rxf_w;
      rf_x = imm_rxf_x;
    }
    jump_o = 0;
  }
  ...;//other instructions
}
```

## 4.2 M<sub>ase</sub> in SENIOR

The overall processor architecture and its instructions are described using a M<sub>ase</sub> FU. It defines the interconnection and clause selection for every M<sub>age</sub> in the data path or control path. A code excerpt from the pipeline definition can be found in Listing 2.

## 4.3 Castle in SENIOR

A code excerpt from the decoder definition in SENIOR is shown in Listing 3 which is used to construct Castle. The code also illustrates the syntax for instructions definitions in the decoder template.

<sup>1</sup> A clause here refers to a collection of statements delimited by a start and end marker, e.g. in C++ this would translate to statements delimited by { and }. In NoCap Common Description (NoCap<sup>CD</sup>) the typical syntax is %NAME {s1; ... sn;}.

## 5 Results

### 5.1 FPGA Flow

The SystemVerilog code generated by *NöGap* was synthesized to a Virtex-4 LX80 speed-grade 12 FPGA. Precision was used for synthesis and Xilinx ISE 10.1 was used for mapping and place & route. The generated code contains no FPGA specific optimizations. The data and instruction memory are not part of the processor design, they are instantiated externally, using the block RAMs present in the Virtex-4. The memories are integrated into the data path of PIONEER using external interface FUs.

Table 1 displays the Look-Up Table (LUT) usage by part for the synthesized result.

**Table 1** LUT usage

Part	LUTs
SENIOR	1211/7508
+data path	800/6136
+sequence path	74/105
total	7508

**Table 2** Comparison for SENIOR in FPGA

Virtex-4	Original	<i>NöGap</i>
Max frequency	75.1MHz	85.02MHz
Slices usage	5928	7508
Critical path	mac_reg(o)→scale→mac_reg(o)→scale→acc→round→sat→acc→round→sat→fwd→mul(i)	mac_reg(o)→scale→mac_reg(i)

The FPGA comparison between original SENIOR and *NöGap* SENIOR is shown in Table 2. In the table, (o) and (i) stands for the pipeline registers at the output and input of the modules.

The timing performance results could make the observant reader surprised to see that the maximum frequency in *NöGap* SENIOR is higher than in original SENIOR, which is heavily optimized and has been designed for a long time. The reason is that original SENIOR have register forwarding implemented which means that forwarding multiplexers are inserted and thus increase the time delay. The critical path is almost the same as the critical path in original SENIOR except without a forwarding multiplexer.

### 5.2 ASIC Flow

The SystemVerilog code generated by *NöGap* was also targeted to an ASIC flow. For this we used a tool chain supplied by Synopsys. The comparison between original SENIOR and *NöGap* SENIOR, for the ASIC flow, is shown in Table 3. All the memories, peripherals, and Direct Memory Access (DMA) units are removed from both versions of SENIORs. The critical path reported by the ASIC synthesis tool is also listed in the table, (o) and (i) stands for the pipeline registers at the output and input of the modules.

**Table 3** Comparison for SENIOR in ASIC

CORE65LPSVT	Original	NoGap
Max frequency <sup>2</sup>	238 MHz	215 MHz
Area	63812 $\mu\text{m}^2$	63878 $\mu\text{m}^2$
Power	3.0511 mW	3.1328 mW
Critical path	scale(i)→round→acc→ round→sat→fwd(o)	scale(i)→acc→round→ sat→mac_reg(i)
Designing time	> 4.5 years	~ 3 man-weeks

The maximum frequency reached for the ASIC design were 215 MHz, which is only 10% less than original SENIOR in 65 nm technology with 100% low power logic. The area and power consumption are almost the same in both SENIORs.

## 6 Conclusion

Architecture Description Language (ADL) tools are valuable tools that can be used to speed up design times for processor designs. The currently available ADL tools, for processor design, assumes a lot about the underlying architecture of the processor and thus limit the design freedom for novel processor architectures. Although NoGap also assumes something about the system being designed it assumes a lot less than any other ADL tool, giving the designer more freedom to explore novel processor architectures. This paper presented a case study, where the DSP processor SENIOR, was reimplemented using NoGap. The design process took under three man-weeks.

The most important conclusion about the hardware generated with NoGap, is that the hardware introduced by NoGap were not in the critical path and thus did not incurred any penalty on the performance of the system.

**Acknowledgements.** Thanks to VR, the Swedish research council, for their funding.

## References

1. Fauth, A., Van Praet, J., Freericks, M.: Describing instruction set processors using nML. In: Proceedings of European Design and Test Conference, ED&TC 1995, pp. 503–507 (1995), <http://dx.doi.org/10.1109/EDTC.1995.470354>, doi:10.1109/EDTC.1995.470354
2. Halambi, A., Grun, P., Ganesh, V., Khare, A., Dutt, N., Nicolau, A.: EXPRESSION: a language for architecture exploration through compiler/simulator retargetability. In: Proceedings of Design, Automation and Test in Europe Conference and Exhibition 1999, pp. 485–490 (2002), <http://dx.doi.org/10.1109/DATE.1999.761170>, doi:10.1109/DATE.1999.761170

3. Itoh, M., Higaki, S., Sato, J., Shiomi, A., Takeuchi, Y., Kitajima, A., Imai, M.: PEAS-III: an ASIP design environment. In: Proceedings of 2000 International Conference on Computer Design, pp. 430–436 (2000), <http://dx.doi.org/10.1109/ICCD.2000.878319>, doi:10.1109/ICCD.2000.878319
4. Karlström, P., Akhlaq, F., Loganathan, S., Zhou, W., Liu, D.: Cycle accurate simulator generator for NoGAP. In: 2010 Asia Pacific Conference on Postgraduate Research, Microelectronics and Electronics (PrimeAsia), , pp. 57–60 (2010), <http://dx.doi.org/10.1109/PRIMEASIA.2010.5604963>, doi:10.1109/PRIMEASIA.2010.5604963
5. Karlström, P., Zhou, W., Liu, D.: Automatic port and bus sizing in NoGAP. In: International Symposium on Systems, Architectures, Modeling, and Simulation, SAMOS 2010 (2010)
6. Karlström, P., Zhou, W., Liu, D.: Implementation of a floating point adder and subtractor in NoGAP, a comparative case study. In: IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, EUC (2010)
7. Karlström, P., Zhou, W., Liu, D.: Operation classification for control path synthetization with NoGAP. In: 2010 Seventh International Conference on Information Technology: New Generations (ITNG 2010), pp. 1195–1200 (2010), <http://dx.doi.org/10.1109/ITNG.2010.142>, doi:10.1109/ITNG.2010.142
8. Karlström, P.A.: NoGAP: Novel generator of accelerators and processors. Ph.D. thesis, Linköping University Linköping University, Computer Engineering, The Institute of Technology (2010)
9. Leupers, R., Marwedel, P.: Retargetable code generation based on structural processor descriptions (1998), <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4520>
10. Liu, D.: Embedded DSP Processor Design, Volume 2: Application Specific Instruction Set Processors (Systems on Silicon), illustrated edn. Morgan Kaufmann (2008), <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0123741238>
11. Rigo, S., Araujo, G., Bartholomeu, M., Azevedo, R.: ArchC: A SystemC-based architecture description language. In: 16th Symposium on Computer Architecture and High Performance Computing, SBAC-PAD 2004, pp. 66–73 (2004), <http://dx.doi.org/10.1109/SBAC-PAD.2004.8>, doi:10.1109/SBAC-PAD.2004.8
12. Zivojnovic, V., Pees, S., Meyr, H.: LISA - machine description language and generic machine model for HW/SW Co-Design. In: Proceedings of the IEEE Workshop on VLSI Signal Processing, pp. 127–136 (1996), <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7123>

# GWDL: A Graphical Workflow Definition Language for Business Workflows

M. Sultan Mahmud, Saad Abdullah, and Shazzad Hosain

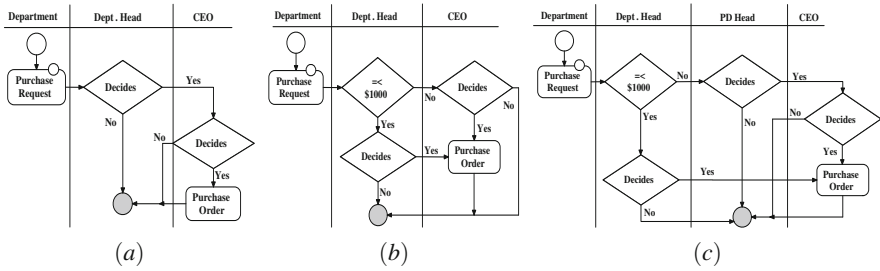
**Abstract.** Workflow Management Systems (WfMS), both open source and commercial, represent workflows either in XML or in other structured languages that cause impedance mismatch with the relational database of application software. This paper presents a new modeling scheme for WfMS, where both workflow data and application data can share a common relational database. However, the main focus is to develop a Graphical Workflow Definition Language (GWDL) so that a process expert, who is often a non-IT expert, can design a workflow visually using “drag and drop” method. We have developed a prototype system, where users define workflow in GWDL, which is then automatically translated to relational tables through a series of transformations.

## 1 Introduction

The cost of business process automation is soaring day by day as business rules change very frequently. Let us take the following example: A startup company has six departments, any purchase order initiated from a department is verified by the department head and forwarded to *CEO* who decides (approve/disapprove) for necessary actions. The workflow is shown in figure 1a. The company has a software system to automate this workflow. Now, assume that the company grows and purchase rule changes: “If purchase order  $\leq 1000$  USD, the department head takes decision, else the *CEO* decides”, shown in figure 1b. The software requires to modify accordingly. Now, let’s say, the company creates a new department, “Procure Department” (*PD*), to look after purchases  $> 1000$  USD. If *PD* Head approves any such purchases then it goes to *CEO* for final approval as shown in figure 1c. Similarly many such changes may arise that have the following consequences:

---

M. Sultan Mahmud · Saad Abdullah · Shazzad Hosain  
Department of EECS, North South University, Dhaka, Bangladesh  
e-mail: [{{sultan068,saad511}}@eeecs.northsouth.edu](mailto:{{sultan068,saad511}}@eeecs.northsouth.edu)  
[shazzad@northsouth.edu](mailto:shazzad@northsouth.edu)



**Fig. 1** Three different versions of a workflow

- Requires software people to make changes in software that increase the cost.
- Requirement change after deployment of software increases the cost sharply.
- The software is not readily available for the new requirement, because software change requires considerable time. Even sometimes when the new software arrives the requirements change again.

Very recently workflow systems, both commercial and open source, are coming up to address these issues in business process automation. Workflow is the sequence of actions used in a process, which is run by more than one involved parties and uses many different resources usually raised from a set of operation rules [11]. Systems that automate workflows are known as Workflow Management System (WfMS). A good WfMS has the following properties:

- Separation of business process from data requirements.
- Easy process designing so that non-IT people can easily design a workflow using “drag and drop” idea.
- The workflow representation should be easily integratable to applications.

Since most of the business applications store data in relational databases, relational representation of workflows is more desirable than other representations. But most of the WfMSs that we find in the literature use XML based languages such as XPDL (XML Process Definition Language) [5], XRL (eXchangeable Routing Language) [14] or other structured languages such as PDL (Process Definition Language) [11], YAWL (Yet Another Workflow Language) [6] etc., which create an impedance mismatch with the existing business applications. Thus our goal is to develop a WfMS that will store workflow definitions in relational database, at the same time process experts can design workflows by “drag and drop” method. First a workflow is defined using a graphical editor, then it is translated to a series of steps into relational tables so that the execution of workflow is faster and has no impedance mismatch.

The paper is organized as the following. Section 2 presents related works, section 3 presents the Graphical Workflow Definition Language (GWDL), section 4 shows the translation process of GWDL to relational representation, and finally section 5 gives the conclusion and future works.



## 2 Related Works

In the literature we find two different types of WfMS models. The first model uses XML based workflow languages such as XPDL (XML Process Definition Language) [5], XRL (eXchangeable Routing Language) [14], YAWL [6] and so on. While XPDL is XML based, XRL and YAWL are petri-net based that use XML for their underlying representations. According to [8], systems that follow this model allow users to define workflows either in textual languages such as JFolder [2], OpenWFE [3], or through a visual interface using “drag and drop” idea such as Bonita [1], JOpera [10], YAWL [6] etc.

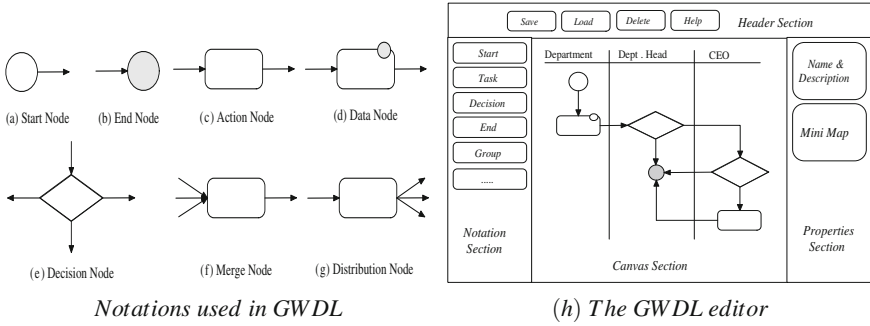
The second WfMS model defines workflows in PDL (Process Definition Language) [7, 11], which is close to natural language, and stores workflow in relational database. Thus workflow data and application data share the same database space and avoid impedance mismatch. This model supports many of the workflow patterns (or basic elements) as identified in [13] that can exist in workflow definitions.

In our research we take the third approach, where we provide a graphical editor for defining workflows as well as store workflow definitions in relational database. Thus our model includes the strengths of graphical editor as well as removes the impedance mismatch between WfMS and application software. Authors of [7, 11] first propose the relational representation of workflows and we extend this model in our case. The main database consists of two layers: *definition* layer and *execution* layer. However, in this paper we focus only with the definition layer.

## 3 Graphical Workflow Definition Language

Like many other visual languages, the Graphical Workflow Definition Language (GWDL) also has a number notations. Using these GWDL notations one can construct a workflow visually. Kiepuszewski in [9] evaluated workflow systems based on the coverage of workflow patterns found in [13], and showed that a standard workflow model should support the common patterns that produce single instance and dead-lock free workflows. In GWDL we have implemented eight out of nine standard patterns using the following seven notations:

1. **Start Node:** Denotes the start of a process that has no input edge, figure 2a.
2. **End Node:** Denotes the end of a process that has no output edge, figure 2b.
3. **Action Node:** This node receives a task from a source node, performs some action on that task and then sends the task to the next node. It contains single input and single output as shown in figure 2c.
4. **Data Node:** This node, shown in figure 2d, depicts a task that uses one or more data form(s). For example, “Purchase Request” data node in figure 1b tells that a request form is submitted with this task.
5. **Decision Node:** This notation, figure 2e, is used to perform the decision related tasks. It contains single input but multiple outputs with multiple results.



**Fig. 2** Notations of GWDL and the GWDL editor

6. **Merge Node:** This notation has multiple inputs and a single output as shown in figure 2f. It can be *OR* merge or *AND* merge [13].
7. **Distribution Node:** It is opposite to merge node. It contains a single input but multiple outputs with same result for parallel split, shown in figure 2g.

A workflow graph has three main components: *nodes*, *edges* and *parties*. Parties participate in a workflow, for example in figure 1c, *Department*, *Dept. Head*, *PD head* and *CEO* are the parties involved in the workflow.

### 3.1 The GWDL Editor

We have developed a Web based graphical editor for GWDL. The GWDL editor, figure 2h, has four sections: *Header*, *Notation*, *Canvas* and *Properties*. The *Header* section consists of four buttons (**New**, **Load**, **Save** and **Help**) for creating, saving and loading/editing a process. The *Notation* section contains all the notations that are used to define a process. Process experts can easily define a workflow by dragging and dropping these notations. The *Canvas* section is basically an empty area where users can draw a process by specifying process name and description from the *Properties* section. It also contains a *mini map* that displays traversable components, objects, important locations, and helps user to observe the whole process.

## 4 GWDL to Relational Representation

The workflow in GWDL is stored into relational tables in two steps. First, the GWDL definition of a workflow is translated to an intermediary language, then the intermediary language is parsed to generate SQL statements that are executed to store workflow definitions into relational tables.

node Id	party Id	description	source Node	destination Node	description	party Id	description
10	1	Purchase request	10	20	Submit purchase request	1	The department
20	2	Check amount	20	30	Less than \$1000	2	Head of the department
30	2	Department Head verifies	30	60	Order for purchase	3	Procure department
...	...	... ..	...	...	... ..	...	... ..

(a) nodes
(b) edges
(c) parties

Fig. 3 Relational tables

### 4.1 The Intermediary Language

We have developed the GWDL editor using an open source library called ‘WireIt’ [4]. This library is extensively modified to extract data from the graphical workflow definition and to generate the intermediary language. For example, table 1 shows the intermediary language for the workflow shown in figure 1. The intermediary language has three main sections: *Info*, *Task* and *Edge*.

**Info:** The *Info* section contains information about the name and description of a workflow. It also contains the involved group’s or party’s *name* and *id*. For example, the first five statements in table 1 belong to *Info* section. As we see, the first statement contains workflow name “purchase” and description “purchase application”, and the next four lines define the four parties involved in the workflow and each one is identified with a number from 1 to 4.

**Task:** This section contains different nodes’ information and the statements start with “Node” keyword, shown in table 1. Each node has its individual *node id*, associated *party id* and *data form id*. For example, the first line of this section in table 1 has node name “purchase request”, *node id* 10, *party id* 1 and *data form id* 100. The next statement has no *data form id* since the node is a control flow node.

**Edge:** This section contains information about edges with “Edge” keyword, shown in table 1. The statement contains two numbers for *source node* and *destination node*, then the edge description followed by a condition if the source node is a decision node.

Table 1 Intermediary Language for the Workflow from Figure 1

[Workflow : <purchase>, <purchase application>],	[Node: <purchase order>, < 60 >, < 4 >],
[Party :<the department>, < 1 >],	[Node: <end process>, < 70 >, < 3 >],
[Party :<head of the department>, < 2 >],	[Edge: < 10 >, < 20 >, <submit purchase request>],
[Party :<procure department>, < 3 >],	[Edge: < 20 >, < 30 >, <less than \$1000>, <yes>],
[Party :<chief executing officer>, < 4 >],	[Edge: < 30 >, < 60 >, <order for purchase>, <yes>],
[Node: <purchase request>, < 10 >, < 1 >, < 100 >],	[Edge: < 30 >, < 70 >, <end the process>, <no>],
[Node: <check amount>, < 20 >, < 2 >],	[Edge: < 20 >, < 40 >, <greater than \$1000>, <yes>],
[Node: <dept. head verifies>, < 30 >, < 2 >],	[Edge: < 40 >, < 70 >, <procure dept. rejects>, <no>],
[Node: <procure dept. approval>, < 40 >, < 3 >],	[Edge: < 40 >, < 50 >, <procure dept. approves>, <yes>],
[Node: <CEO approval>, < 50 >, < 4 >],	[Edge: < 50 >, < 60 >, <order for purchase>, <yes>],
	[Edge: < 60 >, < 70 >, <CEO rejects the order>, <no>]

## 4.2 The Relational Representation

We have developed a parser that translates the intermediary language into relational tables. For example, the *nodes*, *edges* and *parties* are stored in *nodes*, *edges* and *parties* tables respectively as shown in figure 3. There are other relationship tables too according to the design principles found in [7, 11].

## 5 Conclusion and Future Works

In this paper we have presented a new model to define business workflows. In this model the user first creates a process using a visual editor in GWDL, which is translated and stored in relational tables. We used WireIt [4] to translate the GWDL representation of the workflows to an intermediary language, which eventually is translated to the relational representation of the workflows. Thus the model brings both the advantages of the graphical modeling and the power of relational databases. GWDL editor supports most frequently used workflow patterns as discussed in [13]. The challenge will be to support more advanced workflow patterns and find their relational representations.

## References

1. BONITA, <http://www.bonitasoft.com/>
2. JFolder, <http://www.jfolder.com/>
3. OpenWFE, <http://java-source.net/open-source/workflow-engines/openwfe>
4. WireIt - A Java Script Wiring Library, <http://neyric.github.com/wireit/>
5. XPD, <http://www.xpd.org/>
6. Aalst, W.V.S., Hofstede, A.H.M.T.: YAWL: Yet Another Workflow Language. *Information Systems* 30, 245–275 (2002)
7. Amir, P., Michael, H.: Gathering Unstructured Workflow Data into Relational Database Model Using Process Definition Language. In: *Proceedings of the 24th IASTED International Conference on Database and Applications*, Innsbruck, Austria, pp. 32–37 (2006)
8. Garcés, R., de Jesus, T., Cardoso, J., Valente, P.: Open Source Workflow Management Systems: A Concise Survey. In: *2009 BPM & Workflow Handbook*, pp. 333–346 (2009)
9. Kiepuszewski, B.: Expressiveness and Suitability of Languages for Control Flow Modelling in Workflows. Ph.D. thesis. Queensland University of Technology, Brisbane, Australia (2002)
10. Pautasso, C., Alonso, G.: The JOpera Visual Composition Language. *Journal of Visual Languages and Computing* 16, 119–152 (2005)
11. Pourabdollah, A.: A User-Friendly Process Description Language Used in Creating Database Model of Workflows. Master's thesis. The University of Nottingham (2004)
12. Russell, N., Hofstede, A.H.M.T., Mulyar, N.: Workflow ControlFlow Patterns: A Revised View. Tech. rep., BPM Center Report BPM-06-22, BPMcenter.org (2006)
13. Van Der Aalst, W.M.P., Ter Hofstede, A.H.M., Kiepuszewski, B., Barros, A.P.: Workflow Patterns. *Distributed Parallel Databases* 14, 5–51 (2003)
14. Verbeek, H.M.W., Van Der Aalst, W.M.P., Kumar, A.: XRL/Woflan: Verification and Extensibility of an XML/Petri-Net-Based Language for Inter-Organizational Workflows. *Information Technology and Management* 5, 65–110 (2004)

# A Feature Representation and Extraction Method for Malicious Code Detection Based on LZW Compression Algorithm

Yingxu Lai, Hongnan Liu, and Zhen Yang

**Abstract.** Differing in traditional methods which extracted too much features or filtered valuable items, we proposed a feature representation and extraction method based on LZW compression algorithm to detect malicious codes. The compression algorithm not only reduces the number of features, but also is enough to cover malicious codes. In this paper, we described the process of our feature extraction in detail, including 0-data processing, fix-length coding and threshold setting. The experimental results show that our method outperforms other methods based on Bayes and SVM in DR and AR.

## 1 Introduction

According to first half of 2010 Internet Security Report from Rising [1], its “cloud security” system have intercepted 4,221,366 new computer virus samples, in which Trojan, backdoor and worm occupied top three.

The traditional signature-based scanning technology performs well in detecting known malicious codes, but it could not be applied to the unknown ones. So detection of unknown malicious codes has become one of the prime research interests of information security. In general, extracting features is a difficult problem in the area of unknown malicious code detection. Many solutions have been proposed. Kolter et al. [2] used a hex-dump utility to convert each executable to hexadecimal code in ASCII format and produced *n-gram* features by combining each four-byte sequence into a single term. They used information gain to select valued features and tested the performance. Reddy [3] proposed a novel feature extraction method by extracting variable length *n-gram* from binary codes of files, and used

---

Yingxu Lai · Hongnan Liu · Zhen Yang

College of Computer Science, Beijing University of Technology, Beijing 100124, China  
e-mail: {laiyingxu, yangzhen}@bjut.edu.cn

class-wise frequency as feature selection method. Schultz et al. [4] observed three feature extraction methods (DLL, Printable String and Byte Sequences). They reported that the best performing approaches were naive Bayes over the printable strings. Reference[5] introduced a feature extraction method based on API Call sequence. In those methods,  $n$ -gram extracted too many features which affect the detection rate, and API or String method maybe filter items which is value to present malicious codes.

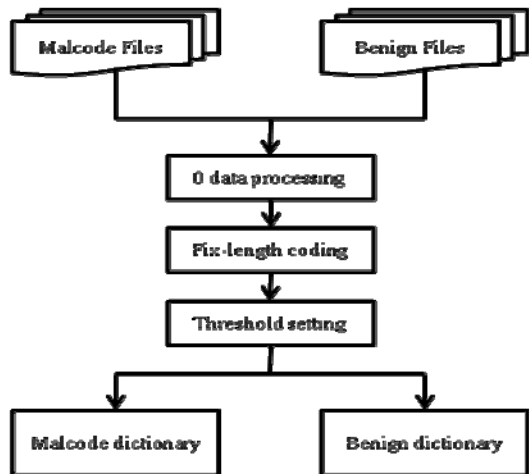
To overcome these disadvantages, we proposed a feature representation and extraction method based on compression algorithm. Compression algorithm not only reduces the number of features, but also is enough to cover malicious codes. In this paper, we are interested in applying LZW compression algorithm as feature representation and extraction method to detect malicious code. Previous work [7] has proved that this method could detect unknown malicious code efficiently. We will introduce the feature extraction process in detail. Experimental results show that the scheme detects malicious codes far more efficient than other known methods.

The rest of the paper is organized as follows. Section 2 states the feature selection method. Section 3 outlines detection model. The experimental results are discussed in Section 4. Finally, conclusion is sketched in Section 5.

## 2 Feature Extraction Method

### 2.1 Frame of Feature Extraction

The process of feature selection is shown in Fig. 1.



**Fig. 1** Flow chart of Feature Selection

We get features from malicious and benign dictionaries which are built by extraction process. In this paper, we adopt the idea of LZW compression algorithm which puts the string appeared firstly into dictionary and replace the string with a code. From the perspective of information theory, compression is to remove redundant information that retain uncertain information and get rid of certain information, which instead of the original redundancy description with closer to the nature information.

In next section, we will discuss each step in Fig. 1.

## 2.2 0-Data Processing

0-data is the data that is presented as 00 in hex, and they are meaningless in ASCII. Many malicious and benign files are win32 executable files which has their own format (PE file format). Due to segment alignment and reserved space at runtime, there are many 0-data in PE files, which are useless to extract feature. We adopt a method to process them shown in Fig. 2. When 0-data reach a certain number, the series of 0-data are considered as segment separation,  $M_{len}$  is the threshold.

---

### Algorithm 1. 0-data processing algorithm

Input: The data stream in PE files

Output: String dictionary

Set a counter and initial it to 0

Set a flag and initial it to true

For each char in data stream

    If flag is true

        If the char is not 0 then Operate it

        Else then counter increase and set flag to false

    Else

        If counter is less than  $M_{len}$

            If the char equal to 0 then counter increase

            Else then set flag is true, operate counter-chars before current char, set counter is 0

        Else

            If the char is not 0 then set flag is true, operate it, set counter is 0

            Else then counter increase

        End If

    End If

End For

---

Fig. 2 0-data processing

## 2.3 Fix-Length Coding

Fix-length coding [6] is a method that changes every character in a file to a given length code. As non-text files do not have obvious word structure, LZW algorithm

could not get excellent effect. We should convert non-text files to text files with word structure, and then use LZW algorithm to extract features. Computer file is composed of Byte, which has a code between 0 and 255. Fix-length coding method is that high Byte and low Byte from data stream are separated, then high Byte and low Byte are replaced with 0~F, output the converted results.

## 2.4 Threshold Setting

Dictionary is created according to the rule of LZW, each item in the dictionary is not limited in length. The worst case is that item includes a long string, which is difficult to store and search. To overcome the disadvantage, a limit of item's length should be set. Fig. 3 shows the process of building LZW dictionary under threshold.

---

**Algorithm 2.** Feature dictionary is built with threshold

**Input:** The sequence of the data from array converted by fix-length coding. Stream = {ch1, ch2, ch3....}

**Output:** String Dictionary

1. For each char in Stream
  2. STRTEMP add the char to its back;
  3. if The length of STRTEMP is over threshold
  4. clear STRTEMP
  5. end if
  6. if STRTEMP is not in the Dictionary
  7. Add STRTEMP to Dictionary
  8. clear STRTEMP
  9. end if
  10. end For
  11. Output the Dictionary to Database
- 

**Fig. 3** The process of feature dictionary built under threshold

## 3 Detection Model

Detection model is shown in Fig. 4. In our previous work [7], the model was also used. In the figure, main modules are described as following.

**Feature selection module:** Extracting features from learned files.

**Feature database:** Storing features created from feature selection model.

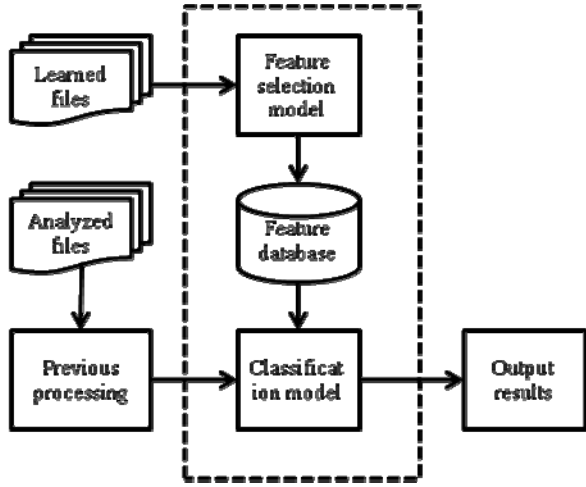
**Previous processing:** include 0-data processing and fix-length coding.

**Classification model:** Using features from feature database to compress analyzed files and classifying files to the class which get better compression ratio.

In this section, we will introduce our classification method. The rule of classification is that using malicious and benign dictionaries to compress analyzed file by which different compression ratio can be gotten, and then according to MDL (Minimum Description Length), classifying the file to the class obtained better compression ratio.



**Fig. 4** Detection model based on LZW compression algorithm



**Definition.** We assume that the size of analyzed file is  $S_{size}$ , the size of analyzed file compressed by benign dictionary is  $B_{size}$ , the size of analyzed file compressed by malicious code dictionary is  $M_{size}$ , compression ratio  $N_{compress}$  and  $V_{compress}$  are defined as following.

$$N_{compress} = \frac{S_{size} - B_{size}}{S_{size}} \times 100\% \quad (1)$$

$$V_{compress} = \frac{S_{size} - M_{size}}{S_{size}} \times 100\% \quad (2)$$

If  $N_{compress}$  is higher than  $V_{compress}$ , analyzed file is classified to benign file class, else to malicious file class.

## 4 Experimental Results and Analysis

Our training set is consisted of two parts, malicious and benign which have 1000 instances respectively. All benign instances are obtained from system32 folder in Windows and the malicious code instances are obtained from VX Heavens (VX HEAVEN, available on <http://vx.netlux.org/>), shown in table 1.

As our previous work [7] has discussed the threshold set at 200 is better. In this paper, we contrast our mentioned method with other malicious detection methods (Bayes and SVM).

**Table 1** The type and number of malicious codes composed

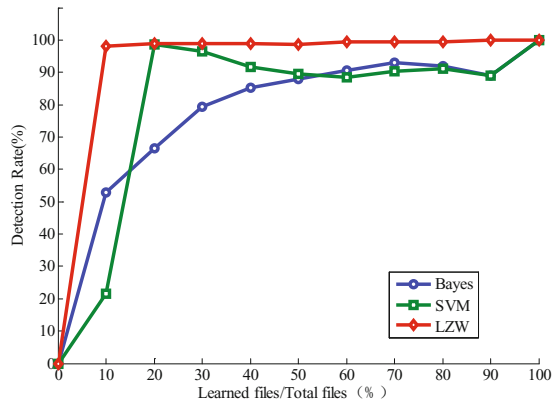
Type	number
Backdoor	94
Exploit	84
DoS	21
Trojan	304
Worm	276
Constructor	16
Flooder	87
Virus	118

For evaluation purposes, we define that Detection Rate ( $DR$ ) is the proportion of malicious codes classified correctly, and Accuracy Rate ( $AR$ ) is the proportion of codes classified correctly to benign or malicious class.

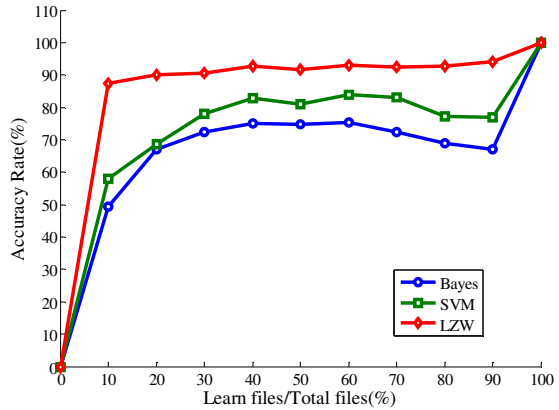
$$DR = \frac{TP}{TP + FN} \quad (3)$$

$$AR = \frac{TP + TN}{N} \quad (4)$$

In [8], author compared several classification methods, and stressed out Bayes based on string has best effect. SVM is also a good method to detect unknown malicious codes. In this paper, our method is compare with Bayes and SVM (RBF kernel, and Penalty factor  $C=100$ ). The experiment result is shown in Fig. 5 and 6.

**Fig. 5** Detection Rate of different methods

**Fig. 6** Accuracy Rate of different methods



From Figure 5 and 6, we know that our method is better than Bayes and SVM method in DR and AR. Although the absence of sufficient instances, it also has a good effect.

## 5 Conclusion

Malicious code is complex, heterogeneity and non-structure, there are many difficulties in feature extraction and expression. Therefore it is hard to deal with malicious codes by the traditional methods. To overcome these issues, we propose a method based on the LZW compression algorithm for unknown malicious codes detection. First of all, features are extracted from normal and malicious codes according to compress algorithm. Then both normal and malicious compression dictionaries are built by the extract feature module. In order to detect unknown malicious codes, normal code dictionary and malicious code dictionary are used to compress an analyzed file by which two compression ratios are gotten. According to MDL theory, the analyzed file is classified into the class which has better compression ratio.

Our method can select effective features with a series of feature extraction methods (include 0-data processing, fix-length coding and threshold setting). The experiment results show that our method outperforms other malicious code detection methods.

As feature selection method based on LZW compression algorithm will create a big dictionary, it occupies very large memory. Our future work will be to solve how to rebuild dictionary when dictionary items reaches a certain length.

**Acknowledgment.** Supported by National Natural Science Foundation of China (No: 61001178), Beijing Natural Science Foundation (No:4102012), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (No:PHR201108016) and Scientific Research Common Program of Beijing Municipal Commission of Education (No:SQKM201210858003).

## References

1. Rising Co. "Rising cloud security" system, Rising First Half of 2010 Internet Security Report (2010), <http://www.rising.com.cn/about/news/rising/2010-07-30/7950.html>
2. Kolter, J.Z., Maloof, M.A.: Learning to Detect and Classify Malicious Executables in the Wild. In: International Conference on Knowledge Discovery and Data Mining, pp. 470–478 (2004)
3. Reddy, D.K.S., Dash, S.K., Pujari, A.K.: New Malicious Code Detection Using Variable Length n-grams. In: International Conference on Information Systems Security, pp. 276–288 (2006)
4. Schultz, M.G., Eskin, E., Zadok, E., et al.: Data Mining Methods for Detection of New Malicious Executables. In: Proc. of the 2001 IEEE Symposium on Security and Privacy, pp. 38–49 (2001)
5. Zhang, B., Yin, J., Hao, J.: Intelligent Detection Computer Viruses Based on Multiple Classifiers. *Ubiquitous Intelligence and Computing*, 1181–1190 (2007)
6. Wang, F.: An improved algorithm based on LZW code. *Journal of Wuhan Polytechnic University*, 65–68 (2009)
7. Lai, Y., Liu, H.-N., Yang, Z., et al.: Research of Unknown Malicious Codes Detection Based on LZW Compression Algorithm. Submitted to *Journal of Beijing University of Technology*
8. Schultz, M.G., Eskin, E., Zadok, E.: Data Mining Methods for Detection of New Malicious Executables. In: *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, pp. 38–50 (2001)

# CASM: Coherent Automated Schema Matcher

Rudraneel Chakraborty, Faiyaz Ahmed, and Shazzad Hosain

**Abstract.** Schema matching has been one of the basic tasks in almost every data intensive distributed applications such as enterprise information integration, collaborating web services, web catalogue integration, and schema based point to point database systems and so on. Typical schema matchers perform manually and use a set of matching algorithms with a composition function by using them in an arbitrary manner which results in wasteful computations and needs manual specification for different domains. Recently, there has been some schema matching strategy proposed with partial or full automation. Such a schema matching strategy is OntoMatch. In this paper, we propose an element level automated linguistic based schema matching strategy motivated by the concept of OntoMatch, with more powerful matching algorithms and definite property construction for matcher selection that produces better output. Experimental result is also provided to support the claim of the improvement.

## 1 Introduction

As the Internet becomes a vast repository of information, often it requires integrating many different sources to answer a single query. For example, let one likes to attend the DEIT 2011 conference at Bali, Indonesia, he likes to stay in a hotel nearby the conference venue and the hotel should be nearby of a public transport facility so that he could visit tourist places easily. To get all the necessary information often he requires visiting several Web sites such as airline ticket reservation sites, hotel reservation sites, tourism sites etc. and gleaning the information to get the correct picture. To address the problem it requires autonomous integration of different sites and the research community has investigated several approaches in this direction.

---

Rudraneel Chakraborty · Faiyaz Ahmed · Shazzad Hosain  
Department of EECS, North South University, Dhaka, Bangladesh  
e-mail: [rudraneel, faiyaz@eeecs.northsouth.edu](mailto:rudraneel,faiyaz@eeecs.northsouth.edu)  
[shazzad@northsouth.edu](mailto:shazzad@northsouth.edu)

One of the approaches is to model the Web as relations [7] so that semantic heterogeneity can be resolved through automatic object identification and consolidation process. However, the key problem of this approach is to find a good schema matching technique, often known as *schema matcher*, so that it can match two different terms from two different sources automatically.

Schema matching is the solution of finding semantic relationship between elements of two schemas. Manually performing schema matchers usually use a set of simple matching algorithms such as, edit distance matcher [10], synonym matcher [3] etc, then combine the individual matching scores using some function such as average, weighted average and so on. This approach has its inherent limitations. For example, let us take two terms *movie* and *cinema*. While the synonym matcher would give a score of 1 in the scale of 0 to 1, which reflects a perfect match; using another matcher, let's say edit-distance matcher would give a score of 0.3. If we take the average of these two scores, it will be 0.65, and if we take 0.8 as a threshold, then the matcher would output a mismatch. So the problem is to decide which set of matchers in which order should we consider to determine the final score of the match?

A number of approaches have been proposed that use machine learning techniques, data mining techniques and so on. However, very recently Bhattacharjee et. al. [5] proposed a new way to decide different matching algorithms from a set of matching algorithms. He proposed OntoMatch [5], a property based matcher, that cleverly decides a subset of matchers from a set of known matchers to increase the quality of match. Though OntoMatch performs well to Cupid [8], COMA [6] and so on, it has several limitations. It uses edit-distance matcher for string similarity, while Jaro-Winkler [11] and Monge-Elkan [9] perform better than Edit-Distance-Matcher. Also there are few limitations in matcher selection approach, which we will describe in next section. In our research we addressed these limitations of OntoMatch and developed a new property based schema matcher called CASM that outperforms OntoMatch [5] and other matchers such as COMA [6].

The paper is organized as the following. Section 2 provides the related works, section 3 gives problem formulation, section 4 describes matcher characterization, section 5 gives CASM implementation, section 6 gives comparative analysis and finally section 7 gives conclusion and future works.

## 2 Related Works

In the literature, we found three types of matching algorithms: instance based matching, representation based or schema matching and usage based or ontology matching. We are only interested in schema matching approaches that only consider schema information, not the instance data. While most of matchers have emerged from the context of a specific application, a few approaches such as COMA [6], Cupid [8], and OntoMatch [5], try to address the schema matching problem in a generic way that is suitable for different applications and schema languages. Cupid [8] is a linguistic based matcher. Systems like COMA [6] uses a set of matchers and

combines the match result with similarity combination function. `OntoMatch` [5] also uses a set of matchers but selects those based on some properties of the matchers. In our research we also select matchers from a set of matchers much like `OntoMatch` do.

### 3 Problem Formulation

Before defining the problem, let us formally define the following few concepts.

**Definition 3.1 (Schema and Term).** A schema  $S$  of a database is a list of attributes  $\mathcal{A}$ , where each  $a \in \mathcal{A}$  has its associated domain  $d \in \mathcal{D}$ . A term  $t$  is either  $a$  or  $d$ .

**Definition 3.2 (Term Similarity).** Let  $\psi'$  be a similarity function<sup>1</sup> such that for any two  $t_1, t_2 \in \mathcal{A} \vee \mathcal{D}$ ,  $\psi'(t_1, t_2) \in [0, 1]$ , where 0 indicates complete dissimilarity and 1 means two terms are identical. Let  $\psi$  be a function and  $\varepsilon$  be a threshold such that  $\psi(\varepsilon, \psi'(t_1, t_2)) \in \{0, 1\}$ , i.e.,  $\psi(\varepsilon, \psi'(t_1, t_2)) = 0$  if  $\psi'(t_1, t_2) < \varepsilon$ ,  $\psi(\varepsilon, \psi'(t_1, t_2)) = 1$ , otherwise, and we say  $t_1$  and  $t_2$  are similar, denoted  $t_1 \sim t_2$ . From here on, we write  $\psi(t_1, t_2)$  as a short hand for  $\psi(\varepsilon, \psi'(t_1, t_2))$  unless specified otherwise.

**Definition 3.3 (Mapping).** Given two schemas  $S_1$  and  $S_2$  the mapping is a list of attribute pairs i.e.  $\{ \langle a_1, b_1 \rangle, \langle a_2, b_2 \rangle, \dots \}$ , where  $a_i \in S_1$ ,  $b_i \in S_2$  and  $a_i \sim b_i$

**Definition 3.4 (Term Matcher).** Given two terms  $t_1, t_2$  a term matcher  $\mu$  returns whether the two terms match or not, based on the values of similarity functions.

**Definition 3.5 (Distance Matrix).** Let  $S_1, S_2$  be two schemas,  $m$  is the number of attributes in  $S_1$ ,  $n$  is the number of attributes in  $S_2$ , then a distance matrix is a two dimensional array  $dm[m][n]$  that stores the value of  $\psi'(t_i, t_j)$ , where  $1 \leq i \leq m$  and  $1 \leq j \leq n$ .

#### 3.1 Schema Matching Problem

Given two schemas  $S_1$  and  $S_2$  the schema matching problem is to find all the correct mappings among the terms of the two schemas. Usually the linguistic matchers use more than one term matcher from a set of term matchers  $\{\mu_1, \mu_2, \dots, \mu_n\}$  to determine the ultimate mapping. In our research effort we focus on selecting the best term matcher(s) based on some properties of the matchers for the schema matching problem. To limit our problem, we choose the following five term matchers, of which we apply one or more term matchers to find the final matching value.

- **Jaro-Winkler:** The Jaro-Winkler [4] distance metric is a measure of similarity which is basically a variant of the Jaro distance metric algorithm. For two terms  $t_1$  and  $t_2$ , the Jaro metric calculates the similarity by identifying the number and the placement of the common characters between the two terms.

---

<sup>1</sup> Such as string edit distance [10], thesaurus at WordNet [3], etc. or a combination of such functions.

- **Abbreviation Matcher (AB):** For two terms  $t_1$  and  $t_2$ , the abbreviation matcher returns value 1 if abbreviation of  $t_1$  is  $t_2$  or abbreviation of  $t_1$  is abbreviation of  $t_2$  or vice versa. Otherwise it returns 0.
- **Monger-Elkan:** Monger-Elkan [9] matcher is an affine (assigns unique cost to each edit operation) variant of Smith-Waterman distance metric. It has unified cost parameters that produces similarity values between 0 and 1.
- **Sound-Matcher:** For two terms  $t_1$  and  $t_2$ , the sound-matcher returns 1 if they sounds similar otherwise returns 0. The matcher is based on the Soundex [12] algorithm, which is a phonetic algorithm that indexes names by their sounds when pronounced in English.
- **Synonym-Matcher:** Given two terms  $t_1$  and  $t_2$ , the synonym matcher returns 1 if the terms are synonymous to each other, else it returns 0.

## 4 Matcher Characterization

The basic goal here is to identify the best matcher  $\mu$  for the term  $t_1, t_2$  from a set of available distinct matchers  $\mu_1, \mu_2, \dots, \mu_n$ . To get the best applicable matcher, we set some properties to matchers. First, the abbreviation matcher, AB matcher ( $\mu_1$ ), is used based on the property that one term is much shorter than the second term. If the AB matcher provides a score less than the threshold or AB matcher is not applicable, i.e. the terms are of similar length, then it applies Jaro-Winkler-Monger-Elkan matcher, JWME matcher ( $\mu_2$ ). The JWME matcher applies two matchers Jaro-Winkler [4] and Monge-Elkan matcher [9] and chooses one that gives the best score. If JWME matcher does not provide score greater than the threshold, it applies synonym matcher, SM matcher ( $\mu_3$ ) and sound matcher, SD matcher ( $\mu_4$ ), if required. Finally, if no satisfactory match result is found, it returns the best match value produced so far.

## 5 CASM Implementation

To verify the concept, we have implemented CASM and used the data available in the UIUC web integration repository [2]. The UIUC repository contains XML schemas, while the CASM works on relational schemas. So we first parse the XML schema to convert it into relational schema, then apply the CASM algorithm. The overall mapping process is divided into three steps:

1. Parse the XML schema using external XML-parser such as JDOM [1].
2. Compute the distance matrix  $dm$ , definition 3.5, using the CASM algorithm given below.
3. Obtain a final mapping based on threshold, matching score and distance matrix by applying the *stable marriage* algorithm.

Based on the properties the algorithm first sorts the term matchers. The order is  $\mu_1, \dots, \mu_4$  in this case as stated in section 4. The algorithm then applies the matchers



---

**Algorithm 1.** Algorithm to select the best applicable matcher

---

**Require:**  $S_1, S_2, bestScore = 0, dm[m][n] = 0$ , where  $m = |S_1|, n = |S_2|$   
 sort the matchers  $\mu_i \in \mathcal{M}$   
**for** each term  $t_i \in S_1$  **do**  
     **for** each term  $t_j \in S_2$  **do**  
          $bestScore = 0$   
         **for** each matcher  $\mu_i \in \mathcal{M}$  **do**  
              $score = \mu_i(t_i, t_j)$   
             **if**  $score \geq \tau$  **then**  
                  $bestScore = score$   
                 **break**  
             **end if**  
         **if**  $score \geq bestScore$  **then**  
              $bestScore = score$   
         **end if**  
     **end for**  
      $dm[i][j] = bestScore$   
**end for**  
**end for**

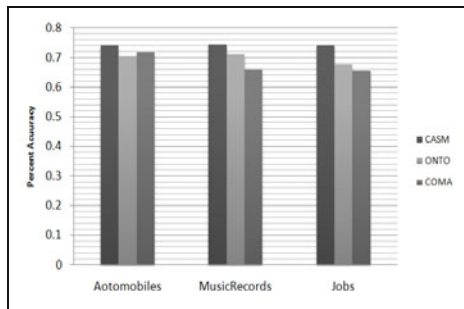
---

successively, stops when it gets a match or assigns the best score among all the scores that are less than the threshold.

## 6 Comparative Analysis

The performance of CASM was compared using BAMB extracted query schemas available in the UIUC web integration repository [2]. We took three types of schemas: Automobiles, MusicRecords and Jobs. Each category contains 40 schemas on an average. We chose one schema and compared it with all others schemas of the same category. The schemas have 3 to 20 attributes and we chose the average one with 12 attributes to compare it with other schemas of the same category. Figure 1 shows the comparison of CASM with two other matchers.

**Fig. 1** CASM is compared with two other matchers OntoMatch and COMA. It uses the area under the precision recall curve (AUPRC), the same technique used by OntoMatch, to measure performance. CASM performs better than both of the matchers in all three data sets.



## 7 Conclusion and Future Works

In this paper we presented a new approach towards ordering term matchers in an efficient way. This new approach selects one/more term matchers from a pool of matchers for automatic schema matching. The matcher CASM performed better than OntoMatch and COMA. We provided the experimental result to support the claim. Also the CASM's selection algorithm showed that only the appropriate matcher(s) was(were) selected thus avoiding wasteful computation.

In future, we like to modify the algorithm to include more term matchers and to use more data sets to increase the confidence of the result. However, the more important future direction would be improve CASM to handle XML schemas.

## References

1. JDOM, <http://www.jdom.org>
2. UIUC Web Repository, <http://metaquerier.cs.uiuc.edu/repository/datasets/bamm/index.html>
3. WordNet, <http://wordnet.princeton.edu/>
4. Jaro, M.A.: Probabilistic linkage of large public health data files. *Stat Med.* 14(5-7), 491–498 (1999)
5. Bhattacharjee, A., Jamil, H.M.: OntoMatch: A monotonically improving schema matching system for autonomous data integration. In: *IEEE IRI, Las Vegas, NV (2009)*
6. Do, H.H., Rahm, E.: COMA - A system for flexible combination of schema matching approaches. In: *VLDB, Hong Kong, China*, pp. 610–621 (2002)
7. Hosain, S., Jmail, H.: An algebraic language for semantic data integration on the hidden web. In: *Proceedings of the 3rd IEEE ICSC 2009, Berkeley, California*, pp. 237–244 (2009)
8. Madhavan, J., Bernstein, P.A., Rahm, E.: Generic schema matching with Cupid. In: *VLDB, Italy*, pp. 49–58 (2001)
9. Monge, A., Elkan, C.: The field matching problem: Algorithms and applications. In: *2nd International Conference on KDDM*, pp. 267–270 (1996)
10. Ristad, E.S., Yianilos, P.N.: Learning string edit distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence* 20(5), 522–532 (1998)
11. Winkler, W.E.: The state of record linkage and current research problems. Tech. rep., Statistical Research Division, U.S. Census Bureau (1999)
12. Zobel, J.: Phonetic string matching: Lessons from information retrieval. In: *SIGIR 1996*, pp. 166–172 (1996)

# Research on Constructing 3D Geological Model of the Construction Layers in Daxing New City Area of Beijing City

Xiao Huang, Yingshuang Wang, Naiqi Shen, Yufeng Liu, and Gang Chen

**Abstract.** The new city in Daxing district is the focus of future development in the southern of Beijing city, establishing 3D geological model of its engineering construction layer can provide a scientific basis for urban planning and construction as well as the land's proper utilization. After collecting the geological drilling data, we establish the geological model based on a 3D geological modeling platform developed by Peking University. The 3D spatial modeling method based on sections is used to analyze two key technical problems about simulation process named high-precision modeling and rapid modeling respectively. In the modeling process, Delaunay triangulation and Kriging interpolation algorithms are used. Furthermore, the manual intervention is achieved through the use of flexible and convenient 3D interactive modeling tools. The application examples indicate that the 3D geological model constructed by this method can better reflect the actual situation, it also possesses the advantages of high precision and fast modeling.

## 1 Introduction

In recent years, with the development of economy and the frequent human activities, urban geological problems become increasingly prominent. The geo-environment is the fundamental basis of urban construction, so deeply understanding its characteristics is helpful to sustainable development of urban construction.

---

Xiao Huang · Yingshuang Wang · Naiqi Shen · Yufeng Liu  
China University of geosciences, Beijing, China  
e-mail: nqshen@cugb.edu.cn

Gang Chen  
Institute of Geological engineering(BIGE), Beijing, China  
e-mail: Chengang2011@163.com

In the past, the traditional geological information was normally shown in two-dimension, for instance, geological maps and profiles are used to analyze geological problems, which might only reflect spatial information limitedly and distortedly. With the development of science and technology and the universal use of computer technology, the three-dimensional geological simulation system comes to the world. It would truly reflect the actual urban geology and maximally enhance the accuracy of geological analysis. Since 1980s, foreign countries have launched many sophisticated 3d modeling software, such as Surpac, GOCAD, 3Dmove etc, which have been widely used in underground mining, petroleum exploration and geological structure analysis[1,2]. China is relatively later in this field but has made several achievements after ten years' exploration. For example, based on Windows NT and 3d graphics libraries OpenGL, Gong Jianhua developed a 3D visualization system Geo3Dvision, dynamically displayed the 3d geographical entity; Xu Nengxiong developed a rock mass structure visualization modeling system which can reconstruct the complex structure of the rock mass model; Dr. Yang Qin developed a 3d geological visual interactive system which can construct geological model mixed up with lens and intrusive masses. Peking University developed a three-dimensional geological modeling platform named GSIS in recent years, it can automatically / semi-automatically construct 3d geological model based on vertical drilling, section, fault investigation and other data. However, the above mentioned groups of modeling software need to be further improved in modeling speed, accuracy, and some other aspects.

Located in the heart of Beijing, Tianjin and Hebei, Daxing New City is the focus of future development in the south of China's capital. Constructing 3d geological model of the engineering construction layer is helpful to control the strata's spatial extension in different directions and analyze the geological conditions. Moreover, it also provides a scientific basis to urban planning and construction, land utilization and prevention of geologic hazards. In this research, based on 3d geological modeling platform developed by Peking University, the model is constructed by section-based method. In addition, some key techniques are discussed to improve modeling precision and modeling speed. The application example indicates that the modeling approach can achieve pretty considerable results.

## **2 General Geology of Daxing New City Area**

Located in the south of Beijing city, Daxing new city is situated at the mid-lower part of aggraded flood and alluvial plain of the Yongding River, the elevation is 32 meters to 52 meters. The land is flat, and slightly sloping from northwest to southeast by an average slope of 1.0 percent to 1.5 percent.

Drilling survey results shows that the upper strata mainly consist of quaternary system artificial soil and loose sedimentary soil of aggraded flood and alluvial plain. The thickness of the layer is 50 meters to 60 meters in the northwest, while increase to 250 meters in the southeast. The lower strata are bedrocks and it consists of Jixian System, Qingbaikou System, Cambrian, Ordovician, Cretaceous and

Tertiary System. The construction of this new city is mainly relevant to the underground soil within 30 meters to 40 meters including artificial fill, pebble, sand, clay and silt, etc. The soil layers change greatly in the spatial distributions and the typical profile is shown in Figure 1.

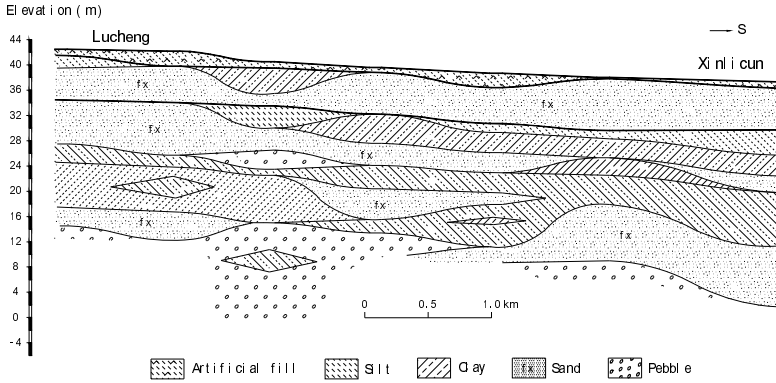


Fig. 1 A geological profile from Lucheng to Xinlicun of Daxing new city.

### 3 Modeling Method

The specific modeling process is shown in Figure 2. First of all, the drilling layout is carried on according to geological data. Synchronously, some virtual drilling should be added to meet the accuracy requirement. And then, we should connect strata and construct 2D geological sections. Finally, the section data is put into 3d geological modeling system, a precise 3d geological model is built after some unreasonable place amended.

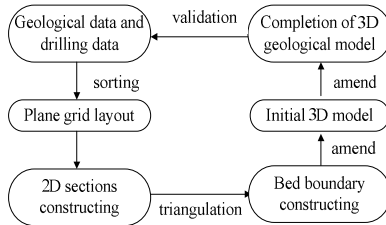
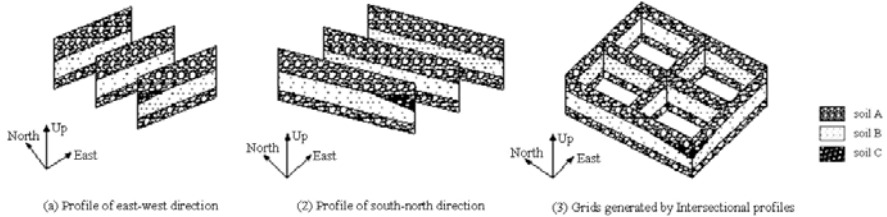


Fig. 2 The flow chart of modeling process.

From Figure 1 we can see that the studying area composes typical multi-layer soil structures, and lots of lenses. As a result, section approach is chosen as the geological modeling method due to the complex geological conditions. Different

from the traditional modeling method based on parallel sections, this work chooses series of netted sections of cross-cutting (Figure 3). Divided into several partitions, each part of the model should be handled independently. This method has a high degree of automation; it can better describe strata spatial distributions in different directions, the strata's pinch-out position can be controlled more accurately, so a more precise model is obtained.



**Fig. 3** The sketch of simple modeling based on the cross section.

## 4 Key Technologies

Borehole data from geological drilling is the main source of 3D geological modeling. These data have the characteristics of complex, heterogeneous, uncertainty, dynamic and multi-source. Thereby, there are some key technical problems should be concerned in the simulation process, as the “high-precision modeling” and “rapid modeling” are two typical examples. After careful analysis, we put forward the following methods to solve them.

### A Using Interpolation Algorithm

The modeling is based on sections, therefore the precision mainly depends on two aspects: one is the degree of geological surveying, just the precision of the geological section itself, the more detailed the description of formation, the higher precision the model will be; the other is the density of the geological cross-sections, the density is bigger, the sections will be bigger in quantity. So it can reflect the actual situation much better and significantly increases the modeling precision. In this modeling process, Delaunay triangulation and Krieger interpolation algorithms are used to improve modeling precision respectively.

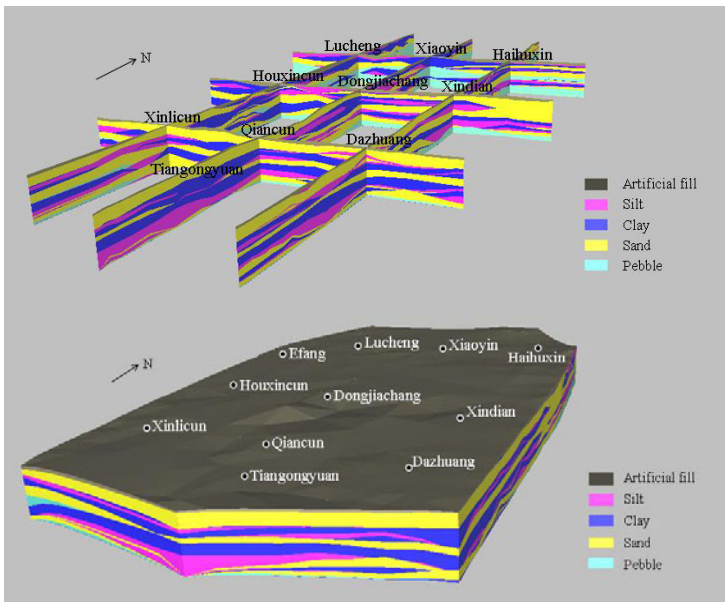
### B Using the Interactive Modeling Tools

Due to the complexity and multi-solutions of geological problems, the aim to set up the entire geological model is unrealistic only counting on the computer automatically. It is necessary to use the flexible and convenient interactive modeling tools to accelerate the modeling speed. In our research, we select the mouse to

directly conduct artificially interactive operations based on software-oriented techniques. On the screen, users can use the mouse to divide model boundary and external constraint and they can also modify the internal nodes. Its basic operation includes model zoom, translation, rotation, and cutting and attribute query etc. The common usage of keyboard and mouse will improve the modeling usability and flexibility, and greatly accelerate the speed of modeling. However, as noted, it is difficult to thoroughly solve the problem of 2D screen pixels which have multi-solutions in 3d space by using the keyboard and mouse, so it still needs some researches in depth.

### 5 Sketch of Model

Daxing new city is located in the south of Beijing city, which stem at Xindiancun in the east, Houxincun in the west, Xiaoyin in the north and Tiangongyuan in the south. It covers 67 square kilometers. Based on the engineering exploration work in Daxing new city and the above modeling methods, we construct the 3D geological model of the construction layer by applying the modeling platform developed by Peking University. According to the accuracy requirement, we do the layout of drilling (Figure 4), and then select a total of 147 boreholes with the depth of 30 meters to 40 meters. After sorting and screening, we import these data into the modeling system. Finally, after several adjustments and amendments, the geological model is constructed as Figure 5 shows.



**Fig. 4** 3d geological model of the engineering construction layers of Daxing new city.

The model achieves the 3D visualization of the engineering geology information, and it faithfully represents the underground geological structures in the study area. Meanwhile, it has the practicability to forecast the geological information of the unknown area or the region with limited data.

## 6 Conclusions

Using the modeling method based on sections can better describe the strata spatial distributions in different directions. As a result, it can control the stratum's pinch-out positions more precisely to attain the lifelike effect. Two key technical problems named "high-precision modeling" and "rapid modeling" are solved by using interpolation algorithm and interactive modeling tools. The models constructed using these approaches will be more precise in details.

In the models of engineering construction layers in Daxing new city, the 3D visualization of engineering geology information is displayed, and it better reflects the spatial distributions of strata in the study area. It also provides a scientific basis to the construction planning and the rationalization of land exploitation.

## References

- [1] Zheng, G.Z., Shen, Y.L.: 3D analysis of geological characteristics and status research of 3D geology modeling. *Advance in Earth Sciences* 19, 218–223 (2004)
- [2] Zhu, L.F., Wu, X.C., Yin, K.L., Liu, X.G.: Study of Management and Service System of Urban 3D Geological Data Supported By 3D GIS. *J. of HUST (Urban Science Edition)* 4, 40–46 (2003)
- [3] Ning, S.N., Li, Y.F., Liu, T.F.: Study on the visualization software theory of 3D geological body. *Coal Engineering* 7, 41–43 (2002)
- [4] Nu, X.J., Liu, X.Q., Yu, C.L.: The outcome report of Beijing urban geological information management and the service system (unpublished)



# Preliminary Study of HGML-Based Virtual Scene Construction and Interaction Display

Yang Wenhui, Xie Yan, and Miao Fang

**Abstract.** In spatial information networks which incorporate the client side dynamic aggregation services of G/S mode, in order to construct virtual scene that serves the real world, it requires extremely high fidelity and ability of real-time interaction of the environment. When dealing with data that have diverse, complex types and tremendous scale, consistent organization, management and data processing are essential. This article centers on introducing: synthesizing virtual reality environments based on HGML, 3D terrain models and 3D entity models, which are constructed through DEM and remote sensing data, and a preliminary research of the implementation and real-time interaction of the dynamic features.

**Keywords:** G/S, HGML, virtual scene construction, styling; interaction display.

## 1 Introduction

HGML (Hyper Geographic Markup Language), based on XML, is a kind of distributed spatial data modeling and visualization language, and was first time put forward by Miao Fang, professor in Chengdu University of Technology, at a conference when he mentioned the technical architecture of the “Digital China”, 2007.

With the emergence of “digitalization” and “intelligentize”, the construction of virtual environment and demonstration of interactive operations (Virtual Display) has widely used in city construction, Medicare, transportation, military, exhibition and many other fields, for it is capable of restoring past, displaying present and creatively imagining the future. Virtual Display is a computer simulated 3D environment, based on LAN and the Internet.

---

Yang Wenhui · Xie Yan

Key Lab of Earth Exploration & Information Techniques of Ministry of Education,  
Chengdu, China

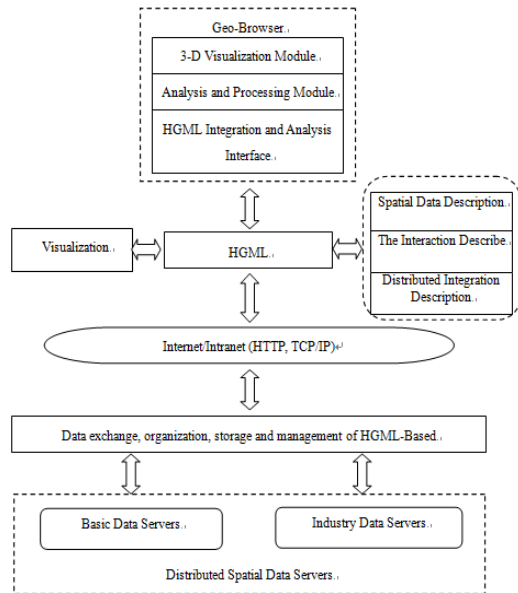
e-mail: ywhui@cduet.edu.cn, xiey521@126.com

It simulates and displays the environment in a dynamic, omnidirectional and 3D approach and users can operate every object with an immersed sense. Since the environment is equipped with hearing, vision and other multimedia functions and the ability of real-time interaction, its data type is very complex and data volume is huge.

The traditional approach of scene construction and interaction display is based on a single and specific data type. Virtual environments created by this approach usually have a low fidelity and lack of interaction. HGML, as a universal standard, is a solution of massive data storage and is able to organize, manage and display data, as well as manage interactions between these data. Moreover, it can integrate data of varied formats, such as texture mapping data, remote sensing image data, DEM data, accurate geographic information data, etc. All the above features increase the fidelity and interaction of the virtual environment dramatically, thus realize visualization.

## 2 Basis of Theory

G/S (Geo-information Browser/Distributed Spatial Data Servers) is a new generation of Network connection mode based on the request-aggregation-service mechanism. G/S mode combines the advantage of both the C/S mode and B/S mode (C/S mode makes full use of client resources, processes client data efficiently. B/S mode has unified client, shows the data boundlessly, etc.), improves the complexity of information processing and the efficiency of information service in B/S mode, compensates the shortfall C/S and B/S mode owned in the spatial information network service application.



**Fig. 1** Data storage, visualization based on G/S mode.

G side processes application logics, and data processing logics are accomplished at S side. S side only optimizes the data management and does not refer to data computation, which improves the efficiency of the servers.

HGML, the core of G/S mode, is capable of conveniently connecting together as a whole the data stored on the distributed servers. There are several advantages of HGML: it enhances the functionality and efficiency of the G side, optimizes data management, cuts back network costs, is favorable of transmission of large-scale modeling data, facilitates data storage, sharing, visualization and analysis, and implements dynamic retrieval, 3D display and interoperability of spatial data, etc., as illustrated in FIG. 1.

The structure of HGML consists of GeoHead and GeoBody. GeoHead describes the information required by the browser and GeoBody comprises the specific information according to different requirements, such as data types, attributes, files and structures, etc

### 3 The Construction of Virtual Scene Based on HGML

Various types of data are organized uniformly and managed by HGML which is capable of constructing virtual reality environment.

#### A Data Analysis

Throughout the process of the construction of Virtual Reality scene and 3D visualization, there are two types of scene data. One is 3D geographical modeling data, another is 3D entity modeling data.

- 3D geographical modeling data is one of the important elements of the geographical environment, and is basis of the construction of Virtual Reality. This article uses DEM and Remote Sensing data to describe 3D geographical models.
- 3D entity modeling data is one of the important elements, it includes artificial and natural building models, etc.

**Fig. 2** HGML effectively organizes various types of data (“The Display Platform of Achievement about Digital City based G/S” in Beijing reconstruction projects in Shifang, shows the renderings of Virtual scene).



According to the analysis about spatial data mentioned above, the fidelity of the scene depends on realistic model and accurate location. With the development of G/S, HGML can organize and exchange large-scale modeling data, and define precise location of 3D model, at the same time, describing 3D geographical model. Above features of HGML effectively accomplish the visualization.

### B The Construction of Virtual Scene

Virtual scene refers to the set of physical environment’s 3D models created using technology of the computer. This article shows flow chart of construction of scene through the analysis of spatial data, as illustrated in FIG. 3.

HGML effectively organizes different types of data, as standard follows:

```

<?xml version="1.0"encoding="UTF-8"?>
<HGML xmlns="http://earth.UStar.com/HGML/1.0">
<Document>
  <Placemark>
    <name>...</name>
    <LookAt>...</LookAt>//Positioning the lens to view objects
    <Model id=" ModleName ">
      <Link>//reference model
        <href>files/ModleName.dae</href>
      </Link>
      <Location>...</Location>//define precise location of model
      <Orientation>...</Orientation>//define rotation
      <Scale>...</Scale>//Degree of change along different direction
    </Placemark>
  </Document>
</HGML>

```

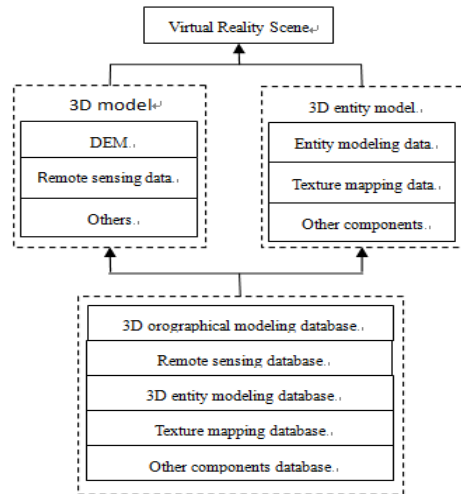


Fig. 3 The flow of 3D scene construction (3D models based on real image are organized together according to the logic relation.)

## 4 The Implementation of Interactions Based on HGML

At present virtual interaction approach is commonly used. This approach has bottle neck in complexity, real-time performance, stability and consistency, and cannot fulfill users' requirements any longer. Current prevalent virtual interaction approaches, in terms of large amount of computation and complex interaction referring high real-time performance, cannot meet the users' requirements, and increase the difficulties of implementing some complex interaction functions, thus stability and consistency cannot meet the users' requirements. With the support of G/S mode, above problems can be solved effectively.

Virtual interaction has two aspects, dynamic characteristics and real-time interaction. Dynamic characteristics comprise wander operation, dynamic landmark, etc.

### A *Roaming*

Roaming refers to the process that the spatial location changes as the viewpoint changes. In roaming, the document of HGML can include any legal elements of HGML, it can be achieved by embedded a specific spatial data into the HGML file. HGML customizes several key elements for roaming.

- `<gx:FlyTo>`: it defines view location and rotation, nested `AbstractView` (`<Camera>` or `<LookAt>`) and `<gx:duration>` and `<gx:flyToMode>` and so on.
- `<gx:SoundCue>`: it defines how to play stereo sound, nested `<gx:playMode>`, enhance the realism.
- `<gx:balloonVisibility>`: it defines the Bubble Box used to sauxiliary explanation about scene and environment, any time or any places.

### B *Dynamic Landmark*

Dynamic landmark refers to the models motion that the motion state associates with time properties (it is often used as the background scene). This technology can be used for GPS tracking, and produce animation effects, such as falling leaves, surging grass, flying aircraft, driving boat, driving car (as illustrated in FIG. 4), etc.

It is associated with time data that all landmarks or 3D models in HGML file, visibility of the corresponding data sets was limited in the period of time or an exact time, then programmers can do Linear interpolation and show the corresponding data.

**Fig. 4** Dynamic landmark

HGML customs several key elements for dynamic landmark, as `<TimePrimitive>` (be extended by the `<TimeStamp>`), `<Placemark>` and `<Model>` and so on.

`<TimeStamp>`: it specifies exact time for necessary landmarks or 3D models.

`<Model>`: it defines model size and proportion and so on, nested several key elements to reference landmarks and specify latitude and longitude and so on. For example, `<Link>`, `<altitudeMode>`, `<gx:altitudeMode>` and `<Location>` and `<Orientation>`.

`<Placemark>`: it defines the location of models, and nested `<TimeStamp>` and `<Model>` and another elements, it can show needed data in different time.

### C Real-Time Interactivity

Real time and interactive is very important in the process of virtual display. With the support of G/S mode, and the loose-coupled distributed data, data retrieval and update are undertaken by various application servers. Not only does data maintenance become more efficient and accessible, but also protects complexity of data and real-time.

HGML integrates other types of data to achieve virtual interactive, for example, JavaScript and Java3D and others, and satisfies the real-time, that according to demand immediately randomly access to information, making the virtual exhibition is more diversity and initiative.

HGML customs several key elements for various types of data, `<link>` is the most important element, it is used to locate the HGML file from Internet and others, it embeds time-based `<refreshMode>` and `<refreshInterval>`. Besides, it also embeds `<viewRefreshMode>` and `<viewRefreshTime>` based on the current lens, these elements are used to load and refresh HGML file according to requirements, and complete real-time interactivity.

## 5 Conclusion

With the support of G/S mode, this article formulated the construction of virtual reality environment based on HGML and the implementation of real-time interaction. The implementation has several features:

- Implemented the client-side aggregation service that the parsing and display of the environment are both processed in G side.
- Environment construction could base on multiple data types, such as 3D terrain, 3D entity models, etc. Virtual reality environments created by this approach not only have a high fidelity, but also capable of serving our life, for example, LBS (Location Based Services) in the Internet of Things.
- The distributed servers of G/S mode with HGML at the core, manage the source data in a distributed approach, and it is a good solution of the bottle neck in the transmission of massive spatial models over the Internet. Therefore, it fulfills the requirements of real-time interaction.

With the development of G/S mode and improvement of HGML, the application domains in which the core technology mentioned in this article can be used will become increasingly extensive.

**Acknowledgment.** The authors wish to acknowledge with appreciation the financial support (Grant No. NSFC 61071121) provided by the National Natural Science Foundation of China.

## References

- [1] Miao, F., Ye, C.-M., Liu, R.: Discussion on digital earth platform and the technology architecture of digital China. *Science of Surveying and Mapping* 32, 157–158 (2007)
- [2] Xiao, Y.U.: Spatial information management —— research of G / S mode (unpublished)
- [3] Guo, X., Miao, F., Wang, H., Liu, R., Wu, Y.: The Research on Digital Tourism Services Platform Based on G/S model Architecture. *Remote Sensing Technology and Application* 24(4) (August 2009)
- [4] Li, K.-R., Miao, F.: The analysis of multi-dimension spatial data subdivision and storage. In: *ESIAT 2009*, vol. 3, pp. 385–388. IEEE Computer Society CPS (2009)
- [5] Chen, F., Fang, M., Wang, W., Wang, H.: Design and implementation of ggearth spatial data service application system. In: *Proceedings of SPIE – The International Society for Optical Engineering, Second International Conference on Earth Observation for Global Changes*, vol. 7471 (2009)

# The Next-Generation Search Engine: Challenges and Key Technologies

Bo Lang, Xianglong Liu, and Wei Li

**Abstract.** This paper analyzes limitations and challenges of state-of-the-art internet search engines, points out the features and key technologies of next-generation search engines. The solutions of relevant key technologies are proposed and discussed, including a new search engine architecture, a multidimensional data model, the distributed data storage and parallel processing techniques, and intelligent service oriented query language. Finally an unstructured data retrieval system is established based on the above ideas.

## 1 Introduction

With the quick development of Internet applications, data on Internet has new features. In addition to web pages, there exist a large amount of images, audios, videos, and data from social networks, deep web, and mobile Internet. Also, the rate of data growth is accelerated and it follows the Moore's Law: doubling every 18 months, of which about 70% are emails, photos, videos, and data of social networks. According to a study of IDC and EMC, in 2011 1,800 EB (1EB = 1,000 PB) digital information will be produced, and the amount of information will increase tenfold from 2005 to 2011 [1].

The rapid growth of the amount of data on Internet has brought new problems to massive information capture, analysis, organization, and retrieval. In the past decades, search engines developed quickly from the first-generation which are based on the earliest directory services, web information display and browsing, to the current second-generation search engines of web-based link analysis and keyword matching [2]. However, nowadays mainstream search engines have shown significant limitations in massive heterogeneous information search. We believe that only search engines that expand the search scope and depth and at the same time improve search accuracy and speed can provide users with high-quality

---

Bo Lang · Xianglong Liu · Wei Li

State Key Lab of Software Development Environment, Beihang University, Beijing, China



search services. Search engines with such abilities, compared with the traditional search engines, need to have several innovations and changes in concept and techniques, and therefore can be called next-generation search engines.

This paper firstly analyzes the challenges of search engines, and then discusses the features and related technologies for the next-generation search engines, including a new architecture and a tetrahedral data model.

## 2 Challenges of the Next-Generation Search Engine

The objective of an Internet search engine is to quickly and accurately locate the information that users need among massive and heterogeneous data. However, to achieve this goal, the present search engines are facing the following challenges:

(1) Supporting multi-type inputs, such as natural language sentences and images, and being capable of understanding users' requirements and giving responds intelligently.

Imagine the situation that someone unfamiliar with Beihang wants to visit this university. He would like to find its geographical position and other related information through search engines. Using current search engines, what he can do is probably to enter "Beihang, map" and other keywords in the browser box. The search results will include a lot of irrelevant contents containing these keywords. If the user can search by entering sentences able to express the complete meaning [3], for example, "Where is Beihang", then the search results include not only a map indicating Beihang's location but also the public transport information, school profile, and other related information, which means that the accuracy of search results will be substantially increased and users' satisfaction will also be greatly enhanced.

Therefore, for many searches, users want to input request through natural language, a picture or other direct and intuitive means, which requires the search engine to possess abilities of intelligent understanding and information associated retrieval.

(2) Supporting accurate retrieval of massive unstructured data like images, videos and audios.

Currently, in addition to web pages on the Internet, there are a large number of audios, videos, images and other multimedia data. Existing search engines mostly retrieve them by keywords. Since audio, video, image and other multimedia data usually have complex structures and they are rich in content and semantics, simply using keywords cannot completely described them, and thus the accuracy of search results might be low. Therefore, how to realize the accurate retrieval of massive videos, images, audios and other unstructured data [10] becomes another challenge that search engines face.

(3) Supporting the blog, micro blog and other social networking data search, and support social cooperative search.

Many people view Internet as a universal expert system. When having problems, they often find the answers by searching Internet. Usually, for searching a correct answer, half an hour or even longer time is spent. In fact, micro blogs

and other social networks contain large amount of information and knowledge. If the search engine can search data of micro blogs and other social networking services (SNS) [7][9], or search with multiple user participations [5] to quickly find the most appropriate experts for answering the question, then users can get the best answers within a shorter time.

(4) Supporting real-time connections to a variety of platforms, such as mobile phones and vehicle navigation, and supporting the retrieval of real-time data.

Mobile terminals like smart phones and vehicle navigation systems can display real-time data, and they can also produce a lot of valuable information [8]. Users expect that mobile phones, navigation systems and other platforms can be seamlessly connected to the Internet, and information can be pushed to users anytime and anywhere.

Current search engines are mostly built on web link analysis and keyword matching. The next-generation search engines, which can comprehensively and accurately search SNS data, Deep Web data, and all kinds of multimedia data, will become the trend.

### **3 Architecture of the Next-Generation Search Engine**

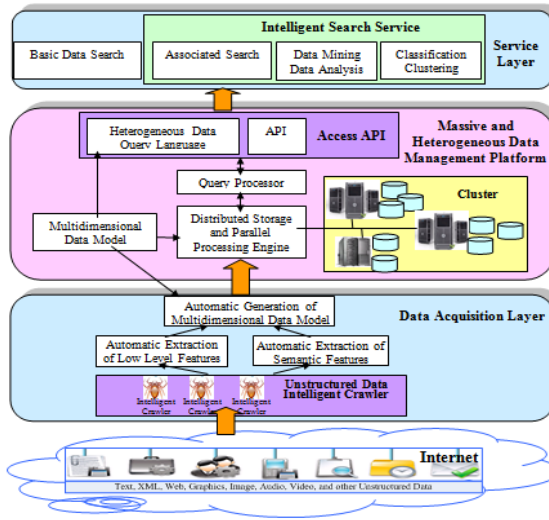
#### ***3.1 Key Technologies of the Next-Generation Search Engine***

In order to address the above challenges that traditional search engines face, the next-generation search engine should possess the following features:

- Providing intelligence services
- Supporting unstructured data
- Supporting social applications
- Integrating diverse platforms.

To have the above properties, the new generation search engine should break through a series of key technologies at levels of data acquisition, data storage and data services. In the data acquisition layer, it needs to expand the scope and depth of data acquisition. Data acquisition includes data capture and data processing: for data capture, through distributed intelligent crawlers, vast amounts of unstructured data can be captured in real time, deep web data can be collected and extracted, and the complex social network information can be gathered and analyzed. The captured data require multi-modal processing, which contains not only keywords extraction, but also multimedia high-dimensional feature extraction and semantic annotation [10][13].

New data modeling technique is needed for organizing and expressing the captured multi-modal data including raw data and their low level features, semantic features, and text descriptions [15]. With such modeling technique, search engines can achieve an integrated organization and expression for different kinds of unstructured data, making data of different modal as a whole, and thus support the upper complex data search services.



**Fig. 1** Architecture of the next-generation search engine

Distributed data storage and parallel processing [14] and effective index [10][12] are the two key technologies for speeding up data searching.

Humanization is the target at search engine service level. Through techniques such as dynamic topic-associated analysis of multi-source data [11][15], user behavior mining [4], pattern matching and data classification / clustering [10], and social network structure mining [5][9], search engines can provide data associated retrieval, personalized services, content-based retrieval, and relationships and message propagation model mining.

### 3.2 Architecture of the Next-Generation Search Engine

The traditional search engine is defined as the web information retrieval system. However, the next-generation search engine will be an Internet-scale information management system, which manages and retrieves web pages, SNS data and other unstructured data to achieve more effective Internet information storage and management and to provide accurate and efficient data search services.

The key of building the next-generation search engine is to establish a multidimensional data model with the capacity of multi-modal data representation, and to build an unstructured data management platform for massive data storage and parallel processing based on such a model. The next-generation search engine cannot be created relying on optimization and extension of the traditional search engines. New technologies are needed, and a variety of information science technologies, including database, multimedia data processing, and artificial intelligence, need to be introduced. Search engine companies who occupy the core technologies.

The next-generation of search engines will take the dominant positions in the market. Therefore, next-generation search engines have provided a great opportunity for the development of both information technology and enterprise.

In a traditional search engine, the main way of data acquisition is web crawling; data organization uses key word index; query processing and execution employ key word matching based on the index; the main way of user's query provided is keyword input. Data organization is the middle layer of the search engine architecture. Current data representation and description methods cannot support complex intelligence services like associated retrieval for the upper layer, and also limit the expansion of data sources of the lower layer. Therefore, the data organization becomes the core problem of traditional search engines, and the important breakthrough in establishing a new generation search engine should focus on a unified, associated and multidimensional data model.

Basing on such a model, the massive and heterogeneous data management platform could be build. The architecture of the next-generation search engine is shown in Fig. 1. In the massive and heterogeneous data management platform, multidimensional data model should be established first to achieve integrated representation of multi-modal unstructured data. Amounts of unstructured data can be stored and operated in parallel through the distributed storage and parallel processing engine. In data acquisition layer, the search engine crawls unstructured data using the distributed intelligent crawlers, extracts low level features and semantic features of the crawled data to automatically generate the multidimensional data model representations, and finally store them into the massive and heterogeneous data management platform. The massive and heterogeneous data management platform externally provides heterogeneous data query language or programming APIs for the search engine service layer to establish basic data search services based on keywords and intelligent search services including associated retrieval.

The key technologies to establish the massive and heterogeneous data management platform includes the tetrahedral data model, massive data storage and parallel distributed processing, unstructured data query language Advanced Query Language (AQL), the low level feature extraction techniques of image, video, audio and other data, pattern matching, and data clustering techniques.

Distributed data storage engine consists of a multi-node cluster [14]. The storage structure based on big table and the distributed file system makes the storage engine have a strong dynamic scalability, and the capability of storing vast amounts of data. Through effective parallel computing models such as MAP-REDUCE [14], multiple data nodes can conduct parallel operations on data, including parallel extraction of low level features, parallel matching, and parallel retrieval.

The query language oriented to intelligent computing services should have basic functions of data update and retrieval. Basic retrieval functions include search by basic attributes, search by semantic features, and search by low level features. In addition, the query language should support associated retrieval between different modal of data and semantic associated retrieval between

different data. It should also have intelligent search functions to support multidimensional data analysis, classification, and clustering. Moreover, it should have extensibility to allow users to customize data operations.

## 4 A Multidimensional Data Model – The Tetrahedral Data Model

Unstructured data, such as text, graphics, image, audio and video, has a non-uniform structure, and is stored as raw data. Therefore, it cannot be understood and processed directly by computers. In order to manage unstructured data, the fundamental approach is to describe the data and then to use the descriptive information to implement data operations. Keywords based on semantic descriptions, descriptions of low-level features, or concept-based semantic descriptions are presently used to describe unstructured data [15]. Unstructured data is composed of basic attributes, semantic features, low-level features and raw data, and there are relationships between the elements of these components:

- Basic attributes: All kinds of unstructured data have attributes such as name, type, author, and time of creation. However, it should be noted here that basic attributes do not include the semantics of the data.
- Semantic features: Special semantic properties are expressed using text, including the intention of the author, subject explanation, and the meaning of low-level features.
- Low-level features: Low-level features include properties of unstructured data acquired by using special data processing techniques, such as color, texture, and shape for images.
- Raw data: Raw data refers to the stored files of unstructured data.

The tetrahedral data model presented in this paper characterizes unstructured data based on these four aspects.

The tetrahedral model is composed of a vertex, four facets and the lines between the facets, as shown in Fig. 2. The vertex represents the unique identifier for the unstructured data; the bottom facet represents the raw data; the three side facets represent the basic attributes, the semantic features and the low-level features separately, and the lines connecting facets represent the associations between the elements on each facet.

**Definition 1.** The data structure of the tetrahedral model

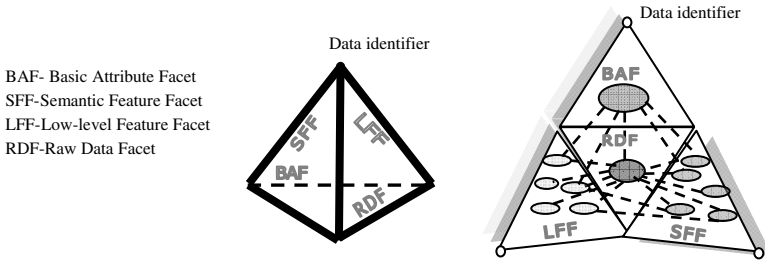
The tetrahedral model is composed of a vertex, four facets and the lines between the facets, and is described by a 6-tuple:

$$Tetrahedron=(V,BA\_FACET,SF\_FACET,LF\_FACET,D\_FACET,CONJS) \quad (1)$$

—V is the vertex of the tetrahedron, and is the intersection point of BA\_FACET, SF\_FACET and LF\_FACET. V uniquely identifies a tetrahedron;

—BA\_FACET (Basic Attribute Facet) is the facet that describes the basic attributes of the unstructured data. The point on BA\_FACET named Basic Attribute represents the set of basic information of the data, including type, time of creation, author, etc.:

$$BA\_FACET = \{Basic\_Attribute\} \tag{2}$$



**Fig. 2** The structure of tetrahedral model

—SF\_FACET (Semantic Feature Facet) is the semantic feature facet of the model. The point on this facet named Semantic\_Feature<sub>j</sub> represents semantic information available in text, such as the subject, the intention of the author, or the meaning of a data object or a low-level feature:

$$SF\_FACET = \{Semantic\_Feature_j \mid j \in [1, m]\} \tag{3}$$

m is a positive integer indicating the total number of the semantic features.

—LF\_FACET (Low-level Feature Facet) represents the low-level feature facet of the model. The point Low-level\_Feature<sub>k</sub> on the facet represents a feature that is obtained from the data using multimedia feature extraction techniques, such as audio frequency pitch for audios; color, texture, and shape of images; and key frame for videos:

$$LF\_FACET = \{Lowlevel\_Feature_k \mid k \in [1, n]\} \tag{4}$$

n is a positive integer indicating the total number of the low-level features.

—RD\_FACET (Data Facet) denotes the facet of raw data, a point Data<sub>l</sub> represents a raw data file:

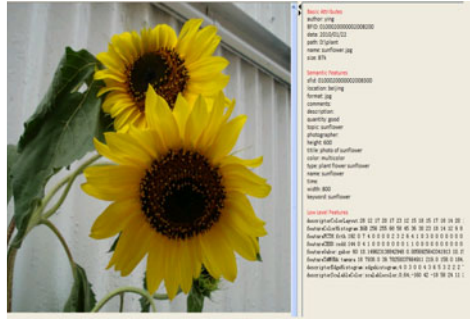
$$RD\_FACET = \{Data_l \mid l \in [1, p]\} \tag{5}$$

p is a positive integer indicating the total number of data files.

—CONJS (Conjunction) is a set that contains all the lines that connect different objects on facets. A line means that there is an association between the two end point objects:

$$\begin{aligned} CONJS = & \{BA\_FACET \times SF\_FACET \cup BA\_FACET \times LF\_FACET \\ & \cup BA\_FACET \times RD\_FACET \cup SF\_FACET \times LF\_FACET \\ & \cup SF\_FACET \times RD\_FACET \cup LF\_FACET \times RD\_FACET\} \end{aligned} \tag{6}$$

**Fig. 3** An example of the tetrahedral model of an image: the raw data on the left: the sunflower image; the basic attributes on the right, including the author, date, image name, image size; image semantic information, including the image format, subject, image quality, color, type, title, keywords, and so on; the low level features shown on the bottom right, including color histogram and texture.



### 5 Examples

We built a massive unstructured data management platform, with text, image, audio, video and other data management functions. All data in the system are described by the tetrahedral model. Fig. 3 is the tetrahedral model of an image.

The system has data retrieval functions based on low level features (namely content-based image retrieval), shown in Fig. 4. In Fig. 4(a), a tomato image is input as the retrieval sample. The low level features of the image are extracted and matched with the features of images stored in system. In order to facilitate users to find the desired image, the system provides the clustering function, which divides the results into 3 categories as shown in Fig. 4(b). Users can quickly locate the desired image in the category they are interested in.

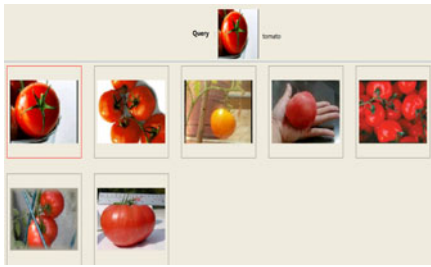
To achieve more precise retrieval, the associated retrieval combining low level features and semantic features is implemented in the system. For example shown in Fig. 5(a), the user gives the sample image and at the same time inputs the key word "tomato". The search results only contain tomato images similar to the sample image, which means that the search accuracy is largely improved.

There are often some relationships between data of different types such as images, text, video and audio. In this system, retrieval based on this relationship is implemented. For example in Fig. 5(b), users have searched the image of a hibiscus flower and they want to learn more about this kind of plant from some textual information. The system can find texts related to the image, like Word documents and web pages, through the relationship between the semantic features of the tetrahedrons in the system. Users can view the contents just by clicking on any text file, for example, a web page as shown in Fig. 5(b).



(a) Example of retrieval by low level features (b) Example of image clustering

**Fig. 4** Examples of retrieval by low level features and image clustering



(a) Associated retrieval for image



(b) Associated retrieval between image and text

**Fig. 5** Two Types of sociated retrieval

## 6 Conclusions

In this paper, with respect to the features of the next-generation search engine, an innovative scheme of the next-generation search engine is proposed, which mainly contains a new search engine architecture with massive and heterogeneous data management platform, and the tetrahedral data model for unstructured data. With the unified multi-modal data representation as a breakthrough, the scheme provides valuable ideas for building the next- generation search engines.

**Acknowledgments.** The work was supported by the Hi-Tech Research and Development Program of China under Grant No.2007AA010301, the Foundation of the State Key Laboratory of Software Development Environment under Grant No.SKLSDE-2009ZX-06, and the National Important Research Plan of Infrastructure Software under Grant No.2010ZX01042-002-001-00.

## References

1. Gantz, J.F., Chute, C., Manfrediz, A., et al.: The Diverse and Exploding Digital Universe - An Updated Forecast of Worldwide Information Growth Through 2011. An IDC White Paper sponsored by EMC (2008)
2. Google, <http://www.google.com>
3. Box Computing <http://boxcomputing.baidu.com/>
4. Yi, X., Raghavan, H., Legetter, C.: Discovering users' specific geo intention in web search. In: Proceedings of the 18th International Conference on World Wide Web, Madrid, Spain (2009)
5. Bao, S., Xue, G., Wu, X., Yu, Y., et al.: Optimizing web search using social annotations. In: Proceedings of the 16th International Conference on World Wide Web, Banff, Alberta, Canada (2007)
6. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vision* 60, 91–110 (2004)
7. Facebook, <http://www.facebook.com>
8. Technology Review, Special Report: Next-Generation Search (2010)



9. Gou, L., Chen, H., Kim, J., Zhang, X., Giles, C.L.: SNDocRank: a social network-based video search ranking framework. In: MIR 2010, pp. 367–376. ACM, New York (2010)
10. Nister, D., Stewenius, H.: Scalable Recognition with a Vocabulary Tree. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2161–2168 (2006)
11. Ngo, C., Ma, Y., Zhang, H.: Automatic Video Summarization by Graph Modeling. In: ICCV, Washington, DC, p. 104 (2003)
12. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-Sensitive Hashing Scheme Based on  $p$ -Stable Distributions. In: Proceedings of the Symposium on Computational Geometry (2004)
13. Chang, S.-F., Sikora, T., Puri, A.: Overview of the mpeg-7 standard. IEEE Transactions on Circuits and Systems for Video Technology 11(6), 688–695 (2001)
14. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM 51(1), 107–113 (2008)
15. Li, W., Lang, B.: A tetrahedral data model for unstructured data management. Science China Information Sciences 53(8), 1497–1510 (2010)

# An Efficient $k$ -Anonymization Algorithm with Low Information Loss

Md. Nurul Huda, Shigeki Yamada, and Noboru Sonehara

**Abstract.** Publishing microdata with preserving privacy led to the paradigms of  $k$ -anonymity. Sensitive information in  $k$ -anonymous microdata cannot be linked to specific individuals with a confidence value of more than  $1/k$ . However,  $k$ -anonymous data loses its importance with loss of precision or information contained in the data. Existing  $k$ -anonymization approaches suffer from high information loss. In this paper, we present a heuristic  $k$ -anonymization algorithm that results in very low information loss compared to existing similar algorithms. Also, the average case complexity of our algorithm is not high. Experimental results show that the information loss in our algorithm is significantly lower than that of the current state-of-the-art algorithm for  $k$ -anonymization.

## 1 Introduction

Microdata containing sensitive information (e.g., disease names) and demographic information (e.g., Age, Sex) can be very useful for research to discover important relationships between demographic properties and sensitive information (e.g., tendency of certain disease for Sex, Age, Location etc.). However, sharing of the data with third parties (e.g., researchers) may cause privacy threats [3][9]. To address these threats, [9] and [8] propose the  $k$ -anonymity model: for every record in a released table there should be at least  $(k - 1)$  other records identical to it along the quasi identifier (QID) attributes.

Existing algorithms for  $k$ -anonymity [6][10][11] result in high information loss and make the data less useful to the third parties. In this paper, we first present a simple greedy algorithm, called **Greedy**, that extracts the largest group of records

---

Md. Nurul Huda · Shigeki Yamada · Noboru Sonehara

National Institute of Informatics, Tokyo, Japan

e-mail: [huda, shigeki, sonehara}@nii.ac.jp](mailto:{huda, shigeki, sonehara}@nii.ac.jp)

of size  $\geq k$  into the anonymous data set based on a condition *clause* which is created by including each attribute of the QID in an order, and causes minimum information loss when put into one equivalent class. The order in which the attributes are included in the clause depends upon cardinalities of the attributes,  $\zeta(A_i)$ . Record extraction stops when the number of remaining records becomes  $< k$ .

Analysis of the Greedy algorithm shows that in most of the cases, the attribute with the minimum cardinality is included in the condition first. Based on this observation, we have developed an efficient heuristic algorithm, called **LowCost**, where the attribute having the minimum attribute cardinality is included into the clause before higher cardinality attributes. Experimental results show that the LowCost algorithm clearly outperforms the Greedy algorithm and the existing state-of-the-art algorithm, Mondrian, in terms of information loss.

## 2 Background and Related Works

Two commonly employed techniques to preserve privacy are suppression and generalization [9]. Suppression excludes some QID attributes or entire records from the microdata, altogether. In generalization a particular detailed value is mapped to a generalized value. Between the two recoding techniques of generalization *local recoding* is more flexible than *global recoding* and has the potential to achieve lower information loss [6].

In [7] the authors have proved that optimal  $k$ -anonymity for multi-dimensional QID is NP-hard, under both the generalization and suppression models. The authors in [1] proposes an optimal algorithm for single-dimensional recoding with respect to the DM metric. Incognito [5] finds an optimal solution using dynamic programming (for global, single-dimensional recoding) by considering all possible generalizations. The current state-of-the-art  $k$ -anonymity algorithm, Mondrian [6], partitions at medians across the dimension with the widest normalized range of values so that approximately half the items fall in each subgroup. A recursive branch halts when a group cannot be divided anymore due to the limitation on group size ( $G_i \geq k$ ).

Data quality of the anonymous data can be measured by the Discernibility Metric (DM) [1] though it does not capture the distribution of records in the QID space. More accurate is the Generalized Loss Metric [4] and the similar Normalized Certainty Penalty (NCP) [11] that are defined for numerical attributes and categorical attributes separately. For space limitation, we skip their descriptions here.

## 3 Proposed Algorithm

In our framework each of the data items can be represented with a numerical value so that they become ordered. The data set is considered sorted by attributes.

### 3.1 Greedy Algorithm

Our proposed Greedy heuristic searches in each of the candidate attributes (the QID) for the largest group of size  $\geq k$  that, if put into one equivalence class, will suffer from minimum information loss on the attribute. Among all of the above groups (one for each of the candidate attributes), the algorithm then takes the group and the respective attribute for which the information loss is minimum across the QID and creates a clause for that attribute (e.g., Sex = ‘Male’) which will return the chosen group. In the next step, it performs the operations described above, but on the selected data set in the previous steps, to include another attribute in the clause from the remaining attributes of the QID. It repeats the process of including an attribute in the clause until all of the attributes are included. Finally, the group that results from the constructed clause is put into an equivalence class through local recoding, and is removed from the original data source. The operation of extracting a group of records (of size  $\geq k$ ) is repeated until the number of remaining records in the original table is less than  $k$ .

By analyzing our Greedy algorithm it was found that the attributes get included in the condition based on their cardinalities. From this observation, we have developed the LowCost heuristics algorithm that performs better than the Greedy algorithm.

### 3.2 LowCost Algorithm

The largest group of records that, if put into one equivalence class, suffers from minimum information loss will be at the attribute with the smallest cardinality. Thus, we create the clause by including the attributes in the order of their cardinalities from the smallest to the largest. The rest of the process is similar to the Greedy algorithm.

For a lower computational cost we also follow some additional heuristics. When the cardinality of an attribute  $A_i$ , expressed as  $\zeta(A_i)$ , in the source data is low (i.e.,  $(n/\zeta(A_i)) \geq k$ ), where  $n$  is the number of records; instead of searching in the whole data set for the largest group on the attribute for the minimum cost, we check the group size of the largest group having a distinct value in the attribute. If the group size  $G_i \geq k$  then a condition can be constructed with the distinct value without further checking. Otherwise, a sequential search by groups of  $k$  records is necessary on the attributes. The above heuristics saves computation time, because if  $(n/\zeta(A_i)) \geq k$ , there will be at least one distinct value on that attribute whose group size on the attribute is  $\geq k$ . Fig.1 presents a pseudo code of the LowCost algorithm.

**Algorithm Complexity:** The average group size for a specific value at an attribute  $A_i$  would be  $n/\zeta(A_i)$ . If  $n/\zeta(A_i) \geq k$  then the searching cost on attribute  $A_i$  will be 1 (largest group at the top). Thus, the search complexity reduces with a decrease in  $\zeta(A_i)$ . If  $n/\zeta(A_i) < k$ , then a sequential search is necessary and the time complexity for searching in the first attribute would be  $(n - k)$ . Search space for the  $2^{nd}$  attribute depends upon the number of records selected by the clause that includes the  $1^{st}$  attribute (cardinality of  $1^{st}$  attribute in the clause). At most  $(n - k)$  comparisons are necessary in the  $2^{nd}$  attribute. In the worst case, total  $d \times ((n - k) + (n - 2k) + (n - (n/k - 1) \times k)) = d \times (n/k) \times (n - (k \times (k + 1)/2))$  comparisons are necessary.

---

```

LowCost(input T) {
  S = T;  $\hat{T}$  = null; n = record_count(T); d = QID_size; A[] = QID;
  while n > k do {
    condition = "";
    sort A[] on |Aj| Asc, where 1 ≤ j ≤ d;
    for j = 1 to d {
      find the largest group Gi in S of size ≥ k so that Gen_Cost_to_one_Eqv_Class(Gi,j) is minimum;
      condition = condition + "Aj between Gi,j(min_value) and Gi,j(max_value)";
      S = "Select * from S where" + condition;
    }
     $\hat{T}$  =  $\hat{T}$  + S;
    T = T - S;
    n = record_count(T); S = T;
  }
  output ( $\hat{T}$ )
}

```

---

**Fig. 1** The LowCost Algorithm for low information loss.

Thus, the worst case complexity of the algorithm is of order  $O(n^2)$ . However, in real data, most of the attribute cardinalities are not high and thus  $n/\zeta(A_i) \geq k$ , e.g.,  $\zeta(\text{Sex}) = 2$ ,  $\zeta(\text{Age}) \cong 100$  and so on. So, the average case computation time will not be high.

## 4 Experimental Evaluation

We compare the information loss in Greedy, LowCost and Mondrian algorithm in both NCP and DM metrics. We considered a three-attribute QID: Date of birth, Sex and Zip code with their domain cardinality of 3653((10 years range)), 2 and 1000 (3 digits) respectively. Random data were generated for the attributes.

The NCP values presented in the graphs are the average per data unit. The NCP cost of a suppressed data item is considered to be equal to 1. Fig. 2(a) and 2(b) compare the NCP for 2500 records and 500 records respectively for varying anonymity parameter  $k$ . For large number of records (Fig. 2(a)), with an increase in  $k$ , NCP in both Mondrian and the Greedy algorithm increases while it remains steady in LowCost. For small number of records (Fig. 2(b)), NCP increases in all three algorithms. The rate of increase is highest in Mondrian followed by Greedy. We see that the NCP cost in LowCost is significantly lower than those in other algorithms for all cases. The high cost in Mondrian is due to the fact that it starts with the largest and most generalized group and then splits the group if the splitting satisfies  $k$ -anonymity. However, in our algorithm we start with the most specific value and find an equivalence class of size  $\geq k$ . Thus, the values are generalized less in our algorithm.

Fig. 3(a) and 3(b) compares DM cost of the three algorithms for 2500 records and 500 records respectively for varying value of  $k$ . All the three algorithms' costs do not change for change in the number of records. A lower value in DM is desirable and the ideal DM value is equal to  $k$ . From the figures, we can see that LowCost and Greedy algorithms outperform Mondrian in DM and the DM values are equal to the minimum/ideal in both LowCost and Greedy algorithms.

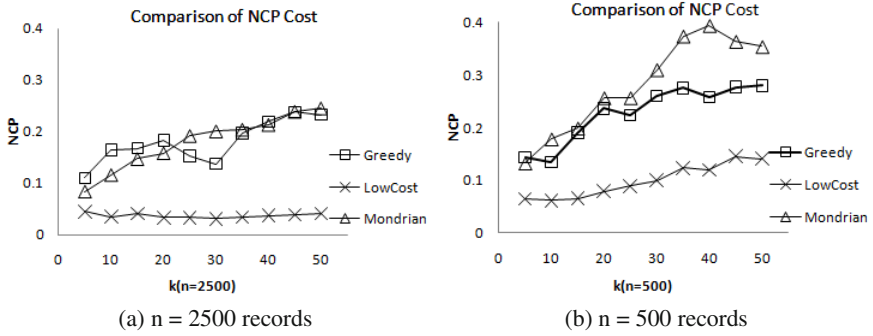


Fig. 2 Comparison of NCP cost for varying  $k$

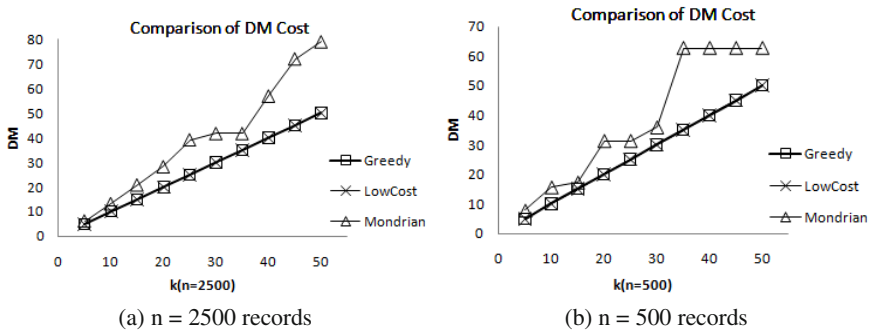


Fig. 3 Comparison of DM cost for varying  $k$

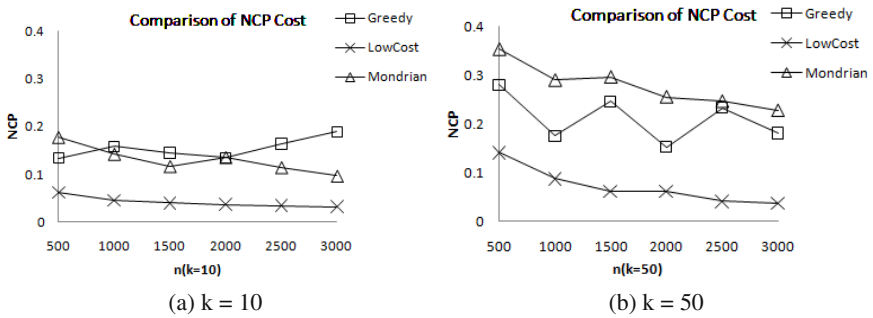


Fig. 4 Comparison of NCP cost for varying  $n$ .

Fig. 4(a) and 4(b) compare the NCP cost for  $k = 10$  and  $k = 50$  respectively for varying number of records,  $n$ . With an increase in  $n$ , NCP cost decreases slowly in all algorithms. For a small value of  $k$  (Fig. 4(a)), Mondrian performs slightly better than the Greedy algorithm. However, for a large value of  $k$  (Fig. 4(a)) the Greedy algorithm performs better than Mondrian. Please note that the graphs show NCP

cost per data unit. A small difference in information loss per data unit will make a large difference on the information loss for the whole data set. We clearly observe significantly lower information loss in the LowCost algorithm compared to the other two algorithms for all settings.

## 5 Conclusion

In this paper, we have presented an efficient heuristic algorithm, called LowCost algorithm, for  $k$ -anonymization that results in very low information loss compared to other well-known algorithms. We used computer generated random sample data and measured the information loss in two metrics (NCP and DM). In both metrics, LowCost algorithm's information loss is much lower than the others. For a small value of  $k$ , Mondrian and the Greedy algorithm have similar NCP cost, but for a large value of  $k$ , the Greedy algorithm outperforms Mondrian. Though the worst case complexity of our algorithm is of order  $O(n^2)$ , the average case complexity is much lower than that. Its complexity decreases sharply with decrease in attribute cardinality. Practical data are found to have very low cardinalities on their attributes. Therefore, the LowCost algorithm can support robustness with practical data.

## References

1. Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., Zhu, A.: Approximation algorithms for  $k$ -anonymity. *Journal of Privacy Technology* (2005)
2. Cormode, G., Li, N., Li, T., Srivastava, D.: Minimizing minimality and maximizing utility: Analyzing method-based attacks on anonymized data. In: *Proc. VLDB*, vol. 3, pp. 1045–1056 (2010)
3. Froomkin, A.: The Death of Privacy. *Stanford Law Review* 52, 1461–1543 (2000)
4. Iyengar, V.: Transforming data to satisfy privacy constraints. In: *Proc. Int. Conf. on Knowl. Discovery and Data Mining, KDD* (2002)
5. LeFevre, K., DeWitt, D., Ramakrishnan, R.: Incognito: efficient full-domain  $k$ -anonymity. In: *Proc. SIGMOD*, pp. 49–60 (2005)
6. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Mondrian multidimensional  $k$ -anonymity. In: *Proc. IEEE Int. Conf. on Data Eng. (ICDE)*, p. 25 (2006)
7. Meyerson, A., Williams, R.: On the complexity of optimal  $k$ -anonymity. In: *Proc. ACM SIGMOD-SIGACT-SIGART Symposium, PODS* (2004)
8. Samarati, P., Sweeney, L.: Generalizing Data to Provide Anonymity when Disclosing Information. In: *Proc. ACM SIGMOD-SIGACT-SIGART Symposium, PODS* (1998)
9. Sweeney, L.:  $k$ -Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 557–570 (2002)
10. Xiao, X., Tao, Y.: Anatomy: Simple and Effective Privacy Preservation. In: *Proc. of VLDB*, pp. 139–150 (2006)
11. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.: Utility-Based Anonymization Using Local Recoding. In: *Proc. Int. Conf. on Knowl. Discovery and Data Mining (KDD)*, pp. 785–790 (2006)

# Design and Realization of Data Gateway for Large-Scale Alternately Supervisory System

Yong Bai, Zhangli Lan, and Yi Zhou

**Abstract.** This paper discusses the research actuality and developing trend of alternately supervisory system. Combining with the characteristic of the system, a pattern of financial computer system data gateway is proposed. It uses the specific message mode to carry on data acquisition with hierarchical and modulation design conception. Application results indicate the pattern can improve system reusability, shorten the development cycle and reduce the maintenance costs.

## 1 Introduction

Considering the financial business system's stability and treating efficiency, present financial alternately supervisory system generally bases on UNIX platform. Meanwhile, based on the requirements of management and service, financial business system will integrate in regions or even in whole country. In order to master system's operation conditions and give a better service to customers, a well-designed and powerful alternately supervisory system becomes an inevitable trend. Mainly in the following aspects:

(1) For developers: at the system trial operation stage, developers can detect problems through the transaction surveillance, then locate and solve the problems by detailed transaction logs, make the system access to stable stage quickly.

(2) For maintainers: at the system stable operation stage, as a normal operation.

(3) For business: by supervising chronological files, can acquire customer's consulting about the reason of the transactions failure in real-time, and improve financial services' level.

---

Yong Bai  
Chongqing Electric Power college, Chongqing, China  
e-mail: byong@163.com

Zhangli Lan · Yi Zhou  
Chongqing Jiaotong University, Chongqing, China

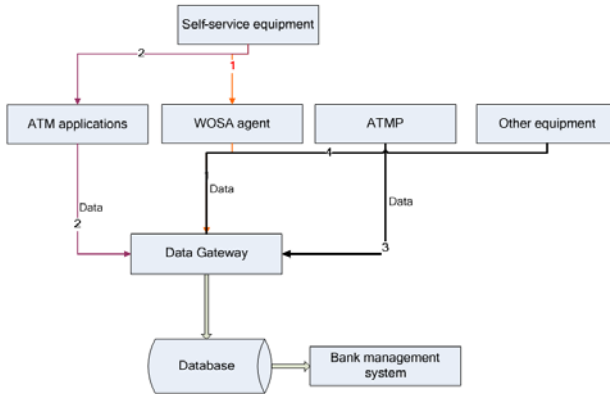


The existing financial alternately supervisory system mainly adopt TCP/IP protocol and Socket programming, transfer data directly from supervising server-side to client-side. As the market demand, financial operations continue to expand, the developing trend is integrating various business’s monitor in a large-scale interactive monitoring system. However, the existing pattern has deficiencies in scalability, compatibility, data integrity, system interoperability.

According to the demands of the large-scale interactive monitoring system, this paper proposes a design pattern of financial computer system data gateway combining with the characteristics of bank’s alternately supervisory system. This pattern uses the specific message mode to carry on the data acquisition with hierarchical and modulation design conception.

## 2 Structure of Alternately Supervisory System

The structure of proposed system uses data gateway that is the date exchange middleware based on service. It provides date service and application integration between heterogeneous systems. It can shield the difference of application protocols and realize the data exchange. The proposed system uses three-layer structure which can be seen in fig.1. The first layer is the client (ATM and other equipments); the second layer is the gateway server which realizes communication and data exchange between heterogeneous systems; the third layer is the data server which is in charge of data management, storage and visit.



**Fig. 1** The structure of alternately supervisory system.

The data gateway process has functions as follows:

(1) Considering the independence of each application system and the expansibility of platform, a unified interface specification can be realized in order to keep that data exchange can be performed between the monitoring system platform and every application systems of banks. Date can be exchanged easily by the unified interface specification between systems even they are heterogeneous.

(2) Splitting the data collection and analysis can share the data collection pressure of monitoring systems and improve the efficiency and security.

### 3 Design and Realization of Data Gateway Module

#### A Data Acquisition

As shown in fig.1, There are four data transmission lines from the bank equipment to the data gateway, in other words, the data gateway collects information through four ways:

The first line is through the WOSA (Windows Open System Architecture) agent program. The WOSA agent program visits the WOSA layer to gain information and transmits it to the data gateway.

The second line is through the ATM application program. ATM application program gets messages by the WOSA layer, and transmits messages to the data gateway. The difference between the first line and the second line is that the second method needs to transform the ATM application program according to the communication interface which is provides by data gateway.

The third line is through the ATMP (ATM prepositive computer). For some reasons, it impracticability by the first two methods, it can access through the ATMP channel. Because the current ATMP has the simple ATM management functions, according to a negotiation message protocol by the two sides, it can transmit the data which ATMP has collected to the data gateway, thus enhances the management and statistical functions.

The forth line refers to NON-WOSA standard equipment. These devices required the manufacturer in accordance with the gateway's data communication interface to collect messages and delivery.

#### B Data Gateway Mode

According to the WOSA standard interface or other custom interfaces, the client sends messages which in accordance with the prior agreement of the Message Protocol and assembled into 8583 packets (International Standard Protocol) to the data gateway. The packets are constructed into these messages using it as the sender's messages, and are sent into the message queue. Then creates a JAVA thread pool to analysis these messages, and creates a JDBC connection pool, finally, transfer data to the database. Data Gateway pattern shown in Fig.2.

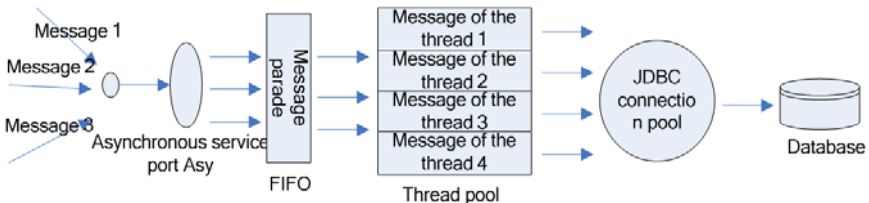


Fig. 2 Data gateway model

This Gateway model which designed in the paper has the following features:

(1) To reduce the pressure on the data gateway, the gateway of the communication part uses the producer-consumer model and the queue priority scheduling model to manage, and to achieve lower the concurrency quantity. As the result, the pressure on the server is reduced.

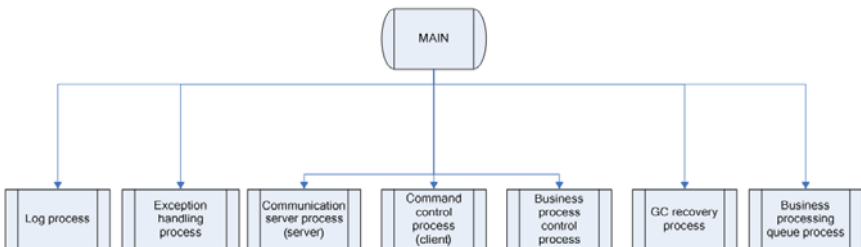
(2) Create Java thread pool in this system. According to the system's environmental condition, it can set the number of threads automatically or manually, to achieve the best effect. Using this thread pool, it can control the number of threads, and let other threads waiting in the line. When a task is finished, take it out of queue and then start executing the front of the queue. If the queue does not have the waiting process, the thread pool is waiting. When a new task needs to start executing, if the thread pool has the waiting threads, it can start, or go into the waiting queue.

(3) The database connection used the JDBC connection pool. Using this pool can improve database operation and ensure safety: reduce the creating link time, because the link can be reused in the pool; solve the multi-threading problems; connection pool can specify the number of maximum connections, so that each link can be achieved the most efficient use.

### C Data Gateway Module

Gateway server as an independent process in the system, through the socket exchanges data form the gateway client to the third party. According to the interactive system features, the bank can be designed for the following modules: communication module, message processing and flow control module, business component, log module, and abnormal handling module. Show in Fig.3.

Communication modules: It can only access TCP / IP communications. Using FIFO and thread pool technology to achieve the high throughput and the high concurrency. In the database, the different terminal custom configuration can be achieved through the access terminal registrant. After the connection access, receive the data stream, and stuff the data stream into the data processing queue.



**Fig. 3** Data Gateway Module

Message processing and flow control module: This module can be divided into 2 parts; one is process controller, the other is grammar parser. In process controller, grammar parsers can be directly called. Its driven mode is in accordance with the configuration's definition, by the analysis of data from left to right, to achieve its function. The grammar parser used JAVACC (Java Compiler) technology to achieve to the message of dynamic analysis and configuration.

Business component: When packet in accordance with the agreement resolves to the data element, it starts the next step processing, and returns data elements. In the process controller, business components are for the call by reflection.

Log module: log module is a text log, using asynchronous recording, with a separate thread responsible. It has the following characteristics: simultaneous operation of multiple log files, the file number can be configured; dynamically adding, remove and replace the log file; good multi-threading support.

Abnormal handling module: It can interrupt the system to debug state; can get abnormal code and address; can get the abnormal code in which the thread of the context of structure (ie register information); can quickly generate a abnormal detailed report, the function is convenient collecting the exception of information of the program for the author; can support the restoration exception for proceedings to continue running the state.

## 4 Conclusion

In order to compare the querying speed of two languages, we realized executing queries in both MDX and SQL. Table 1 shows the average execution times of different types of queries; compare MDX with SQL on runtime comparison.

A database of 60MB is tested, including 22 dimension tables, 8 fact sheets on CPU of Genuine Intel(R) T2050. Table 1 shows that MDX is faster and more efficient than SQL, and the grammar is more suitable for multidimensional operation, the only insufficient of MDX is it has to spend time and space in generating cube. As the database becomes larger, multi-dimensional inquires enables users to analyze data better. In the trend of financial analysis transferring from 2 to multi-dimensional, MDX language will get more extensive application.

The gateway model proposed in this paper has been applied in a company. The company's interactive monitoring and controlling system adopts three-layer construction model. Data exchange between 2 heterogeneous platforms realizes with the data gateway. The system's reusability is improved greatly. When financial business changed, its business logic, database and the client need not change. It uses the way of specific message to collect data assembled in 8583 and then transfers to the packet data gateway. So the data collection does not need to be analyzed too much, and the development cycle of this system would be shortened when a business increases. Thanks to model design concept, when new demands made by the customer are submitted, only the corresponding model need to be changed, and the maintenance costs would be reduced.

In short, application results indicate that this pattern can improve the system's reusability, shorten the development cycle and reduce the maintenance costs.

**Acknowledge.** This work is supported by Natural Science Foundation Project of CQ CSTC (2009BA6070), Scientific and Technological Project of CQ CSTC (2011BA6102), Science and Technology Research Project of Chongqing Education Commission (KJ110405), Science and Technology Achievement Project of Chongqing Education Commission, Construction Projects in CQ Science and Technology (City Section 2010 (94)), Open Foundation of Laboratory (CQ) 2011(CQSLBF-Y11-8), and Chongqing Transportation Commission.

## References

- [1] He, Z., Jin, O., He, J., et al.: Web-based self-service terminal Remote Monitoring System. *Computer Technology and Development* 16(4), 119–121 (2006) (in Chinese)
- [2] Young, C., Juang, W.L., Michael, J.D.: Real-time Intranet-controlled virtual instrument multiple-circuit power monitoring. *IEEE Transactions on Instrumentation and Measurement* 49(3), 579–584 (2000)
- [3] Shi, Y., Zhang, Y.: Financial business systems, real-time transaction monitoring to achieve universal. *Hei Longjiang Science and Technology Information* (12), 23 (2009) (in Chinese)
- [4] Zhang, L., Liu, Y., Chen, Y.: Data communication network management system design and implementation. *Computer Applications and Software* 25(1), 133–135 (2008) (in Chinese)
- [5] Qu, Y., Zhou, L., Li, X., et al.: Bank of computer information system security technical specifications. *Electronics Industry Press* (2001) (in Chinese)

# Research on Mass Geospatial Raster Data Processing Based on Map/Reduce Model

Fang Yin, Min Feng, and Jia Song

**Abstract.** With the development of Earth observation technologies, geospatial raster data such as DEM and remote sensing data experienced explosive growth in last forty years, providing TB or even PB amount of data to research projects. However, storing and processing the large amount of data can be very challenge to many projects, especially these without access to high performance computers. Distributed computing can integrate existed computing and storage resources through the Internet, providing an alternative way to facilitate these projects on handling the raster data. This study presents a distributed DEM processing approach developed based on the Map/Reduce, which has been widely adopted in cloud computing applications. The approach allows users to store and process the raster type DEM data in a distributed storage regime to utilize computing and storage capabilities combined from many average computers, e.g., PC. The approach has been implemented in a prototyped system, which is developed by utilizing the Apache Hadoop. The prototype has been deployed in the experimental environment, and then 90-m resolution DEM for China and a DEM hillshade model have been ingested into the prototype to evaluate the prototype.

## 1 Introduction

Geospatial raster data is playing an important role in Earth science research as well as social activities. However, the amount of remote sensing data acquired has

---

Yin Fang

College of Earth Science and Resources, Chang'an University, Xian, China

e-mail: yinf@chd.edu.cn

Feng Min · Song Jia

Institute of Geographical Sciences and Natural Resources Research,

Chinese Academy of Science, Beijing, China

e-mail: {fengm, songj}@lreis.ac.cn

increased greatly to GB, TB or even PB level with the rapid development of earth observation technologies, which is challenge to mass remote sensing data storage, management and distribution. Traditional method of image storing and processing cannot meet the needs of accessing remote sensing data quickly and accurately for relevant research. It is therefore, an important problem for scientists to manage and process massive geospatial image data, for the purpose of making use of current remote sensing data for further study.

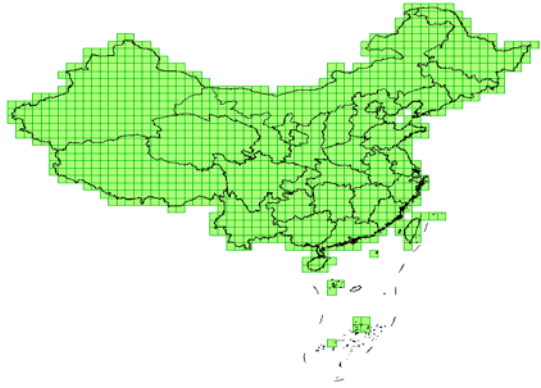
With the rapid development of web and computer technology, especially more and more universal applications of distributed computation, many research institutions worldwide carried out related research on mass data management and made significant progress. For example, NASA JPL developed RASCHAL system for global image mosaic (storage capacity of 40 TB, 15 TB used) [1]. Reference [2] presented GeoImageDB for managing multi-source image data. Moreover, some researchers use cluster high performance platform to improve the efficiency of in RS applications [3].

The cloud computing, together with cloud storage and Map/Reduce distributed computation model provide a new chance to store and process data with great capacity. Map/Reduce programming model requires expressing the solutions with two functions: map and reduce. A map function takes a key/value pair, executes some computation, and emits a set of intermediate key/value pairs as output. A reduce function merges all intermediate values associated with the same intermediate key, executes some computation on them, and emits the final output. More complex interactions can be achieved by pipelining several MapReduce compounds in a workflow fashion. Google sets a good example for managing 70.5 TB data in which image data account for 70 TB, using cloud technology comprised of GFS, Map/Reduce model and so on [4] [5] [6]. The following Apache Hadoop (including HDFS, Map/Reduce programming model, and etc) promotes greatly cloud computing applications [7] [8]. A data set is stored as a set of files in HDFS, which are in turn stored as a sequence of blocks (typically of 64MB in size) that are replicated on multiple nodes to provide fault-tolerance. The idea of map and reduction gave a new chance to storing and processing geospatial raster image.

## **2 Raster Data and Data Processing**

Each raster data file has spatial extent as a basic geospatial data type. Raster data files with different spatial coverage can be seen spatial splits of global scope. Spatial division can become one split method for distributed storage, if the size of single data file is suitable for study. Using DEM based on 30\*30m data spacing downloaded from the Global Land Cover Facility (GLCF) for an example (Figure1), China was partitioned to 1131 tiles by 1 in latitude and 1 in longitude, and with each tile size is 2.8MB. Therefore, DEM files in China can be stored in 1131 data blocks and the computing task be mapped into 1131 subtasks.

**Fig. 1** The China is mainly divided into 1131 tiles by 1 in latitude and 1 in longitude, and with each tile size is 2.8MB. Therefore, DEM files over China can be stored into 1131 data blocks and computing task be mapped into 1131 subtasks



### 3 Distributed Computing Framework

#### A *Distributed Storage Structure*

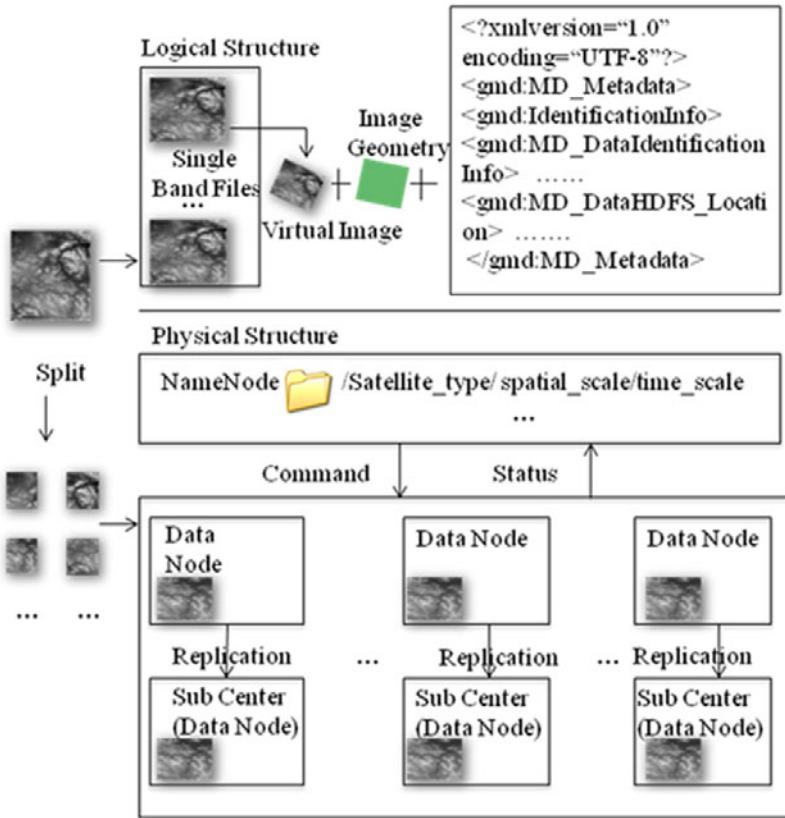
The storage framework has two parts: logical and physical structure (Fig.2). Logical structure mainly organized metadata of the data files to enable users to find images quickly.

Single band file was the basic management unit in our structure, which was the main difference from traditional metadata management methods of remote sensing raster images. Physical structure applied mapping and reducing strategy to data blocks in HDFS to enhance the distributed operation of massive raster data. Each band file has a unique id (abbreviated as *band\_id*), considering that current remote sensing raster data are composed of several bands with different wavelength range and spectral resolution [10]. The band files with the same spatial and time scale are organized into a virtual image (abbreviated by *image\_id*) through establishing association between *band\_id* and *image\_id* (Figure2).

Each virtual image has a metadata record comprised of spatial geometry (identified by Image Geometry in Fig.2), time property, and attribute information. Image Geometry is abstract on spatial information of virtual image expressed by geometric features, which consists of two parts: geometry feature and spatial reference information. Because each virtual image has a certain coverage area, the geometry feature chooses the polygon element to convey effective coverage area of virtual image. Therefore, Image Geometry can support complex feature polygon topology calculation with vector data.

Attribute information was described in XML files, considering that XML shows great flexibility for metadata expression. XML description document was designed following Geographic Information Metadata Standard, in order to accurately describe remote sensing data from multi-sources. The document contains bands information, image quality, cloud cover percentage and etc [10]. Moreover, the *MD\_DataHDFS\_Location* element was added to the XML document to represent the storage directory in HDFS.





**Fig. 2** Cloud Storage Framework

The basic principle of physical storage was data striping. The global scope was divided to different parts according to spatial extent of virtual images. Each part has a storage directory in HDFS and several subdirectories named by time. The band files with same time were organized in the same subdirectory were divided into same 256\*256 data blocks. The block was stored in distributed data nodes. A mapping file was created in <key1 (minx, miny, maxx, maxy), key2, value> records for all the blocks (Figure3), where key1 was spatial extent of block, key2 was the block number in data file, and value was the storage directory in HDFS. And then the task was mapped into subtasks combining geospatial scope and Map/Reduce model. It was noted that single image file would be chosen as a data block if spatial extent and data size was suitable.

### ***B Distributed MapReduce Computing***

The section discusses the MapReduce-based process of hillshade computation of DEM files. Let us start our description by defining the problem. Let G be the

spatial extent of study region composed of objects  $o_i, i=1, \dots, |G|$ . Each object  $o$  has two attributes  $\langle o.id, o.S \rangle$ , where  $o.id$  is the object's unique identifier and  $o.S$  is the object's spatial extent. The three main phases of the process are: (1) create  $\langle \text{key}/\text{value} \rangle$  configuration file for job in Hadoop, (2) implement mapper interface and extend `MapReduceBase` for running job, and (3) implement reducer process for merging job results.

(1) Create configuration file

The MapReduce  $\langle \text{key}/\text{value} \rangle$  input pairs were created according to study region. Each pair is  $\langle o.id, o.S \rangle$ , where  $o.S$  has the same value with `key1` in mapping file (Section III.A). Therefore, each Mapper is related to a data block. The configuration file had 1131  $\langle \text{key}/\text{value} \rangle$  input lines in our example. Each Mapper task would correlate with single DEM file.

(2) Map task implementation

Mapper interface was extended to implement hillshade model computing. Each Mapper applied `FSDDataInputStream` class to read DEM file in byte stream format, started a Java Process that executed `gdaldem hillshade` command to handle data stream, and then called `FSDDataOutputStream` to store the result.

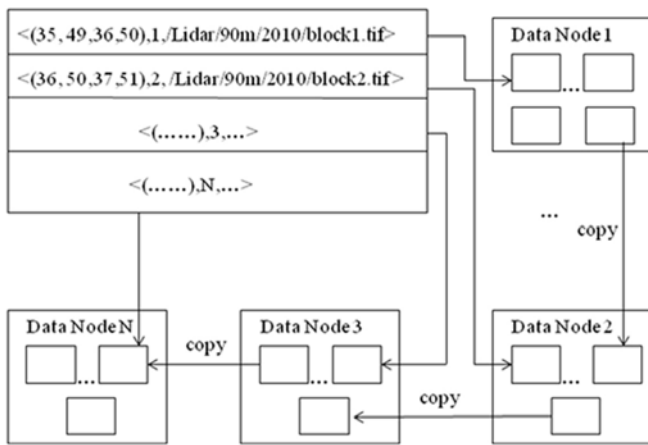


Fig. 3 Physical storage

(3) Reduce task implementation

Reducer interface was implemented to collect results. When all Mapper tasks were completed, Reducer executed `gdal_merge` command to merge the output files of Mappers.

## 4 Prototype System

### A Experimental Environment

Hadoop made it possible for constructing high performance platform using PC computers. Our experimental environment consisted of one main node (name node in Hadoop) and ten sub nodes (data nodes). All the nodes had same configuration such as 2.5 GHz T9300 CPU, 2GB memory, 150 GB hard disk and 100 Mb/s Ethernet card, Ubuntu10.0.4 operating system, and installed 1.6 JDK and HDFS of hadoop-0.20.2 with replication set to 3.

### B Applications

The Hadoop operation is in the way of command (bin/start-all.sh .e.g.), which made it inconvenient for the communication between the application and the Hadoop. REST services were developed to start Hadoop in SSH (Secure Shell), get the number of Hadoop nodes, access status and progress of Hadoop running job, and so on.

Web interface was designed to select study region flexibly by drawing a rectangle. Then, the REST services were called to interact with the Hadoop (Figure4).



Fig. 4 Web interface

## ***C Results Analysis***

The computing time of each image cost 10 seconds and 1131 images cost 3.14 hours in a PC.

Beijing region covering 4 images was tested and cost 21 seconds. The computation of each image increased 2 times in our cloud cluster based on ten PC nodes, 1132 images cost 1.5 hours.

If spatial resolution of images was 1m, the data size of China would reach 32.4TB, which could be store and process in our cloud cluster as long as increasing data nodes to enable enough space.

## **5 Conclusion**

The paper presented a practical framework for mass raster data storage and computing using latest HDFS and Map/Reduce model of Apache Hadoop. The framework shows great flexibility in actual deployment and operation. Anyone can set up his distributed computing platform using PCs. Our method achieved efficient integration of computing and storage resources, compared to other distributed environment. As the constant development of cloud computing, our framework was believed to achieve further progress in mass spatial data management. Web applications were developed to compute hillshade of DEM files efficiently, based on this solution. The applications had been practiced in Data-sharing Network of Earth System Science Project in IGSNRR and had been proved feasible for integrating geospatial data, models with computing resources.

The solution also has some shortcomings, e.g. split strategies used was not very mature for various raster data. Moreover, data storage with TB-level has not been completed because of capacity limit of hard disk in our computers. More and more nodes would be added to the cluster platform to solve the problems caused by TB-level data In future work.

The research is funded by China National Scientific Data Sharing Program of Ministry of Science and Technology of PRC: Earth System Scientific Data Sharing. Especially thanks to the leader of this research: Pro. SUN Jiulin and all of other program members.

## **References**

- [1] Hulen, H., Graf, O., Fitzgerald, K., Watson, R.: Storage Area Networks and the High Performance Storage System. In: Tenth NASA Goddard Conference on Mass Storage Systems (April 2002)
- [2] Wang, M., Gong, J.Y., Li, D.R.: Design and Implementation of Large-Scale Image Database Management System. Editorial Board of Geomatics and Information Science of Wuhan University 28, 294–300 (2003), doi: CNKI:SUN:WHCH.0.2003-03-007

- [3] Jin, H.Z.: Research and Design of Cluster-Based Job Management System Oriented to Spatial Data Processing. Wuhan University (May 2005), doi: CNKI:CDMD:2.2006.034413
- [4] Ghemawat, S., Gobiuff, H., Shun-Tak, L.: The Google File System. In: SOSP 2003: Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, vol. 37 (December 2003), doi:10.1145/945445.945450
- [5] Barroso, L.A., Dean, J., Holzle, U.: Web search for a planet: The Google cluster architecture: The Google Cluster Architecture. *IEEE Computer Society* 23, 22–28 (2003), doi:10.1109/MM.2003.1196112
- [6] Lammel, R.: Google's MapReduce programming model—Revisited. *Science of Computer Programming* 70, 1–30 (2008), doi:10.1016/j.scico.2007.07.001
- [7] [http://hadoop.apache.org/common/docs/r0.20.0/hdfs\\_user\\_guide.pdf](http://hadoop.apache.org/common/docs/r0.20.0/hdfs_user_guide.pdf)
- [8] [http://hadoop.apache.org/common/docs/r0.20.0/hdfs\\_user\\_guide.pdf](http://hadoop.apache.org/common/docs/r0.20.0/hdfs_user_guide.pdf)
- [9] Wu, L.: Principles, Methods and Applications of Geographic Information Systems. Science Press, Beijing (2001)
- [10] Feng, M., Zhu, Y.Q., Zhang, M.Z., Zhao, H., Yu, M.L.: Design and Realization of Multi-source Remote Sensing Images Sharing Platform, vol. 10, pp. 102–108, doi: CNKI:SUN:DQXX.0.2008-01-018

# A Parallel-Enabled Frequency Offset Estimation Scheme for Optical Coherent Receivers

Zhi-Yu Li, Xiao-Bo Meng\*, Xian Zhou, Xue Chen, Yang-Yang Fan, and Hai Zhu

**Abstract.** PADE, a digital frequency offset (FO) estimator for optical coherent receivers, is accurate and has a wide stable-running estimation range. But its operation requires making estimation in one symbol period. For very high-speed demand of optical transmission system, this would be difficult to realize with the existing hardware. In order to solve the problem, we propose an FO estimator scheme in which PADE can be operated group by group, where a group stands for a group of consecutive symbols. This grouping mechanism would make parallel operation possible for FO estimation, and reduce the hardware speed requirement. At the same time, the average calculation complexity for each symbol becomes much lower. Considering of practice, we give an initialization scheme for the new estimator. The simulation results show that, with feasible group length, the new FO estimator is effective and have the same performance as PADE.

## 1 Introduction

Optical coherent Polarization Multiplexed (PM-) QPSK transmission has become a promising scheme due to its high spectral efficiency for high capacity, and the facility to compensate the channel dispersion with all digital processing in electronic domain[1]. The frequency and phase offset between a free-running local oscillator (LO) and carrier can also be compensated with digital algorithms. For FO compensation, all digital feed forward estimators [2, 3] are widely used. Compared with the  $m^{\text{th}}$  power FO estimator, the complexity of PADE [2] with pre-decision is obviously much lower. Due to removing data phase with pre-decision, PADE can theoretically track large FO range of  $[-Rs/2, +Rs/2]$ , which is much

---

Zhi-Yu Li · Xiao-Bo Meng · Xian Zhou · Xue Chen · Yang-Yang Fan · Hai Zhu  
State Key Lab of Information Photonics and Optical Communications,  
Beijing University of Posts and Telecommunications, Beijing 100876, China  
e-mail: xbmeng\_315@163.com

\* Corresponding author.

wider than  $[-Rs/8, +Rs/8]$  of the  $m^{\text{th}}$  powermethod, where  $R_s$  is symbol rate. But PADE requires an estimation operation when each symbol is input, and the estimation and compensation must be completed in the very symbol period.

In order to decrease the speed requirement, by improving on PADE we propose a new FO estimator scheme which can be operated group by group, where a group stands for a group of consecutive symbols, namely Grouped-PADE (G-PADE). G-PADE makes parallel operation possible for FO estimation and lowers the average calculation complexity for each symbol almost linearly. The following sections are organized as follows: In Section 2, we analyze the principle of G-PADE, compare the complexity of G-PADE and PADE, and propose an initialization scheme for G-PADE. In Section 3, the simulation model is described. In Section 4, we give the simulation results. At last we conclude the results.

## 2 Operation Principle

### 2.1 Introduction of PADE

The block diagram of PADE is shown in Fig.1. The FO estimation is operated in Fig.1(a), and the compensation is shown in Fig.1(b). We present the improved estimator G-PADE derived from PADE, which is suit for parallelization. In section 2.2, principle of G-PADE is described, from which the principle of PADE can be understood well.

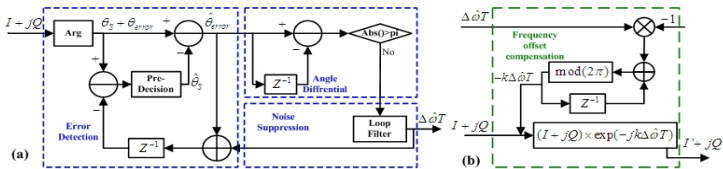
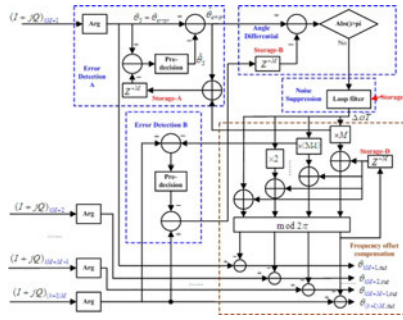


Fig. 1 Principle of PADE, (a) Estimation and (b) Compensation

### 2.2 Principle of G-PADE

For  $M$  consecutive symbols, G-PADE just do once estimation for the first symbol, and the estimation result belonging to the first symbol is applied to all the  $M$  symbols in the group. In optical coherent receiver, at FO estimator, the phase of the  $k^{\text{th}}$  received symbol can be expressed as  $\theta_k = \theta_{S,k} + \Delta\omega t_k + \theta_{n,k} = \theta_{S,k} + \theta_{err,k}$ , where  $\theta_{S,k}$  is data modulation phase.  $\theta_{n,k}$  is the phase fluctuation, and  $\Delta\omega t_k$  is the phase error.



**Fig. 2** Principle of G-PADE

The principle of G-PADE is shown in Fig.2. It mainly includes five parts: “Error Detection A”, removes the data phase of the first symbol in each group; “Error Detection B”, removes the data phase of the  $M^{th}$  symbol of each group; “Angle Differential”, gets the differential value between the error phase of the first symbol in each group and the last symbol in the former group, which includes  $\Delta\omega T$  and the phase caused by ASE noises; “Noise Suppression”, removes the noises and gives the accurate FO estimation; “Frequency Offset Compensation”, compensates all the  $M$  symbols of the same group using the estimation result of the first symbol.

### 2.3 Complexity Improvement of G-PADE

Since there is only once FO estimation in each group in G-PADE, the average complexity of G-PADE, is much lower than PADE. Table.1 gives the comparison of the average calculation complexity for each symbol in PADE and G-PADE. The “Arg” operation should be realized in look up table (LUT), and can not be simplified or cancelled in both methods. When  $M > 1$ , the complexity of G-PADE in addition, multiplication, comparison and pre-decision becomes much lower than PADE as  $M$  increases.

**Table 1** Average complexity of PADE and G-PADE for each symbol ( $M$  is group length)

	LUT	Addition	Multiplication	Comparison	Pre-decision
PADE	2	6	3	1	1
G-PADE	2	$2 + 7/M$	$1 + 1/M$	$1/M$	$2/M$



## 2.4 Initialization of G-PADE

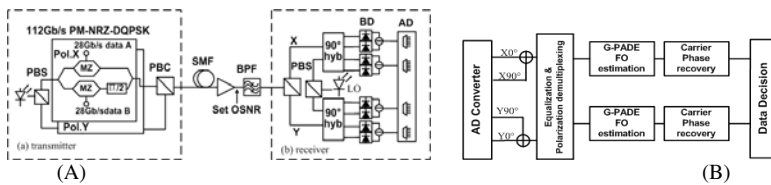
The same as PADE, G-PADE needs an initial set of FO. We adopt the traditional  $m^{\text{th}}$  power FO estimator [3], 4<sup>th</sup> power for (D)QPSK, which can give G-PADE an exact enough initialization. The 4<sup>th</sup> power method does not need initialization. It only needs hundreds of symbols to estimate a roughly right FO value when the receiver starts to work. This roughly right value can be exacter as the length of the symbols gets bigger, because the impact of noises can be removed more.

For (D)QPSK system, when the real FO goes beyond, the wrong estimated value and the real FO have a relation that  $|\Delta\omega T - \Delta\hat{\omega} T| = i\pi / 2$ , where  $i$  is integer. So we can change the wrong estimated value to the right one if we can judge whether the estimated value is right. This judgment can easily be made by BER aiding.

So, for all possible real FO value, we have a way to give G-PADE an initialization, which can be described as follows: (1) Before G-PADE starts to work, use 4<sup>th</sup> power to estimate FO of a section of symbols and get BER of them. (2) Judge whether the estimation is right. (3) If right, give G-PADE an initialization with the estimated value; if wrong, change the estimated value using the  $|\Delta\omega T - \Delta\hat{\omega} T| = \pi / 2$ , and give G-PADE the changed value.

## 3 Simulation Model

G-PADE for 112Gb/s PM-NRZ-DQPSK transmission system is simulated. We use VPI to simulate the transmitter, optical link and the front-end of the receiver, which is shown in Figure.3(A). The simulation condition is as follows: OSNR 16.5dB, CD 100ps/nm and PMD 10ps. Both transmitter laser and local laser linewidth are 1MHz.



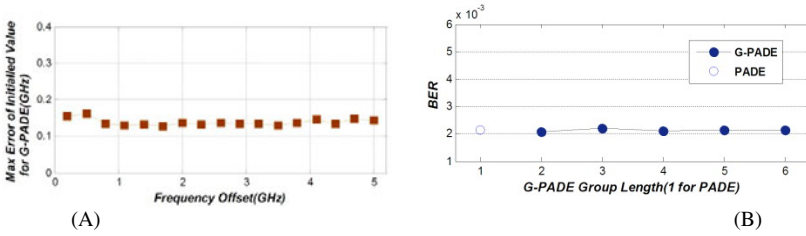
**Fig. 3** (A) VPI PM-DQPSK transmission system model (a) transmitter, and (b) front-end of receiver; (B)offline PC processing structure

The digital samples are then processed using MATLAB on PC. The processing structure is shown in Figure.3(B). The signals are first equalized and polarization de-multiplexed, and then down sampled to 1 sample/symbol.

G-PADE and phase recovery are used to remove the frequency and phase offset between LO and carrier, after which the data is decided and decoded. The initialization of G-PADE is as described in section 2.3.

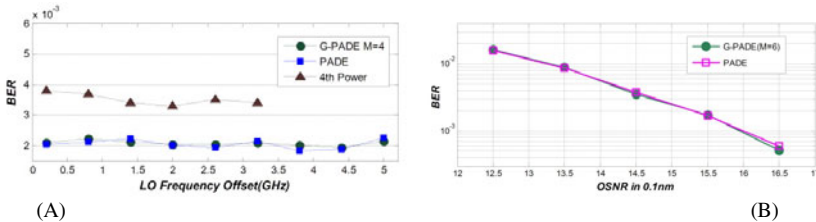
### 4 Simulation Results and Analysis

Fig.4(A) shows the maximum possible initialed error for G-PADE using 4<sup>th</sup> power method under various FO. Fig. 3 tells that with fixed group length, the maximum possible initialed errors of 4<sup>th</sup> power are nearly the same under different FO. When the value of real FO is negative, the results are similar with Fig. 4(A). The length 1000 can make the max error less than 0.2GHz. This means the way to initialize G-PADE as we described in section 2.3 is feasible.



**Fig. 4** (A)max possible error of initialed value by 4th power FO estimator (with block length 1000)under possible FO; (B)BER of PADE and G-PADE(with group length 2 to 6) at 16.5dB OSNR and 2.5G FO

Fig.4(B) shows the BER comparison between G-PADE and PADE after correct convergence using the way as we described in section 2.3. The group length of G-PADE is from 2 to 6. We can see that as group length increase from 2 to 6, BER of G-PADE keeps almost the same, about just over  $2 \times 10^{-3}$ , and this BER is just the same as BER of PADE. This means for G-PADE of group length 2~6, the performance does not go worse compare to PADE.



**Fig. 5** (A)BER of PADE and G-PADE (with group length 4) under possible FO; (B) BER of G-PADE and PADE (with group length 6) under different OSNR

Fig.5(A) shows the BER comparison of G-PADE (take group length 4 as an example), PADE and 4<sup>th</sup> power method under different FO. As shown in Fig. 5(A), the 4th power method with higher complexity has higher BER than PADE and G-PADE. Additionally, the range of its estimation can not be over 1/8 symbol rate, which is 3.5GHz in the simulation system. Those are the reasons why 4th power is not considered as the appropriate FO estimator. As FO increases, BER of G-PADE keeps staying around  $2 \times 10^{-3}$ , and curve of PADE looks the same, keeping stable around  $2 \times 10^{-3}$ . This means G-PADE has the same performance under different FO, so does PADE. Additionally, the two curves are at the same level, which again proves the correctness of Fig.4(B).

Last, Fig.5(B) gives the BER-OSNR curve of G-PADE(M=6) and PADE with only FO and laser linewidth impact. The FO is 2.5GHz and the linewidth of both transmitter and local laser is 1MHz. The performance of G-PADE is just the same as PADE under different OSNR from 12.5dB to 16.5dB.

## 5 Conclusion

We propose an improved FO estimation scheme for optical coherent receiver: G-PADE with 4<sup>th</sup> power initialization by BER aiding. The grouping mechanism makes parallel operation possible for FO estimation, and linearly reduces the request for the hardware speed, which offers a possibility for hardware implementation using existing devices. We also give an initialization scheme to make G-PADE suit for [-5GHz, +5GHz] of frequency offset, which is proved to be effective. Using this initialization method, for 112Gb/s PM-DQPSK system at 16.5dB OSNR, G-PADE converges correctly on condition that the group length is no more than 6, and no visible penalty is found. This means the hardware speed requirement of G-PADE can be one sixth of PADE. Additionally, the stable-running complexity of G-PADE decreases much compared with PADE, this is a benefit for hardware scale and power consumption under stable-running.

## References

1. Raybon, G., Winzer, P.J.: 100Gb/s Challenges and Solutions. In: Optical Fiber Communication Conference (OFC 2008), OTuG1, San Diego, CA, February 24-28 (2008)
2. Li, L., Tao, Z., Oda, S., Hoshida, T., Rasmussen, J.C.: Wide-range, Accurate and Simple Digital Frequency Offset Compensator for Optical Coherent Receivers. In: Optical Fiber Communication Conference (OFC 2008), OWT4, San Diego, CA, February 24-28 (2008)
3. Leven, A., Kaneda, N., Koc, U.-V., Chen, Y.-K.: Frequency Estimation in Intradynic Reception. IEEE PTL 19(6) (March 15, 2007)

# Process-Driven Integrated Product Design Platform

Bo Zhao, Jun Luo, Rongmei Nie, Xiaojun Hu, and Hui Zhao

**Abstract.** In order to realize effective process management and application integration used in the whole product development process, some advanced techniques or methods have been proposed and successfully applied in product design. But in distributed environment, how these engineering tools can effectively access right data in right format at the right time inevitably becomes a very important factor in product development. So a flexible integrated design platform for process integration has been established, which consists of representation layer, process control layer, services layer and repository layer. Based on analyzing three patterns of organizing data, context data model is introduced, which organizes data based on process. The mechanism of interaction between context data model and process unit is described in detail. The model-view-controller (MVC) design pattern is adopted in our design platform which separates data model from its various views. Web services are developed to realize the dynamic data integration between applications. The design platform based on the key techniques is developed as a prototype to prove the concept.

## 1 Introduction

The development of information technology and the global economy strongly drive the transformation of enterprise of design and manufacture. Enterprise should have a rapid response to the needs from market. Nowadays the advance methods and techniques, such as: CAx(computer-aided technology), PDM (Product Data Management) and DFX (Design For Assemble: DFA, Design For Manufacturing: DFM and etc.), have been successfully applied in product design. But plenty of heterogeneous data which generated by different applications or systems exists in product lifecycle throughout the enterprise. Moreover, the product

---

Bo Zhao

Beijing Institute of Astronautical Systems Engineering, Beijing, China

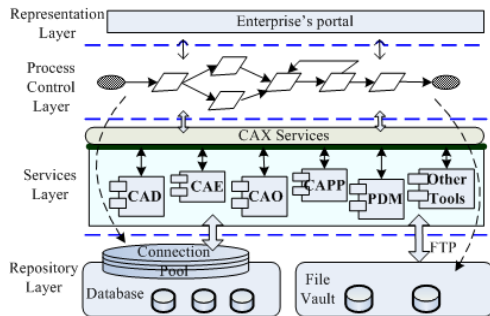
e-mail: zzzbbb163@163.com

development process involves multidisciplinary knowledge, multi-designers, multi-parts of company. So the simple use of CAx applications ceases to be an important factor in different competitors. Therefore how multi-participant collaboratively finish design or analyse and how applications can access right data in right format at the right time determine whether a specific project can be successful or not. So it is necessary to establish a flexible integrated design platform which can integrate different applications and support designers' work.

Some of valuable works have been researched by different person or groups, For example, integrated web-based PDM system [1], collaborative product development environment [2], process integration framework for product design [3], process integration and resource sharing of different design software [4], the model for application integration [5], and etc.

## 2 Architecture

**Fig. 1.** For optimizing the design process and smoothly exchange data, document and knowledge, we establish a extensible process-driven integrated design platform[3], which consists of four tiers, namely, representation layer, process control layer, services layer and repository layer. Every layer will be ex-plained in more detail below.



### (1) Representation layer

Web-based interface for users is created, which allows every authorized user to visit the enterprise's portal in everywhere. The model-view-controller (MVC) design pattern is used to design modular and interactive user interfaces, through which the data model can be separated from its various representations. This pattern will be introduced in the following key techniques.

### (2) Process control layer

This layer is the functional core of this architecture, which realizes most functions of process integration, namely, mapping enterprise's business flow, regulating the logic sequence of actions in process, optimizing every phase of process, encapsulating business logic, managing users and their roles, and maintaining security of system. In this layer, as data organization is based on process, the services provided by applications are invoked by process engine, resources are configured by activities from process and human are organized by process, all elements involved in product design are process-driven. So this architecture is process-centralized.

### (3) Services layer

In distributed environment, all applications and engineering tools can be regards as service providers and service applicants. Some operations and functions

can be developed as web services. There is a service registration centre which records the information of services provided by applications. All the requestors can find the required services which have been registered.

#### (4) Repository layer

It stores all data in process, which has the capability of retaining the data's integrity and security while being concurrently accessed. Meta-data are read and written in database through connection pool, while files are stored in file vault. Knowledge of context is also stored in database in a regular format.

For establishment of the four-tier architecture, we research four key techniques, which will be introduced as follows.

### 3 Key Techniques

#### A *Patterns of Organizing Data*

In order to effectively manage and utilize various data in development process, there are some following patterns for organizing data [6]:

##### (1) Organizing data based on product structure

In this pattern, Product structure is the framework of data organization and management. Due to the dynamic change of product structure tree in development process, the operation that adding, deleting or modifying a node would influence the data on this node and the relevant nodes. Moreover, there are many different views in product development process, such as design view, assembly view, and etc., so the organization may change for the special view, and express the static relations about product information in a certain view. In other words, product data simply stack on structure tree.

##### (2) Organizing data based on folder

In this pattern, file folder is regarded as the container which reflects the product structure. The structure of folder is decided by product structure, which can differentiate and manage different information about different parts. The consecution and rationality of the folder directly influence the convenience of data's access.

##### (3) Organizing data based on product lifecycle

A.McKay[7] proposed a framework for product data, in which data is divided into three kinds, namely, specification, definition and actual data. The three kinds of data correspond with the different phases of product development. Organizing data based on product lifecycle express generation order of data in different phases, which can be regarded as process-oriented in part. However, this pattern also classifies data from product structure, which groups information in different folders at different phases.

##### (4) Organizing data based on process

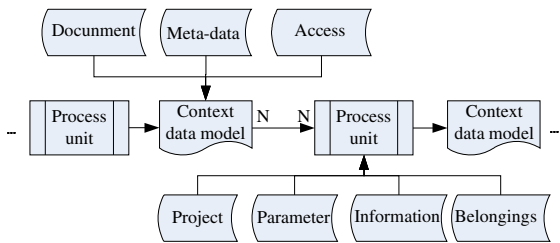
The three pattern mentioned above only orient a certain view of product data. Due to the complexity and dynamic of product and its development, organizing data based on product structure or lifecycle is a static method comparatively, which weaken the relations of different information generated in process. So for

the dynamic process, we present context data model. Context data model should not only describe many kinds of data type, but also reflect complex relations in process, through which the steps of process can exchange information more fluently and engineering applications will collaboratively work.

**B Mechanism of Interaction between Context Data Model and Process Unit**

Organizing data based on process reflects the close relation between data and process. Generally speaking, the process of product development can span a widely period, from customers’ needs, conceptual design, and detailed design to manufacturing. The development process can be divided into three levels according to different granularity from macroscopy to microcosm: the project level, the workflow level and the design process level, which could be called as project flow, workflow and design process. No matter how size of the granularity of the nodes, there are some cells which hold the independent functions in process. These cells can be called process unit. Process unit specifies how context data model transfer. The logic permutation and combination of units decides the flow direction of data. Context data model materializes the input and output of unit, and decide which unit will be executed in branch, who will run it, when and how it runs. The network constructed by process units defines the framework of data’s flow and transformation, while the speed and direction of data flow are decided by itself.

**Fig. 2.** The Mechanism of interaction between context data model and process unit is depicted in Fig. 6. The task of process integration is defining the topology based on process units and constructing the container for context data model. So the product development process is the instantiation process for data model, in which data experience a course from generation, transformation, and enrichment to perfectness.



**C Model-View-Controller (MVC) Design Pattern**

Models are those components of application systems that hold specific information, in which data stores as a certain format. Views deal with everything graphical, which request data from a Model and display data. Views are closely associated with a Controller. Each Controller-View pair has one model, whereas

each model may have many Controller-View pairs. Controllers are the bridges which connect Models and Views.

For example, XML file which represents a class of Context model could be presented to requestors as the particular format. Therefore, if the interface needs to change, the View-Controller pairs can be rapidly modified without changing the Model.

## ***D Dynamic Data Integration Based on Web Services***

In the design process, applications need to exchange data with design platform or other applications at run time. So, an effective method should be presented to satisfy the requirement. A web service is a software module performing a discrete task or set of tasks that can be found and invoked over a network including and especially the World Wide Web. The developer can create a client application that invokes a series of web services through remote procedure calls (RPC) or a messaging service to provide some or most of the application's logic. Web services can be written in any language and run on any platform. Therefore, it can be used to realize the dynamic data integration between applications with the following steps:

Step 1: the development of web services. Some application's logic is developed as Web services. Take PDM system for example, we write the program by Microsoft Visual C#. And users can upload or obtain product data through web pages. But it is impossible for an application to exchange data with PDM through web pages. So some interfaces should be developed to facilitate the data availability between applications and PDM. We encapsulate some functions which can be only achieved through web pages before as web services by invoking or combining the methods which exist in the business logic layer of PDM. These functions include checking in or out documents, searching, obtaining attribute data of a part or a component, etc.

Step 2: the registration of web services. All developed web services should be registered so that they can be found.

We use Web Services Description Language (WSDL) to describe the entrance, the interface, the input and the output of a web service. And then, we register it in UDDI (Universal Description, Discovery & Integration) registration centre which contains basic information of all developed web services.

Step 3: finding and invoking web services. After service provider register its services, consumers can lookup the required service by service information in registration centre which will provide the location and the WSDL file of service provider. A service broker which is created for each application can remotely invoke the methods of web services as the prescriptive format. They communicate with the SOAP (Simple Object Access Protocol) messages which are actually XML documents. SOAP messages are transported by HTTP between a web service and the calling application. Thus, web services can normally work through different platforms.



## 4 Conclusion

Based on the flexible architecture and the key techniques mentioned above, we have developed a web-based prototype for an institute which supports the design of air filter for engine. The system consists of ten steps for design, and integrates some applications, such as: MatLab, Pro/ENGINEER, ANSYS and other tools we developed. So the key techniques and architecture are confirmed by the prototype. And we will develop another system for a specific product; meanwhile, the theory about process integration, process modeling and architecture will be improved.

## References

- [1] Wang, L., Xu, Z., Ruxin, N.: Key Techniques for the Development of Web-Based PDM System. *Journal of Beijing Institute of Technology* 15(3), 269–272 (2006)
- [2] Jianshou, K., Zhang, Y., Wang, H.: Study on Integration of Distributed PM and Workflow in Collaborative Product Development Environment. *China Mechanical Engineering* 14(13), 1122–1125 (2003)
- [3] Bo, Z., Yan, Y., Ruxin, N., et al.: Key techniques about the process integration for product design. In: *IEEM 2007: IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 568–572 (2007)
- [4] Hu, Y., Fan, Y., Tong, X.: Process integration & resource sharing based on XML & process ontology for MDO. *Computer Integrated Manufacturing Systems* 13(6), 1076–1082 (2007) (in Chinese)
- [5] Harikumar, A.K., Lee, R., Yang, H.S., Kang, B.: A model for application integration using Web services. In: *Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science (ICIS 2005)*, pp. 468–475 (2005)
- [6] Wang, B.: Research on Theory, Technology and Application of Digital Product Development Process Management. Ph. D. dissertation of Beijing Institute of Technology, pp. 32–34 (2004) (in Chinese)
- [7] McKay, A., Bloor, M.S., de Pennington, A.: A Framework for Product Data. *IEEE Transaction on Knowledge and Data Engineering* 8(5), 825–838 (1996)

# Segmentation of Audio Visual Malay Digit Utterances Using Endpoint Detection

Mohd Ridzuwary Mohd Zainal, Aini Hussain, and Salina Abdul Samad

**Abstract.** An endpoint detection algorithm is utilised for segmentation of audio video Malay utterances. An audio visual Malay speech database of subjects uttering numerical digits is used. Synchronization between video frames and audio signals is taken into considerations for audio visual speech processing. The proposed system is able to group together the individual syllables that make up each of the uttered Malay digits.

## 1 Introduction

Using visual speech to augment automatic speech recognition is not very surprising since human speech perception naturally uses both types of information to understand speech correctly. A review done by Potamianos et al. has listed a multitude of research on bimodal speech recognition systems dating from 1954 [1]. The review proves that there is a lot of interest surrounding this topic. It is worth noting that among those studies, there is very little effort done on automatic speech recognition centred on Malay language words and pronunciations, specifically audio visual bimodal systems.

Nevertheless, recent publications and literature shows some increase of interest for automatic Malay speech recognition. Al-Haddad et al., Reem and Raja, for example, used fusion methods of techniques including dynamic time warping [7],[8], hidden Markov models [8],[9], and artificial neural networks [6] to automatically recognize audio only speech of isolated Malay digits. Noraini and Kamaruzaman took a different approach and focused more on studying the language itself. Their work includes segmentation, labelling, and modelling of Malay pronunciation variation so that a more customized recognition system can be developed [3],[4].

One of the main components in speech processing systems is the endpoint detection or speech detection. This is where an automated speech processing system

---

Mohd Ridzuwary Mohd Zainal · Aini Hussain · Salina Abdul Samad  
Universiti Kebangsaan Malaysia, Malaysia

differentiates which part of the given input is speech and which part is non-speech activity. In this paper, we study an application of an endpoint detection technique to audio visual Malay speech utterance input. The method was suggested by Li et al. in 2002 [5].

## 2 Audio Visual Malay Digit Database and Segmentation

The Malay language is part of the Austronesian language and is spoken by many in the South East Asian region and is the official language of Malaysia, Indonesia and Brunei. Malay language is an agglutinative language. This means that the words in the Malay language are formed by joining syllables. The syllabic structure in Malay language is well-defined and can be unambiguously derived from a phone string. All the syllables in the Malay language are pronounced almost equally which makes it a non-tonal language [4].

Research for audio video Malay speech processing requires database samplings and recordings for analysis, system development and testing. In this work, the database that is used consists of audio video recordings of volunteers uttering digits from zero to nine in the Malay language. Fig. 1 shows a sample frame of the recording and the dynamics of the mouth region when saying the word 'satu' ('one' in Malay). Table 1 gives the Malay words and phoneme representation of the uttered digits. The database is recorded in the format of 720 x 576 video resolution, 25 Hz video frame rate, 16-bit audio resolution and 48000 Hz audio sampling rate.

The purpose of segmentation is to prepare audio video segments that can readily be used as inputs in audio video speech processing systems. Either the audio stream or the video stream can be used to determine how to segment the combined stream. However, we chose to use the audio stream since there are more instances where the audio speech signal changes but the images in the video frames remains the unchanged such as when a person finished uttering a digit but did not change his lips shape to its initial position.

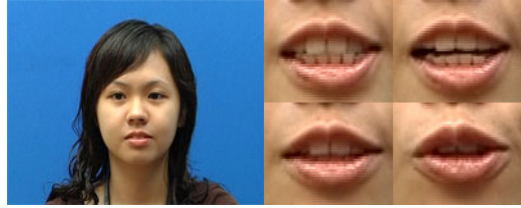
The audio video stream is separated into its audio and video streams for two different processes. Then the audio stream is downsampled to 8000 Hz and is run through an endpoint detection algorithm. This algorithm detects where a digit utterance starts and ends. The position of these endpoints is then used on the video stream to determine which frames an endpoint occurs. A frame where a beginning endpoint occurs is called a start-frame while the frame where the ending endpoint occurs is an end-frame. The video stream is then segmented for every pair of start and end-frames.

Synchronization between video frames and audio signals is very important for audio visual speech processing. For this reason, we did not choose the exact endpoints detected for segmenting the audio signal. The audio signal is first framed to match every video frame. In our case, the video frame rate is 25 Hz and the audio signal rate is 8000 Hz. Therefore, every video frame corresponds to 320 samples of audio signal, which is a single audio frame. Audio signal segmentation begins at the first sample of the start-frame and ends at the last sample of the end-frame.

**Table 1** Digits in English and Malay Language

Digits	English	Malay	Phoneme
0	Zero	Kosong	/koson/
1	One	Satu	/satu/
2	Two	Dua	/dua/
3	Three	Tiga	/tiga/
4	Four	Empat	/empat/
5	Five	Lima	/lima/
6	Six	Enam	/enam/
7	Seven	Tujuh	/tujuh/
8	Eight	Lapan	/lapan/
9	Nine	Sembilan	/sembilan/

**Fig. 1** (Left) A single image frame from a video in the database. (Right) Mouth changes when uttering the digit 'Satu' (one). The images from left to right and top to bottom coincides with the letters 'S', 'A', 'T', and 'U', respectively.



### 3 Endpoint Detection

The short-term energy of the signal is calculated using the one-dimensional short-term energy in the cepstral feature equation

$$g(t) = 10 \log_{10} \sum_{j=n_t}^{n_t+I-1} o(j)^2 \quad (1)$$

where  $o(j)$  is the audio stream samples,  $t$  is the frame number,  $g(t)$  is the frame energy in decibels,  $I$  is number of samples per frame and  $n_t$  is the number of first sample for every frame. The equation transforms the audio signal into a short-time energy signal with

$$N = L / I \quad (2)$$

sample points, where  $L$  is the number of samples in the original audio signal and  $I$  is the number of samples per frame used to calculate energy previously.

The endpoint detection is based on a robust speech endpoint detection proposed by Li et al. in 2002 [5]. Their work is based on the ramp-edge detection proposed by Petrou and Kittler [2] which was an extension of the Canny edge detector that was originally used for detecting edges in images. The endpoint detector is a filter given by the function

$$\begin{aligned}
 f(x) = & e^{Ax} [K_1 \sin(Ax) + K_2 \cos(Ax)] \\
 & + e^{-Ax} [K_3 \sin(Ax) + K_4 \cos(Ax)] \\
 & + K_5 + K_6 e^{sx}
 \end{aligned} \tag{3}$$

where  $A$ ,  $s$ , and  $K_i$  are filter parameters. The filter has a width of  $(2W+1)$  and the actual filter coefficients are given by

$$h(i) = \begin{cases} f(i), & \text{for } -W \leq i \leq 0 \\ -f(-i), & \text{for } 1 \leq i \leq W \end{cases} \tag{4}$$

where  $i$  is an integer.

The filter parameters are calculated so that the filter is antisymmetric and is of finite extent going smoothly to zero at its ends. The resulting filter has to have a maximum amplitude of  $|k|:f(x_m) = k$  where  $x_m$  is defined by  $f'(x_m) = 0$  and  $x_m$  is in the interval  $(-w,0)$ . Several examples of the filter parameters for different values of  $W$  have been given in [2]. According to [5], these given parameters can be applied to different filter sizes by calculating new  $s$  and  $A$  values using the equations

$$s_{new} = s(W/W_{new}) \tag{5}$$

$$A_{new} = A.s_{new} \tag{6}$$

where  $W_{new}$ ,  $s_{new}$ , and  $A_{new}$  are new  $W$ ,  $s$ , and  $A$  values, respectively.

Once we have the filter parameters, we can apply the filter just like a moving average filter with the equation

$$F(t) = \sum_{i=-W}^W h(i)g(t+i) \tag{7}$$

where  $g$  is the short-term energy feature from (1) and  $t$  is the current frame number.

## 4 State Transition and Video Endframe Detection

Five constants which are the upper threshold,  $TU$ , the lower threshold,  $TL$ , the length of a word ending from  $TL$ ,  $L_{end}$ , the minimum syllable length,  $L_{min}$ , and the maximum syllable length,  $L_{max}$ , are used to determine the endpoints from the filter output. These constants are determined by experimentation. When the filter output moves upwards across  $TU$ , the frame number is marked as a word start point. Then the transition across  $TL$  is searched for, and a word end point is marked at  $L_{end}$  frames after that.

If the output crosses  $TU$  again before  $L_{end}$  frames, it means that the next syllable of the word spoken is found and the word end point is moved to  $L_{end}$  frames after the next  $TL$  crossing. If the total length from a word start point to end point is less than  $L_{min}$  the points are rejected as word endpoints. If the total length from a word start point to the word end point in a single syllable is larger than  $L_{end}$  then the endpoints are also discarded.

Whenever any endpoint in the audio stream is confirmed, video frame that corresponds to that particular point is marked similarly. In other words, a video start frame is marked where its speech filter output frame is marked as a word start point and a video end point is marked where its speech filter is marked as a word end point.

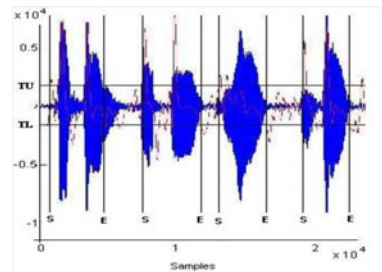
A word segment is taken from the first audio sample from the video start frame to the last audio sample from the video end frame regardless of the position of the audio endpoints within the video end frames. This is chosen so that the segments do not lose its audio video synchronization in relation to its absolute audio sample time and video frame time. In order to avoid significant non-speech events to be included in the final audio video segment, all audio samples outside of the audio endpoint boundaries are converted to audio silence.

## 5 Results and Discussions

The endpoint detection algorithm is tested on a Malay digit utterance audio visual database consisting of 47 native and non-native speakers. Each person uttered digits consecutively from zero to nine for 20 times. The speech rate varies from 39 words per minute (wpm) to 89 wpm with an average of about 61 wpm. The gap between uttered digits also differs among speakers with lengths between 50 ms to 700 ms. Difference of speech style and speed gives varying results. Therefore the results are categorised to three groups; slow speakers, medium speakers, and fast speakers. Fig. 2 shows a sample result of the endpoint detection and segmentation for the utterances of the digits 'Kosong', 'Satu', 'Dua' and 'Tiga'.

The low detection rate for the fast speech group is anticipated because of the closeness of the spoken digit between the words. Most of the errors are caused by two different words that are detected as one. The medium speech speed gives the best results, which means that the algorithm is suitable for this particular speech rate. The low detection rate for the slowest speech speed is due to the nature of the speaker giving a slight pause between syllables in a word. Most of the errors in this group are caused by syllables of a single word that are detected as different words. Changing the threshold values can raise the detection rate but may cause the system to be speaker dependent.

**Fig. 2** Endpoint detection for the utterance of 'Kosong', 'Satu', 'Dua', and 'Tiga'. S is the label for word start points and E is the label for word end points



## 5 Conclusion

The segmentation of audio visual Malay digit utterances is achieved using a speech endpoint detection method. The algorithm managed to group together the syllables that make up a word. Although the algorithm has low performance for speech with high speech rates, it is useful as a low complexity word segmentation algorithm for systems that use medium speech speed or for discontinuous speech.

## References

1. Potamianos, G., Neti, C., Luetttin, J., Matthews, I.: Audio-visual automatic speech recognition: An overview. *Issues in Visual and Audio-visual Speech Process* (2004)
2. Petrou, M., Kittler, J.: Optimal edge detectors for ramp edges. *IEEE Trans. Pattern Anal. Mach. Intell.* 13, 483–491 (1991)
3. Seman, N., Jusoff, K.: Acoustic pronunciation variations modeling for standard Malay speech recognition. *Computer and Information Sci. J.* 1(4), 112–120 (2008)
4. Seman, N., Jusoff, K.: Automatic segmentation and labeling for spontaneous standard Malay speech recognition. In: *Int. Conf. Adv. Comput. Theory and Engineering*, pp. 59–63 (2008)
5. Li, Q., Zheng, J., Tsai, A., Zhou, Q.: Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Trans. Speech Audio Process.* 10, 146–157 (2002)
6. Sabah, R., Ainon, R.N.: Isolated digit speech recognition in Malay language using neuro-fuzzy approach. In: *3rd Asia Int. Conf. Modelling and Simulation*, pp. 336–340 (2009)
7. Al-Haddad, S.A.R., Samad, S.A., Hussain, A.: Automatic Recognition for Malay Isolated Digits. In: *3rd Int. Colloq. Signal Process. and Its Appl.*, Melaka, Malaysia, March 9-11 (2007)
8. Al-Haddad, S.A.R., Samad, S.A., Hussain, A., Ishak, K.A.: Isolated Malay digit recognition using pattern recognition fusion of dynamic time warping and hidden markov models. *Am. J. of Appl. Sci.* 5(6), 714–720 (2008), doi:10.3844/ajassp.2008.714.720.
9. Al-Haddad, S.A.R., Samad, S.A., Hussain, A., Ishak, K.A., Noor, A.O.A.: Robust speech recognition using fusion techniques and adaptive filtering. *American Journal of Applied Sciences* 6(2), 290–295 (2009), doi:10.3844/ajassp.2009.290.295

# Data Mining in Clinical Decision Support Systems

Liljana Aleksovska-Stojkovska and Suzana Loskovska

**Abstract.** Data mining is an emerging methodology in the knowledge discovery area, which has a vast potential to be used in the health care. This paper presents the possibility for using data mining methods in extracting patient specific rules from the individual data collected for asthma patients, which are used to support the clinical decision process.

## A

association rules, 3

Asthma, 3

## C

clinical decision support systems, 3

## D

data mining, 3

## P

peak flow, 3

## S

school-age children, 3

## 1 Introduction

Clinical decision making is a complex process, which applies general medical knowledge on patient individual data and generates patient specific decisions. The medical institutions collect and store huge amount of patient related data,

---

Liljana Aleksovska-Stojkovska

MAK-System Corp., 2720 River Rd, Suite 225, Des Plaines, IL, USA

PhD student at „Ss. Cyril and Methodius – Skopje“, Macedonia

e-mail: Liljana.A.Stojkovska@gmail.com

Suzana Loskovska

Faculty of Electrical Engineering and IT

University „Ss. Cyril and Methodius – Skopje“, Skopje, Republic of Macedonia

e-mail: suze@feit.ukim.edu.mk



including medical history, current diseases, symptoms, lab results, allergies, treatment plans etc. Unfortunately, much of this data is never analyzed at all and it may take weeks for the human analysts to discover useful information, which is often “hidden” and not readily evident [1]. This phenomenon is best described with the popular expression “We are drowning in data, but starving for knowledge” [1].

With such motivations, the scientists are looking into data mining, as an emerging area of the computational intelligence, which is offering new theories, techniques and tools for analysis of large data sets [2]. Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships, which provide a clear and useful result to the data analyst [3]. It is a relatively new methodology, developed in the mid-1990s, which has been successfully applied to other areas, such as financial institutions, marketers, retailers and manufacturers [4]. There is also a vast potential for applying data mining in health care, including: evaluation of treatment effectiveness, healthcare management, customer relationship management, etc. [4].

Data Mining is a multi-disciplinary approach, including methods from computer science, data visualization and statistics [3]. In general, the data mining tasks can be classified to: (1) tasks of description, which aim to find human-interpretable patterns and associations in set of data and (2) tasks of prediction, which aim to foretell the outcome of specific point of interest and may be applied to the construction of decision models for supporting procedures such as diagnosis and treatment planning, which once evaluated and confirmed may be embedded within the clinical information system [3].

The purpose of this paper is to present the potential for applying data mining in clinical decision support. Specifically, we are building a CDSS for managing asthma in school-age children and are looking into the data mining methods to identify patterns in the patient’s medical data and extract rules, which will be used in conjunction with the general knowledge base to support the decision process.

The paper is organized as follows. Section 2 describes the CDSS for managing asthma in school-age children with focus on the knowledge base. It further presents how a data mining method can be used to extract patient specific rules from the patient’s electronic data. The section 3 provides conclusions and directions for future work.

## **2 Using Data Mining Methods to Expand the Knowledge Base in Clinical Decision Support System**

This section describes the idea of using the data mining methods to extract rules from the patient’s electronic record and expand the knowledge base in a clinical decision support system for managing asthma in school-age children.

## ***2.1 PedAst – CDSS for Managing Pediatric Asthma in School Age Children***

PedAst (Pediatric Asthma) is a CDSS intended to meet the following functions:

- Provide means for capturing and storing patient's medical information
- Open communication channel between the patient, school and doctor office
- Assist the decision making process by assessing the patient's state based on the entered data and generating reminders and recommendations.

PedAst is a web-based modular system composed of the following modules:

**Doctor module** – to be used in the doctor's offices by the medical practitioners for following and assessing the patient's state and providing support when making diagnoses and prescribing medications.

**School module** – to be used by the nurses in the schools, for following the patient's state and providing communication channel with the doctor's office.

**Patient's personal module** – to be used by the child's caregivers for capturing and storing data from the regular peak flow measurements, medications received during the day, daily activities, symptoms, unexpected reactions etc.

**Core module** – composed of inference machine and repositories for the knowledge base and the patient's electronic medical records.

The data entered by the doctors, school nurse practitioners and patients is to be integrated in a central electronic medical record, stored in the core module. This data is to be processed by the inference engine, which by applying the rules from the knowledge base on the patient specific data will generate appropriate feedback (i.e. alert message, diagnostic or therapeutic recommendations).

The system is expected to improve significantly the treatment and the management of asthma in the school-age patients, by allowing continuous monitoring of the medical conditions of the patients and quick intervention if the results are not satisfactory.

## ***2.2 The Knowledge Base in PedAst***

The knowledge base in PedAst contains medical knowledge in a form of logical rules (IF – THEN). An example of a simple rule would be: "If the peak flow reading is less than 50 percent of the normal value (RED zone), then generate an alert message for the patient to go straight to the emergency room". The rules in the knowledge base can be generated through a knowledge base editor by the human medical experts or extracted from the large set of data in the patient's electronic medical records. The innovative approach that we are proposing here is to have a general knowledge base applicable to all the patients and individual knowledge base specific to an individual patient. The need for keeping separate knowledge

bases for the general patient population and for individual patients is that asthma, as many other chronic diseases, has some general characteristics applicable to the majority of the patients, but also there are many specific facts that determine the management of the disease based on the patient's individual characteristic. While the general rules and common facts about the disease can be extracted from the medical books, web-sites, human expert knowledge etc., the specific rules about an individual patient must be extracted from its own data. The whole idea of having an individual patient knowledge base is to analyze the patient specific data about the disease collected over the period of time and determine a pattern. Based on this, specific rules can be generated, which can be applicable only to that individual patient. For example, a patient may have severe pollen allergies which are exacerbating his/her asthma symptoms during the spring. Therefore, a specific rule can be generated applicable to this patient, that he needs an increased dose of the asthma medication during the spring months. Based on this rule, the system will generate an alert message before the pollen season starts and appropriate preventive action can be taken before the patient experiences any symptoms.

### ***2.3 Association Rules Mining Method***

The following example will illustrate the application of data mining in the CDSS for managing asthma in school age children. Table 1 represents a sample of daily log for a child with severe asthma, who needs a bronchodilator and regular peak flow measurements. This data is collected and entered into the system by the patient's caregiver or medical professional at the school and stored into a central database. Over the time, the data accumulates and it is interesting to see how different factors impact the course of the patient's disease. The goal of the patient data analysis is to identify the relationships among the collected clinical variables and determine the factors that lead to worsening or improving of the asthma symptoms in order to establish a better control over the disease.

Using the terminology of the Association Rules Mining Method, each row of the table is itemset  $I = \{i_1, i_2, \dots, i_n\}$  and the entire table represents the transactional set or database [5]. While the typical Association Rules method associates a Boolean variable to each item: "1" if the item is present in the transaction, otherwise "0", in our case each item can have multiple responses. Association rule is an implication of the form  $X \Rightarrow Y$ , where  $X, Y$  are sets of some items in  $I$ , and  $X, Y$  do not have common items. The set of items  $X$  is called antecedent (left-hand-side or LHS) and the set of items  $Y$  is consequent (right-hand-side or RHS) of the rule [5].

While the number of potential rules is huge, we are interested only in generating the rules with a real world value, i.e. the rules that satisfy specific constraints. First of all, there is a syntactic constraint, which restricts the items that can appear in a rule: we may be interested only in rules that have specific items  $I_x$  as antecedent and specific items  $I_y$  as consequent [5]. In our case, we are interested in the rules where "Peak Flow Reading" is different than Green" and "Asthma Symptoms" different than "None". In other words, our "goal is to find factors that are leading to worsening of the asthma symptoms in order to control them.

**Table 1** Daily Log of Asthma Patient

Date	Time	Peak Flow Reading	Asthma Symptoms	Physical Activity	Whether condition	Medication and Dose
1-Sep-10	Morning	Green	None	None	Sunny	Ventolin; 1 puff
1-Sep-10	Noon	Red	Wheezing, Cough	Vigorous	Sunny	Ventolin; 2 puffs
1-Sep-10	Evening	Yellow	Shortness of breath	Mild	Sunny	Ventolin; 2 puffs
2-Sep-10	Morning	Green	None	None	Rain	Ventolin; 1 puff
2-Sep-10	Noon	Red	Cough	Vigorous	Sunny	Ventolin; 2 puffs
2-Sep-10	Evening	Red	Wheezing	Vigorous	Rain	Ventolin; 2 puffs

To select which of the interesting rules are worth consideration, we have to identify the rules that satisfy user-specified minimum threshold on the rule's statistical significance and rule's strength, determined by the following constraints [6]:

- Support of a rule  $\text{supp}(X)$ , corresponds to statistical significance and indicates how frequently its items appear in the database
- Confidence of a rule  $\text{conf}(X \Rightarrow Y)$  corresponds to strength of the rule and indicates the probability that if the left hand side appears in a T, also the right hand side will.

In our example, we are going to evaluate the rule:

$(\text{PA}=\text{"vigorous"}) \Rightarrow (\text{PFR}=\text{"Red"})$ , where PA stands for Physical Activity and PFR is Peak Flow Reading. The support of the rule  $(X \Rightarrow Y)$  is the percentage of rows that have  $\{(PA=\text{"vigorous"}), (PFR=\text{"Red"})\}$ , which according to Table 1 is  $3/6=0.5$ . The support of the antecedent X is the percentage of rows that have  $\{(PA=\text{"vigorous"})\}$ , which is  $3/6=0.5$ .

The confidence is calculated with the following formulas:

$$\text{conf}(X \Rightarrow Y) = \text{supp}(XUY) / \text{supp}(X) \quad (1)$$

$$\text{conf}((\text{PA}=\text{"vigorous"}) \Rightarrow (\text{PFR}=\text{"Red"})) = 0.5 / 0.5 = 1 \quad (2)$$

This means that the rule is 100% true. From this, the following verbal rule can be generated: "If the patient has vigorous physical activity, the peak flow reading is in the Red zone". While exercise is one of the common triggers of asthma attacks, it can not be applied directly to all asthma patients and that is why this rule has to be stored in the individual patient database. Having this information, the system can generate a warning to the patient to limit the physical activity and the doctors can recommend an increased bronchodilator dose before the physical activity.

The Association Rules method extracts the rules in two steps [6]:

- Frequent Itemset Generation – identify all itemsets X, whose  $\text{support}(X) > \text{minsupport}$ , where  $\text{minsupport}$  is a user specified threshold value

- Rule Generation – generate rules ( $X \Rightarrow Y$ ) from the frequent itemsets, whose  $conf(X \Rightarrow Y) > minconf$ , where  $minconf$  is a specified minimum confidence that has to be satisfied by the rules.

While the second step is pretty straight forward, the frequent itemsets generation is computationally expensive, since it involves searching all possible itemsets ( $2^n - 1$ ), where  $n$  is number of items. To overcome this problem, many effective algorithms have been developed, such as Apriori [7]. It uses a breadth-first search strategy to count the support of itemsets and a candidate generation function to exploit the downward closure property of support.

### 3 Conclusions and Future Work

This paper presents the idea of using data mining methods to extract rules from the patient individual data collected on a regular base for asthmatic children. These rules are stored in an individual patient knowledge base, which is supposed to be used in combination with the general knowledge base by the clinical decision support system for managing asthma in school-age children.

To illustrate the application of data mining in generating patient specific rules, we used the well-known data mining method – Association Rules. Despite the broad application of this method in many areas and the extensive research being performed on developing and improving the Association Rules algorithms, there are still some major drawbacks of this method, such as: huge number of discovered rules, obtaining non interesting rules and low algorithm performance [8]. Keeping in mind these problems, we are in process of exploring the other available data mining methods in order to identify the one that will best fit our system. At the same time, our goal is to identify the area of possible improvements and eventually create an enhanced data mining algorithm, which will be used not only by our system, but also beneficial for the KDD in general. Another question that our future research needs to answer is how to evaluate the generated rules.

The possibility of applying data mining in the clinical decision support and the healthcare in general promises to make a significant impact in the medicine and it well deserves the research effort to uncover its full potential.

### References

1. Swaroop, P., Golden, B.: Data Mining: Introduction and a Health Care Application (unpublished)
2. Jeffrey, A.K., Kern, J.A., Kernstine, K.H., Tseng, B.T.L.: Autonomous Decision Making - A Data Mining Approach. *IEEE Transactions On Information Technology in Biomedicine* 4(4), 274–284 (2000)
3. Bellazzi, R., Zupanb, B.: Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics* 77, 81–97 (2008), <http://www.intl.elsevierhealth.com/journals/ijmi>

4. Koh, H.C., Tan, G.: Data mining applications in healthcare. *J. Healthc. Inf. Manag.* 19(2), 64–72 (2005)
5. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. In: *SIGMOD 1993*, pp. 207–216 (1993)
6. Kulkarni, S.: Association Rules in Data Mining (September 2010), <http://www.authorstream.com/Presentation/sushiltry-108428-association-rules-data-mining-science-technology-ppt-powerpoint/>
7. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: *Proceedings of the 20th International Conference VLDB*, Santiago, Chile, pp. 487–499 (September 1994)
8. Moreno, M., Segrera, S., López, V.: Association Rules: Problems, solutions and new applications. *Actas del III Taller Nacional de Minería de Datos y Aprendizaje*, Tamida, 317–323 (2005)

# Spatial Influence Index Effects in the Analysis of Epidemic Data

Fatima F. Santos and Nelson F.F. Ebecken

**Abstract.** In the context of epidemic analysis, the aim of this paper is to present the development of a spatial influence index that considers the spatial neighborhood relation, expressed by distance and direction bands in the analysis of non-spatial attribute values in order to predict an epidemic evolution. The studied case is the evolution of aids propagation in the city of Rio de Janeiro, Southeastern of Brazil. The prediction of this epidemic data, using the spatial influence index has shown an improved performance when compared to the traditional approaches without this information. Despite some instability of the model, it is able to forecast the epidemic evolution. We conclude that aids evolution analysis should take into account object proximity and socio economic indicators to obtain useful knowledge.

**Keywords:** data mining, geographical information systems, aids, epidemic evolution.

## 1 Introduction

The idea that health is associated to the environment and people's life conditions is very old. However, only after the development of the social medicine, in the century XIX, systematic researches gave subsidies to that theory. The document known as Lalonde report [1], positions the health issue from a sociopolitical, technical, economical and biological perspective through the so called determinants of health: people's conditions and lifestyles, environmental situation and biological development.

Reliable monitoring of the evolution of aids epidemic has become a politically vital task. The assessment of properties and processes of aids is a major issue in

---

Fatima F. Santos · Nelson F.F. Ebecken  
Federal University of Rio de Janeiro – Brazil  
e-mail: fatimaferrao@uol.com.br, nelson@ntt.ufrj.br

the Brazilian Management of Health Program. The management and control of aids is a complex multidisciplinary task requiring the adequate approaches and techniques. One of the most overwhelming important information in aids analysis is people's sexual preferences. Another important question is movement and contamination, since multiple agents come from different neighborhoods in a dynamic movement around the city. Usually, the information about movement is rare in healthy data bases. One motivation of this study is to develop a predictive model based on static information, considering neighborhood where each one with HIV virus lives.

Possible demographical, socio economical and cultural differences influenced the evolution of aids rates in the last decades. The improvement of health indicators reflects two important tendencies of socio economical and political dynamics: from one side, we have got the progresses gotten in the city, reflected in their social and sanitarians indicators [2]; from the other one, we have the regional heterogeneousness in social classes, that are expressed in deep social dissimilarities that prevent the city to sustainability grow.

The increased availability of digital maps and spatially enabled applications has created an unprecedented opportunity to incorporate geographic factors into decision making process and analysis. For the exposed reasons, the opportunity of using more appropriate computational representations to capture the properties of the aids proliferation was observed[2]. The geographical information systems (GIS) offer a wider group of data and algorithms structures capable of representing the great diversity of space conceptions and mainly, the movement of the epidemic in the geographic space[3].

The main objective of this work is to (1) develop spatial influence index, as an alternative neighborhood specifications for better understanding evolution of a phenomenon in geographical space (2) select the best set of variables to a predictive model (3) develop a model to predict the proliferation of the epidemic (4) analysis the effects of spatial influence index in aids data.

## 2 Problem Description

The studied case is the evolutionary phases of the HIV infection in the city of Rio de Janeiro, Brazil. The current epidemic we face is quite complex, resultant of the existence of regional sub-epidemics or defined according to the nature of the different social interactions. Therefore it is necessary to evaluate it under diverse and complement perspectives, renewing and refining permanently their analysis instruments.

On account of the countless and deep mutations in the evolutionary phases of the HIV infection, the detailed exam of the epidemic tendencies should mandatorily combine data originated from HIV cases already registered - a consolidated picture of the past with those originated from life style as socio economic conditions, sanitary level, sexual relationships and others. Considering this motivation, the main objective of this study is to develop a predictive model of evolutionary phases of the HIV infection based on health determinants.



Related to the spatial analysis, two cuts were established. The first one, aiming to study the incidence rate temporal evolution according to the municipal district traditional division into neighborhoods. The second, through others characteristics of the residential neighborhoods of the studied cases: population size, poverty concentration measured by the rate of heads of the family with month income lower than the minimum wage and others.

Data was provided by the Notification Damage Information System and Hospital Internment Information System of the Health Ministry. Data was supplied by the Municipal Secretary of Health including all aids notifications in the municipality of Rio de Janeiro from 1982 to 2007. All notified cases of persons thirteen years old or over, with one year diagnosis in the period 1982 to 2007, were considered. Aiming to reduce the influence of notification delay on temporary tendencies, the year of 2007 was the last considered for this analysis.

### 3 Spatial Influence Index

#### 3.1 Measurement of Proximity

The proximity matrix, widely applied in spatial statistics, is probably the most popular measure in literature. In the spatial proximity matrix  $W$ , each element,  $w_{ij}$ , represents a measure of the proximity between areas  $A_i$  and  $A_j$ . Criteria for  $w_{ij}$  value calculations may be based on the distance between the centroids of two areas or on the shared boundaries between  $A_i$  and  $A_j$ , or on a combination of both. The boundary criterion considers  $w_{ij}$  equal to 1, if  $A_j$  shares boundary with  $A_i$ , or 0 otherwise. Thus, for a set of  $n$  areas  $\{A_1, A_2, \dots, A_n\}$ ,  $w_{ij} = 1$ , if  $A_i$  shares a common side with  $A_j$ . Otherwise,  $w_{ij} = 0$ , as Figure 1, [4].

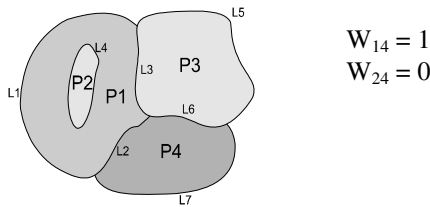


Fig. 1 Example of boundary criterion.

Once the spatial proximity criterion is defined, the spatial dependence of a data set may be determined. A simple way to measure the variation of data spatial tendency is to calculate the average value of the neighbors. This calculation provides a first approximation of spatial variability. The spatial moving average, although it presents spatial patterns and tendencies, does not measure the spatial dependence, i.e. it does not assess the attribute variation in relation to spatial distribution of the areas or how the values are correlated in the space. The most used concept

to evaluate this correlation is spatial autocorrelation, which measures how much the observed value is dependent of the values of the same attribute in neighbor locations. Countless techniques, each with its own strong and weak points, are used to measure spatial dependence. The Global Moran’s Index (I) and the Geary’s Index (C) are the two most known measures[5].

Considering a certain proximity matrix W, Global Moran’s Index(I) is expressed by Equation 1, where n is the number of regions;  $y_i$  is the value of the attribute in region i;  $\bar{y}$  is the attribute average value in the regions under consideration;  $w_{ij}$  are the elements of the spatial proximity matrix.

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left( \sum_{i=1}^n (y_i - \bar{y})^2 \right) \sum_{i \neq j} w_{ij}} \tag{1}$$

Equation (1) may be simplified [ $N(\mu = 0$  and  $\sigma^2 = 1)$ ] and W changed in such a way that the sum of the elements in each line of the spatial proximity matrix is equal to 1, according equation (2).

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{2}$$

Moran’s Index (I) is a measure of the spatial autocorrelation used to detect distancing from a random spatial distribution. Each attribute deviation from the mean is multiplied by the deviations of the neighborhood, obtained from the spatial proximity matrix that represents the spatial dependence of the involved areas. The index evaluates whether the connected areas present greater similarity in relation to the indicator under study than expected from a random pattern. The null hypothesis is of complete spatial randomness. As a correlation coefficient, values usually vary from -1 to +1, scaling the existent degree of spatial correlation. Positive values indicate a direct correlation and negative values an inverse correlation. Small values denote little correlated regions and high values strong correlated regions.

Autocorrelation modeling has countless benefits. A simple way to show those benefits is through regression equation. If dependent variables  $Y_i$  are supposed to be auto correlated, that is,  $Y_i = f ( Y_j )$ , for every i different from j, the regression equation should be modified to  $Y = a W y + b X + e$ , where W is the proximity matrix. After introducing the proximity matrix term W, the residual error will be less influenced by the spatial autocorrelation, consequently reducing the difference between real and expected values.

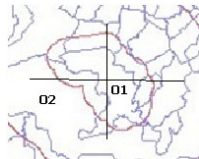
### 3.2 Definition of Spatial Influence Index

Global spatial autocorrelation indexes such as Moran's (I), supply a single value as a measure of the spatial association for the whole data set, characterizing the whole region under study. However, usually a spatial autocorrelation analysis limited to specific regions allows a better understanding of the phenomenon. In this way, local spatial association indicators are used, defining a specific value for each object, allowing a decomposition of the global index of spatial association.

The index that is being proposed was developed based on the concept of local spatial correlation. However, it introduces two characteristics of spatial relations: direction and distance. The index measures the regular modification of a non-spatial attribute while it moves away from an area following a certain direction. To reach this goal, it considers the spatial neighborhood relation, expressed by distance and direction bands in the analysis of non-spatial attribute values of a neighborhood. In this way, the spatial influence is valued through a metrics that takes into account neighborhood relations in the attribute space and in the physical space. The direction and distance characteristics may also be combined by logical operators to express a more complex neighborhood relation and consequently to obtain more specific results in spatial data mining tasks.

Distance spatial relation is intuitive. A previously defined criterion of minimum and / or maximum distance was considered. The distance between polygons was classified in a distance band. Thus, the classification in a certain band, takes into account the distance between centroids. The distance bands are one thousand meter, followed by six, ten, twenty and thirty thousand meters.

Otherwise, the direction spatial relation is not so simple. In order to define the direction spatial relation of an object O2 in relation to an object O1, a point that represents the object O1 (centroid) was considered. The polygon centroid was considered as the origin of a virtual system of coordinates, whose quadrants and planes define the directions, as shown in Figure 2. In order to define O2 direction more than half O2 points should be in the respective plane area. There is not just one direction between two objects, thus, the more exact was considered. Between south and southeast direction, for example, southeast was chosen. At the beginning of the research, the identification of the direction and the calculation of the distance of each object (polygon) in relation to all others in the region under study were stored in a database, since those values are static.



**Fig. 2** Virtual system of coordinates to define distance and direction between objects.

### 3.3 Spatial Influence Index Framework

The proposed spatial influence index, identified as  $IFd$ , is a measure of spatial association calculated by object.  $IFd$  can be calculated in relation to any non-spatial attribute, provided that the attribute is numerical. To calculate  $IFd$ , any polygon distant up to thirty thousand meters from  $O$ 's centroid in the standard specific direction  $d$ , is considered a neighbor of the object  $O$ , according Equation 3. Directions are standardized in zero, forty five, ninety, a hundred and thirty five, a hundred and eighty, two hundred and twenty five, two hundred and seventy and three hundred and fifteen degrees, and the respective indexes identified as  $IF0$ ,  $IF45$ ,  $IF90$  until  $IF315$ .

$$IF = \frac{(y_i - \bar{y}) \sum_{j=1}^n W_{ij} (y_j - \bar{y})}{\sum_{j=1}^n (y_j - \bar{y})^2} \quad (3)$$

where:

$n$  is the total number of polygons;

$y_i$  is the value of the attribute considered in polygon  $i$ ;

$y_j$  is the value of the attribute considered in polygon  $j$ ;

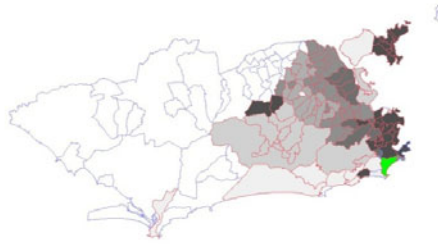
$\bar{y}$  is the average value of the attribute;

$w_{ij}$  are the polygons neighbors to polygon  $y_i$  with  $w_{ij}$  equal to 1, for every polygon  $y_j$  whose distance to  $O_i$  centroid is less than thirty thousand meters in a direction  $d$  considered.

Each grey shade, in Figure 3, shows the set of polygons considered in the calculation of  $IFd$ , in each respective direction. The straight line in Figure 3 indicates the polygons in the ninety degrees direction from an imaginary coordinate axle with coordinates  $x$  and  $y$  (0,0) in object  $O144$  and with maximum distance equal to thirty thousand meters. These polygons were considered in the calculation of  $IF90$ , considering a particular attribute of object  $O144$ .

Regular modifications of non-spatial attributes may be identified through regression analysis, where the independent variable ( $x$ ) measures the distance between objects  $O2$  and  $O1$ . The dependent variable ( $y$ ) measures the difference between the values of a non-spatial attribute for  $O2$  and  $O1$  objects. When the absolute value of the correlation coefficient is significant, it denotes a spatial tendency for the specific attribute from  $O1$  object.

Spatial tendency analysis may be useful in the data pre-analysis step. In this case, linear regression may be considered to map the spatial relation between areas, based on the premise that the influence of certain phenomenon in its neighborhood is usually linear or may be transformed into a linear model as, for example, exponential regression. Linear regression calculation through distance bands in pre-defined directions and an attribute, identify correlations between this attribute and the geographical space, since regression analysis satisfies a minimum correlation coefficient.



**Fig. 3** Spatial influence index IF90 of object O144, Dark grey shade indicates considered polygons.

Spatial Influence Index proposed is an alternative measure of spatial relation neighborhood, based on distance and direction, to an attribute under consideration. In this case, each studied area will probably have a neighborhood direction that is more significant than the others. The value of the spatial influence index in this specific direction will be an important input in a spatial and temporal predictive model.

Considering that neighborhood definition through direction and distance are variables that rarely change, a tree structure was defined to store spatial relation data as presented in Table 1.

**Table 1** Tree structure with spatial relation data by object.

Spatial relation data			
Key	Neighbour	Distance	Direction
O1	O2	1000	0°
...	...	...	...
O1	O144	22500.5	315°

Considering *max-dist* and *dist* as real numbers and direction *d*, the neighborhood concept was defined for a geographical database with distance *max-dist* and direction *d* as:  $O_{neighborhood} = \{ (o1, o2, dist, d) \mid o1, o2 \in DB, o1 \text{ dist } o2 \leq dist\text{-max} \text{ and } o1 \text{ } d \text{ } o2\}$ .

IFd is calculated over a set of all the objects connected to object O through neighborhood that satisfies the selection criterion: distance and direction. After identifying the set of objects that satisfies this criterion, the index value is calculated.

## 4 Results

Ideally, a model for predicting evolution of epidemic would take as input the relevance measure of each variable to discriminate object of the study. Different

methods have been used to identify the data dimension as factor analysis, principle component analysis and others. In this paper, the method proposed by [6] was applied. This method consists basically in the relevance measure of each variable to discriminate the system of the study. The relevance of the component  $j$  is defined as Equation 4.

$$R_{(x_j)} = \frac{\sum_{i=1}^N \|(x_i) - (x_i')\|^2}{N} \tag{4}$$

where:

- $x_i$  is the value of the output neuron, after artificial neural network (ANN) training, for each standard presented to ANN.
- $x_i'$  is the value of the output neuron, when attribute  $j$  is replaced by its average value, calculated for all the standards presented to ANN.
- $N$  is the number of the standards.

The ANN output is the rate of contamination in the neighbor. The term  $x_i'$  is the output vector when component  $j$  is replaced by its average. The result is a map of the relevance of the input variables, which measures how much the ANN output changes when an attribute is replaced by its average. The most important variables have high values of relevance.

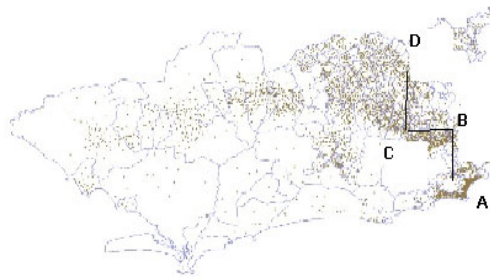
Several sanitary and socio economic indicators such as population size and poverty concentration were considered in the analysis of relevance. The attributes with higher values of relevance were: (1) rate of families with their own house; (2) rate of heads of the family with more than eight years of study; (3) rate of heads of the family with less than two salaries (4) rate of heads of the family with superior graduate; (5) rate of heads of the family is a woman. The rate calculation considered the sum of cases of contamination by HIV virus, chronologically ordered, for each object O. Finally, values of IFd from IF0 to IF315 (for attribute rate of population infected) were considered as input data for each period of time (month-year). As an example, considering object O144 and attribute rate of population infected, values for IF90 are 6.31 and 6.90 (1988 and 1999, respectively), as shown in the Table 2.

**Table 2** Neighborhood's IFd

Neighborhood's O144			
Object	Direction	1988	1999
O144	IF90	6.31	6.90
...	...	...	...
O144	IF225	0.90	0.15

These values express a strong spatial influence in this direction. On the contrary, IF225, for the same period of time, are 0.90 and 0.15, which means a weak spatial influence, for the attribute considered, in this direction.

Through the values of  $IF_d$  for each standard direction (considering attribute *rate of population infected*) for each object, and period of time, it was possible to draw the trajectory from A (first case of aids) to D. It was possible through the identification of the higher values of  $IF_d$ , which express the main influence areas of the evolution (since we considered rate of population infected). Note that rate of population infected, and respective  $IF_d$  are different concepts. One neighbor with high values for attribute rate of infection and low values for  $IF_d$ , probably exists for a hot spot or a new focus of epidemic, since region in direction  $d$  considered has no effect on it. In this case, it is an heterogeneous area for the attribute considered (in this example, rate of population infected). From A to B and from B to C, through analysis of  $IF_d$ , it was possible to identify the movement of aids. This algorithm was repeated until point D. This is not an automatic, but an iterative method and the user must decide when to stop; usually when the value of  $IF_d$  does not change significantly. Figure 4, from A to D, represents the direction movement.



**Fig. 4** Direction of epidemic using values of  $IF_d$ .

#### ***4.1 Predictive Spatial Temporal Model***

One of the objectives of this study is to implement a robust predictive model, i.e. a model that produces results under a wide range of operating conditions. Ideally, this predictive model should use as input data, determinants of health: people's socio economic conditions, lifestyles and environmental situation, such as sanitary conditions. An appropriate source of such models is machine learning methods such as neural networks (ANN) and genetic algorithms, because they are designed for analyzing poorly structured domains.

A computational system was used to define two ANNs of multi-layer perceptron type, with backpropagation algorithm. The first ANN considered in its input layer, (1) rate of contamination (for each period of time month-year), (2)  $IF_d$  based on rate of contamination (for each month-year) and the other attributes considered relevant as presented in the second paragraph. For the second ANN, a well known spatial correlation index, Moran's Index (I), was used instead of  $IF_d$  to compare the results. The other inputs were the same. In computer science data mining, it is expected that the proposed method goes through some performance

evaluation. The considered cycle for time series analysis was twelve months. The output neuron is rate of contamination for each object (neighbor) for a month in advance. For each ANN, five best results are shown in Tables 3 and 4. The ANN with *IFd* in the input layer had a better performance.

**Table 3** Predicted Rate for contamination by neighborhood (ANN's output layer) with *IFd* in the input layer.

	Rate year.1	Rate year.2	..	..	Rate year.5
Data Mean	9.391317	9.391317			9.391317
Data S.D.	6.808094	6.808094			6.808094
Error Mean	0.018791	0.001765			0.004036
Error S.D.	0.185951	0.106722			0.061593
Abs E. Mean	0.133079	0.06581			0.0398
S.D. Ratio	0.027313	0.015676			0.009047
Correlation	0.999633	0.999878			0.999959

**Table 4** Predicted rate for contamination by neighborhood (ANN's output layer) with Moran's Index (*I*) in the input layer.

	Rate year.1	Rate year.2	..	..	Rate year.5
Data Mean	9.391317	9.391317			9.391317
Data S.D.	6.808094	6.808094			6.808094
Error Mean	0.357959	0.428495			0.295187
Error S.D.	5.837753	5.810031			5.767635
Abs E. Mean	5.059816	5.037736			5.043155
S.D. Ratio	0.857472	0.8534			0.847173
Correlation	0.515874	0.521542			0.531349

## 5 Conclusion

This work presented the proposal of a spatial influence index, which generates a neighborhood concept through spatial relation characteristics of distance and direction between objects. The use of the suggested spatial influence index in data mining tasks resulted in a better performance when compared to one of the most popular spatial correlation index. The performance of the index proposed was measured through the reduction of assessment and test errors in an artificial neural



network used to predict the contamination rate. Such improvement is explained by the implicit inclusion of spatial characteristics through *IFd* neighborhood particularly expressed by the proposed index.

The available data is part of the long term Health Ministry Aids monitoring program whose intention is to serve as much for the population under greater risk to HIV exposure as for the general population, indisputably more vulnerable nowadays than in the origins of the epidemic activity.

The scenery of epidemic changed in the last years. This study, clearly show, the direction of aids proliferation from south region of the city (higher IDH) through north region (lower IDH). Although aids is transmitted mainly through sexual relations, it has its main causes in socio economic aspects. Reduction of rate of infection could not depend on the control of one sector or technology. It was not possible, until now, to deploy effectively the control of the disease in poor regions, the main affected. It is necessary to act differently with different neighbors, which must have preferential and rigorous action.

The objective of this paper is to show the development and performance evaluation of an spatial influence index (*IFd*) and the results obtained in temporal series analysis.

**Acknowledgments.** The authors thank CAPES, CNPq and FAPERJ for financial support.

## References

1. IBGE (2007),  
<http://www.ibge.gov.br/home/estatistica/populacao/tabuadevida/2007/default.shtm> (accessed January 24, 2009)
2. Lalonde, M.: A new perspective of the health of Canadians: a work document. In: Proceedings of Ottawa Health Conference, Ago (1978)
3. Miller, H.J., Han, J.: Geographic data mining and knowledge discovery. Taylor & Francis, London (2001)
4. Osei, F.B.: Spatial Statistics of Epidemic Data: The Case of Cholera Epi-Demiology in Ghana. D.Sc. thesis. University of Twente, ITC Prin-Ting Department, The Netherlands (2010)
5. Druck, S., Carvalho, M.S., Camara, G., Monteiro, A.V.M.: Spatial. Analysis of geographic Data, Brasília, Embrapa (2004) ISBN 85-7383-260-6

# EPON's Research on Transmission of Sampling Value in Process Level of Digital Substation

Qiu Ming-yi and Xu Xi-dong

**Abstract.** Non-conventional instrument transformers and optical fiber communication are utilized in secondary system of digital substation. Sampling value signals are transmitted to Merging Unit through proprietary links and Merging Unit transmits the signals to IED equipments by means of IEC 61850-9-1 or IEC 61850-9-2. Merging Unit consists of multiple communication interfaces, connecting voltage and current signals input from non-conventional instrument transformers. In this paper, Ethernet Passive Optical Network (EPON) is proposed to connect non-conventional instrument transformers with Merging Unit. Meanwhile, the definition of OLT-Merging Unit and ONU-Instrument Transformer is put forward. Under circumstance of synchronization between Merging Unit and non-conventional instrument transformer, TDMA scheme is adopted in transmission of sampling value and performance of such data transmission scheme is analyzed. The use of EPON for sampling value's transmission is compatible with the Ethernet data transmission from Merging Unit to IED. Finally, the deployment of EPON in digital substation is suggested.

## 1 Introduction

With the application and improvement of Standard IEC 61850, non-conventional instrument transformer and intelligent primary equipment, digital substation technology has gradually come to maturity and relative coverage would be more extensive. In IEC 61850, substation automation system (SAS) is logically defined as three levels: process level, bay level and station level [1].

---

Qiu Ming-yi

College of Electrical Engineering, Zhejiang University, Hangzhou, China  
e-mail: sjtuqmy@163.com

Xu Xi-dong

College of Electrical Engineering, Zhejiang University, Hangzhou, China  
e-mail: xxd@zju.edu.cn

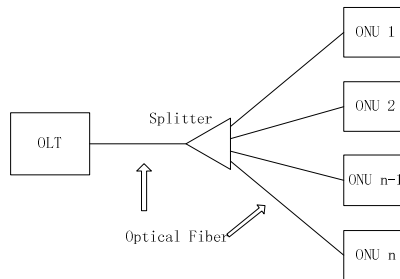
Merging Unit (MU) is the main equipment in process level, which receives sampling value outputs from non-conventional instrument transformer [2-3]. Sampling value signals are sent to the IED equipments from MU by means of IEC 61850-9-1 or IEC 61850-9-2. [4-6] proposed new methods using FPGA to realize MU. A novel method using ARM-based controller was introduced to implement the main functions of MU in [7]. The concept of MU is firstly mentioned in Standard IEC 60044-7/8 and MU is expanded in IEC 61850-9. IEC 61850-9 defines a new physical unit in SAS, namely MU, whose inputs consist of seven current signals and five voltage signals from instrument transformers. Sampling value signals are encapsulated into standardized frame format and transmitted to secondary protective and control equipment from MU. Optical fiber and copper wire can be used to connect instrument transformers with MU.

Based on the existing research on MU and digital substation, EPON is proposed to transmit sampling value in process level of digital substation. OLT-Merging Unit (OLT-MU) and ONU-Instrument Transformer (ONU-IT) are defined to realize data transmission. The performance of sampling value transmission in process level using EPON is analyzed. At last, the deployment of EPON in digital substation is presented, including backup and redundancy issues.

## 2 EPON

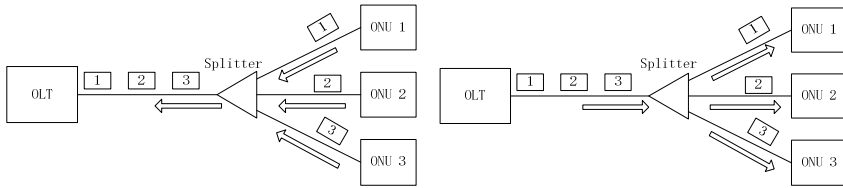
EPON is based on Ethernet technology, using point to multi-point architecture and passive optical fiber transmission. EPON combines low-cost, high-bandwidth Ethernet equipment with widespread fiber-optic network technology.

The typical architecture of EPON is tree topology, depicted in Fig.1. EPON system consists of Optical Line Terminal (OLT), Optical Network Unit (ONU) and Optical Distribution Network (ODN).



**Fig. 1** Basic Structure of EPON.

Transmission of signals is bidirectional. In the upstream direction, signals sent from ONU1 would only arrive at OLT and the other ONUs cannot communicate with ONU1 directly. To avoid the collision among all ONUs, Time Division Multiple Access (TDMA) is adopted in the upstream direction. Fig.2 shows the upstream and downstream transmission mode of EPON.



**Fig. 2** Upstream and Downstream Transmission Mode of EPON

OLT gathers various signals from superior network and sends to the end users, exactly to ONU. On the other hand, signals collected from ONU will be transmitted to the superior network via OLT. The superior network and OLT could be connected via switch or router. ONU receives signals sent by OLT and send messages to OLT similarly. ONU is able to connect other terminal equipments as long as proper communication interfaces are designed to connect different equipments. ODN plays a role in connecting OLT and ONU and provide the transmission path between OLT and ONU. In EPON system, ODN is mainly composed of optical fibers and passive optical splitters. Point to multi-point network topology is implemented by splitters, along with the protocol layer control of EPON.

IEEE 802.3ah [8] specifies the relationship between protocol layer of EPON and OSI reference model layer. The protocol layer model of EPON is based on IEEE 802.3 Ethernet hierarchy model.

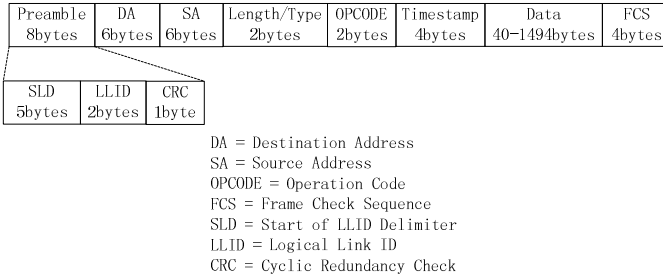
### 3 Epon in Substation

Ethernet has been the most sophisticated LAN technology at present, regarded as the mainstream in LAN realm. In addition, EPON is modified based on the IEEE 802.3 protocol and Ethernet frame format of EPON would guarantee compatibility. EPON provides a rate of 1Gbps, which meets the requirement of the multiple-channel current and voltage signal sampling in digital substation. With advantage of low loss and high bandwidth, optical fiber has a long transmission distance up to 10km. Electromagnetic and lighting interference can be prevented and transmission quality is guaranteed.

#### A Equipment Definition

According to IEC 61850-9, MU receives the multi-channel current and voltage signals. Instrument transformer consists of conventional and non-conventional ones. The design of MU is flexible at least for now, including analog and digital inputs. The multiple input signals require external clock synchronization and sampling values are synchronized via timestamp. MU send sampling value packets to IED equipments through IEC 61850-9-2 process bus.

In this paper, EPON is presented to achieve transmission of sampling value. Compared to the conventional transmission mode, network transmission is nearer to the field apparatus, exactly to instrument transformers. Sampling value signals are transmitted from instrument transformers to MU via EPON. The EPON frame format is shown as Fig.3.



**Fig. 3** Frame Format of EPON

Several equipments are defined in this paper which adopt EPON technology, i.e. OLT-Merging Unit (OLT-MU) and ONU-Instrument Transformer. ONU-Instrument transformer mainly refers to ONU-Current Instrument Transformer (ONU-CT) and ONU-Voltage Instrument Transformer (ONU-VT). OLT-MU combines the function of OLT with MU, including following functions: OLT-MU synchronizes ONU equipments by broadcast. OLT-MU receives sampling value packets sent by ONU-CT and ONU-VT. Sampling value packets are transmitted from OLT-MU to IED equipments via IEC 61850-9-2 process bus. OLT-MU offers management functions over ONUs, such as Time Slot allocation. ONU-CT is a combination of ONU and CT. ONU-VT is a combination of ONU and VT. Following functions are included: ONU-CT and ONU-VT convert high power signals to low power signals. ONU convert sampling value into EPON frame format signals. The frame signals are transmitted to the OLT-MU. ONU equipments are synchronized to OLT-MU.

## ***B Analysis about Data Transmission***

Different from the MU defined in IEC 61850, OLT-MU requires few interfaces because one optical network interface can receive multiple electric signals sent by several instrument transformers. Therefore, EPON-based transmission network saves the optical fiber cabling and the scalability is much better.

Each ONU is required to avoid collision when transmitting data in the upstream direction because of the point to multi-point EPON architecture. In order to solve the collision problem, TDMA transmission scheme is adopted in the upstream direction. Because of pre-allocating bandwidth, TDMA is more suitable for real-time periodic data transmission compared to CSMA/CD method [9].

In order to achieve more accurate synchronization, IEC 61588 PTP protocol is recommended in EPON systems and IEC 61588 is based on Ethernet. The PTP protocol includes clock offset correction and path delay correction, realizing the microsecond to sub-microsecond accuracy [10]. Hence, IEC 61588 synchronization will satisfy the timestamp precision requirements of sampling value in process level.

According to the point to multi-point transmission mode of sampling value, the following analysis is presented.

Considering 50Hz sine wave,  $f=50\text{Hz}$ ,  $T=20\text{ms}$ . Supposing sampling frequency  $f_s=1000\text{Hz}$ , sampling point in a cycle is  $N=1000/50=20$ . The time interval of sampling point is  $t=1\text{ms}$ . Assuming that ONU-CT outputs tri-phase current signals and ONU-VT outputs tri-phase voltage signals, OLT-MU receives six channel signals in all. Since the time interval of sampling point is 1ms, six-channel signals should be transmitted to OLT-MU in this interval and signals avoid collision. The time interval of 1ms can be divided in to seven time slots. The first time slot is 0.4ms and the second to seventh time slot are 0.1ms. The first time slot is used to synchronize between OLT-MU and ONU-CT/ONU-VT. The second to seventh time slots are used to transmit the tri-phase current and voltage signals. The second time slot transmits the A-phase current sampling value. Sampling value  $I_A$  is converted to Ethernet frame and sent to OLT-MU in this time slot. If frame packets of  $I_A$  is 128bytes=1024bits, the required transmission rate is  $R=1024\text{bits}/0.1\text{ms}=10.24\text{Mbps}\ll 1\text{Gbps}$ , which satisfies the transmission requirements. Similarly, B-phase current and C-phase current signals are transmitted in the third and fourth time slot. The tri-phase voltage signals are transmitted from the fifth to seventh time slot.

Supposing the sampling point in each cycle is  $N=80$ , the time interval of sampling point is  $t=0.25\text{ms}$ . In that way, the time interval is also divided into 7 time slots. The first time slot is used to synchronize. The next six time slots are used to transmit current and voltage signals. Each time slot for transmission of current and voltage signals is 0.025ms. The required transmission rate is  $R=40.96\text{Mbps}$ , which is satisfied.

The adoption of EPON simplifies the architecture of communication line and reduces the number of interfaces of MU. At the same time, EPON supports high sampling rate because of the TDMA transmission mode.

#### 4 Deployment of EPON in Substation

The topology between OLT-MU and ONU-Instrument Transformer is point to multi-point. The following deployment is proposed. The tri-phase ONU-CT and ONU-VT are connected to the OLT-MU directly via optical fiber and splitter.

The trunk fiber and OLT-MU are more important in transmission system, so backup protection or redundant configuration should be paid attention to.

Considering the trunk backup and redundancy, OLT-MU is single-port equipment and connects double trunk optical fibers. Once the main trunk optical fiber fails, the backup optical fiber takes action. Such configurations can response to the

fault of the trunk optical fiber and do nothing on the condition that the port of OLT-MU breaks down.

Considering trunk and port backup and redundancy, OLT-MU is dual-port and each port connects trunk optical fiber. Each pair of port and trunk optical fiber is independent from each other. Redundant configurations can response to both the trunk optical fiber fault and port fault of OLT-MU.

## 5 Conclusion

With development of EPON in recent years, the application of EPON in industrial automation will come into reality. The transmission efficiency of data link layer is improved due to the compatibility of Ethernet.

The use of EPON changes the transmission mode of sampling values in digital substation. The point to multi-point topology accords with the connection between MU and instrument transformers. Transmission of sampling values adopts TDMA mode and the details of TDMA is analyzed above, which is feasible theoretically. At the same time, accurate synchronization is essential to TDMA. Consequently, IEC 61588 is introduced for synchronization and IEC 61588 PTP protocol is compatible to Ethernet. Finally, several deployment scheme of EPON is recommended.

## References

- [1] IEC 61850 Ed1, Communication Networks and Systems in Substations, 14 Parts (2003-2005), <http://www.iec.ch>
- [2] Li, J.-H., Zheng, Y.-P., Gu, S.-D., Xu, L.: Application of Electronic Instrument Transformer in Digital Substation. *Automation of Electric Power Systems* 31(7), 94–98 (2007)
- [3] Zhang, J., Yuan, Z., Li, Y., Guo, Z.: A new method to realize the relay protection of AOCT following IEC61850. In: *International Conference on Power System Technology, PowerCon 2006, Chongqing, October 22-26*, pp. 1–5 (2006)
- [4] Yin, Z.-L., Liu, W.-S., Yang, Q.-X., Qin, Y.-L.: New Method for Implementing the Synchronizaion of Merging Unit According to the IEC 61850 Standard. *Automation of Electric Power Systems* 28(11), 57–61 (2004)
- [5] Zhang, H., He, J.H., Kilmek, A., Bo, Z.Q.: Design of a real-time substation communication system for integrated protection. In: *2008 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America, Bogota, August 13-15*, pp. 1–5 (2008)
- [6] Yin, Z.L., Liu, W.S.: A novel FPGA-based method to design the merging unit following IEC 61850. In: *2004 International Conference on Power System Technology, PowerCon 2004, November 21-24, vol. 1* (2004)
- [7] Liu, J., Li, K., Yang, H.: The design of a merging unit of electronic transformer based on arm. In: *42nd International Universities Power Engineering Conference, UPEC 2007, Brighton, September 4-6*, pp. 712–716 (2007)

- [8] 802.3ah-2004 IEEE Standard for Information Technology- Telecommunications and Information Exchange Between Systems- Local and Metropolitan Area Networks-Specific Requirements Part 3: Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks (2004)
- [9] Wang, Z., He, K., Li, K., Wang, H., Sun, D.: Hard Real-Time Communication Based on Shared Ethernet. In: 4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008, Dalian, October 12-14, pp. 1-5 (2008)
- [10] IEC 61588:2009(E) IEEE Standard for a precision clock synchronization protocol for networked measurement and control systems (2009)



# Intelligent Clinical Information Systems for the Cardiovascular Diseases

Nan-Chen Hsieh, Huan-Chao Keh, Chung-Yi Chang, and Chien-Hui Chan

**Abstract.** Medical informatics has become an extensive field of research and many of these approaches have demonstrated potential value for improving medical quality. The clinical information system (CIS) is primarily used to enhance the quality, safety, and efficiency of patient care, as well as operational and surgical workflow. The aim of this study was to develop a web-based cardiovascular CIS that can be used as a tool for tracking individual and broad population-based surgical care information. We used UML technique to analyze the special charts and workflow of the creation of the registries. The built CIS allowed groups of potential patients to be selected for investigatory clinical trials based on their data, and surgeons can provide reasonable conclusions and explanations in uncertain environments.

## 1 Introduction

In the past, surgeons analyzed patients' surgical risk and evaluated their surgical outcomes by using their own experience or by reviewing the relevant literature. Recent studies have shown that computerized medical case monitoring and

---

Nan-Chen Hsieh

Department of Information Management, National Taipei University of Nursing and Health Sciences, Taiwan, ROC

e-mail: nchsieh@gmail.com

Huan-Chao Keh · Chien-Hui Chan

Department of Computer Science and Information Engineering, Tamkang University, Taiwan, ROC

e-mail: keh@cs.tku.edu.tw, emmacc@gmail.com

Chung-Yi Chang

Division of Cardiovascular Surgery, Department of Surgery Heart Center, Cheng-Hsin General Hospital, Taiwan

decision support systems can be used to assist surgeons in the diagnosis of disease, optimize surgical operations, aid in drug therapy and decrease the cost of medical treatment[1]. Traditional paper-based CIS has been evolved in response to changes in clinical work practices and technology. In order to enhance data manipulating and sharing ability, the emergency of computer-based CIS was used to store and present patient and treatment related data[2]. The Loyola Open-Heart registry[3] is an early stages fully operational database that designed to input, modify, verify, maintain, update and analyze its raw data. For commercial CISs, Dendrite (<http://www.e-dendrite.com/>) provides a sophisticated clinical outcomes database management system with the customizable clinical datasets ability, and include fundamental analytic and risk modeling functions. MetaVision (<http://www.imd-soft.com/>) designed the CISs that support comprehensive patient information for clinical workflow, and the surveillance scheme notifies surgeons when a set of pre-defined conditions has been met.

The core of a cardiovascular CIS is the computerized registry [3] which assisting surgeons and nurses in the improvement of clinical outcomes. When the registry is more complete, its value is greater. Complete registries are beneficial for monitoring and tracking patients. Each patient in the registry database has their individual file folder, and each file folder contains data on patients' preoperative characteristics, risk factors, details of the surgical procedure, physical characteristics of the heart and postoperative physiological and laboratory findings. Moreover, the CIS should provide a search function for the registries, extract the most relevant data for the patient, and accumulate and summarize the data using disease-specific or population-based information.

The information collected can be used by surgeons and patients to evaluate whether or not the surgical procedures are likely to be successful. Similarly, information on surgical mortality and postoperative morbidity is also important for the selection of low-risk patients for an operation and for counseling patients about the risks of undergoing surgical operations.

The research objectives of this study include in the following:

- To develop the cardiovascular registries, for collection, storage, analysis and interpretation of the data from patients with cardiac-related diseases.
- To create a four-layer architecture of a CIS and apply intelligent decision-making for the assistance of surgical operations.
- To describe the clinical and technical aspects of the CIS and outline its capabilities for clinical decision support.

## **2 The CIS Architecture**

It is important to establish a CIS that allows surgeons, laboratory technicians and nurses to offer, and patients to obtain, complete medical services. All of above information can be constructed by fully electronic patient health records. For example, outpatient department system includes the referral note and outpatient medical record which written in a SOAP format note. SOAP is an acronym stand

for subjective, objective, assessment and plan which is written to improve communication among all those caring for the patient to display the assessment, problems and plans in an organized format. Aside from patient data maintenance, this study uses statistics and data mining to assist surgeons' research and understanding of the factors than can affect cardiovascular treatment methods, as well as to establish a knowledge base for the comparison of biochemical data of patients at the preoperative versus postoperative stages. The CIS architecture described in Fig. 1:

1. The *Monitoring Layer*: The medical plan and the patient's case management system are developed at this layer and the relative information is stored.
2. The *Surveillance Layer*: The relative decision models are implemented at this layer. The key variables and specific instances required for the decision influencing model are acquired at the model construction layer.
3. The *Model Construction Layer*: The clinical decision analysis model and the ensemble classifiers are developed at this layer. The basic aim of this layer is knowledge discovery.
4. The *Patient Integration Layer*: A hospital information system (HIS) is generally composed of several subsystems. An enterprise master patient index (EMPI) will create a unique identifier for each patient and maintain a mapping to the identifiers used in each record's respective system. To create a master index can be used to obtain a complete and single view of a patient.

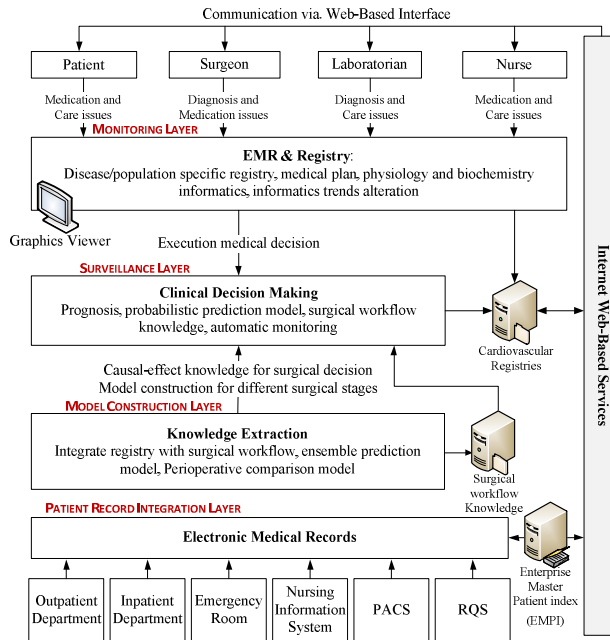


Fig. 1 The cardiovascular CIS architecture.

### 3 The Design of CIS with Surgical Workflow

Aside from surgical registries, surgical workflow is an abstraction of the surgical procedure. The buildings of surgical workflow models are useful for several users including surgeons, nurses, care givers, and medical engineers. The study of surgical workflows is a novel approach for the describing of surgical interventions. According to a rigorous surgical workflow analysis, the retrieved data can be used to optimize the surgical procedures, to guide the development of new surgical assist systems, to compare different surgical approaches, or for exploratory data analysis. Surgical workflow provides high amount of potentially useful information. The choice of suitable sampling of information restricts, and various surgical questions could be answered by the help of surgical workflow knowledge. For example, a simple question is to use surgical workflow analysis to find answers to typical questions arising during optimization of surgical interventions: “Does using intraoperative surgical method A is less mortality then surgical method B?”, “Surgeon A has fewer complications than surgeon B for similar cardiac surgeries.

For cognitive system engineering, clear and accurate definition of modeling objectives and work domains is a crucial initial step [4]. The surgical workflow could be decomposition into a fact hierarchy (Fig. 2), and the surgical process can be recorded according to the preoperative, intraoperative or postoperative, three different surgical timelines. Each of these aspects can be examined at different levels of granularity, and can be focused on roles, devices, tasks, systems, states, information or knowledge. In this study, the surgical work domain that we try to model is related to the Cardiac surgical treatment performed in different surgical timeline for the individual patient. This necessary information, which includes the data shall be recorded, major surgical steps and their sequencing, was collected by interviews with surgeons and studies the different surgical timeline’s reports. Our objective aimed at making the knowledge used by surgeons. That is, the development of knowledge base is a crucial and required step in modeling the cardiac surgical domain [5].

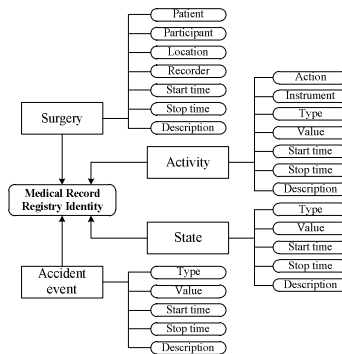


Fig. 2 Decomposition of the surgical workflow into a fact hierarchy

## 4 Intelligent Prediction of Postoperative Morbidity

The use of machine learning algorithms for the building of predictive data mining models has become widely accepted in medical applications. Various models including BNs, ANNs, and SVMs have been tested in a wide variety of clinical and medical applications [6]. In contrast to ordinary models try to learn one hypothesis from the training dataset, ensemble models try to construct a set of hypotheses and combine results to use, which is a machine learning paradigm which multiple models are trained to solve the same problem [7].

The development of this prediction model proceeds as in Fig. 3. During the training process, the same training data set is used for all individual models in order to reduce the diversity among individual models, keeping in mind that, in an ensemble model, it is important to construct appropriate training data sets that maintain good balance between accuracy and diversity among individual models. The model selection scheme, designated a stacking scheme, is a mixture of stacking and cross-validation that is chosen in order to improve overall classification by combining models trained on randomly generated subsets of the entire training set. For the implementation of models, we chose BayesNet, MultilayerPerceptron and SMO as base models in WEKA, and StackingC as ensemble method. All the default parameters in WEKA were used.

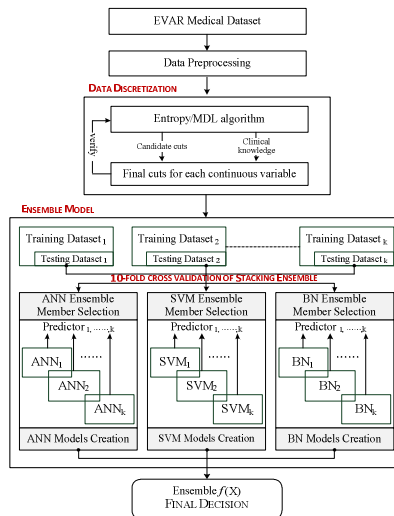


Fig. 3 Proposed architecture for ensemble model development

## 5 Discussion and Conclusion

After we developed the CIS, surgeons realized the value of CISs, and highlighted the benefits that might be faced by hospitals who implement the systems. Some of

the conclusions for this study are: (1) Integration of existing systems: We integrated CIS with legacy hospital information system, such as HIS, PACs, Laboratory, etc. This poses opportunities to provide well medical care for patients. (2) Structured registry information: Cardiovascular registries enabled surgeons easier to maintain and to search relevant information. The result information is relevant and cross-reference, making fewer mistakes would be made due to illegible data entry. (3) Easy acquire patient data: CIS provided convenient access to medical records at ubiquitous care. Internet web-based access improves the ability to remotely access medical data. (4) Improved prescription and patient safety: The developed CIS assisted surgeon to prognosis patients' results before surgery, which could help to assess the risk of patient. (5) Electronic-based patient record maintenance: By the help of CIS, surgeons can efficient maintain the required special chart, and keep the most recent information for their patients.

**Acknowledgment.** This research was partially supported by National Science Council of Taiwan (NSC 99-2410-H-227-002-MY2).

## References

1. Uğuz, H.: A Biomedical System Based on Artificial Neural Network and Principal Component Analysis for Diagnosis of the Heart Valve Diseases. *Journal of Medical Systems*, 1–12 (2010)
2. Fraenkel, D.J., Cowie, M., Daley, P.: Quality benefits of an intensive care clinical information system. *Critical Care Medicine* 31, 120–125 (2003)
3. Wright, J.G., Bieniewski, C.L., Pifarre, R., Gunnar, R.M., Scanlon, P.J.: A database management system for cardiovascular disease. *Computer Methods and Programs in Biomedicine* 20, 117–121 (1985)
4. Rasmussen, J., Pejtersen, A.M., Goodstein, L.P.: *Cognitive Systems Engineering*. John Wiley & Sons (1994)
5. Chandrasekaran, B., Josephson, J.R., Benjamins, V.R.: What are ontologies and why do we need them? *IEEE Intelligent Systems* 14, 20–26 (1999)
6. Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: Current issues and guidelines. *International Journal of Medical Informatics* 77, 81–97 (2008)
7. Polikar, R.: Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6, 21–45 (2006)

# Forecast the Foreign Exchange Rate between Rupiah and US Dollar by Applying Grey Method

Tien-Chin Wang, Su-Hui Kuo, Truong Ngoc Anh, and Li Li

**Abstract.** Indonesia is the largest archipelago in the world and also the largest nation in South East Asia. As a big country which is rich of natural resources, there have been many foreign investors from Asia and Europe taking the opportunities to invest in this country. This study applied the GM (1,1) model [6,9,10,11] to predict the exchange rate between Rupiah and US Dollar from 2010/01/01 to 2010/10/18. The results presented that the average accuracy of the forecasting model exceeds 99.65%.

## 1 Introduction

This research collected 207 lots of exchange rate between Rupiah and US Dollar for nearly one year from 2010/01/01 to 2010/10/18 on the website of financial analysis engineering. Grey theory, advanced originally by Professor Deng(1982)[6], which has been generally recognized and applied by many academicians, chosen Grey prediction as an ability forecasting means because of having relatively low data requirements, and a GM model can be constructed from a sample of just four pieces of data. The forecast method is significant by using the transformed Grey rolling modeling mechanism, which is a metabolism technique that updates the input data by discarding old data for each cycling. This rolling

---

Tien-Chin Wang · Su-Hui Kuo · Truong Ngoc Anh  
Department of International Business, National Kaoshiung University of Applied Sciences,  
Kaohsiung, Taiwan  
e-mail: tcwang@cc.kuas.edu.tw

Li Li  
Department of Urban Planning & Economic Management, Harbin Institute of Technology,  
China  
e-mail: Liszsz@hit.edu.cn

modeling mechanism provides a means to guarantee input data are always the most recent values. An expression introducing comparison of rolling modeling data and primitive data of forecasted results show their average residual error different from rolling modeling GM (1,1) Furthermore, this study presents methodologies for projecting the most correctly predicts of the exchange rate between Rupiah and US Dollar by analyzing the precision of the Grey forecasting model.

The whole aim of the accurate prediction the exchange rate between Rupiah and US Dollar is to learn more about currency crises in Indonesia to make sure that Indonesia is a good foreign exchange market for foreign investors to invest in this country.

The Grey forecasting is based on GM (1,1) model. It is a time series prediction model encompassing a group of differential equations adapted for parameter variance and a first-order differential equation [5,16]. GM (1,1) can be denoted by the functions [12,16, 17,18].

## 2 Methodology

Grey theory, developed by Deng in 1982[6], is suitable for short-term prediction and does not rely on a statistical method. The Grey forecasting method has been successfully applied in many areas of research including finance, engineer, agriculture, and management. Grey generating, Grey relational analysis, Grey prediction, Grey decision, and Grey controller are the mainly methodology of Grey system theory.

The prediction method is significantly by applying the transformed Grey rolling modeling mechanism, which is a metabolism technique that updates the input data by discarding old data for each cycling. This rolling modeling mechanism provides a means to guarantee input data are always the most recent values. In the prior study (Yao and Chen, 2001) [3], this research is applied the general GM (1,1) with the technique of rolling modeling to forecast the foreign exchange rate between Rupiah and US Dollar for nearly one year from 2010/01/01 to 2010/10/18. An expression introducing comparison of rolling modeling data and primitive data of forecasted results show as Figure 1. The average residual error different rolling modeling GM (1,1) (i.e. Method 1: choose first six continuous data to predict the 7th of output value, 2nd to 7th consecutive data to predict the 8th output value and thereafter, Method 2: Predict the 8th of output value by adopting first seven consecutive data, 2nd to 8th consecutive data to forecast the 9th output value and thenceforth. Furthermore, the study presents methodologies for projecting the most accurately predicts of the exchange rate between Rupiah and US Dollar by testing the precision of the Grey forecasting model.

### A *Gray GM (1,1) Prediction Model*

Grey prediction is applied in this paper in order to construct a forecasting model. Professor Deng [6] proposed the Grey system theory to build a Grey model for



forecasting. Their results displayed the ability of Grey system theory to effectively deal with incomplete and uncertain information.

**B Accumulated Generation Operation (AGO)**

Accumulating obtained systematic regularity discrete time-series data

$$x^{(0)} = (x^{(0)}(1), x^{(0)}(2), \dots, x^{(0)}(n)) \tag{1}$$

$x^{(1)}$  is  $x^{(0)}$  one-order accumulated generating operation (AGO) sequence, that is

$$x^{(1)} = \left( \sum_{k=1}^1 x^{(0)}(k), \sum_{k=1}^2 x^{(0)}(k), \dots, \sum_{k=1}^n x^{(0)}(k) \right) \tag{2}$$

**C Inverse-Accumulated Generating Operation (IAGO)**

$$\hat{x}^{(0)}(k) = x^{(1)}(k) - x^{(1)}(k-1)$$

**1. Grey Derivatives:**

$$z^{(1)} = 0.5 x^{(1)}(k) + 0.5 x^{(1)}(k-1) \tag{3}$$

**2. Gray Difference Equation Derivatives:**

The first order differential equation of GM (1,1) model is  $dx/dt + ax = b$ , where  $t$  denotes the independent variables in the system,  $a$  represents the developed coefficient,  $b$  is the Grey controlled variable, moreover  $a$  and  $b$  denoted the parameters requiring determination in the model. When a model is constructed, the differential equation is  $x^{(0)}(k) + az^{(1)}(k) = b$ , including  $k = 2, 3, \dots, n$ , where  $a, b$  denoted standby substantial number, this differential equation  $x^{(0)}(k) + az^{(1)}(k) = b$  is called as GM (1,1) model.

$$Y_N = BA, \quad B^T Y_N = B^T BA, \quad A = (B^T B)^{-1} B^T Y_N$$

Furthermore, accumulated matrix  $a$  and  $b$  are as below expand equations:

$$a = \frac{\sum_{k=2}^n z^{(1)}(k) \sum_{k=2}^n x^{(0)}(k) - (n-1) \sum_{k=2}^n z^{(1)}(k)x^{(0)}(k)}{(n-1) \sum_{k=2}^n [z^{(1)}(k)]^2 - \left[ \sum_{k=2}^n z^{(1)}(k) \right]^2} \tag{4}$$

$$b = \frac{\sum_{k=2}^n [z^{(1)}(k)]^2 \sum_{k=2}^n x^{(0)}(k) - \sum_{k=2}^n z^{(1)}(k) \sum_{k=2}^n z^{(1)}(k)x^{(0)}(k)}{(n-1) \sum_{k=2}^n [z^{(1)}(k)]^2 - \left[ \sum_{k=2}^n z^{(1)}(k) \right]^2} \tag{5}$$

**3. Whitening Equation:**

$$x^{(1)}(k) = \left[ x^{(1)}(1) - \frac{b}{a} \right] e^{a} e^{-ak} + \frac{b}{a} = \left[ x^{(1)}(1) - \frac{b}{a} \right] e^{-a(k-1)} + \frac{b}{a}$$

$$x^{(1)}(k+1) = \left[ x^{(0)}(1) - \frac{b}{a} \right] e^{-ak} + \frac{b}{a}$$

**4. Utilize Inverse-accumulated generating operation (IAGO) equation as below:**

$$\hat{x}^{(0)}(k+1) = x^{(1)}(k+1) - x^{(1)}(k) = (1 - e^{-a}) \left[ x^{(0)}(1) - \frac{b}{a} \right] e^{-ak} \tag{6}$$

**3 Data Analysis and Results**

**A Raw Data**

Raw data of quality of exchange rate between Rupiah and US Dollar from 2010/01/01 to 2010/10/18 as following Fig. 1:



**Fig. 1** Raw data of exchange rate between Rupiah and USDollar from 2010/01/01 to 2010/10/18

**B Rolling Modeling to Forecast the Exchange Rate**

In this primary research (Yao and Chen, 2001) [2], using the general GM(1,1) with the formula of rolling modeling to forecast the exchange rate between Rupiah

and US Dollar from 2010/01/01 to 2010/10/18. An expression introducing comparison of rolling modeling data and primitive data of forecasting results lay out as Fig.2. The average surplus error different rolling modeling GM (1,1).

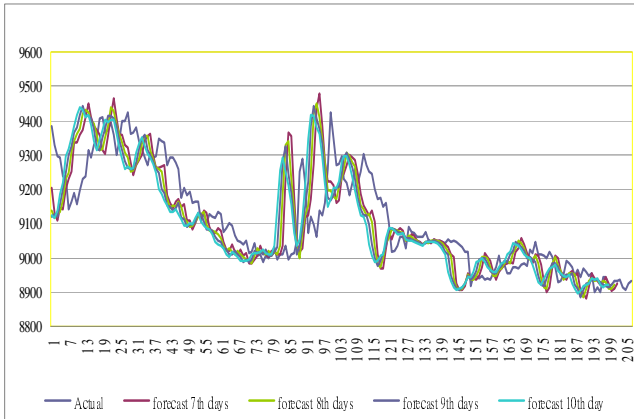


Fig. 2 The average surplus error different rolling modeling GM (1,1)

### C Average Residual Error

According to used the general GM (1,1) with the way of rolling modeling to predict the quality of exchange rate between Rupiah and US Dollar from 2010/01/01 to 2010/10/18. The actual output valuation of average residual error from forecasted value and fundamental value, the summarization the survey outcomes as Fig 3.

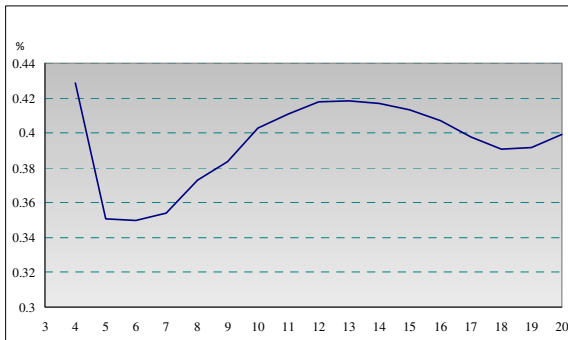


Fig. 3 Average residual error of the exchange rate between Rupiah and US Dollar from 2010/01/01 to 2010/10/18

Using the general GM(1,1) with the approach of rolling modeling to forecast the exchange rate between Rupiah and US Dollar from 2010/01/01 to 2010/10/18. An expression introducing comparison of rolling modeling data and primitive data of forecasted results describe as from Fig.1, Fig.2, Fig.3. The average residual error different rolling modeling GM(1,1). The results show that the Grey forecasting model exhibits highest prediction accuracy average accurate rate at the average residual error of the Grey forecasting model is almost over 0.35%. The above statistics confirm the efficiency of the proposed forecasting model. Specially, 99.65% is on inputting 6 consecutive data in rolling model for amount of exchange rate between Rupiah and US Dollar, their average residual error is lower than 0.35%. In such a way, the forecasting method for significantly by applying the transformed Grey model is the most accurately predict the output values of trading.

## 4 Conclusions

This paper used Grey forecasting method to predict the exchange rate between Rupiah and US Dollar from 2010/01/01 to 2010/10/18. The results presented that the average accuracy of the forecasting model exceeds 99.65%. The model thus clearly has high prediction validity and is a viable goal for forecasting exchange rate in Indonesia. Moreover, its forecasting shows that the exchange rate currency in Indonesia is more and more flexible, and it will make a new plan for foreign country to invest in many types in the future. With the highest prediction accuracy and average residual error 0.35% of the Grey forecasting model, this means that the Grey forecasting is the most efficient method to accurate predictive the trend of exchange rates in Indonesia.

**Acknowledgement.** This research was supported in part by the National Science Council of the Republic of China under the grant NSC99-2410-H-151-020.

## References

1. Yao, A.W.L., Chen, J.H.: The optimal parameters design for Grey forecasting of electric demand control. In: Proc. of the 2001 Conference on Control Systems, pp. 137–143. IEEE Control Systems Society, Taipei (2001)
2. Alessandro, P., Schinasi, G.J.: EMU and International Capitals Market: Structural Implications and Risks, IMF Working Paper (1997)
3. Chiang, J.S., Wu, P.L., Chiang, S.D., Chang, T.J., Chang, S.T., Wen, K.L.: Introduction of Grey System Theory. Gao-Li Publication, Taiwan (1998)
4. Deng, J.L.: Control problems of Grey systems. *Systems and Control Letters* 5, 288–294 (1982)
5. Deng, J.L.: Grey system fundamental method. Huazhong University of Science and Technology Wuhan, China (1982)
6. Deng, J.L.: Grey prediction and decision. Huazhong University of Science and Technology, Wuhan, China (1986)
7. Deng, J.L.: Introduction to Grey system theory. *The Journal of Grey System* 1(1), 1–24 (1989)

8. Deng, J.L.: *The Course on Grey System Theory*, p. 91. Huazhong University of Science & Technology Publish House, Wuhan (1990)
9. Deng, J.L.: *The Essential Methods of Grey Systems*. Huazhong University of Science and Technology Press, Wuhan (1992)
10. Hsu, C.I., Wen, Y.U.: Improved Grey prediction models for trans-Pacific air passenger market. *Transp. Plann. Technol.* 22, 87–107 (1998)
11. Hsu, L.C.: The comparison of three residual modification model. *J. Grey System, Assoc.* 4(2), 97–110 (2001)
12. Lin, C.-T., Hsu, P.-F.: Forecast of non-alcoholic beverage sales in Taiwan using the Grey theory. *Asia Pacific Journal of Marketing and Logistics* 4(14), 3–12 (2002)
13. Lin, C.-T., Yang, S.-Y.: Forecast of the output value of Taiwan's opto-electronics industry using the Grey forecasting model. *Technological Forecasting & Social Change*, 177–186 (2003)
14. Renn, J.C., Chen, W.J.: Control of a servo hydraulic positioning system using state space controller with Grey forecasting. *JSME*
15. Tseng, F.M., Yu, H.C., Tzeng, G.H.: Applied hybrid Grey model to forecast seasonal time series. *Technol. Forecast. So.* 67, 291–302 (2001)
16. Wen, J.C., Huang, K.H., Wen, K.L.: The study of a in GM(1,1) model. *J. Chin. Inst. Eng.* 23(5), 583–589 (2000)
17. Whyman, P.: The impact of Economic and Monetary Union on British business. *European Business Journal* 13(1) (Quarterly 2001)
18. Wu, J.H., Lauh, C.R.: A study to improve GM(1,1) via Heuristic method. *J. Grey System* 10(3), 183–192 (1998)
19. Yeh, M.F., Lu, H.C.: A new modified Grey model. *J. Grey System* 1(8), 209–216 (1996)
20. Cheng, S.-R., Chen, J.-C., Wu, W.-H., Liu, Y.-T.: Forecasting Equity Fund Performance via GA. *International Journal of Innovative Computing, Information and Control* 4(2), 333–339 (2010)
21. Mussa, M.: The exchange rate, the balance of payments, and monetary and fiscal policy under a regime of controlled floating. *Scandinavian Journal of Economics* 78, 229–248 (1976)
22. Rapach, D., Wohar, M.: In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance* 13, 231–247 (2006)
23. Sweeney, R.: Mean reversion in G-10 nominal exchange rates. *Journal of Financial and Quantitative Analysis* (2006)
24. Faust, J., Rogers, J., Wright, J.: Exchange rate forecasting: The errors we've really made. *Journal of International Economics* 60, 35–59 (2003)

# Simulating Patent Knowledge Contexts

Jussi Kantola and Aviv Segev

**Abstract.** Patent users such as government, inventors, and manufacturing organizations strive to identify the directions in which the new technology is advancing. The organization of patent knowledge in maps aims at outlining the boundaries of existing knowledge. This article demonstrates the methodology for simulating alternative knowledge contexts beyond the border of existing knowledge. The process starts with extracting knowledge from patents and applying self-organizing maps for presenting knowledge. The knowledge extraction model was tested earlier on patents from the United States Patent and Trademark Office. A demonstrator tool is then used to perform “what-if” type of analysis/simulation on the clusters in the dataset to see alternative knowledge contexts for the new knowledge “entity”. This may open up new directions and help to plan for the future. The demonstrator tool has been tested earlier on other datasets. The proposed knowledge context simulation shows promise for the future development and applications.

## 1 Introduction

Government services attempt to forecast main research areas that would be beneficial to fund. Similarly, researchers try to map knowledge and identify possible gaps that would be relevant to the advancement of science. The extraction of relevant information from patents allows the analysis of main research areas and the mapping of the current topics of interest. The creation of such a service, which allows analysis of patents over time, will provide decision makers with a top level overview of the direction of new inventions. In addition, the service could support knowledge seekers in identifying worthwhile research tasks. A knowledge map service can enable a researcher to identify the need for specific research directions considered “hot”. In addition, research and government

---

Jussi Kantola · Aviv Segev

Department of Knowledge Service Engineering, KAIST - Korea Advanced Institute of Science and Technology, 291 Daehak-ro, Yuseong-gu, Daejeon, Korea  
e-mail: {jussi, aviv}@kaist.edu

institutions providing funding will be able to preplan with a longer horizon and divert research funds to necessary fields. Knowledge maps of patents can assist in the classification of directions of research in the past and in the attempt to predict future discovery directions.

The patent service is unique compared to other knowledge based services because of the requirement to identify whether similar knowledge exists as opposed to the need to locate knowledge. Contemporary knowledge based services are based on using existing information, while the patent support service is required to assist in identifying similar domains and patterns that would result in the rejection of a patent request. Furthermore, patents in different countries are not classified under one classification system.

The premise of the patent system lies in its mutual benefit to both the inventor and the public. In return for full public disclosure, a patent offers certain rights to an inventor for a limited period of time, during which the inventor may exclude all others from making, using, importing, or selling his or her invention. The patent is published and disseminated to the public so that others may study the invention and improve upon it. The constant evolution of science and technology, spurred by the monetary incentive the patent system offers to inventors, strengthens the economy. New inventions lead to new technologies, create new jobs, and improve our quality of life.

The work analyzes patents to create an outline of knowledge. The research aims at building a simulation model that predicts the identification of new critical research areas that can exponentially speed up the overall research in specific fields. The patent project analyzes patents from the United States Patent and Trademark Office. The patent analysis process is the following: Existing patents -> Patent knowledge extraction -> Knowledge representation using Self-Organizing Maps -> Knowledge representation analysis / simulation. The first step includes parsing existing patents text. In the analyzed cases the entire patent description was used. Alternative methods include parsing the patents according to dates or according to topics. The model includes three major steps: patent knowledge extraction and knowledge representation using self-organizing maps. The patent knowledge extraction extracts key features from each patent. The knowledge representation creates an evolving map using the self-organizing map technique to represent the patents research topics. The last step involves the analysis /simulation of the knowledge representation map evolution.

One benefit of using simulation is the insight gained into the importance of variables and their interaction [2] [8]. Knowledge elements in the existing patents resemble those variables. Another benefit is the possibility to experiment with new policies before their implementation [2] [8], thus saving both money and time. Knowledge contexts can be tested before implementation. Simulation allows answering “what-if” questions that are important when new systems are being developed [2] [8]. These benefits seem attractive for knowledge context simulation as well.

The next section describes the Self-Organizing Maps. Section 3 describes the SIMU\_SOM demonstrator tool. Section 4 describes knowledge extraction process and section 5 describes patent knowledge context simulation approach Section 6 presents a discussion and some concluding remarks.

## 2 Self-Organizing Maps

The Self-Organizing Map (SOM) is a two-layer unsupervised neural network that maps multidimensional data onto a two dimensional topological grid [6]. The data are grouped according to similarities and patterns found in the dataset, using some form of distance measure, usually the Euclidean distance. The results are displayed as a series of nodes on the map, which can be divided into a number of clusters based upon the distances between the clusters. Since the SOM is unsupervised, no target outcomes are provided, and the SOM is allowed to freely organize itself, based on the patterns identified, making the SOM an ideal tool for exploratory data analysis [3].

According to Kaski and Kohonen [5], exploratory data analysis methods, such as SOM, are like general-purpose instruments that illustrate the essential features of a data set, such as its clustering structure and the relations between its data items. The SOM perform visual clustering of data [3]. More information about the methodology of applying self-organizing maps is provided by Back et al. [1]. The most commonly used method for visualizing the final self-organizing map is the unified distance matrix method, or U-matrix [11]. The U-matrix method can be used to discover otherwise invisible relationships in a high-dimensional data space. It also makes it possible to classify data sets into clusters of similar values. Feature planes, representing the values in a single vector column, are used to identify the characteristics of these clusters [3]. This helps in explaining the meaning of the SOM.

## 3 SIMU\_SOM Demonstrator

To display the interactive approach on SOMs, the SIMU\_SOM demonstrator tool was constructed at Tampere University of Technology, Finland by Vesanen, Toivonen and Visa. SIMU\_SOM is described in [4]. The demonstrator allows the user to interactively view and comprehend the structure of the constructed SOM and to perform sensitivity analysis on the map. The prototype was coded in the Linux operating system with the Perl Toolkit (Perl). The SIMU\_SOM application takes a constructed SOM vector file as its input and draws the map for it. The vector file contains the numerical results of the whole group. Figure 1 shows a sample SIMU\_SOM screenshot to illustrate the elements of the demonstrator.

For each variable in the dataset a slider is created and presented on the right side of the window. In this case, patent context elements are the variables. The user can analyze the effect of each variable on the map position using sliders to change parameter values. A pointer, a white ball, changes its position to the closest matching node of the map as the user changes the values of the variables. The pointer shows the simulated map position of the patent application. The closest match is determined using the smallest Euclidean distance between the map node vector and the current value vector of the sliders. The labels that belong



to the current position of the pointer are shown below the map. In Figure 1, the arrows can be interpreted as follows: the starting point of an arrow is the current map position and the end of each arrow is the simulated optional target map position. This means that a person can roughly see where incremental increases in the level of context match put the new knowledge entity.

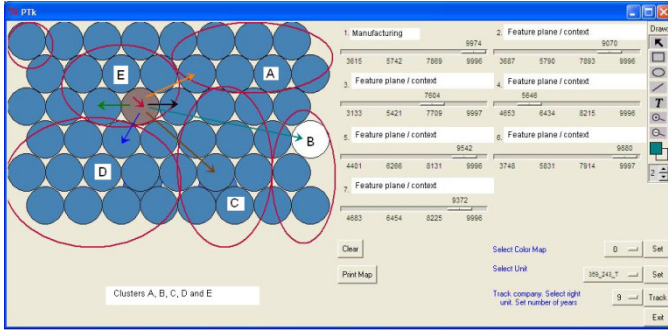


Fig. 1 The SIMU\_SOM demonstrator has the SOM and feature plane sliders [4].

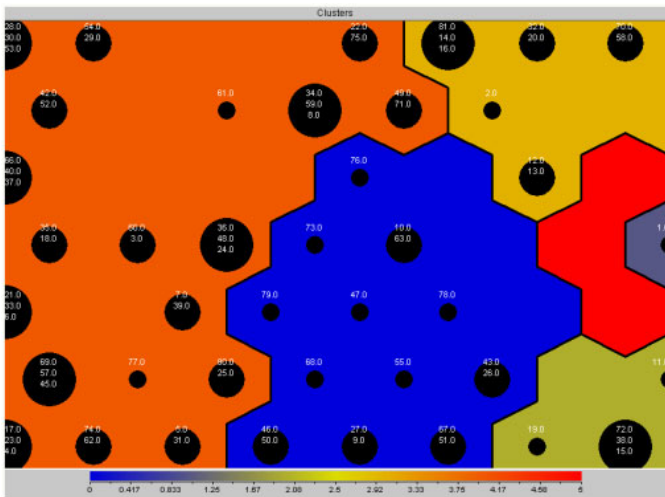
## 4 Patent Knowledge Extraction

Each claim is analyzed separately through the Domain Representation process. To analyze the claims, a context extraction algorithm can be used. To handle the different vocabularies used by different information sources, a comparison based on context is used in addition to simple string matching. For each document the context is extracted by the Patent Knowledge Extraction and then compared with the ontology concept by the Patent Domain Representation.

We define a context descriptor  $c_i$  from domain DOM as an index term used to identify a record of information [7], which in our case is a patent. It can consist of a word, phrase, or alphanumeric term. A weight  $w_i \in \mathbb{R}$  identifies the importance of descriptor  $c_i$  in relation to the patent. For example, we can have a descriptor  $c_1 = \text{Length}$  and  $w_1 = 2$ . A descriptor set  $\{\langle c_i, w_i \rangle\}$  is defined by a set of pairs, descriptors and weights. Each descriptor can define a different point of view of the concept. The descriptor set eventually defines all the different perspectives and their relevant weights, which identify the importance of each perspective.

By collecting all the different viewpoints delineated by the different descriptors, we obtain the context. A context  $C = \{\{\langle c_{ij}, w_{ij} \rangle\}_i\}_j$  is a set of finite sets of descriptors, where  $i$  represents each context descriptor and  $j$  represents the index of each set. For example, a context  $C$  may be a set of words (hence DOM is a set of all possible character combinations) defining a patent and the weights can represent the relevance of a descriptor to the patent. In classic Information Retrieval,  $\langle c_{ij}, w_{ij} \rangle$  may represent the fact that the word  $c_{ij}$  is repeated  $w_{ij}$  times in the patent.

The Patent Knowledge Extraction process uses the World Wide Web as a knowledge base to extract multiple contexts in multiple languages for the textual information. The algorithm input is defined as a set of textual propositions representing the claim information description. The result of the algorithm is a set of contexts - terms that are related to the propositions in multiple languages. The context recognition algorithm was adapted from [10] and consists of the following three steps: 1) Context retrieval: Submit each parsed claim to a Web-based search engine. The contexts are extracted and clustered from the results. 2) Context ranking: Rank the results according to the number of references to the keyword, the number of Web sites that refer to the keyword, and the ranking of the Web sites. 3) Context selection: Assemble the set of contexts for the textual proposition, defined as the outer context. The algorithm then calculates the sum of the number of Web pages that identify the same descriptor and the sum of number of references to the descriptor in the patent. A high ranking in only one of the weights does not necessarily indicate the importance of the context descriptor. For example, high ranking in only Web references may mean that the descriptor is important since the descriptor widely appears on the Web, but it might not be relevant to the topic of the patent. The external weight of each context is determined according to the number of retrieved Web references related to the concept and the number of references to the concepts in the patents. In addition, the Term Frequency/Inverse Document Frequency (TF/IDF) method analyzes the patent from an internal point of view, i.e., what concept in the text best describes the patent. The patent knowledge extraction is described in more detail in [9]. The experiments included a set of 81 patents randomly selected from the United States Patent and Trademark Office. Each patent included a vector with 43 top ranking context extracted values. The tool used to create SOMs was eSom2. The tool suggested six different clusters, Figure 2. Each cluster is represented by a different color.



**Fig. 2** The dataset of 81 patents formed six context clusters.

One example of context classification is displayed in Figure 3, which presents the self-organizing map feature plane *Manufacturing*. We can see that *Manufacturing* is relevant to only one patent and slightly relevant to 20% of the patents according to context matching index.

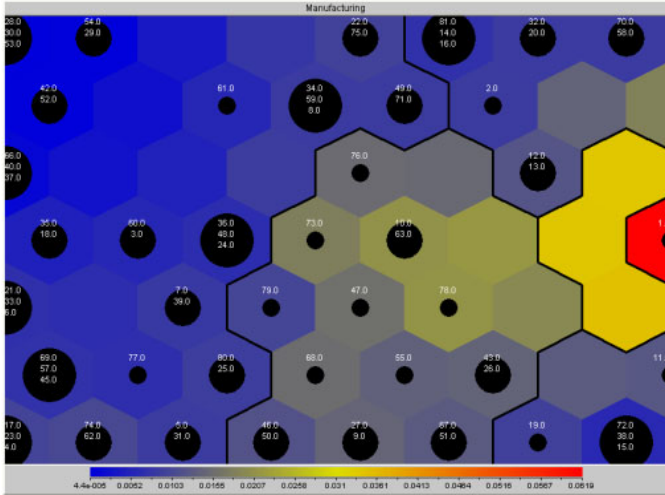


Fig. 3 Manufacturing feature plane of the SOM

## 5 Patent Knowledge Context Simulation

To visualize the values of a single variable of the SOM, it is possible to change the coloring of the map. The user can, for example, select a color map where red represents big values of a variable and blue represents low values. These colors further help the user to understand what the different positions on the map represent on a single variable level. The user can also keep track of the change in position on the map. Each change is marked to represent the previous change in position. The user can select how many changes are tracked. In the SIMU\_SOM demonstrator it is also possible to select with the mouse a desired position on the map and see the context match values that constitute that selected position. The context match values for the selected position on the SOM can be seen on the sliders when selecting a node on the map with the mouse.

By knowledge context simulation we refer to a “what-if” type of analysis of alternative knowledge contexts with the SIMU\_SOM demonstrator. The basic idea is to show the approximate effect of optional knowledge contexts. This kind of simulation provides an idea of where incremental changes in the new knowledge content would place the new knowledge entity. In simulation, the individual moves the sliders (context elements) she is willing to change or develop. The amount of movement in sliders resembles the amount of change in knowledge; a

slight increase signifies a slight change and a large change signifies a major change required in the new knowledge entity. In Figure 1, this can be interpreted as follows: the starting point of an arrow is the current knowledge context position of a new patent application and the end of each arrow is the simulated alternative knowledge context position of a new patent application. This simulation gives an idea of how much change is required in the new knowledge entity to achieve a desired knowledge context cluster of the future. The amount of desired change in context that the simulation indicates refers to the effort required in the real world. In practice, this means changing technology plans, investment plans, design plans, etc. The course of the simulation helps the individual to recognize and understand the meaning of context elements and especially to recognize such elements that are the most meaningful in the path towards desired knowledge context in the future. Context simulation can be an eye-opener for the participants. The aim is to involve participants in the course of taking action that will lead to the personal experience required to internalize something new. This is more than just providing optional vision for the future. Participatory methods usually result in good commitment and motivation.

Knowledge context simulation may save the enterprise time, money, and resources as the impact of patent plans can be roughly estimated on the computer screen first, instead of experimenting in the real world directly. Knowledge context simulation can take the quality and the meaning of planning to a new, targeted level by showing what kinds of development paths could be taken. Do the current patent content and plans meet the vision of the company? Context simulation can help in selecting those actions to which the organization and individuals can commit.

## 6 Discussion and Future Work

The patent service model described in the paper allows a self-organizing map to be created on the boundaries of existing knowledge. The model shows promise in extending the field of patent service. This paper describes a work-in-process, and we are currently working on validating and expanding the data set of the proposed joined approach. We are in the process of verifying the results predicted by SOM by tagging the patents according to their year and evaluating which contexts have become realities for those patents that are a few years old. We may be able to say something about the predicting power of context SOMs and about the “lead time” from patent context to reality. In this work, SOM produces results based on existing patents. Therefore a question arises on how we can forecast something that lies ahead in the future, based on the existing dataset embedded in the SOM. When we use context time-series in SOM we may see tendencies that show us a way that leads to “out-of-scope” knowledge or to knowledge that is beyond existing knowledge. If we can get some additional hint or clue on what will be important in the future, we individuals and our organizations can be proactive in many ways. We believe that the usefulness of the proposed patent context simulation is in the increased change to see what will be important in the future.

We are suggesting that knowledge context simulation provides increased understanding of the patent contents and patenting plans and a way to improve the existing patent applications and the overall patenting plan. Future work involves evaluating the patent search model against patents over a timeline to evaluate change in knowledge. We will finish the on-going verification, as discussed above, and explore in what other ways context simulation can be implemented on existing knowledge bases. Another direction is to extend the model to multiple languages. In addition, the demonstrator described in this paper needs further development to suit well to patent context knowledge simulation.

**Acknowledgments.** This research was partially supported by the Korean Government IT R&D Program of MKE/KEIT (10035166, Development of Intelligent Tutoring System for Nursing Creative HR).

## References

- Back, B., Sere, K., Vanharanta, H.: Managing complexity in large data bases using self-organizing maps. *Accounting Management and Information Technologies* 8(4), 191–210 (1998)
- Banks, J., Carson II, J.S., Nelson, L.B., Nicol, D.M.: *Discrete-event system simulation*, International edition. International Series in Industrial and Systems Engineering. Prentice-Hall, Upper Saddle River (2005)
- Eklund, T., Back, B., Vanharanta, H., Visa, A.: Using the self-organizing map as a visualization tool in financial benchmarking. *Information Visualization* 2(3), 171–181 (2003)
- Kantola, J., Piirto, A., Toivonen, J., Vanharanta, H.: Simulation with Occupational Work Role Competences. In: *Proceeding of Conference on Grand Challenges in Modeling and Simulation (GCMS 2009)*, Istanbul, Turkey, July 13-16 (2009)
- Kaski, S., Kohonen, T.: Exploratory data analysis by the self-organizing map: structures of welfare and poverty in the world. In: Apostolos, P.N., Refenes, Y.A.-M., Moody, J., Weigend, A. (eds.) *Neural Networks in Financial Engineering*, pp. 498–507. World Scientific, Singapore (1996)
- Kohonen, T.: *Self-Organizing Maps*. Springer, Leipzig (2001)
- Moore, C.: Descriptors. In: *Encyclopedia of Library and Information Science*, vol. 7, pp. 31–45. Marcel Dekker (1972)
- Pegden, C.D., Shannon, R.E., Sadowski, R.P.: *Introduction to Simulation Using SIMAN*, 2nd edn. McGraw-Hill, New York (1995)
- Segev, A., Kantola, J.: Knowledge Discovery System for Patents. In: *7th International Conference on Knowledge Management (ICKM 2010)*, Pittsburgh, Pennsylvania, USA, October 22-23 (2010)
- Siegel, M., Madnick, S.E.: A metadata approach to resolving semantic conflicts. In: *Proceedings of the 17th International Conference on Very Large Data Bases*, pp. 133–145 (1991)
- Ultsch, A.: Self organized feature planes for monitoring and knowledge acquisition of a chemical process. In: *The International Conference on Artificial Neural Networks*, pp. 864–867. Springer, London (1993)

# Apply Data Mining on Location-Based Healthcare System

Chen-Yang Cheng, Ja-Hao Chen, Tsuo-Hung Lan, Chin-Hong Chan, and Yu-Hsun Huang

**Abstract.** In recent years, mostly hospital substantially applied Zigbee to establish wireless sensor networks for avoiding human errors and decreasing potential healthcare safety problems. Current researches mostly focused on patients' physiological signals like their heart beats and body temperatures monitoring. However, the patient localization was seldom mentioned, which made the healthcare personnel fail to notice that patients might step into potential dangerous areas and be in danger. For that reason, the main idea of this research is to propose a two-step clustering localization algorithm combining the ZigBee Wireless Sensor Network. With this technology, a real-time monitoring healthcare system could be constructed to enhance the accuracy of tracking patients' location.

## 1 Introduction

In recent years, "patient safety" of the healthcare service quality has been an important issue. Besides adapting regular measures to do the patient safety management, major hospitals also try to apply technologies such as Zigbee to establish wireless sensor networks for avoiding human errors and decreasing potential healthcare safety problems[1,2]. Wireless sensor networks could also help to improve the management procedure and save the labor cost. This comprehensive platform would help to relieve the burden of caring patients[3,4].

Most of the wireless sensor healthcare systems would sense patients' bio-signals and upload them to the servers. The healthcare personnel could observe patients' health condition more precisely with all the long-term recorded data[3,4], but the patient localization were seldom mentioned. In fact, the information of location could be analyzed for researching medical behaviors and be used to avoid patients stepping into potential dangerous areas or any inappropriate contact between male and female patients. Although there were researches about the location algorithm, but the improvements to these researches are still required. Any inaccuracy of patient location could not only lead to unrecovered errors, but make the healthcare personnel fail to notice that patients might be in danger.

Accordingly, this research is to propose a two-step clustering localization algorithm combining the ZigBee wireless sensor network to establish healthcare system concentrating on the patient location and their bio-signal collecting. If the patient faints, the healthcare personnel could sense the abnormal bio-signal and locate this patient through the location algorithm to achieve the effectiveness of real-time monitoring. The rest of the paper is organized as follows. In the Section 2, the healthcare related literatures, concepts of location algorithm and cluster analysis will be discussed. Then, this system and its construction will be illustrated in the Section 3. The experiment and examination to the system will be demonstrated in the Section 4. Finally, the Section 5 will be the conclusion and follow-up researches.

## 2 Indoor Location Algorithm

In general, accurate localization requires algorithm using received signal from wireless devices. Localization algorithm can be divided into three major categories: triangulation[5], scene analysis[6], and proximity [7]. Most algorithms have their reliability issues because they strongly rely on variance Radio Signal Strength Indication (RSSI) measurement [8,9]. Recently, several advanced localization algorithms are proposed using reference tags concept [10,11]. The LANDMARC system adapts all known locations and applies reference tags on them for data collection [12]. In this system, the RSSI errors between all those reference tags and traced tags are calculated, and then the K-nearest reference tags are selected to estimate the locations of the traced items. Kalman system collects RSSI through reference tags to establish a distributive regional signal map. This system uses the Kalman filter repeatedly to estimate the locations of the targets [13]. According to above problem, this research considered two-step clustering, it helps to find a suitable group for calculating.

Some researches pointed out that the result of clustering was more precise by adapting both the Ward's method[3,4] and the k-means method synchronously[3,4]. In the first stage, once the number of clusters was calculated then it couldn't divide anymore. Sometimes it will cluster bad observation. The second stage was to improve the disadvantage that the individuals no longer moved in the Ward's method. In k-means method, the initial groups would affect the result. Accordingly, if the best amount of clusters could be calculated through the Ward's method and the cluster number could be the initial groups for the K-means method.

## 3 System Analyses and Design

### 3.1 Two-Step Location Algorithm

This research proposed a two-step location algorithm, it overcome the VIRE method which make too many mean less reference tags. Following are the procedures of adapting this algorithm:

Step 1: The selected readers and the amount of tags 1 should be set as a sensing area for the experiment.

Step 2: Four actual areas with real Reference tags should be divided into  $n \times n$  zones. There will be  $(n+1)^2 - 4$  reference tags in the environment.

Step 3: RSSI will be given to each virtual tag through the linear interpolation. The formulas are the following:

$$S_k(V_{a,b}) = S_k(V_{a,b}) + \frac{S_k(R_{X_i, Y_i}) - S_k(R_{X_{i+1}, Y_{i+1}})}{n-1} \tag{1}$$

$$a = X_{Ri} + \frac{X_{Ri} - X_{Ri+1}}{n-1}, b = Y_{Ri} + \frac{Y_{Ri} - Y_{Ri+1}}{n-1} \tag{2}$$

$S_k(V_{a,b})$  will be the RSSI value when the virtual tags are on the coordinate (a,b),  $S_k(R_{X_i, Y_i})$  is the RSSI of real reference tags,  $R_{X_i}$  and  $R_{Y_i}$  means on coordinate (Xi, Yi) and the k is the number of readers.

Step 4: If the difference of RSSI values between the region and tracking tag is smaller than a threshold, here will mark as a hot point. Each readers has own hot-point, and all the hotspots of other readers should be overlaid on one another. After that, the overlap should be saved.

Step 5: Use (3) for calculating E value, when the different of RSSI between virtual tags and tracking tag is small, then it got bigger E value. And using (4) to get the weighted value.

$$E_{(a,b)} = \sqrt{\sum_1^k (S_k(V_{a,b}) - T_k)^2} \tag{3}$$

$$W_{(a,b)} = \frac{\frac{1}{E_{(a,b)}^2}}{\sum_1^k \frac{1}{E_{(a,b)}^2}} \tag{4}$$

Step 6: After calculating the weight clusters of all the Reference tags in the sensing environment, the result should be applied to the Ward's method. Through this stage, the most appropriate cluster number to this cluster and the initial clustering result will be revealed.

$$D_{Ward} = \sum_{n=1}^d \sum_{m=1}^t (w_{nm} - \overline{w_m})^2 \tag{5}$$

$D_{Ward}$  is the number of cluster,  $w_{nm}$  means the m weight on d cluster,  $\overline{w_m}$  is the mean of m weight.

Step 7: The cluster number and the initial clustering result will be adapted in the second stage, the K-Means method, in the two-stage clustering method. The final clustering result will be unveiled.

Step 8: The values in the highest referencing value among all clusters should be used in the real location. These values will be the source for showing the numbers and the data of guiding points. The final locating point(X,Y) will be calculated through the (6) and the difference between the coordinate of real tracking tags and the anticipated coordinate will be calculated in the (7).



$$(X, Y) = \sum_1^k w(a, b) * (a, b) \tag{6}$$

$$e = \sqrt{(X - X_0)^2 + (Y - Y_0)^2} \tag{7}$$

### 3.2 System Design

When designing the system, patients' needs and potential requirements would be the references. After generalizing these requirements, this research proposes a basic healthcare monitoring system based on the Zigbee sensor network. This system obtained these functions such as Patient's identification, Patient's real-time location management, Patient's physiological information monitoring, Event management and Patients behavior research. The structure of this system will be consisted of the Hospital Environment layer, the System Layer and Zigbee packets layer. In Hospital environment layer, Zigbee equipment installed on the environment including the receivers and the sensors on the patients. In Zigbee packets layer, these sensors will deliver the patient's information within the sensing area to the system. And in the System layer, here will receive the information sent by the sensors wore by the patients for calculating and monitoring. The overall system structure is illustrated in Fig. 1.

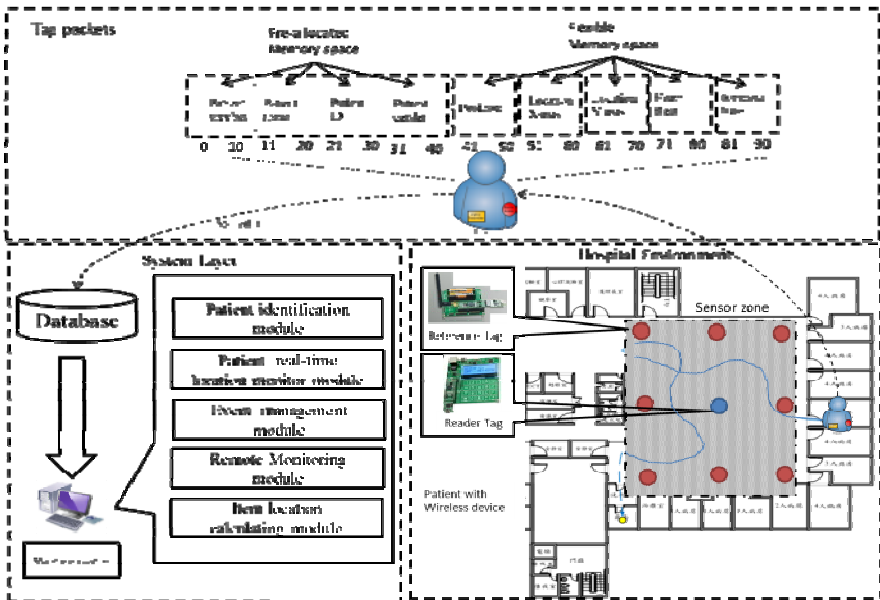


Fig. 1 The System Structure

## 4 Conclusion

This research proposes a two-step location algorithm with the Zigbee wireless sensor network. The error distance was about one meter by calculating through the two-step clustering localization algorithm, and the accuracy was 90% high when the range of error distance was within three meters. This system was better than other location algorithms and could locate patients accurately and promptly. Many improvements might be applied to this system in the future, for example, a smaller heartbeat sensor will be designed in the future to allow the patient wearing them comfortably. All the records and data were saved in the database, but more cases are needed for analyzing the behavior pattern.

## References

1. Prado, M., Reina, J., Roa, L.: Distributed intelligent architecture for falling detection and physical activity analysis in the elderly. Paper presented at the Proceedings of the Second Joint EMBS/BMES Conference (2002)
2. Anderson, J.: Beyond Mobile Coronary Care A Telehealthcare Case Study. Paper Presented at the Institution of Engineering and Technology Healthcare Technologies Network (2007)
3. Ondrej, S., Zdenek, B., Petr, F., Ondrej, H.: ZigBee Technology and Device Design. Paper presented at the Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, ICN/ICONS/MCL (2006)
4. Lee, J.-S.: An experiment on performance study of IEEE 802.15.4 wireless networks. Paper Presented at the 10th IEEE Conference on Emerging Technologies and Factory Automation, ETFA 2005 (2005)
5. Savvides, A., Han, C.-C., Strivastava, M.B.: Dynamic fine-grained localization in Ad-Hoc networks of sensors. Paper Presented at the Proceedings of the 7th Annual International Conference on Mobile Computing and Networking, Rome, Italy (2001)
6. Sanhae, K., Jungwoo, L., Myungsik, Y., Yoan, S.: An improved TDoA-based tracking algorithm in mobile-WiMAX systems. In: 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, September 13-16, pp. 561–565 (2009)
7. Hightower, J., Borriello, G.: A Survey and Taxonomy of Location Systems for Ubiquitous Computing, vol. 34 (2001) doi: citeulike-article-id:4235661
8. Bekkali, A., Sanson, H., Matsumoto, M.: RFID indoor positioning based on probabilistic RFID map and kalman filtering. Paper presented at the Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, WiMob 2007 (2007)
9. Want, R., Hopper, A., Falcão, V., Gibbons, J.: The Active Badge Location System. ACM Transactions on Information Systems (TOIS) 10, 91–102 (1992)
10. Hightower, J., Want, R., Borriello, G.: SpotON: an indoor 3D location sensing technology based on RF signal strength (2000)

11. Wang, C., Wu, H., Tzeng, N.: RFID-based 3-D positioning schemes. Paper presented at the IEEE INFOCOM (2007)
12. Ni, L.M., Liu, Y., Lau, Y.C., Patil, A.P.: LANDMARC: Indoor Location Sensing Using Active RFID. Paper presented at the Proceedings of the First IEEE International Conference on Pervasive Computing and Communications (2003)
13. Bekkali, A., Sanson, H., Matsumoto, M.: RFID indoor positioning based on probabilistic RFID map and kalman filtering. Paper Presented at the 3rd IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (2007)

# Using Fuzzy Logic in Virus-Mediated Gene Therapy

Arie H. Tan and Fabian H. Tan

**Abstract.** The increasing complexity of scientific procedure has led to the use of linguistic expressions, based on subjective judgment, for assessing different aspects of the process, especially uncertainties. For instance, in gene therapy, there are uncertainties associated with errors that can be incurred during the actual insertion process of the gene, although these are seldom addressed in overall process performance. Furthermore, one often finds inadequacies in traditional methods used to quantify these uncertainties. Therefore, an approach has been developed that uses fuzzy set concepts to attain the goal of appraising the overall efficacy of a certain method of virus-mediated gene therapy. The analysis was done through the mathematical and graphical interpretation of linguistic terms for these appraisals. This method can be used to further reduce the possibility of errors in the generation of viral vectors for use in this field of research.

## 1 Introduction

Although the method of gene therapy that is currently being explored with viral vectors is relatively accurate and safe, errors in manufacture of the vectors can still occur given the current state of research on the method. Such errors, typically uncertain in nature, can vary, but their identification can be of one of three primary methods: deterministic, non-deterministic probabilistic, and fuzzy logic. This paper presents a method based on fuzzy set theory to assess the efficacy of virus-mediated gene therapy, so as to present a form of quality control when generating the appropriate vectors to be used in gene therapy.

## 2 Possible Causes of Errors

In the field of gene therapy, one of the many methods used to insert the gene is the usage of viruses, which are basically packages of genes embedded into a protein capsid; normally, it protects the viral genome as it enters the cell. However, it can

also be used in gene therapy, by substituting the viral genes with the desired target genes. In doing so, one can insert the genes into the cell efficiently without any of the risks incurred by other methods. The risks involved in the use of viral vectors are often divided into four categories: Safety, low toxicity, cell-type specificity, and genomic stability. Safety is the primary issue here, because of the fact that the virus could potentially cause disease, although the risk for this is minimized by the deletion of viral replication genes. The resulting virus generally requires a helper virus to be capable of reproducing; however, considering the nature of the deletion, there is a level of uncertainty as to whether the methods are viable. Also, even if the virus turns out to be harmless, the immune response mediated against it can be devastating [1]. This is the primary issue of the paper, which is emphasized by the comparison between binding efficiency/safety and safety/therapeutic value.

The issue of low toxicity partially falls under the first factor, by the fact that the virus must not physiologically affect the organism in question, even if its genome has rendered it pathologically harmless. The last two factors can be pooled into the issue of efficiency: the first, cell-type specificity, is often invoked since the defective cells might be of a specific tissue; aiming for these cells minimizes any risk involved with other healthy cells. More important for efficiency, however, is genomic stability. Some viruses persist under certain conditions that cause its genome to become relatively unstable. The mutations that may result may cancel out the benefits of the gene therapy and sometimes even do more harm to the individual than good. These uncertainties are the ones that our assessment aims to detect so that further research can minimize or counteract them.

The three primary causes of error can be divided as follows: (1) errors in incorporation efficiency, (2) errors interfering with safety of the vector (i.e. promoting virulence), and (3) errors in therapeutic efficiency. Typically, errors in incorporation and binding efficiency arise from failure of cell specificity, inadequate receptor binding capacity or corruption of the capsid of the virus preventing intake by the cell [2]. Errors pertaining to virulence could result from either a reversion of the deactivated replication genes, incorrect byproducts from defective vectors or even unwanted or unnecessary interference of the cellular genome [3]. Finally, errors in therapeutic value derive from both of these factors, along with the possibility that the gene alone might not be sufficient for correcting the disorder. In our study, these factors serve as our universes of discourse.

### **3 Assessment of Gene Therapy Performance**

Table 1 lists positive influences of efficiency of binding and the negative influences of the potential virulence of the vector that enable us to determine the overall efficacy of the vector. For example, one possible relationship between virulence and these two factors, based on present knowledge of the mechanics behind viral vectors is given on Tables 2 and 3.



**Table 3** Events affecting Therapeutic Value and Virulence

Efficiency	Virulence											
	<i>Not Therapeutic</i>		<i>Lightly Therapeutic</i>		<i>Somewhat Therapeutic</i>		<i>Fairly Therapeutic</i>		<i>Therapeutic</i>		<i>Highly Therapeutic</i>	
Highly Virulent	V1	V2	V1	V2	V1			V2				
	V3	V4	V3	V4	V3		V3	V4				
Virulent	V1	V2	V1		V1			V2				
	V3	V4	V3	V4	V3		V3	V4				
Fairly Virulent	V1	V2	V1		V1			V2				
	V3	V4	V3	V4			V3	V4				
Somewhat Virulent	V1		V1									
	V3		V3	V4								
Lightly Virulent	V1		V1									
	V3		V3									
Harmless												
	V3											

Utilizing the relationships above, the user can make observations on the eight different factors and thereby arrive at the following relationships based on these observations:

**Table 4** Efficiency/Virulence Related Events

Efficiency-Related Elements	Efficiency	Virulence
E1	Very Efficient	Virulent
E2	Fairly Poor	Fairly Virulent
E3	Efficient	Highly Virulent
E4	Poor	Somewhat Virulent

**Table 5** Virulence/Therapeutic Related Events

Therapeutic-Related Elements	Therapeutic Value	Virulence
V1	Not Therapeutic	Virulent
V2	Fairly Therapeutic	Slightly Virulent
V3	Therapeutic	Highly Virulent
V4	Very Therapeutic	Somewhat Virulent

What follows is a procedure designed to help quantify overall performance. For the first example in our study, we use the set assignments shown in Table 4. Please note that these values are given subjectively by the assessors or evaluators who employ linguistic expressions. The process to generate each of the individual matrices is to take the individual fuzzy sets/criteria (e.g. “Virulence” or “Therapeutic”), then input these with the corresponding membership values [1]. Since two values are produced as a result, the *minimum* of these two values is taken to be used in the final result. Once this process is completed for all other tables, they are combined; to do this, one takes the *maximum* value of the four cells that occupy the same position[1, 6]. Once the two main matrices are generated, they are finally combined to generate the overall membership function by taking their minimum. This gives rise to the composition of the rules pertaining to efficiency and those pertaining to therapeutic value [1, 6].

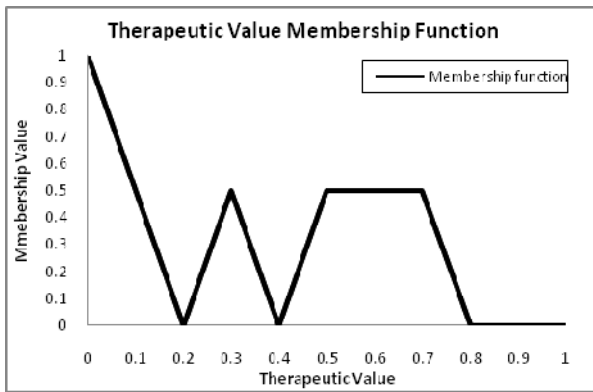
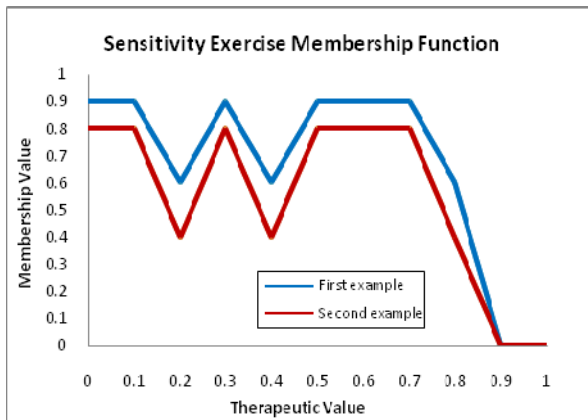


Fig. 1 Graph of the final membership function for the above example.

#### 4 Sensitivity of Linguistic Criteria

These membership values are selected based on human judgment, thus it is particularly important to analyze the sensitivity of the analysis based on the opinions of researchers. So to do this, we assign different membership values. Using the values given in Table 5, we get an overall performance set as shown on the right hand portion of Fig.2. A steeper judgment by the researchers can lead to the performance set as illustrated in Fig. 2. In any case, as long as the researchers agree on the general area of the membership values, this method of deciding the membership function is a rather robust one.





**Fig. 2** Altered values for membership function

## 5 Conclusion

In summary, the use of fuzzy logic in the generation of our viral vectors for gene therapy is potentially a useful approach; this is especially true when the results must be judged in a subjective fashion, such as via color of the solution. Note, however, that this is best applied for the actual research and manufacture of the vectors, all of which take place before any form of testing, first on animals, then in clinical trials. Once we reach the clinical trial stage, the important factor at stake is the effectiveness of the treatment on the trial groups versus the control group, and the techniques commonly used for analysis would be statistical in nature. Nonetheless, knowing that we now have the start of a new means to manufacture the viral vectors more efficiently, we can utilize fuzzy logic to help us manufacture our viral vectors more rapidly than ever before, thanks to the increased possibilities via automation of synthesis. These vectors can even be made in a personalized fashion, needing only a DNA sample of the patient to determine exactly what the person needs and to synthesize the resulting treatments.

## References

1. Tan, A., Tan, F., Bali: Proceedings of the 2011 International Conference on Data Engineering and Internet Technology, March 15-17, p. 1047. IEEE, Indonesia (2011)
2. Ziello, J.E., Huang, Y., Jovin, I.S.: Cellular Endocytosis and Gene Delivery. *Molecular Medicine* 16, 222–229 (2010)
3. Cattoglio, C., Facchini, G., Sartori, D., et al.: Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* 110(6), 1770–1778 (2007)
4. Zadeh, L.A.: Fuzzy sets. *Inform. and Control* 8, 338–353 (1965)
5. Hadipriono, F.C.: Fuzzy Set Concepts for Evaluating Performance of Constructed Facilities. *J. Perf. Constr. Fac.* 2(4), 209–225 (1988)
6. Hadipriono, F.C.: Assessment of falsework performance using fuzzy set concepts. *Structural Safety* 3, 47–57 (1985)

# Towards Using Cached Data Mining for Large Scale Recommender Systems

Swapneel Sheth and Gail Kaiser

**Abstract.** Recommender systems are becoming increasingly popular. As these systems become commonplace and the number of users increases, it will become important for these systems to be able to cope with a large and diverse set of users whose recommendation needs may be very different from each other. In particular, large scale recommender systems will need to ensure that users' requests for recommendations can be answered with low response times and high throughput. In this paper, we explore how to use caches and cached data mining to improve the performance of recommender systems by improving throughput and reducing response time for providing recommendations. We describe the structure of our cache, which can be viewed as a prefetch cache that prefetches all types of supported recommendations, and how it is used in our recommender system. We also describe the results of our empirical study to measure the efficacy of our cache.

## 1 Introduction

Recommender systems have become increasingly commonplace. Recommender systems are being used in a variety of domains such as recommending music we may like [10, 14], things we might like to buy [1], and friends we may know [7]. There have also been many recommender systems targeted towards specialized domains such as software engineering [4, 9, 11, 12] and medicine [18]. While there has been a lot of work in the academic community on various aspects of recommender systems such as recommendation algorithms [15, 21] and implications of social networks in recommender systems [8, 20], there has been very limited work that has explored the use of caches and cached data mining to improve the performance of recommender systems by increasing throughput and reducing response time for providing recommendations. This will be of particular concern as these

---

Swapneel Sheth · Gail Kaiser

Department of Computer Science, Columbia University, New York, NY 10027

e-mail:  [{swapneel,kaiser}@cs.columbia.edu](mailto:{swapneel,kaiser}@cs.columbia.edu)

recommender systems become even more popular and their user and fan base grow. With a large number of users, there are two specific issues that recommender systems would have to deal with - how to generate recommendations efficiently from a large set of data and how to provide these recommendations efficiently to a diverse set of users, where each user's requirements for recommendations are different from the others.

In this paper, we describe how we use cached data mining to answer users' queries and provide recommendations in a very efficient way. We describe our background and motivation in the next section. Section 3 describes in detail the recommendations provided by our system and how we use cached data mining. Section 4 describes our empirical study and results. Finally, we conclude the paper with a discussion of the related work in Section 5.

## 2 Background and Motivation

We are working with researchers at Columbia University's Center for Computational Biology and Bioinformatics (C2B2), particularly its MAGNet (Multiscale Analysis of Genomic and Cellular Networks) Center to explore new ways in which researchers in computational biology and bioinformatics can collaborate to share data, analyze results, and share knowledge. Our approach is based on social networking metaphors for collaborative work where users can ask questions such as: Who likes movies that I like?; What food and wine pairings go well together?; What book would I like given that I like this book?

Our implementation of this approach is a system called "genSpace" [13], which uses collaborative filtering to provide recommendations to users. genSpace is a plugin to an open-source Java-based platform for integrated genomics called "geWorkbench" [5]. Using geWorkbench, researchers in computational biology and bioinformatics can load in data sets such as DNA, protein, and gene sequences. They can then run complex analysis tools such as filtering, normalization, clustering, and pattern detection. There are over 50 such analysis tools supported by geWorkbench, and each tool has many different runtime parameters. Choosing the right tool to use and the sequences in which to use these tools (workflows) can be very daunting, especially to new users. One substantial way that we diverge from and expand upon the collaborative filtering provided by popular websites is that we address the *ordering* among related activities conducted in sequence, i.e., as a **workflow**. E.g., a common workflow in geWorkbench is to run the ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) analysis [3] followed by the MINDy (Modulator Inference by Network Dynamics) analysis [17]. Issues stemming from this ordering concern are, however, outside of the scope of this paper.

genSpace aims to flatten the learning curve and enable users to quickly become productive. In particular, for users who do not know where to start, it recommends the most popular three tools and workflows. For users already familiar with using one or more tools in their standalone form outside geWorkbench, it recommends the most popular workflow that starts with or includes a particular tool and the best tool

to run next given that you've just run a particular tool. In order to achieve this, we log users' activities as they use geWorkbench and send the logs to a central server, where data mining and collaborative filtering techniques generate these and other kinds of recommendations.

Currently, our genSpace recommender system is modest in size. Our database has about 10000 rows of data from around 150 distinct users. Since we anticipate a significant increase in usage when geWorkbench soon introduces a Web-based client, we wanted to study how our system would respond to and/or if it could cope with a large increase in the number of users and user data.

In this paper, we discuss how we use cached data mining for providing recommendations to users in genSpace. We also describe an empirical study highlighting their benefits and improvements to the response time and throughput to user queries.

### **3 Cached Data Mining and genSpace Recommendations**

#### ***3.1 Recommendations in genSpace***

In genSpace, we support two different kinds of recommendations - static and dynamic.

##### **3.1.1 Static Recommendations**

Static Recommendations are those recommendations that do not depend on the current activity of the user. Typically, such recommendations follow a "pull" model where a user explicitly asks for these recommendations. Examples of such recommendations include the top tools, the top workflows, and the most popular workflow that includes or starts with a particular tool.

##### **3.1.2 Dynamic Recommendations**

Dynamic Recommendations are those recommendations that do depend on the current activity of the user. Typically, such recommendations follow a "push" model where the system automatically pushes these recommendations to the user. Examples of such recommendations include suggesting the best analysis tool to run next based on what the user has done so far and suggesting popular superflows (workflows that include the user's current workflow).

All these recommendations are generated using data mining to derive patterns and trends from the user data.

#### ***3.2 genSpace Caching***

genSpace has a server-side cache that supports pushing or pulling recommendations to/from the users. It can be viewed as a prefetch cache that prefetches all types of recommendations supported by the system. It is not a traditional cache where items

are added to the cache when they are requested and there exist notions of cache hits, cache misses, cache replacement policies and so on. Every recommendation that we need will be present in the cache and we won't need to go to the database for any information. Due to this, we do not have the problem of a cache miss and we do not need to worry about cache replacement and by definition, our hit rate and recall is 100%. When the genSpace server starts up, the genSpace cache is generated using a combination of SQL queries and stored procedures from our SQL database backend that stores all the user data. The cache is periodically re-generated as needed - currently, every day. If we did not have a cache, we would have to run the SQL queries on demand every time a user request came in for recommendations. We would also have to re-run the same query multiple times if different users asked for the same set of recommendations.

We also address the problem of concept drift [19] where workflows performed by users six months may not be so relevant today. E.g., after publication of major findings that involved a form of analysis that was previously rare or after upgrading to a new geWorkbench release that integrates additional tools or even for no known reason, many users shift their usage patterns. We use an exponential time-decay formula [6] to weigh recent user data more heavily. This weighting is done each time the cache is generated.

After weighting the data, the static recommendations are computed and stored in the cache. We build an index for each analysis tool found in the log data, to represent the following information: the number of times this tool has been used, the number of times this tool has been used as a workflow head, the most popular tool before and after this tool in workflows, the most popular workflows containing this tool, and all workflows that include this tool. This cached index uses hashing based on the tool name to give us constant time lookup for tool-specific information. Finally, a tree-based index of popular workflows aids in the dynamic recommendations. All these parts together comprise the genSpace Caching System.

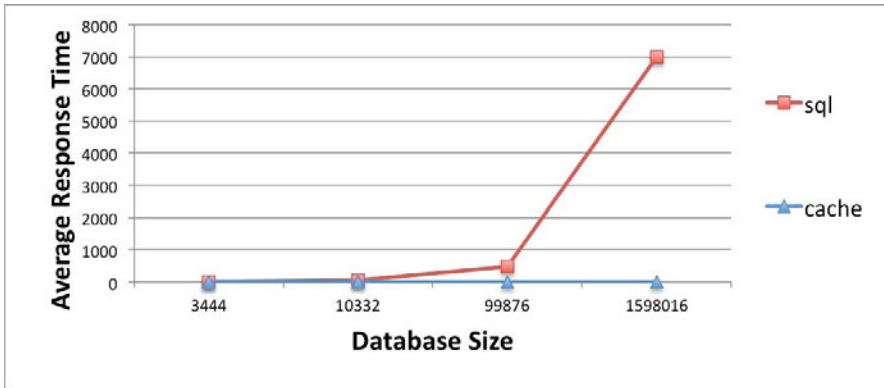
genSpace usually gets around 10-20 new logs every day and due to this, we re-generate our cache every day. As the number of users for our system increases, concept drift may take place on shorter timescales and we may need to re-generate the cache more often to deal with it. The re-generation frequency is easily configured on our server and will be ramped up as needed. However, more studies need to be done to measure and fully understand the effect of concept drift on cache re-generation and this is part of our future work. The next section contains some empirical results on the time required to re-generate our cache.

Finally, due to the structure of the genSpace cache, it can only support the currently existing types of recommendations. If we wanted to support additional types of recommendations, the cache would need to be augmented with the appropriate information. E.g., we currently don't support providing recommendations based on the file-type on which the analysis tools are run. To support this, our cache would need to store information regarding the file-types for the analyses.

## 4 Empirical Study

The genSpace cache has already been deployed in our production system although it may not be needed currently due to the modest number of users. In order to understand the prospective real-world improvements due to our cache, we carried out an empirical study. For our study, we varied the number of rows of our database (in the range of around 3500, 10000, 100000, and one million) and measured its impact on average response time and throughput to user queries for recommendations. We simulated 1000 concurrent users requesting recommendations. We also compared these results to the results obtained if we did not have a cache and used SQL queries instead every time for generating recommendations.

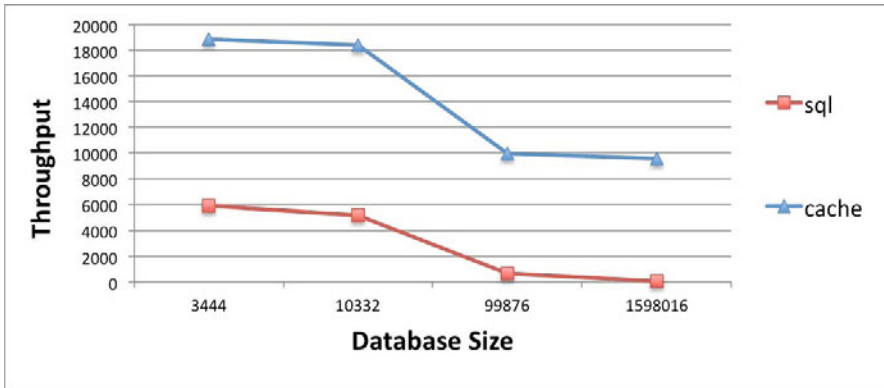
We used Apache JMeter [2] for load testing our server and measuring performance. The genSpace server, including the cache, is implemented in Java. Our server and client machines were common Windows XP machines with no non-essential system processes running and had more than 2GB of surplus RAM available.



**Fig. 1** Database Size vs. Average Response Time, for “Get Most Popular Workflow Heads”

Figure 1 shows the plot of the database size (in number of rows) versus the Average Response Time for a recommendation that gets the most popular workflow heads, i.e., tools at the start of a workflow. The red line with squares as data points shows the response time when using SQL queries-on-demand and the blue line with triangles as data points shows the response time when using our cache. As shown in the figure, as the size of the database increases, the response time using SQL queries-on-demand increases by a large amount. Meanwhile, the response time using our cache remains roughly constant. This shows that as the database size increases, using SQL queries-on-demand is not practical whereas using the cache enables us to answer users’ queries in roughly the same time regardless of the database size.

Figure 2 shows the plot of the database size (in number of rows) versus the Throughput for a recommendation that gets the most popular tools in the system. The red line with squares as data points shows the response time when using SQL



**Fig. 2** Database Size vs. Throughput, for “Get Most Popular Tools”

queries-on-demand and the blue line with triangles as data points shows the response time when using our cache. From the graphs, we see that the cache outperforms the SQL queries-on-demand approach by a factor of at least 3 to as much as 200 as database size increases.

Most of the static and dynamic recommendations mentioned in Section 3 were part of the empirical study and our results were similar to the ones shown above and generally show that using the cache improves the throughput and reduces the response time.

We also measured how long our cache generation process takes. As mentioned earlier, we currently re-generate our cache every day and it might be necessary to re-generate our cache more often. In our study, it takes around ten seconds to generate the cache for a database that has around one million rows and about 100 seconds for a database that has around ten million rows. Thus, even if our database size increases by a large amount, we can still manage to re-generate the cache periodically as often as needed.

## 5 Related Work

To the best of our knowledge, there is very little in the published literature discussing caches for recommendation systems; in fact, we found exactly one paper that discusses this. Qasim et al. [16] propose a general solution using active caches for providing recommendations in all types of recommender systems. Active Caches are caches that can answer neighborhood queries for recommendations, i.e., similar queries to a given query and act as limited query processors. Due to this, the approach proposed by Qasim et al. is limited to neighborhood queries for recommendations and will not work well, in general, for all kinds of queries and focusing on just neighborhood queries may not improve overall system performance by a significant amount. As recommender systems become increasingly popular, there

might exist a very diverse user base that is interested in different kinds of recommendations from the system.

In fact, as mentioned in their paper, due to overheads of caching, the system might actually perform worse than having no cache. Our genSpace solution, on the other hand, is not limited to neighborhood queries for recommendations and works well for all kinds of recommendations supported by our genSpace system. This is because our system, unlike the one mentioned by Qasim et al., is a prefetch cache that prefetches all recommendations; all user recommendations can be answered using the cache, rather than just the neighborhood ones. Of course, as our system evolves and new types of recommendations are added, we would need to enhance our cache to support those as well.

Further, Qasim et al., in the experimental section of their paper, focus on the Hit Ratio, Recall, and Efficiency of computing the cache. While these metrics are important, we feel it would more meaningful to see what this translates to, from a user's point of view. A typical user is not directly concerned about hit ratio and recall; rather, he is usually directly concerned with the latency and response time for these recommendations. Our empirical study shows that using caches in genSpace has significantly improved the throughput and reduced the response time for recommendations, thus improving the overall user experience. Also, as we use a prefetch cache that prefetches all types of recommendations supported by the system, by definition, the hit ratio and recall for our system is 100%.

## 6 Conclusion

We have described how we use prefetch caching in our genSpace recommender system. We have also described the structure of our cache, which can be viewed as a prefetch cache that prefetches all types of supported recommendations, and our empirical study that shows the advantages of using our cache, which improves throughput and reduces response time for recommendations. We believe that the use of such caches will prove very beneficial to recommender systems, particularly as the number of users of such systems grow and the system needs to support the diverse needs of its users, where different users are interested in very different kinds of recommendations from the system and the recommendations they request do not overlap.

**Acknowledgements.** The authors would like to thank Aris Floratos, Kiran Keshav, and Zhou Ji for their guidance and assistance with genSpace. We would also like to thank Cheng Niu, Joshua Nankin, Eric Schmidt, and Yuan Wang for their assistance in the implementation of the genSpace cache and in the empirical study to measure its efficacy. The authors are members of the Programming Systems Lab, funded in part by NSF CNS-0905246, CNS-0717544, CNS-0627473 and CNS-0426623, and NIH 1 U54 CA121852-01A1.



## References

1. Amazon.com, <http://www.amazon.com>
2. Apache: Jmeter, <http://jakarta.apache.org/jmeter/>
3. Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A.: Reverse engineering of regulatory networks in human B cells. *Nature Genetics* 37(4), 382–390 (2005)
4. Begel, A., Phang, K.Y., Zimmermann, T.: Codebook: discovering and exploiting relationships in software repositories. In: *ICSE 2010: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering*, pp. 125–134. ACM, New York (2010), doi: <http://doi.acm.org/10.1145/1806799.1806821>
5. Califano, A., Floratos, A., Kustagi, M., Watkinson, J.: geWorkbench: An Open-Source Platform for Integrated Genomics, <http://www.geworkbench.org>
6. Cohen, E., Strauss, M.: Maintaining time-decaying stream aggregates. In: *Proc. of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pp. 223–233 (2003)
7. Facebook, <http://www.facebook.com>
8. Geyer, W., Dugan, C., Millen, D.R., Muller, M., Freyne, J.: Recommending topics for self-descriptions in online user profiles. In: *RecSys 2008: Proc. of the 2008 ACM Conference on Recommender Systems*, pp. 59–66 (2008), doi: <http://doi.acm.org/10.1145/1454008.1454019>
9. Holmes, R., Ratchford, T., Robillard, M.P., Walker, R.J.: Automatically recommending triage decisions for pragmatic reuse tasks. In: *Proceedings of the 24th IEEE/ACM International Conference on Automated Software Engineering*, pp. 397–408 (2009)
10. Last.fm, <http://www.last.fm>
11. McCarey, F., Cinnéide, M., Kushmerick, N.: Rascal: A recommender agent for agile reuse. *Artificial Intelligence Review* 24(3), 253–276 (2005)
12. Murphy, C., Kaiser, G.E., Loveland, K., Hasan, S.: Retina: Helping Students and Instructors Based on Observed Programming Activities. In: *Proc. of the 40th ACM SIGCSE Techn. Symp. on CS Education*, pp. 178–182 (2009)
13. Murphy, C., Sheth, S., Kaiser, G., Wilcox, L.: genSpace: Exploring Social Networking Metaphors for Knowledge Sharing and Scientific Collaborative Work. In: *1st International Workshop on Social Software Engineering and Applications (SoSEA)*, pp. 29–36 (2008)
14. Pandora Radio, <http://www.pandora.com>
15. Park, Y.J., Tuzhilin, A.: The long tail of recommender systems and how to leverage it. In: *RecSys 2008: Proc. of the 2008 ACM Conf. on Recommender Systems*, pp. 11–18 (2008), doi: <http://doi.acm.org/10.1145/1454008.1454012>
16. Qasim, U., Oria, V., Wu, Y.F.B., Houle, M.E., Özsu, M.T.: A partial-order based active cache for recommender systems. In: *RecSys 2009: Proceedings of the Third ACM Conference on Recommender Systems*, pp. 209–212. ACM, New York (2009), doi: <http://doi.acm.org/10.1145/1639714.1639750>
17. Wang, K., Saito, M., Bisikirska, B., Alvarez, M., Lim, W., Rajbhandari, P., Shen, Q., Nemenman, I., Basso, K., Margolin, A., et al.: Genome-wide identification of post-translational modulators of transcription factor activity in human B cells. *Nature Biotechnology* 27(9), 829–837 (2009)
18. WebMD Symptom Checker, <http://symptoms.webmd.com>

19. Widmer, G., Kubat, M.: Learning in the presence of concept drift and hidden contexts. *Machine Learning* 23(1), 69–101 (1996)
20. Zanardi, V., Capra, L.: Social ranking: uncovering relevant content using tag-based recommender systems. In: *RecSys 2008: Proc. of the 2008 ACM Conf. on Recommender Systems*, pp. 51–58 (2008),  
doi: <http://doi.acm.org/10.1145/1454008.1454018>
21. Zhang, J., Pu, P.: A recursive prediction algorithm for collaborative filtering recommender systems. In: *RecSys 2007: Proc. of the 2007 ACM Conference on Recommender Systems*, pp. 57–64 (2007),  
doi: <http://doi.acm.org/10.1145/1297231.1297241>

# BAX-SET PLUS: A Taxonomic Navigation Model to Categorize, Search and Retrieve Semantic Web Services

Jaime Alberto Guzmán Luna, Ingrid Durley Torres Pardo,  
and Jovani Alberto Jiménez Builes

**Abstract.** This paper proposes a taxonomic navigation model to address the problem of categorizing, searching and retrieving Semantic Web Services (SWS) in BAX-SET PLUS, a SWS categorization, search and retrieval multi-agent system that includes a knowledge module represented by a taxonomic model, an OWL-S extension implemented and a semantic search module for a SWS repository.

## 1 Introduction

Currently, the Web enables us to access remote systems and execute the services these systems have to offer using the technology that supports them [1]. A problem that must be faced is that the end user of those services may be a human being, who sometimes turns out to be an expert in a specific domain, but not an expert in topics regarding Informatics. Thus, the authors propose BAX-SET PLUS, a multi-agent system [2], that from a semantic taxonomy (ontology) representing users domain, allows Web service classification, search and retrieval [1] through a navigation process by means of the concepts of the aforesaid taxonomy. As an example and as a trial case of developed system, authors propose a plastic-arts ontology; a domain that is totally foreign to technological service considerations.

This document describes BAX-SET PLUS's main features, so it has been organized as follows: Section 2 presents a system overview that details the modules

---

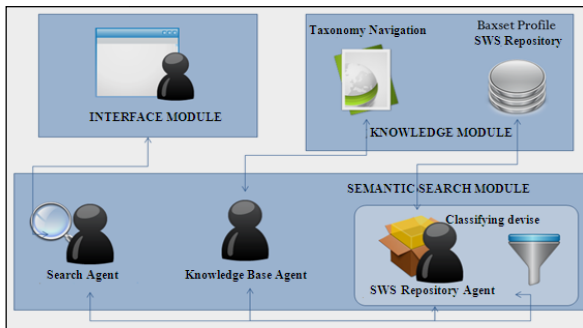
Jaime Alberto Guzmán Luna · Jovani Alberto Jiménez Builes  
SINTELWEB research group, Universidad Nacional de Colombia, Medellín Campus,  
Medellín, Colombia  
e-mail: {jaguzman, jajimen1}@unal.edu.co

Ingrid Durley Torres Pardo  
Fundación Universitaria “Luis Amigó”, Medellín, Colombia  
e-mail: ingrid.torrespa@amigo.edu.co

that form it; section 3 details the most significant aspects of the knowledgebase module representing and managing the knowledge implemented in the semantic taxonomy (ontology of art); as well as, Web-service-ontology-related knowledge (*BaxSetProfile* ontology). Section 4 describes the SWS search and retrieval process implemented. Section 5 explains the interface module's main functions. Finally, section 6 compiles and presents conclusions and future work.

## 2 Architecture and Generalities

A BAX-SET PLUS system includes three basic modules (see Fig. 1):(1) A Knowledge Module is the one which has knowledge related to a semantic taxonomy and the SWS that the system handles. This module is formed by (a) an ontology instantiated in a specific domain (b) the knowledge base in SWS (*BaxSetProfile* ontology), corresponding to OWL-S specification [3, 4]. (2) A Semantic Search Module is the reasoning process to carry out a semantic service classification, search and retrieval. This module is formed by (a) a classifying devise, the search agent along with the base agent of the context and the SWS repository agent. The automatic SWS classification device that works from heuristics that automatically assigns a category (Concept) to a service starting from the degree of similarity with other previously classified services.



**Fig. 1** Architecture of the BAX-SET PLUS System, with its main component

(b) A Search Agent coordinates the Context Agent and the SWS Repository Agent to first perform the search of the concepts contained in a user's search and its related concepts in a Semantic taxonomy (this work is performed by the Context base agent). Later, with these results, they conduct an SWS search classified under these concepts (work performed by the SWS Repository Agent). (3) An Interface Module acts as a bridge between users and the system, allowing the former to navigate through the Semantic Taxonomy.

### 3 Representation of a Knowledge Module

This module includes a taxonomy navigation model, a Plastic Arts ontology and the ontology corresponding to *BaxsetProfile* as an SWS specification (see Fig. 2).

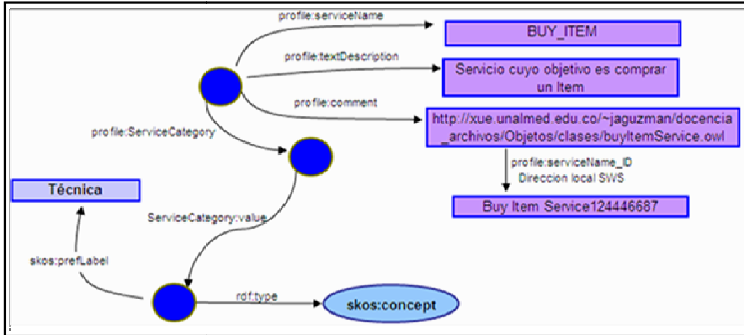


Fig. 2 Example of an SW “BUY\_ITEM” association and the ontological concept, “Technique”

To represent knowledge in the ontology of the arts, they used the steps suggested in the methodology to create ontologies, “Ontology Development 101” [5]. As a result of the methodology used, were as follows: (1) the domain is a matter of art and its ambit is related to the concepts surrounding the syllabus for the plastic arts program at the School of Architecture at Universidad Nacional de Colombia. (2) The use is to implement a semantically enriched thesaurus through formal knowledge [6, 7]. (3) Reusing controlled vocabularies is applicable in this case since there was an ontology named SKOS-Core in literature [8], an ontology created precisely to model concepts under thesaurus schemas. (4) The most important terms in the domain, were extracted from the SKOS-Core ontology: Concept (*skos:Concept*), a concept is defined as “a unit of thought that can be defined or described”, the schema of the concepts (*skos:ConceptSchema*), a schema of concepts is nothing more than a “collection of concepts”, the only descriptor or preferred label (*skos:prefLabel*), non-preferred descriptors (*skos:altLabel*), hidden descriptors (*skos:hiddenLabel*) which represented final users’ mistake when writing one of the labels of each concept (*skos:definition*) and its scope note (*skos:scopeNote*). (5) Concept relations were represented using the following properties retaken from SKOS-Core: *skos:narrower* and *skos:broader*, which represent respectively hierarchy relations aimed at specific and general concepts, and *skos:related* which represents the relation of semantic association between concepts. (6) 209 concepts were identified in the plastic-arts domain; they almost always form the hierarchal tree.

Another component of this module is represented in the ontology of the service; it associates with a specific semantic described in OWL-S, an ontology that contains the fundamental elements that characterize a service and allows the description of the capabilities that support it.

### 4 Semantic Search and Retrieval Module

In the BAX-SET Plus model, the Semantic Search Module is directly associated with processes such as classification, search and retrieval of SWS. In the first case, the classification process is directly associated to the SWS repository which allows their categorization. This categorization according to the model can be performed in two ways (i) manually, in which user chooses from among all the concepts of a domain taxonomy (ontology), the user deems the best to represent the function of the service. Once the concept to be associated with the SW is identified, this is assign in the category property (*Profile:serviceCategory* ). Besides generating a copy of some properties of SWS in order to store it in the repository: a unique identifier that indexes the service is assigned, the public name of the service (*profile:serviceName*), the description of the function of the service (*profile:textDescription*) and the web address of the service (*profile:comment*) and finally, it relates the unique identifier of the associated concept. (ii)The second alternative is completely automatic (even though, there is a need to require categorizations prior to the process). This is achieved through heuristics, associated to the specification of a proper SWS which implements three levels of granularity, in which the total process of categorization, which are as follows: a category level, a service level and a parameter level (see Fig. 3). Under this approach, heuristics is formed by three similarity measures among the different components of the problem where each of the measures corresponds to a category level.

1. Category level	2. Service level
$P(s:c) \sim \sum_{A \subset S} (-1)^{ A +1} \prod_{x \in A} sim(s,x)$ <p>Where, <math>P(s:c)</math> is the probability that category <math>c</math> may be the appropriate classification for a service, <math>s</math> found within an <math>S</math> repository. This probability is estimated through the comparison of <math>s</math> with all the services classified under <math>c</math>.</p>	$sim(op, op') = sim(I_{op}, I_{op'}) \cdot sim(O_{op}, O_{op'})$ <p>Where <math>P</math> is a set of all the parameters present in all the services contained in <math>S</math>, and <math>OP</math> a set of service operations, it is possible to specify that <math>I_{op}, I_{op'}</math>, <math>O_{op}</math> and <math>O_{op'}</math> are a set of input parameters (<math>I_{op}, I_{op'}</math>) and output parameters (<math>O_{op}, O_{op'}</math>) of <math>op</math> and <math>op'</math> respectively.</p>
3. Parameter level	
$sim(t, t') = \left(1 - \frac{\alpha}{T}\right) \cdot \frac{ d-d' }{(d+d')} \cdot \left(\frac{1}{\min(d,d')}\right) \cdot \left(1 - \frac{\max(d,d')-1}{T}\right)$ <p>Where <math>T</math> is considered as the set of all taxonomic concepts, and <math>t</math> and <math>t'</math> as concepts that <math>\in T</math>; and <math>t0</math>, the common ancestor closest to <math>t</math> and <math>t'</math> in <math>T</math>, and <math>d = dist(t, t0) + 1</math> and <math>d' = dist(t', t0) + 1</math>, the number of levels (plus one) between <math>t, t'</math> and <math>t0</math> in concept hierarchy.</p>	

Fig. 3 Heuristics for automatic categorization

Finally, for the process of search and retrieval comes after the following steps: (1) *Searching*: a user formulates the search as semantic terms (concepts) through the interface agent of the visualization module who asks the Search Agent the task of finding SWS's related to these semantic terms. (2) *Identification of concepts and their relations*: the Search Agent takes the semantic terms that make up the search and asks the Search Agent of the knowledge base that determines if these terms coincided with the existing concepts in the knowledgebase. If when the search concludes, the Agent finds that there is one or more of these concepts; it

then starts a search for each one of them looking for the semantic relations it has with other concepts aimed at finding other concepts semantically relevant with the user's search. After finding these new concepts, the Knowledge Base Agent returns to the Search Agent an RDF document [9] with a list of the triplets related to information about the concept. In case the terms that make up the user's original search do not coincide with the existing concepts in the context base, the agent returns a fault in the request. (3) *Identifying documents*: if the Search Agent receives a RDF document from the Agent of the Knowledge Base, this Search Agent requests the SWS Repository Agent to look for all the documents semantically framed with the concepts that make up this list. In case there are documents that fulfill what has been requested, the SWS Repository Agent returns a new RDF document to the Search Agent with a list of the RDF triplets containing relevant information for each SWS found. (4) *Unifying Results*: when the Search Agent gets the RDF document from the SWS Repository Agent with the SWS list, this agent returns that RDF document to the Interface Agent and it returns the RDF document with the list of concepts coming from the contextual base Agent and this one is in charge of showing them on the model screen.

## 5 Interface Module

The interface module includes the Interface Agent as its main element which is in charge of interacting with system users. This Agent selects which information is to be shown to users. This process includes: (1) Definition of the information to be seen: To perform this task, the following basic elements are defined: the domain that describes which elements will be visualized, the source that describes where the instances for the elements have been taken, and the source in a single Visualization Service. The interface Agent dynamically generates for each service on the list of returned services a triplet that will become part of a temporal instance which will show how information will be drawn from Repository Services. (2) Definition of the interface with the results: With this information, the Interface Agent moves on to building the interface which will show users the data contained in the Visualization File. To do so, the Interface Agent generates an interface which not only includes the information it wishes to show, but also the elements required so that a user may interact with the interface.

## 6 Conclusions

The Taxonomic Navigation Model for SWS Search (BAX-SET PLUS) is the ideal frame to enable materializing one of the paradigms proposed for the semantic web project, in other words, semantic management for data access. Practically, all authors have been evidencing the need to have automatic mechanisms to delimit search result spaces. BAX-SET PLUS users are the first benefitted, for they require tools that allow efficient resource exploitation. Moreover, and referring to the validation of the "prototype" model, the results of this project in classification and later in search and presentation of SWS's in the realm of plastic arts, they are

to the turn into important cultural benefits for society, a society which in most cases lacks or just barely knows web technology. One of the future contributions that aim to be included in BAX-SET PLUS is represented by a learning Module which will establishing relations between semantic description relations of their own services and the taxonomic (ontology).

**Acknowledgments.** This paper is supported by the Research Project entitled: “A Web Service Semantic Retrieval Systems in a Taxonomic Navigation Model”, supported by DIME research group at Universidad Nacional de Colombia at Medellin Campus. In partnership with the FundaciónUniversitaria“Luis Amigó” Medellinwiththe project “A System Recovery Bilingual Learning Objects in the area of Engineering”.

## References

1. Benatallah, B., Hacid, M.-S., Rey, C., Toumani, F.: Request Rewriting-Based Web Service Discovery. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 242–257. Springer, Heidelberg (2003)
2. Jennings, N.R.: An Agent-Based Approach for building Complex Software Systems. *Communications of the ACM* 44(4), 35–41 (1999)
3. OWL-S: OWL Services Coalition. Semantic markup (2003); OWL-S, <http://www.daml.org/services/owl-s/0.9/owl-s.pdf> (Consultadojunio 2010)
4. OWL: Ontology Web Language. W3C Recommendation (February 10 2004), <http://www.w3.org/2004/OWL/> (Consultado Julio 2010)
5. Noy, N.F., McGuinness, D.L.: *Ontology Development 101: A guide to creating your first ontology*, Disponible en <http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html> (Consultado Julio 2010)
6. Gil Urdiciani, B.: *Manual of Documentary Languages. Manual de Lenguajes Documentales*. Noesis, Madrid (1996)
7. Gómez-Pérez, A., Fernández-López, M., de Vicente, A.: *Towards a Method to Conceptualize Domain* (1996)
8. SKOS-Core (2006), [http://www2.ub.es/bid/consulta\\_articulos.php?fichero=13perez2.html](http://www2.ub.es/bid/consulta_articulos.php?fichero=13perez2.html) (Consultadojunio 2010)
9. RDF. Semantics W3C Recommendation (February 10, 2004), <http://www.w3.org/TR/rdf-mt/> (Consulted July 2010)



# Text-Transformed Image Classification Based on Data Compression

Nuo Zhang and Toshinori Watanabe

**Abstract.** Image data analysis technology occupies an important position in processing multimedia information. Due to the wide usage of digital images, automatic classification technology with high capacity is necessary for processing enormous number of digital images. In this paper, we propose an automatic classification technique which representing text-transformed image based on data compression. Images are first transformed into texts, which are then divided into segments and replaced by characters. Then, instead of using texts themselves, the similarity between compressibility vectors of texts are used in the classification step, in which we focus on the compressibility of the text string. Finally, the effectivity of the proposed method is verified in our experiments.

## 1 Introduction

There is an increasing trend towards the digitization of image and the formation of adequate archives. With the growing size of image libraries, there is a need for efficient tools that can analyze images and represent it in a way that can be efficiently searched. Image classification technique is widely used in many fields, including scientific research, commerce application, etc. By adopting the image classification technique, images can be searched more accurate [1]. For a large number of digital images, a good classification method can improve efficiency.

Different from most methods, Weiming Hu et al. proposed a novel framework for recognizing pornographic Web pages is described [2]. In which, image and text data are processed by a continuous text classifier and a discrete text classifier respectively,

---

Nuo Zhang · Toshinori Watanabe  
Graduate School of Information Systems  
The University of Electro-Communications  
1-5-1, Chofugaoka, Chofu-shi  
Tokyo, Japan  
e-mail: [zhang@sd.is.uec.ac.jp](mailto:zhang@sd.is.uec.ac.jp)

then an algorithm that fuses the results from the image classifier and the discrete text classifier. Unfortunately, it has to be fused by another method since it adopted many different kind of algorithms which intrinsically need many parameters.

A method named PRDC (Pattern Representation Scheme using Data Compression) [3] is developed to represent multimedia data to uniformly perform analysis lately. PRDC represents the feature of data as compressibility vectors. The measurement for all kinds of data is the distance among the vectors. PRDC can be combined with clustering and classification methods. PRDC is based on data compression and is basically good at processing text data. When processing media data, PRDC still needs to be improved by using other complicated algorithms. In this study, we propose a simple image representation and classification method based on the same consideration of PRDC. In our proposal, dictionary space is used to feature image and convert image to text. The obtained text is analyzed instead of the image itself. The proposed method provides a novel representation for modeling image content.

## 2 Proposal

In this paper we propose an image representation and classification method only based on the PRDC. We transform image into text based on PRDC first and then classify the text-transformed image. By processing image as text file, the proposed method can use the similarity of compressibility vectors for text-transformed image representation. Moreover the dictionary space is constructed based on the image feature, which makes it possible to implement image classification.

### 2.1 PRDC Based Image Representation and Classification

Text compression is used by PRDC to find most frequently repeated and longest feature in text data. In order to adopt this advantage of PRDC, images need to be converted to text data. But the size of image data is too big to directly convert each pixel to a character. Besides that the extracted features will become too much and some of which are redundancy. Hence, we consider to divide an image to segments and cluster them. After which, each cluster is replaced by a character and the converted image is called text-transformed image. In this study, grayscale images are used in the experiments of proposed method. To obtain the text-transformed images, all the images are changed to 256 grayscale in a pre-process step. Then a step function is used on these grayscale images to reduce the 256 gradation into 16.

Every pixel of the image is considered as one character (ASCII code) of the text. Each grayscale image is made into segments with length  $L$ . The PRDC is used to compress the segments into compressibility vectors. The dictionaries used for compression are constructed by compressing the pre-processed images with LZW method, from a small number of randomly chosen images.

On these compressibility vectors, clustering with k-means is performed to get clusters of segments. It is considered that the segments belong to the same cluster have similar properties. Therefore we can replace them by one character (ASCII code), from which we get the text-transformed image.

Now we classify the text-transformed images based on the PRDC. The PRDC is used again to compress the text-transformed image to obtain compressibility vectors. And the dictionary is constructed by compressing the text-transformed image with LZW method. In the same way, clustering is performed on the compressibility vectors. Finally, images are classified by the proposed method.

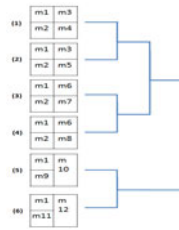
### 3 Experiments

Experiments with using artificial images and real-world images were carried out. In experiment of using the artificial image we proved the principle confirmation of image classification ability of the proposed approach. And also we showed the relation between complexity of the image and the segment length  $L$ . In the experiment of using the real-world image, the expression ability of the proposed approach is verified by comparing the recovered image and original image. A well-known benchmark dataset named Caltech101 was used to verifying the proposed approach. The results are evaluated by measures of recall, precision and F-measure.

#### 3.1 Experiment 1

The proposal was examined by using artificial images as follows. First for confirming image representation and classification ability of the proposed approach, images with some same regions were intentionally generated. Based on these images, the classification ability of the proposed approach was confirmed. Then the relation between complexity of image and segment length  $L$  in text was examined. The segment length of the proposed method for representing image with different complexity was examined.

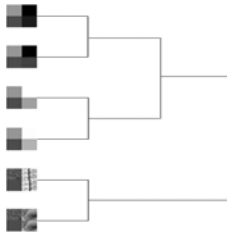
As shown in Fig. 1 we made 6 images with the same size, and made them into regions m1-m12. The images of m1-m12 were different images of one color cut from some nature images. The (1) and (2) contained 3 same regions, which means that they were supposed to be in one cluster. The same assumption could be made for (3) and (4), (5) and (6). We made different images of m1-m12 with different patterns as shown in Fig. 2. By using the artificial images and the assumption, we could show if the proposal could classify images and relation of segment length and classification ability.



**Fig. 1** Relation of artificially generated image.



**Fig. 2** Artificial image pattern.

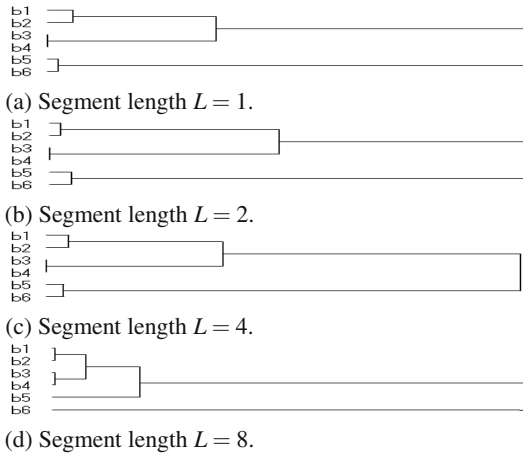


**Fig. 3** Dendrogram of artificial images.

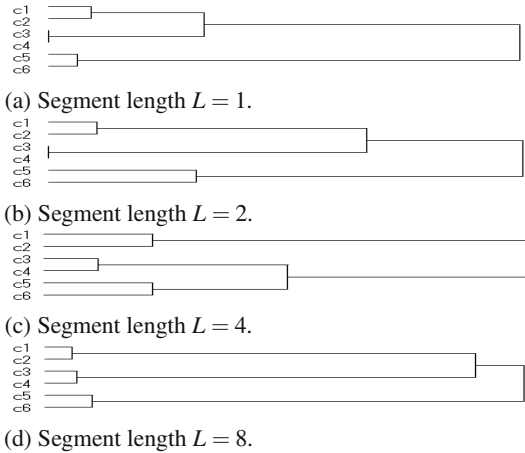
First for experiment of similarity of images, we performed classification for 6 images like shown in Fig. 1 with  $L = 4$ . The classification result was shown in Fig. 3. For this simple case, the proposed method was proved to be able to represent and classify images.

Second, for images like in Fig. 2 with complicated pattern, we used segment length of  $L = 1, 2, 4, 8$ . The result of dendrogram was shown in Fig. 4, 5 and 6. The  $b_i, c_i$ , and  $d_i (i=1-6)$  indicated the different pattern of images.

From the results we could see that the images classification ability was proved. For the relation of  $L$  and images complicity, for images with simple properties like in Fig. 2(b), good result was shown at  $L = 1, 2 \dots 8$ . This result was similar to it in Fig. 1. And even long segment ( $L = 8$ ) gave similar result to the assumption. On the other hand for images with complicated properties like in Fig. 2(d), good result were given when  $L = 1, 2, 4$ . The result was not good at  $L = 8$ . A big value of  $L (= 8)$  did not provide good classification ability for images with complex properties. While the smaller  $L (= 1, 2)$  provided good classification ability for images with complicated properties.



**Fig. 4** Dendrogram of image with simple property (pattern b).

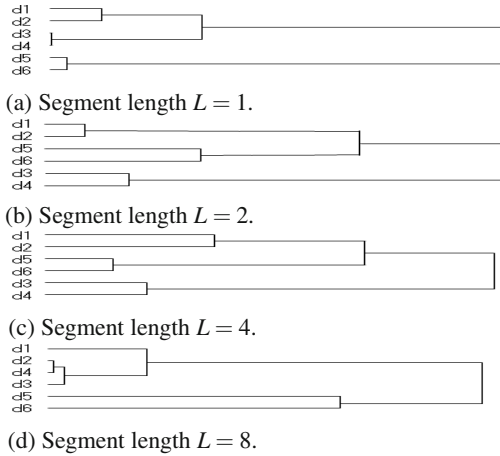


**Fig. 5** Dendrogram of image with complex property (pattern c).

### 3.2 Experiment 2

We used nature images in this experiment. For the simplicity, we chose the images with similar objects and the directions of the objects were similar as well. Three types of images of flower, elephant and tiger, 30 of each were used. We used the objects cut from the background to make the classification for the simplicity, since there are already image extraction methods for this process.

The F-measure from flower and elephant were at 0.93. and 0.74 separately. For tiger, 62% was classified correctly. By using the main objects which were cut from the images, the proposed method showed its ability for the classification of nature images.



**Fig. 6** Dendrogram of image with complex property (pattern d).

## 4 Conclusions

In this study, we proposed an algorithm for image representation and classification. In the proposed algorithm, images are transformed to texts and represented by compressibility vectors. The experiments were performed with artificial images and also nature images with simple and complicated properties separately. The experiment results showed that the proposed method was able to represent image data. The proposed method showed its efficiency in classification of text-transformed images.

**Acknowledgements.** This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 22500122, 2010.

## References

1. Richard, P.E.H., Duda, O., Stork, D.G.: Pattern classification, 2nd edn. John Wiley and Sons (2001)
2. Weiming Hu, Z.C.Z.F., Wu, O., Maybank, S.: Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6), 1019–1034 (2007)
3. Toshinori Watanabe, K.S., Sugihara, H.: A new pattern representation scheme using data compression. *IEEE Trans. PAMI* 24(5), 579–590 (2002)

# Numerical Analysis of Touch Mode Capacitive Pressure Sensor Using Graphical User Interface

Moon Kyu Lee, Jeongho Eom, and Bumkyoo Choi

**Abstract.** Design through numerical methods requires experimental verification of micro electrical mechanical system (MEMS) processes and numerous parametric analyses. In particular, to evaluate the performance of capacitive pressure sensor, electrical calculations are necessary along with structured numerical analysis. For this reason, pre- and post- processors are necessary that would create models and manage results through a dedicated graphic user interface for sensor analysis. Therefore, this study is designed to make efficient and convenient numerical methodology of touch mode capacitive pressure sensor (TMCPS) using graphical user interface (GUI), appropriate for measuring tire pressure, and analyze its capability and linearity responding to geometric design parameters. Design parameters are size of diaphragm, thickness of diaphragm and gap distance between diaphragm and insulator. All of the analysis process is controlled by GUI program. Analysis results show capacitive-pressure graph, regression curve and pressure-contact area graph using calculation and finite element analysis. This research shows negative association between the thickness and the gap for optimum diaphragm. Optimum diaphragm is 2.5mm by 2.5mm size, 25 $\mu$ m thickness, 4.5  $\mu$ m gap within 20~40psi which means the range of tire pressure.

## 1 Introduction

To produce micro-sensors, complicated MEMS processes are generally used. To develop sensors that operate in a specific environment, numerous trials and errors are gone through in determining the design variables of the sensors and related temporal and economic losses are huge. Therefore, numerical methods have been used to design micro-sensors in the past. However, design through numerical methods requires numerous parametric analyses. Manually producing many models

---

Moon Kyu Lee · Jeongho Eom · Bumkyoo Choi

Department of Mechanical Engineering, Sogang University, Seoul, Korea

e-mail: {Abraham, honim, bkchoi}@sogang.ac.kr

in relation to the verification and analyses is inefficient. In particular, to evaluate the performance of capacitive pressure sensor, electrical calculations are necessary along with structured numerical analysis. Therefore, pre and post-processors are necessary that would create models and manage results through a dedicated graphic user interface for sensor analysis.

One of major application areas of capacitive pressure sensors is real time monitoring of vehicle tire pressure, i.e., measuring changes in vehicle tire pressure to see if 30psi which is appropriate vehicle tire pressure is maintained. Since insufficient tire air pressure reduces the handling safety and driving and braking performance of vehicles and increases fuel consumption, it is necessary to monitor vehicle tire pressure [1]. Then, capacitive pressure sensors are frequently used because they are less affected by external temperatures and highly sensitive to pressure changes [2][3]. Capacitive pressure sensors can be divided into touch mode and non-touch mode pressure sensors. The voltage change graph of non-touch mode capacitive pressure sensors relative to pressure changes has lower linearity [4] than touch mode capacitive pressure sensors (TMCPS) [5]. In this respect, in this study, we examined sensor variables that would have the most excellent linearity and sensitivity in the measurement range of the sensors when TMCPSs are used.

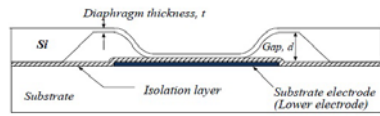
This paper presents a technique to analyze TMCPSs through graphic user interface (GUI). This analytical method is advantageous in that it enables mechanical analysis and electrical analysis to be conducted with one interface through linkage between a commercial tool of finite element method (FEM) and computational code. Therefore, this study is purposed to introduce a numerical linkage analytical method for designing TMCPSs and design the sensors using interface environments for users.

## 2 Analytical Method

The TMCPS to be designed in this study includes a diaphragm and an insulator. The diaphragm is a silicon plate structure that is transformed by external pressure to come into contact with the insulator. The insulator is a plate structure with electrode patterns that enable the occurrence of electrostatic capacity. If the diaphragm is mechanically transformed by external pressure, changes in the electrostatic capacity will occur simultaneously. Therefore, to analyze the TMCPS, mechanical analysis and electrical analysis should be conducted simultaneously. However, mechanical analysis and electrical analysis cannot be conducted simultaneously in general commercial analysis programs. In this regard, in this study, an ABAQUS-MATLAB linking analysis program, typical commercial program, was designed. In this program, ABAQUS implements finite element analysis as a mechanical analysis processes and MATLAB not only conducts modeling for the analysis but also implements electro-dynamic calculation processes for the results.



**Fig. 1** Touch mode capacitance pressure sensor (TMCPS)



As shown in Figure 1, this finite-element analysis model is composed of a diaphragm and an insulator. The diaphragm is a silicon plate with an elastic modulus of 150Gpa and the insulator is made of rigid glass. Table 1 shows the physical properties of the silicon. Important variables in this analysis are the size and thickness of the diaphragm and the space between the insulator and the diaphragm. In this study, analysis was conducted through the changes in the three variables that can be identified in Table 2. The range of pressure to be measured was determined to be 20~40psi which is the range of tire pressure of general passenger vehicles. The size of the diaphragm is 2.5×2.5mm<sup>2</sup>, the thickness of the diaphragm is 15~60µm and the space between the insulator and the diaphragm is 0.5~50µm. This analysis was conducted on multiple thicknesses and spaces prepared at intervals of 5µm and 0.5µm respectively.

All the edges of the diaphragm are fixed. The diaphragm is non-linearly transformed relative to imposed pressure. The finite element method analyzes the deformation by pressure using ABAQUS (ver. 6.9). The capacitance values are drawn through information on non-linear changes in the distance between the diaphragm and the insulator obtained by mechanical analysis and a relational expression applied with Eq. (1).

**Table 1** The material properties of diaphragm and insulator

Material properties	Specific value
Young`s modulus of the bulk silicon, E, (GPa)	150
Poisson`s ratio of the bulk silicon, ν	0.3
Permittivity of vacuum, ε <sub>v</sub> , (pF/mm)	8.85E-03
Dielectric constant of air, ε <sub>a</sub>	1
Dielectric constant of insulator, ε <sub>i</sub>	11
Thickness of insulator, t <sub>i</sub> , (µm)	0.32

**Table 2** The range of design variables

Design vairables	Specific value
Size of square diaphragm, a, (mm <sup>2</sup> )	2.5 × 2.5
Thickness of diaphragm, t, (µm)	15 ~ 60
Gap distance between diaphragm and insulator, d, (µm)	0.5 ~ 5

$$C = \iint \frac{\epsilon_0 \epsilon_a \epsilon_i dx dy}{\epsilon_a t_i + \epsilon_i (d - \omega(x, y))} \tag{1}$$

where C is capacitance in the unit of pF.  $\epsilon_0$  is permittivity in a vacuum and  $\epsilon_a$  and  $\epsilon_i$  are the dielectric constants of the air and the insulator respectively.  $t_i$  is the thickness of the insulator and  $d$  is the initial space between the diaphragm and the insulator.  $\omega(x, y)$  represents deflection at certain points of the diaphragm [4].

In this study a graphic user interface (GUI) was made using MATLAB (Fig. 2). This GUI tool controls all the processes of analysis of the sensor and is composed of four steps in total. The first step is a step of the formation of an input file for the beginning of the finite-element analysis. The input file includes modeling parameters, physical property values and constraint information. The second step is a step of mechanical analysis using ABAQUS. As shown in Fig. 3, the ABAQUS command processor implements modeling and analysis based on information in the input file. The third step is a step to form an output file through the results of the mechanical analysis. The last step is to draw the electrostatic capacity with the MATLAB mathematical calculation method. Simultaneously, a graph of changes in the section of the diaphragm relative to pressure changes as shown in Fig. 4a and a graph of capacitance relative to pressure as shown in Fig. 4b are displayed.

Through this process, quite some data on changes in the variables were collected and the tendency of the data can be grasped through a post-processor. To design TMCPs variables appropriate for tire pressure measurement, the sensitivity and linearity of electrostatic capacity relative to pressure changes in a range of 20~40psi were calculated. In this study, the tendency was found and the most appropriate variable values were selected through comparison between calculated values.

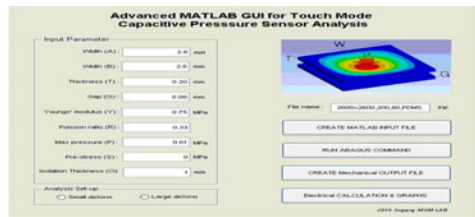


Fig. 2 Graphical user interface

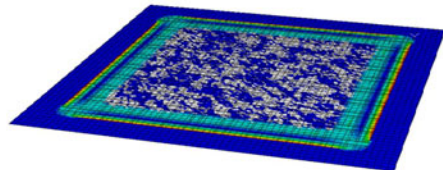
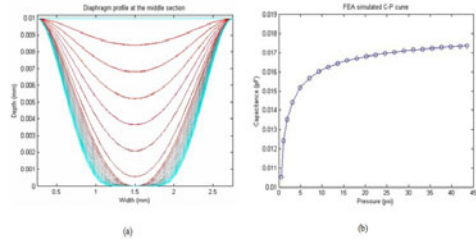


Fig. 3 Captured image of analyzed model

**Fig. 4** The result of pre and post processor (a) deformation of diaphragm (b) capacitance using MATLAB calculation



### 3 Analyzed Results

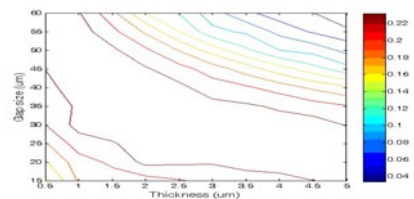
Through the results of calculation of the sensitivity and linearity, the values of the variables appropriate for tire pressure measurement and a tendency between the analytical variables and the sensitivity/linearity were drawn.

Changes in the sensitivity in relation to changes in the thickness and space showed a clear tendency. As shown in Fig. 5, points that had a certain degree of sensitivity showed a tendency of the thickness and the space being inverse proportional to each other. That is, to maintain a certain degree of sensitivity, if the thickness increases, the space should be decreased and if the thickness decreases, the space should be increased.

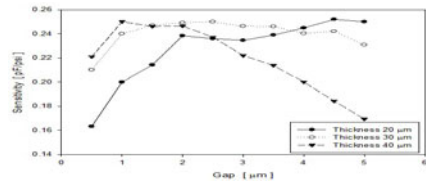
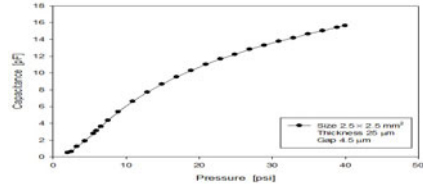
Fig. 6 shows changes in linearity relative to changes in the space when the thickness of the diaphragm is constant. The graph that can be identified in Fig. 6 shows that the sensitivity is the highest at a space of 2.5 $\mu$ m when the thickness is constant at 30 $\mu$ m. Graphs of all the thicknesses in this analysis showed the same tendency.

As identified in Fig. 5, the graphs of the results of the analysis showed a tendency that to maintain a certain degree of sensitivity, when the thickness increased, the space decreased and when the thickness decreased, the space increased. Therefore, in this paper, the space values that showed the highest values of sensitivity by thickness were found and the relationship between the thickness and the space was shown. The results showed a tendency that as the thickness increased, the space decreased.

Finally, in this study, a shape of the diaphragm was drawn that has the sensitivity and linearity the most suitable to the measurement of vehicle tire pressure in a range of 20~40psi within the range of the variables of the analysis conducted. As shown in Fig8, it was indicated that a diaphragm in a size of 2.5 $\times$ 2.5mm<sup>2</sup> with a thickness of 25 $\mu$ m and a space of 4.5 $\mu$ m was the most suitable to the measurement of tire pressure within the range of variables in this analysis (Fig. 7).



**Fig. 5** Nomogram of sensitivity for gap size and thickness

**Fig. 6** Relationship of gap vs sensitivity**Fig. 7** Relationship of pressure vs capacitance graph for optimum case

## 4 Conclusion

In this study, first, capacitive pressure sensor analysis was conducted using this program made by the MATLAB GUI. Through this process, the efficiency and convenience of this program were identified. Second, through the results of analysis of capacitive pressure sensors, in this study, the tendency of the linearity and sensitivity in the capacitance-pressure graph relative to analysis was identified. Finally, through the results of analysis, in this study, the thickness and space of the diaphragm in a size of  $2.5 \times 2.5 \text{ mm}^2$  suitable to the measurement of tire pressure were drawn

- In this study, a proportional tendency existing between the size of TMCPS diaphragms and the sensitivity of the pressure-electrostatic graph was identified.
- In this study, it was identified that diaphragms in a size of  $2.5 \times 2.5 \text{ mm}^2$  with a thickness of  $25 \mu\text{m}$  and a space of  $4.5 \mu\text{m}$  were the most suitable in measuring vehicle tire pressure in a range of 20–40psi within the range of variables used in this analysis.
- In this study, in the analysis of the TMCPS, the efficiency and convenience of this program made through the MATLAB GUI were identified.

**Acknowledgments.** This work was for development of the intelligent-tire has been supported by the Ministry of Knowledge Economy. (No. 10033705).

## References

1. Oh, J.G., Lee, J.Y., Choi, B.K., Kim, J.G.: SAW Based passive radio tire pressure monitoring sensors. KSAE, 1083–1088 (2009)
2. Ko, W.H., Wang, Q.: Touch mode capacitive pressure sensor. Sensor and Actuation, 242–251 (1999)

3. Ko, W.H., Wang, Q.: Touch mode capacitive pressure sensor for industrial applications, pp. 284–289. IEEE (1997)
4. Wang, Q., Ko, W.H.: Modeling of touch mode capacitive pressure sensor and diaphragms. *Sensor and Actuators*, 230–241 (1999)
5. Shaikh, M.Z., Kodad, S.F., Jinaga, B.C.: Modeling and simulation of MEMS characteristics: a numerical integration approach. *JJTIT*, 415–418 (2008)

# Multithreading Embedded Multimedia Application for Vehicle Blackbox

Jajin Koo, Jeongheon Choi, Youjip Won, and Seongjin Lee

**Abstract.** Recent introduction of Vehicle Black Box system enabled the user to use the system to collect driving information such as location, time, and status of the vehicle, which can be analyzed to find the cause of the accident. In this paper, we provide an efficient way of using the data generated from Vehicle Black Box system. We analyze the performance of proposed multi threaded system by comparing it with single threaded implementation of the system. The paper shows that the processing time, and initialization time of the multi threaded system is 39%, 233% faster than single threaded Vehicle Black Box System, respectively. We also show that response time of multi threaded system for single control message is about 10% faster than single threaded system.

## 1 Introduction

Recent introduction of Vehicle Black Box system enabled the user to use the system as supporting material in the event of vehicle accidents and exploit the system to enhance driving experience. Main purpose of the vehicle black box is to collect driving information such as location, time, and status of the vehicle, which can be analyzed to find the cause of the accident. Modern systems also capture video and audio data along with other status information, which increases overhead in the embedded system.

Chet et al. [1] designed the black box as warning system to prevent users from exceeding speed limits, and Kassem et al. [2] focused on minimizing the system in hardware level, while maintaining similar level of functionalities as other black box systems. Shui et al. [3] developed black box system based on ARM+DSP to monitor and record environmental information while users drive the vehicle; this system introduces exploiting video recorder on black box system. Some of recent

---

Jajin Koo · Jeongheon Choi · Youjip Won · Seongjin Lee

Hanyang University, Seoul Korea

e-mail: [space0215, rhythmical, yjwon, james}@ece.hanyang.ac.kr](mailto:{space0215, rhythmical, yjwon, james}@ece.hanyang.ac.kr)

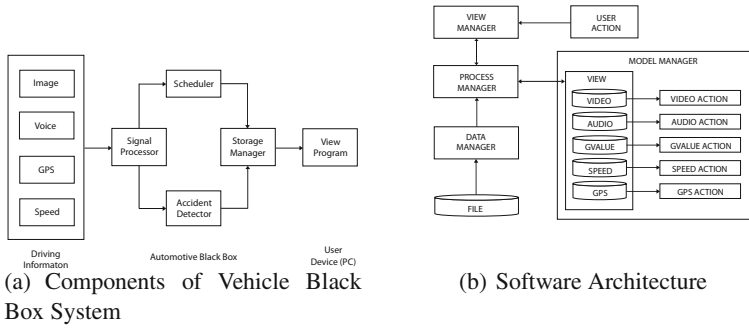


Fig. 1 Components and Software Architecture of Vehicle Black Box System

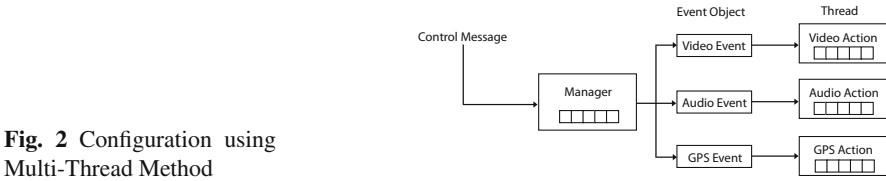


Fig. 2 Configuration using Multi-Thread Method

work on the field [4, 5] presents multi-threading programming model on distributed environment to enhance the processing speed.

Most of the system disregards the efficiency in manipulating the generated data and processing them to form useful information. Typical black box system captures video data, audio data, pressure, and GPS data. In this paper, we develop a new viewer system that exploits efficient processing techniques to illustrate and present generated data. Another important issue in black box is to utilize resources while enhancing response time and processing time; to provide the remedy in enhancing response and processing time, we exploit multi-thread in implementing the system.

## 2 Vehicle Black Box System

Main purpose of Vehicle black box system to capture real-time driving information to provide data which can be used in analyzing cause of accidents. Most existing vehicle black box system employees Digital Video Recorder (DVR) to record scene or same view as the driver. Recent vehicle black box system not only captures the scene, but also captures speed of the vehicle, direction, acceleration and break operation, external pressure, and many more. The black box system exploits vehicle sensors and Global Positioning System (GPS) to collect such data.

Fig. 1(a) illustrates the components used in vehicle black box system with Driving Information, Automotive Black Box, and User Device. Driving Information is considered as generated data while the vehicle is in active mode. It collects images,

voice, GPS, speed through sensors attached to Automotive Black Box. Collected general driving data is stored in Automotive Black Box. It manages data collected by the Driving Information. We separate general driving data and accident driving data according to policy of Storage Manager and saves the data to the external storage device. User Device is a front-end user interface that presents stored Driving Information to the user.

### 3 System Modeling

This paper focuses to efficiently configure Viewer which presents collected data. The Viewer not only generates appropriate views to the user with the collected data, but also decodes and encodes the video data, reflects user control information, and manipulates GPS information in real-time fashion. In process of generating several views for specific purposes, single thread application cannot efficiently deal with synchronization issues. And, it has to suffer from slow response time when user messages coincide with multiple other system generated messages.

We design Viewer that presents the data collected by black box and Fig 1(b) illustrates the software architecture. First component in the viewer is Data Manager (DM) which processes data; DM loads driving data from storage device (SD card, HDD) and analyzes a file header to distinguish video, audio, GPS, and etc. DM transmits the data to each Viewer. Second is Viewer Manager (VM) which is implemented with Model-View Control (MVC) software design pattern. VM collects control messages from user and acts as a agent and interface between user and program. Process Manager (PM) sends control messagees from VM and parses data collected from DM for each Viewer. Model Manager (MM) creates and manages the Viewer for each data; it also processes control messages issued by user.

### 4 Multi-thread Technique

As Section 3 describes there are synchronization and response time issue in single-threaded implementation of the system. In this paper, we propose multi-threaded black box system that provides remedy to both issues. Fig. 2 illustrates how Manager make use of thread according to different events from three Viewers (video, audio, and GPS data.) When Viewer Manager receives Control message, it calls Event Object that runs on separate thread.

Control message is processing unit which drives each viewer. User requests (play, stop, pause, and etc) and application generated requests (file open, close) are examples of control messages. The Manager is responsible for creating and managing each viewer, and storing the control messages. Both Manager and Viewer thread runs on user level thread and they both keep queue to store received control messages. Viewer threads processes the data which are collected by black box. Event object is kernel mode synchronization object for Manager and Viewer thread. When black box system is activated, it first creates appropriate environment for the system to run thread as shown in Fig. 2. Next, the Manager thread waits for messages



from user or the system itself, and the Viewer thread also waits for messages from the Manager thread. When the user or the system issues messages, the Manager thread stores messages in the queue, and each Event Object, which is associated with each Viewer, is set as signaled. When the Event Object is set to signaled, the Manager thread delivers the messages to each Viewer thread, and Viewer thread executes individual requests using each queue message, until all messages from the queue are processed. After all messages are processed, the Event Object resets its status to *Not-signaled* and waits until the state changes to *Signaled*.

## 5 Experiment

We run and test the black box system under Intel(R) Core 2 Duo E8200 and 2Gbyte of DDR2 Memory under Windows XP 32bit. We develop the Viewer system that presents Vehicle Black Box System data, implemented with MFC Microsoft video C++ 6.0. Since the main objective of the paper is to produce a system with better response time, we implemented the system to run on single thread and multi threaded environment. As described in Section 4 the threads used in the system are user level threads provided by the Operating System. In order to handle thread synchronization, we exploit event synchronization.

### 5.1 Processing Time

In order to measure processing time, we compare processing time for given size of data between the single threaded system and multi threaded system. Data used in this experiment is video, audio, GVALUE (pressure of the vehicle), speed, and GPS of the vehicle. A unit of data is denoted as Pack. The Pack consists of one minute of video data, one minute of audio data, and sixty consecutive GVALUE, Speed, and GPS which are collected every second. In order to measure the processing time, we introduce another unit called Group of Data (GoD). For example, GoD of one corresponds to ten frames of video data, ten frames duration of audio data, single GVALUE, single Speed, and single GPS data.

We range the GoD value from 1 to 60 and measure the response time and measure the response time of single threaded system and multi threaded system. Fig. 3(a) illustrates the difference between the two systems. In case of the single threaded system, Viewer processes video, audio, GVALUE, Speed, and GPS information sequentially. In other words, FFmpeg decodes video and audio data; then Viewer draws GVALUE and Speed graph; finally, GPS information is processed. Under such constraints, each Viewer has to wait for data processing time, I/O delays, and Memory access delays. When the system is running under multi thread environment, independent run of each thread removes processing delays. As number of GoD increases the effect becomes clear; when the number of GoD is 60 meaning 600 frames of video data, multi threaded system is about 39% faster than single threaded system.

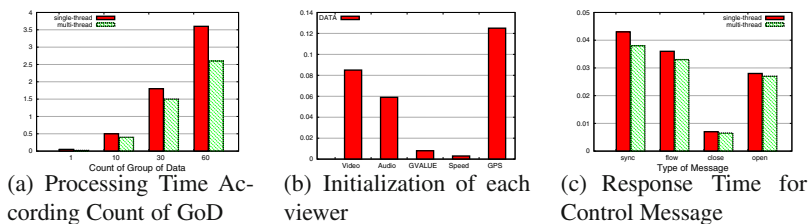


Fig. 3 Processing Time and Initialization Time

### 5.2 Initialization Time

Next, we measure the initialization time of the two systems. It is important to reduce the initialization time as it decides the user experience. It accounts of time the system is loaded and setting the Event Object to *Not-Signaled* state. It excludes the time spent in loading metadata from a file to memory, and separating the data corresponding to each Viewers. Assuming there are loaded data from files on each Viewer, initialization time measures the necessary time for memory allocation, time spent in data parsing. Fig. 3(b) illustrates initialization time for each Viewer. GPS Viewer uses about 0.12 seconds, Video Viewer consumes about 0.08 seconds, and Audio Viewer consumes about 0.06 second in initializing the Viewer. Unlike GPS, Video, Audio Viewers, which has to initialize different codecs and connect to GPS server, GVALUE and Speed only has to retrieve data from the sensor in Automotive Black Box in Fig. 1(a). GVALUE and Speed Viewer consumes 6.5msec and 0.3msec, respectively.

We measured the sum of initialization time difference between the two systems. It shows that multi threaded system (120 msec) is 233% times faster than single threaded system (280 msec). Note that the system initialization time is same as longest time consuming operation, which is GPS Viewer, since all other Viewers has to wait until GPS Viewer is ready to synchronize with other data on the Viewers.

### 5.3 Response Time

Third experiment measures the response time for external control messages of single threaded system and multi threaded system. The control messages used in the experiment is as follows: SYNC (Move position), FLOW (Synchronization message), CLOSE (Exit the program), and OPEN (Start the program). Fig. 3(c) illustrates the difference between the two systems. We measure the response time of each control messages. Similar to processing time experiment (Fig. 3(a)) multi threaded system shows better response time than single threaded system.

The response time of SYNC message, which updates position information, is reduced about 100 msec; FLOW message, which synchronizes the data between each viewer, is reduced about 2 msec. Response time for open and close message is almost same for both systems but slighter shorter for the multi threaded system. The

result shows that effect of multi threading in response time is not too great compared to processing time. Note that the result of Fig. 3(C) is response time of single message. Therefore, effect of multi threading in the real system can be significant when there are multiple messages. Proposed system successfully addresses the issue in response time in handling control message and the system resource utilization.

## 6 Conclusion

As the data generated by the embedded system is increasing, the system has to efficiently manage the utilization resources. Vehicle Black Box system is one of such area, which requires to process large amount of data. In this paper, we provide an efficient way to manipulate and manage various data in a single system with short processing and response time. We propose multi threaded system that independently processes data through separate data viewer. We analyze the performance of proposed system by comparing it with single threaded implementation of the system. The paper shows that the processing time, and initialization time of the multi threaded system is 39%, 233% faster than single threaded Vehicle Black Box System, respectively. We also show that response time of multi threaded system for single control message is about 10% faster than single threaded system.

**Acknowledgements.** This work was supported by Mando Corporation and the NRF (National Research Foundation) grant funded by the Korea government (MEST) (No.R0A-2009-0083128).

## References

1. Chet, N.C.: Design of black box for moving vehicle warning system. In: Proceedings. Student Conference on Research and Development, SCORED 2003, pp. 193–196 (2003)
2. Kassem, A., Jabr, R., Salamouni, G., Maalouf, Z.: Vehicle black box system. In: 2008 2nd Annual IEEE Systems Conference, pp. 1–6 (2008)
3. Shui, Y., Zhen, D., Hong-Bin, C., Jin-Zhong, C., Lei-Ting, C.: Design of black box of vehicle system based on arm+dsp. In: Application Research of Computers 2008 (February 2008)
4. Tian, L., Ming Fan, S., Zhao, W.: Application and research of multi-thread technology in spaceborne sar imaging. In: International Conference on Space Information Technology (2010)
5. Qin, Z., Gu, J.: Multi-thread technology based autonomous underwater vehicle. In: 2010 8th IEEE International Conference on Control and Automation (ICCA), pp. 898–903 (2010)

# An EMD-Oriented Hiding Scheme for Reversible Data Embedding in Images

Chi-Yao Weng, Hung-Min Sun, Cheng-Hsing Yang, and Shiuh-Jeng Wang

**Abstract.** In 2006, Zhang and Wang presented the Exploiting Modification Direction (EMD) method for embedding secret message. The EMD approach is a steganographic approach in which only pixel in  $[1, -1]$  is changed after a message is hidden. This investigation presents a novel reversible data hiding scheme that is based on the modulo function and histogram shifting algorithm. The message can be extracted using a critical function and the original image can be recovered without loss way by histogram shifting. The histogram in this scheme is formed by exploiting the difference between the predicted of pixels and those of their neighboring pixels. The peak point and zero point in the distribution in the histogram are found and the value between the peak point and the zero point is shifted by  $\pm 1$ ; then, the message is hidden at the peak point using the critical function. Experimental results reveal that the developed method has a high capacity without causing noticeable distortion. In one-level hiding, the proposed method maintains a large image quality of 50dB. Furthermore, multiple level embedding can be utilized to provide high capacity.

**Keywords:** Exploiting Modification Direction, Reversible Data Hiding, Histogram Shifting, multiple level embedding.

---

Chi-Yao Weng

Department of Computer Science and Engineering, National Sun Yat-sen University

Hung-Min Sun

Department of Computer Science, National Tsing-Hua University, Hsinchu, Taiwan 300

Cheng-Hsing Yang

Department of Computer Science, National Pingtung University of Education, Pingtung, Taiwan 900

Shiuh-Jeng Wang

Department of Information Management, Central Police University

e-mail: sjwang@mail.cpu.edu.tw

## 1 Introduction

Steganography is a branch of the field of data hiding. It involves embedding secret data into covering multimedia, such as images, video, and text, for communication between a receiver and a sender [1]. The two crucial factors are the transparency of the cover image and the data embedding capacity. Transparency is critical to perfect steganography as secret messages must be undetectable [2, 3]. The capacity of embedding data is reflected by embedding ratio for the multimedia, which is the number of hidden bits in each cover pixel [4]. The embedding ratio should be as large as possible, but embedding distorts the image. Accordingly, the goal of steganography is to develop an algorithm that provides high embedding efficiency and a high embedding ratio simultaneously.

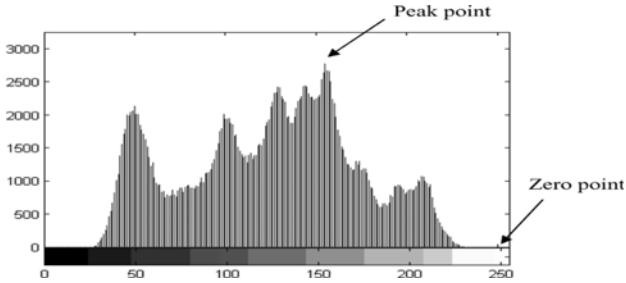
Numerous reversible data approaches have been developed [5-8]. Two basic techniques, the difference expansion (DE) method and the histogram method, can be used to recover the context. In 2003, Tain [5] proposed a method that in which an embedded message was added to the pixel differencing. Since then, other approaches have been proposed to improve on the DE method [6]. In 2006, Ni *et al.* [7] presented a reversible data hiding method that utilizes the zero and peak points in an image histogram to embed a message and achieve reversibility. In their approach, secret data will be hidden when the pixel at the peak point. Accordingly, the approaches pay more attention to increasing the number of peak points to increase embedding capacity [8]. This study presents a novel reversible data hiding approach that is based on the modulo function and predicted error. The predicted error is computed using the predictor method and the current pixel. Notably, the side-match predictive method that was proposed by Yang is adopted [9]. The confidential information is hidden at the peak point of the predicted error using the modulo function [10]. Therefore, a large peak point means that many bits can be embedded. Our approach outperforms previously presented methods.

## 2 Literature Review

The histogram-based scheme that was presented by Ni *et al.* in 2006 exploits the peak point of an image histogram to hide a message in the image [7]. Firstly, a statistical process is applied to a cover image to create an image histogram. Fig. 1 displays a histogram of the 'Lena' image. Then, the peak and zero points in the image histogram are found. The peak point is one whose pixel value occurs most frequently in the image and has the highest peak in the histogram. A zero point is one where the pixel value does not occur in the image.

Secondly, histogram shifting is applied, to shift/modify the pixel values between the peak and zero points. From this distribution of pixel values  $x_i$ ,  $x_i \in [0, 255]$ , a peak point and a zero point are found, and the values within range [peak point+1, zero point] or [zero point, peak point-1] are shifted toward the zero point or equivalently away from the peak point. Shifting produces an empty space next to the peak point in the image. The maximum data embedding capacity depends on the size of the peak point.

Finally, the secret data bit-stream  $s_i \in [0,1]$  is hidden at peak point. For each pixel value that equals the value of peak point, the secret bit is evaluated. If the secret bit equals “0”, then the pixel value remains unchanged; otherwise, the value is changed by 1. In this way, the stego-image is constructed.



**Fig. 1** A Histogram of Lena Image

### 3 Proposed Scheme

This investigation proposes a reversible data hiding scheme that predicted error image is utilized to execute the histogram shifting algorithm and modulo function. The histogram shifting algorithm uses the peak point and shifts neighboring pixel values by 1, then, the secret information will be concealed at peak point. The characteristics of the empirical histogram scheme are carefully observed and the shifting is found to be performed only toward one side, the left or the right of the peak point, by one unit. The value is shifted on both sides of the peak point in the histogram shifting method for hiding data. Then, the modulo function is adopted to hide the secret information in the direction of modification of the peak point. This subsection will elucidate the presented approach, and then the embedding and extracting algorithm in detail.

#### 3.1 Generation Predicted Error Image

In the proposed approach, the embedding process involves calculating a predicted error image between the original image and the predicted image, and then hiding the secret information in the peak point of the predicted error image. Accordingly, increasing the number of peak points increases capacity.

Here, the side-match scheme, which was proposed by Yang *et al.* [9], is utilized to compute the predicted error because there have high relation between current pixel and its neighbor pixel in a natural image. Fig. 2 displays the concept of the side-match. From Fig. 2, the predicted pixel  $P_{i,j}$  value can be obtained by the values of neighboring pixels  $P_{i,j-1}$ ,  $P_{i-1,j-1}$ ,  $P_{i-1,j}$ , and  $P_{i-1,j+1}$ . The details of the error prediction algorithm are as referred to [9].

$P_{i-1,j-1}$	$P_{i-1,j}$	$P_{i-1,j+1}$
$P_{i,j-1}$	$P_{i,j}$	

**Fig. 2** The concept of the side-match prediction.

### 3.2 Embedding Algorithm

The secret information is concealed at the peak point. The algorithm for concealing information in a gray-scale image is as follows.

1. Apply the side-match prediction method, to predict value and calculate the predicted error  $D$  between predicted value and current value.
2. Generate the histogram  $H(x)$  with  $x \in [-255, 255]$  from all predicted errors  $D_{i,j}$  with  $i, j \in [512]$ .
3. Find a peak point  $Pk$  and zero points ( $Z_1, Z_2$ ) that satisfy  $Z_2 < Pk < Z_1$  from the generated histogram. The peak point is divided into  $n$  blocks with two pixels, denoted  $(Px, Py)$ .
4. Shift the histogram as follows.
  - 4.1  $D'_{i,j}$  is set to  $D_{i,j} + 1$ , if  $D_{i,j} \in [Pk + 1, Z_1 - 1]$ .
  - 4.2  $D'_{i,j}$  is set to  $D_{i,j} - 1$ , if  $D_{i,j} \in [Z_2 + 1, Pk - 1]$ .
5. Embed the secret message into the predicted error  $D_{i,j}$  and produce out the new predicted error  $D'_{i,j}$  as follows.
  - 5.1 Get the secret bit string  $s_i$  and divide it in each of three bits, and three bits is converted into a decimal value  $S_i$ , where  $S_i = [0,7], i = 1,2,\dots,n$ .
  - 5.2 Divide the peak point into non-overlapping blocks, each block with two peak points, denoted  $(Pk_{xi}, Pk_{yi}) | i = 1,2,\dots,n$ . Notably, if the total number of peak points is odd, then the new last peak point ( $Pk'_{oi}$ ) is given by  $Pk'_{oi} = Pk_{oi} + s_i, s_i = [0,1]$ . Then, embed the secret message  $S$  in a block using the critical function  $f$ , which is defined as,

$$f(Pk_{xi}, Pk_{yi}) = (Pk_{xi} + Pk_{yi} \times 3) \bmod 9 \tag{2}$$

- 5.3 Apply the following rules to ensure the conditions of set  $S = f(Pk_{xi}, Pk_{yi})$  after the  $(Pk_{xi}, Pk_{yi})$  is changed. After this modification is made, a new block  $(Pk'_{xi}, Pk'_{yi})$  is obtained.

- (1)  $S = f(Pk_{xi} - 1, Pk_{yi}), Pk'_{xi} = Pk_{xi} - 1, Pk'_{yi} = Pk_{yi}$ .
- (2)  $S = f(Pk_{xi} - 1, Pk_{yi} + 1), Pk'_{xi} = Pk_{xi} - 1, Pk'_{yi} = Pk_{yi} + 1$ .
- (3)  $S = f(Pk_{xi} - 1, Pk_{yi} - 1), Pk'_{xi} = Pk_{xi} - 1, Pk'_{yi} = Pk_{yi} - 1$ .
- (4)  $S = f(Pk_{xi}, Pk_{yi} + 1), Pk'_{xi} = Pk_{xi}, Pk'_{yi} = Pk_{yi} + 1$ .
- (5)  $S = f(Pk_{xi}, Pk_{yi} - 1), Pk'_{xi} = Pk_{xi}, Pk'_{yi} = Pk_{yi} - 1$ .
- (6)  $S = f(Pk_{xi} + 1, Pk_{yi} - 1), Pk'_{xi} = Pk_{xi} + 1, Pk'_{yi} = Pk_{yi} - 1$ .

$$(7) S = f(Pk_{xi} + 1, Pk_{yi}), Pk'_{xi} = Pk_{xi} + 1, Pk'_{yi} = Pk_{yi}.$$

$$(8) S = f(Pk_{xi} + 1, Pk_{yi} + 1), Pk'_{xi} = Pk_{xi} + 1, Pk'_{yi} = Pk_{yi} + 1.$$

$$(9) S = f(Pk_{xi}, Pk_{yi}), Pk'_{xi} = Pk_{xi}, Pk'_{yi} = Pk_{yi}.$$

6. Obtain the new predicted error  $D'_{i,j}$  from the summation of the block value  $(D_{i,j}, D_{i,j})$  and  $(Pk'_{xi}, Pk'_{yi})$ . Notably, if the number of peak points is odd, then, then the new last one has a predicted error  $D'_{i,j} = Pk'_{oi} + D_{i,j}$ .
7. Finally, from left to right and top to bottom, perform the inverse of predicted error algorithm, and obtain the embedded pixel value  $P'_{i,j}$ .
8. Output the stego-image  $I'$ , peak point and zero points,  $Pk$ , and  $Z_1, Z_2$ .

Notably, the proposed scheme can be used to perform in cases of multiple-level data hiding by repeating the full embedding algorithm.

### 3.3 Extraction Algorithm

Extraction is similar to the embedded procedure except when the image is a stego-image. The extraction procedure is described below.

1. Apply side-match prediction method to each pixel in the stego-image  $I'$  from left to right and from top to bottom.
2. Generate predicted error value  $D_{i,j}$  from the stego-image of pixel  $P'$ .
3. Use  $(Pk, Z_1)$  and  $(Z_2, Pk)$  to extract the secret data and recover the predicted error values, as follows.
  - (1) If  $D'_{i,j}$  equals  $Pk+1$  or  $Pk-1$ , then, recover the predictive error as  $D_{i,j} = Pk$ , and keep the value into a temporary secret queue  $SQ$ .
  - (2) If  $D'_{i,j} \in [Z_2, Pk-2]$ , then the predictive error can be recovered as  $D_{i,j} = D'_{i,j} + 1$ .
  - (3) If  $D'_{i,j} \in [Pk+2, Z_1]$ , then the predictive error is recovered as  $D_{i,j} = D'_{i,j} - 1$ .
4. Recover the predictive error values  $D_{i,j}$  as original pixel value  $P_{i,j}$ .
5. Divide the  $SQ$  into many pairs,  $(Pk'_x, Pk'_y)$ . Notably, if the total number of  $SQ$  is odd, then the last secret bit string is  $Pk'_o$ .
6. Extract the secret message  $M$  by applying the modulo function in Eq. (2), and transfer  $M$  into a secret bit-string with three bits.

## 4 Experimental Results

This section presents the performance of our proposed method. Experiments are carried out on Four gray-level images with size  $512 \times 512$  are used. The secret bit



in our algorithm was generated randomly in C++. To evaluate the image quality achieved using our scheme, the peak-signal-to-noise-ratio (PSNR) is estimated and the MSE is determined as  $PSNR(db) = 10 \times \log_{10} \left( \frac{255^2}{MSE} \right)$ .

Tables 1 present results concerning image quality and embedding capacity for various levels cover images. According to Table 1, the proposed approach yield images of high quality and maintains acceptable quality when multiple level embedding is employed. Each pixel value is only changed by 1 when the proposed embedding algorithm is used, and the maximum MSE is 1. The approach outperforms previous approaches in terms of capacity, but maintains the same PSNR value, as seen in Table 2.

**Table 1** The resultant of PSNRs and hiding capacity (bits) in our approach using versus hiding level for four test images.

Images Level		Jet	Boat	Gold	Lena
1	PSNR	50.29	50.57	50.74	50.50
	Bits	88,186	48,049	41,100	59,656
2	PSNR	47.90	48.65	47.87	48.54
	Bits	144,055	89,567	78,591	109,597
3	PSNR	43.82	44.62	44.09	44.39
	Bits	186,631	124,439	109,425	148,972

**Table 2** Comparisons of Hiding capacities (bits) and average PSNRs with various approaches.

	Ni et al.'s method [8]	Li et al.'s method [10]	Our approach
Ave. PSNR	48.30	48.47	48.43
Jet	17,042	79,363	144,055
Boat	10,567	49,681	89,567
Gold	5,336	48,977	78,591
Lena	5,760	60,785	109,597

## 5 Conclusions

This study presents a reversible data hiding technique that is based on histogram shifting algorithm and the modulus function. In the proposed method, side-match predictor is utilized to predict error values and increase the height of peak points. The secret data are hidden at the peak point in the histogram. The modulus

function is adopted to improve image quality, because the peak point is shifted by  $[-1, 1]$ . According to the experimental results, the proposed scheme outperforms some previous methods, such as the histogram approach of Ni *et al.*, the scheme of Li *et al.*, and the method of Tsai *et al.*, in terms of both capacity and image quality. Furthermore, the presented scheme can be applied on multiple levels embedding to obtain high capacity, and ensure that the image is affected imperceptibly.

**Acknowledgements.** This research was partially supported by the National Science Council of the Republic of China under the Grant NSC 98-2221-E-015-001-MY3-, NSC 99-2221-E-153 -003, and NSC 99-2918-I-015-001-.

## References

1. Artz, D.: Digital steganographic: hiding data within data. *IEEE International Comput.* 5(3), 75–80 (2001)
2. Yang, C.H.: Inverted pattern approach to improve image quality of information hiding by LSB substitution. *Pattern Recognition* 3(3), 488–497 (2008)
3. Wang, C.M., Wu, N.I., Tsai, C.S., Hwang, M.S.: A high quality steganographic method with pixel-value differencing and modulo function. *The Journal of Systems and Software* 81(1), 150–158 (2008)
4. Yang, C.H., Weng, C.Y., Wang, S.J., Sun, H.M.: Adaptive data hiding in edge areas of images with spatial LSB domain systems. *IEEE Trans. Inf. Forensics Security* 3(3), 488–497 (2008)
5. Tianusing, J.: Reversible data embedding using a difference expansion. *IEEE Trans. Circuits Systems Video Technology* 13(8), 890–896 (2003)
6. Thodi, D.M., Rodríguez, J.J.: Expansion embedding techniques for reversible watermarking. *IEEE Trans. Image Processing* 16(3), 721–730 (2007)
7. Ni, Z., Shi, Y.-Q., Ansari, N., Su, W.: Reversible data hiding. *IEEE Trans. Circuits Systems Video Technology* 16(3), 354–362 (2006)
8. Li, Y.C., Yeh, C.M., Chang, C.C.: Data hiding based on the similarity between neighboring pixels with reversibility. *Digital Signal Processing* 20(4), 1116–1128 (2010)
9. Yang, C.H., Tsai, M.H., Wu, M.H., Jen, C.C.: Side-match approach for improving histogram-based reversible data hiding. In: *National Computer Symposium (NCS 2009)*, Taiwan (2009)
10. Zhang, X., Wang, S.: Efficient steganographic embedding by exploiting modification direction. *IEEE Communications Letters* 10(11), 781–783 (2006)

# Gait Control for Guide Dog Robot to Walk and Climb a Step

Manabu Kosaka

**Abstract.** It is useful to develop robots that can be used instead of guide dogs because the number of the guide dogs is significantly less than that of visually impaired humans demand. This paper describes desired capacity for locomotion such as walking speed and climbing steps that the guide dog robots should have in the practical use in daily life, and proposes the way to stabilize the posture of the robot by controlling roll, pitch and yaw angle detected by a gyro-sensor mounted in the trunk. The guide dog robot is manufactured and verified the practicability by experiments of walking and climbing a step.

**Keywords:** Gait control, Climbing a Step, Guide Dog Robot.

## 1 Introduction

In order to improve the Quality of Life (QOL) of visually impaired humans, textured paving blocks, music guidance at pedestrian crossing and so on have been developed in public space. However, the QOL degraded by impaired sight is more severe than auditory difficulties and language disability because vision has much more information than sound. Especially in the case of lethal accidents such as traffic accidents and so on, the lack of visual information might make it difficult to prevent the accidents. Therefore, visually impaired humans have difficulties in daily life [1].

There are guide dogs and white walking-sticks as the means of walking assistance. The guide dogs are more useful than white walking-sticks because they can predict and prevent the danger to the visually impaired humans. However, it costs about \$20,000 and 1.5 years to educate the guide dogs [1]. Furthermore, only 40%

---

Manabu Kosaka

Dept. of Mechanical Engineering, Kin-Ki University, Osaka, Japan  
e-mail: kosak@mech.kindai.ac.jp

dogs can acquire the ability to work as the guide dogs. Therefore, only one fifth of the people who want the guide dogs can get them [1]. Therefore, guide dog robots alternative to the guide dogs are helpful for social welfare.

The guide dog robots require a lot of functions. There are many conventional works about the functions [2]-[6]. Kotani et. al. developed a wheel robot and studied human interface, how to estimate the self-position of the robot itself, and how to find the way to a goal [2], [3]. Tsujita et. al. studied gait control and so on for quadruped robot [4]-[6]. However, there are no reports about guide robots climbing steps in public space.

This paper describes specifications of gaits for guide robots to be useful in public space, and develops a quadruped robot to satisfy the specifications,

In section 2, specifications of gaits for guide robots are considered. In section 3, a mechanism of a quadruped robot is designed to satisfy the specifications to walk on flatland. In section 4, a gait control suitable for the specification is proposed. In section 5, a mechanical design of a quadruped robot to satisfy the specification of climbing steps is done using simulations. In section 6, it is verified by some experiments that the developed quadruped robot satisfies the specifications.

## **2 Specifications about Gait of Guide Dog Robot**

In order to assist visually impaired humans on urban area instead of guide dogs, it is required to walk as fast as the humans and climb as high steps as humans can climb. The walking speed of healthy people on flatland is 2-3[km/h] [7]. However, visually impaired humans walk slower than healthy people, and the speed of their walk might be less than 2[km/h]. The height of the steps on urban area is restricted to be lower than 230[mm] by Building standard law [8]. Most of the guide dogs are Labrador retriever of which height is about 70[cm], length is about 150[cm], and weight is about 25[kg] [7]. Because there is possibility that guide dog robots collide with human, smaller body and lighter weight of the guide dog robot than the guide dog lead to be safer. Therefore, the following specifications about gaits of the guide dog robot are set:

- 1) Walking speed on flatland is 2[km/h].
- 2) Height to climb is 230[mm].
- 3) Body and weight are less than guide dogs.

## **3 Mechanism Design of a Robot to Satisfy the Specification of Walking Speed on Flatland**

Guide dogs are quadruped, and quadruped robot is considered to have suitable moving mechanism for guide dog robot because of the specification of climbing 230[mm] with the size and the weight which are smaller than those of guide dogs.

Therefore, quadruped mechanism is adopted for guide dog robot. Servo motors for commercial use are adopted for actuators in joints because they are cheaper and easier to get than the motors for special use. In this study, servo motors with rate torque 3[Nm] are used.

The robot has joints which guide dogs mainly use for walking, and the other joints of dogs are omitted for the sake of mechanism simplicity.

The design objective of this robot is to do bending and stretching exercises. Using the 3[Nm] rate torque motors, the design objective is satisfied by setting the weight of the robot 4 [kg] and the length of the leg 230[mm] which satisfy the specification 3).

## 4 Gait Control

Trot gait is adopted for this robot because the guide dogs use this gait when they guide humans. A gyro sensor is mounted in the body for detecting current posture. A posture control works online to decrease the difference between current and reference posture. When the difference between current and reference posture becomes large and the body tilts, the posture control make the leg that is the nearest to the center of gravity contact to farther place than the place that has been planned to be contacted, which leads to recover the irregular posture. For example, when both of the errors of the roll and pitch angles become large to plus, the robot will fall down to the right front direction. In this case, the right foreleg is place to the farther position in order to avoid falling down.

## 5 A Mechanism Design for the Specification of Climbing Steps

Trot gait in which three legs are always touching ground is adopted for this robot gait when this robot climb the steps because it is better to avoid falling down than to move quickly when visually impaired human is guided.

In order to expand the movable range of the center of gravity wider to y-axis direction, two shoulder joints were added as shown in Fig. 1 as patched circular

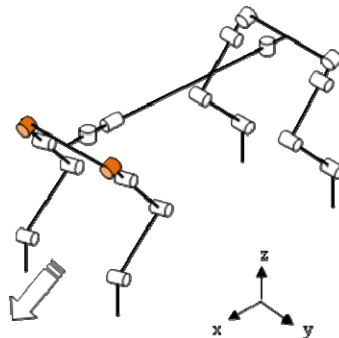
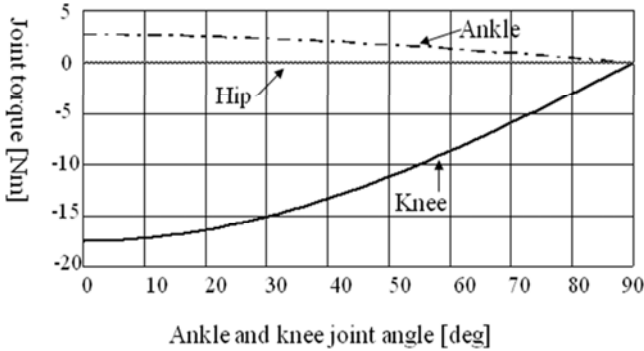


Fig. 1 Adopted Mechanisms



**Fig. 2** Joints torques and joints angle

cylinder. By using this mechanism, simulations with various gait patterns were done. In the best simulation, the robot achieved to climb the step. Therefore, this mechanism was adopted to manufacture the robot.

Next, it is confirmed that the robot has enough torque to climb the step. When the robot climbs the step, three feet always touch at ground because crawl gait is adopted. Therefore, the robot has enough torque to climb the step if the robot can stand up with two legs. The maximum torque is calculated when the robot stands up with two legs with the posture that require the biggest torque. Fig. 2 shows the torque calculation result. From Fig. 2, the maximum torque is 1.72[Nm]. The rate torque of the adopted servo motor is 3[Nm], so it is enough to climb the step.

## 6 Experiments

In this section, the effectiveness of the robot was verified by experiments.

### 6.1 Walking on Flat Land

Fig. 3a showed the posture of the robot body when the robot walked without the posture control. From Fig. 3a, the roll angle vibrated, the vibration became larger, and finally the robot fell down. Fig. 3b showed the posture of the robot body when the robot walked with the posture control. From Fig. 3b, the posture control worked well because the vibration of the roll angle was suppressed and the robot persistently kept walking. The walking speed was 2.3[m/s] that achieved the specification 1).

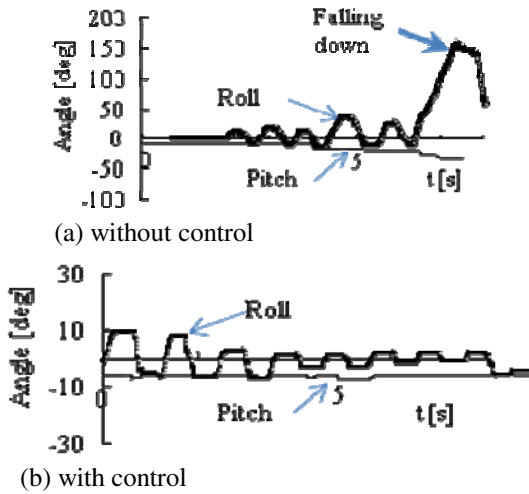


Fig. 3 Posture without/with the posture control

### 6.2 Climbing the Step

Fig. 4 shows scene in which the robot was climbing the step of the height 230[mm]. First, the robot contacted the right foreleg over the step. Next, the robot contacted the left foreleg over the step. Then, the robot walked using crawl gait



Fig. 4 Posture with both back legs over the step

with both elbows flex, and both back legs came near to the step. The left back leg contacted over the step. At this time, the body contacted on the step in order to avoid fell down when the left back leg swung. Finally, both of the back legs were on the step and the robot achieved climbing the step.

Fig. 5 showed the y-axis displacement of the center of gravity of the robot with/without the posture control. From Fig. 5, the maximum displacement was 38 [mm] in cases without the posture control. On the other hand, the maximum displacement decreased to 16[mm] by using the posture control, which lead to more stable step climbing.

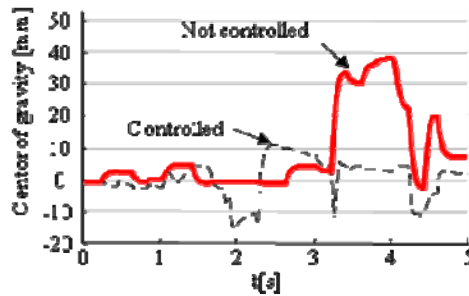


Fig. 5 The center of gravity

## 7 Conclusions

Specifications of moving ability on urban space and body size and weight for robots used instead of guide dogs were considered. A robot satisfying the specifications was designed. Current body posture was detected by a gyro sensor mounted in the body, and a posture control method was developed to reduce the difference between current and reference posture. It was verified by experiments that the posture control stabilized the robot walking and climbing, and the robot satisfied the specifications.

The limitation of the robot is climbing speed that is so slower than a real dog. Future research is to speed up the motion of the robot.

## References

- [1] Matsui, S.: Guide dog handbook, Bungeishunjuu, pp. 16–17 (2002)
- [2] Tsujita, K., Toui, H., Tsuchiya, K.: Dynamic Turning Control of a Quadruped Locomotion Robot using Oscillators. *Journal of Advanced Robotics* 19(10), 1115–1133 (2005)
- [3] Tsujita, K., Tsuchiya, K., Onat, A.: Decentralized Autonomous Control of a Quadruped Locomotion Robot using Oscillators. In: *Artificial Life and Robotics*, vol. 5, pp. 152/158. Springer (2003)



- [4] Tsutiya, K., Thujita, K.: Gait control of quadruped walking robot based on central pattern generator model. *Journal of Robotics Society of Japan* 20(3), 21–24 (2002)
- [5] Shi, Y., Kotani, S., Mori, H.: A Route Comprehension Support System for a Robotic Travel Aid. *Systems and Computers in Japan* 37(9), 87–99 (2006)
- [6] Mori, H., Kotani, S., Saneyoshi, K., Sanada, H., Kobayashi, Y., Mototsune, A., Nakata, T.: The Matching Fund Project for Practical Use of Robotic Travel Aid for the Visually Impaired. *Advanced Robotics* 18(5), 453–472 (2004)
- [7] Honda, M.: Guide dog as the best partner. *Gendai-shoseki*, pp. 2–36 (2002)
- [8] Building Standard Law: Cord 23

# Artificial Neural Network-Based Lot Number Recognition for Cadastral Map

Dave E. Marcial, Ed Darcy Dy, Silvin Federic Maceren, Erivn Rygl Sarno

**Abstract.** This paper discusses the implementation of an artificial neural network in detecting lot numbers in a cadastral map. Specifically, the image processing techniques used in the study are binarization, connected component labeling, and image resizing. A feed-forward with backpropagation artificial neural network was then implemented in the training and learning activities of the system. Fifty different maps with an average of 31 lot numbers at size 832 x 768 pixels were tested and it yielded an average of 90% detection after the 50th training. A 5808 x 4256 pixels map with 237 lot numbers was tested fifty times and gave an average of 84.78% detection rate. The satisfactory percentage of learning and detection based from the different test scenarios have supported that the implementation of the methods and algorithms is sufficient and accurate.

## 1 Introduction

In today's technology, computers are still having a hard time mimicking some of the capabilities in human expertise. Our machines have difficulty reliably reading handwriting. Although the creation of Artificial Neural Network (ANN) made this tasks possible [1].

Images require a lot of computational resources and complicated algorithms to be identified by a computer. ANNs however, does overcome these obstacles. The study focused on cadastral maps as object for the processing and detection of lot numbers using the ANN approach. Knowing that ANN has the ability to learn and adapt, and seeing that the numbers in cadastral maps are not uniformly written due to different handwriting and may be written horizontally or vertically, it was found that an exemplary reason to conduct a study and use ANN in locating numbers in an image-based cadastral maps.

---

Dave E. Marcial · Ed Darcy Dy · Silvin Federic Maceren · Erivn Rygl Sarno  
Silliman University, Hibbard Avenue, Dumaguete City, Negros Oriental, Philippines  
e-mail: demarcial@su.edu.ph, {eddarcydy, ryujen\_raven9}@yahoo.com,  
rygl119@yahoo.com.ph

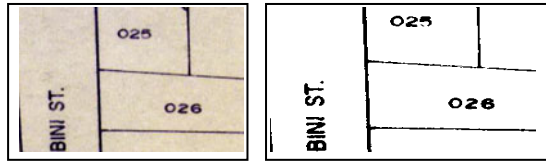
## 2 Methods and Implementation

The succeeding sections present a discussion on the processes conducted during the implementation of ANN in detecting lot numbers in a cadastral map.

### 2.1 Image Processing

There are three (3) major activities done during the image processing, these are: binarization, connected component labeling, and nearest neighbor image scaling.

In the binarization process, the scanned and colored image of the cadastral map was converted into black and white using thresholding algorithm for better and faster processing, as illustrated in Fig. 1. If the pixel (RGB) value is more than 129, the pixel will be turned black, on the other hand if the pixel (RGB) value is less than 130, the pixel will be turned white. The threshold value used in the study was based from a series of trial and error method.



**Fig. 1** It illustrates the original image and the image after the binarization process.

The connected component labeling process worked by scanning an image, pixel-by-pixel from top to bottom and left to right in order to identify connected pixel regions, i.e. regions of adjacent pixels which shared the same set of intensity values [2].

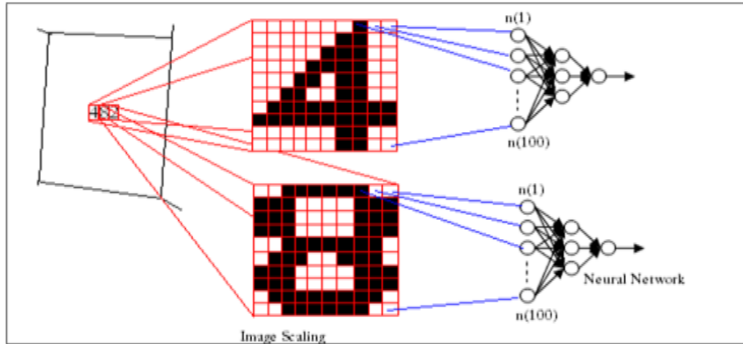
A connected component labeling algorithm is used to detect independent images in a cadastral map such as numbers, letters and other known characters, as shown in Fig. 2. These detected images were used for further processing until only lot numbers were detected.



**Fig. 2** It shows the detected images after processing the connected component labeling of lot numbers in the cadastral map.

In the process of the nearest neighbor image scaling, the algorithm needs the horizontal and vertical ratios between the original image and the scaled image.  $w_1$  and  $h_1$  were the width and height of an image, whereas  $w_2$  and  $h_2$  were the width and height of the new image [3], as shown in Fig. 3. A nearest neighbor image

scaling algorithm is used in order to resize the detected images to normalize the image for further processing. In the neural network processing, a fixed amount of input neurons (100 neurons) is specified, and so with the dimension of the images to be processed which is relatively fixed at 10 x 10 pixels.



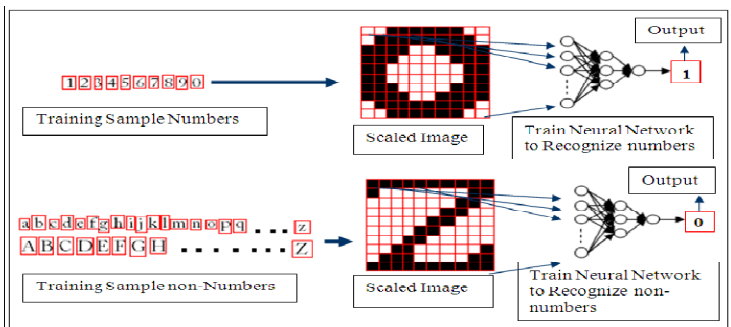
**Fig. 3.** It illustrates the image scaling process.

### 2.2 Backpropagation Training Process

Backpropagation learning is used in training the neural network, as illustrated in Fig.4. It is a common method of teaching artificial neural networks, and is a supervised learning method that is most useful for feed-forward networks, networks that have no feedback, or simply, that have no connections that loop [4].

The learning rate used 20% learning rate in the artificial neural network training since most common range of learning rates used are 15-30%. After training, the weights of the trained neural network were needed which were later used in the running phase.

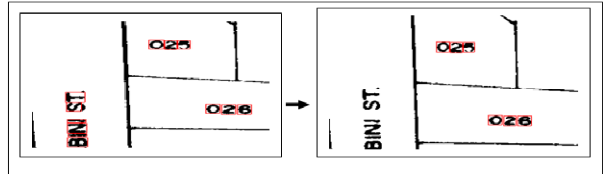
If the neural network needs more training in order to have a desirable output, the error value is necessary. The error value is the difference between the desired output and the actual output of a feed-forward neural network. When the error value is too high, backpropagation algorithm is called for the neural network to learn and should have a lesser error value in the next iteration called epoch.



**Fig. 4** It shows the training phase where two training samples were provided for the neural network.

### 2.3 Neural Network Running Phase

In the running phase, it is assumed to have the neural network's weights which can differentiate number and nonnumber images. While running the trained neural network, it gave an output of 1 if the detected image is a number and an output of 0 if the detected image is not a number. The detected images were then labeled as detected numbers if the neural network output is 1. See Fig. 5 for the illustration.



**Fig. 5** It shows the result after the running phase.

## 3 Findings

Two (2) different applications were developed, one is for training the neural network and the second one is the detection of lot numbers. The application is developed using Microsoft Visual C# 2008 Express Edition.

There were five (5) scenarios of testing conducted during the evaluation phase. This was done to evaluate specific components of accuracy in the implementation of the algorithms used by way of checking the percentage of learning and detection. Fifty (50) different maps with an average of 31 lot numbers at size 832 x 768 pixels were tested and it yielded an average of 90% detection after the 50th training. A 5808 x 4256 pixels map with 237 lot numbers was tested fifty times and gave an average of 84.78% detection rate.

Based from the testing scenarios conducted, the following are the findings:

- The number of detections is affected by several variables, namely: the training sets, the noise, and the algorithms used. It was found out that too few training samples results a few detection. On the other hand, when a sufficient set of training samples is met, adding more training samples will have little to no effect. It can also be said that having unfiltered training sets, causes less detections.
- There was a noticeable decrease in the error value and an increase of numbers that is detected in the succeeding iteration. This is the result of the ANN learning from the samples fed to it.
- The greater the number of noise in the scanned cadastral map, the lesser the detection of lot numbers.
- The efficiency and quality of the algorithm used affects the number of detections since it is the foundation and framework of the whole system.

- Most unwanted detections are the result of noises. The implementation often considers the letter ‘l’, ‘I’, and ‘i’ and the letters ‘O’ and ‘o’ as numbers since it looks like the numbers ‘1’ and ‘0’ respectively. Sometimes it also detects small dots and classifies it as a number since it may be recognized by the system as a number 1. The system also failed to detect numbers too close to each other and numbers too close to the lot border line, since the connected component labeling algorithm detects these faults as a single image.

## 4 Conclusion

There are various processing techniques which might give the same output of detecting characters in cadastral maps or other similar subjects as discovered. However, it was found out that the use of thresholding, connected component labeling and image scaling is sufficient in detecting characters in cadastral maps. Backpropagation learning ANN for the implementation of character detection in a cadastral map is a sufficient solution.

At 130 threshold value, 20% learning rate in the artificial neural network training, and a maximum iteration of 2000, and based from the thorough analysis and the high percentage of detection, it is concluded that the study was efficiently implemented ANN in detecting lot numbers in a cadastral map.

## 5 Recommendation

Upon completion of the study, the following are the recommendations to further improve the results of this research:

1. The threshold value used in the binarization process needs to be reviewed so that it can handle different contrasts of different cadastral maps.
2. There is a need to implement an algorithm to carefully handle cases wherein the numbers are close to each other or are too close to the map’s lot borders.
3. To improve accuracy, additional algorithms is needed to properly handle detections of the letters ‘.’, ‘l’, ‘I’, ‘i’, ‘o’, and ‘O’.

In addition, here is a list of possible topics that would help improve the results of this study, and as follows:

1. Recognition in different file format, e.g. .pdf, .doc, and so on. The algorithms used in this research can be used in the detection of characters inside these particular documents.
2. Integration with the global positioning system (GPS) and availability in portable electronic devices, e.g. cellular phones, palmtops, etc. The parties of the transaction can easily locate the object of the sale, as well as, where they are currently located and the distance of the land for sale from where they are currently in. This can greatly aid the survey of lands, as well as, facilitate real property transactions.

3. A study on the removal of unwanted detections and the removal of noises using other algorithms.

Lastly, it is further recommended that in order to fully utilize the results of this study, a full blown character recognition research be made with the output of a computer application that immediately locates lot numbers in cadastral maps.

## References

1. Fyfe, C.: Artificial neural networks and information theory, 1.2 edn. The University of Praisley (2000)
2. Bălan, I.: Using Mathematical Morphology to Detect the Imperfections of the Printed Circuit Boards. *Journal of Applied Computer Science* 1(2) (2008)
3. Tech-Algorithm.com. Nearest Neighbor Image Scaling (2007),  
<http://tech-algorithm.com/articles/nearest-neighbor-image-scaling/>
4. Gołda, A.: Principles of training multi-layer neural network using backpropagation (2005),  
[http://galaxy.agh.edu.pl/~vlsi/AI/backp\\_t\\_en/backprop.html](http://galaxy.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html)  
(Last modified: September 6, 2004)

# The Novel Virtual Reality Fixture Design and Assembly System (VFDAS)

Qian Cao and Qiang Li

**Abstract.** The primary objectives and aims of this paper were to develop an interactive VR system entitled Virtual Reality Fixture Design & Assembly System (VFDAS), which would allow fixture designers and assembly engineers to complete the design process for modular fixtures within the virtual environment (VE). This required the identification of what types of VR scenario and interactivity in VFDAS are needed to achieve the functionality in terms of computer aided fixture design (CAFD) and solve the actual problems arising from the fixture design process, such as: fixture location determination and assembly interference check. The VFDAS system would need to comprise the functionality of fixture element selection, fixture layout design, assembly planning, essential analysis and so on. More specifically, it would be useful for VFDAS to be used to select modular fixtures from VR libraries, spatially manipulate each element to assemble on the arbitrary workpiece users supply in preference, detect the assembly collision in advance and implement design optimization and verification etc. Furthermore, the feasibility of using interactive VR technology to facilitate the fixture design and assembly process has to be evaluated. The advantages of VR that could be used to support the conventional fixture design have to be identified.

**Keywords:** Virtual reality, fixture design, exact collision detection, interactive physics simulation, VFDAS.

---

Qian Cao  
Experimental Media Studio  
School of City Plan,  
Central Academy of fine arts, Beijing, P.R. China  
e-mail: caoqian@cafa.edu.cn

Qiang Li  
School of Mechanical, Materials and Manufacturing Engineering,  
University of Nottingham, University Park,  
Nottingham, NG7 2RD, United Kingdom  
e-mail: liaql@nottingham.ac.uk



## 1 Introduction

Virtual Reality (VR) was emerging as a potential and useful application around thirty years in relation to numerous fields of interest. Virtual reality technology has been adopted in diverse research fields including 3d graphics, video games development, mechanical engineering, construction engineering, manufacturing engineering, ergonomic analysis, 3d scientific visualization, medical surgical research, education and cognitive mapping study etc [1].

Due to today's heavy, growing competition environment, manufacturing companies have to develop and employ new emerging technologies to increase productivity, reduce production costs, improve product quality, and shorten lead time [2]. The domain of Virtual Reality (VR) has gained great attention during the past few years and is currently explored for pragmatic uses in various industrial areas e.g. CAD, CAM, CAE, CIM, CAPP and computer simulation etc.

On the other hand, a fixture is a device that locates, holds and supports workpieces in position and orientation during machining processes. Fixtures play a significant role in assuring production quality, shorten production cycle time and decreasing the production cost. Owing to the trend towards reducing lead time and human effort devoted to fixture planning, the computerization of fixture design is urgently required. Consequently, computer aided fixture design (CAFD) has been investigated and become an important role of computer aided design/manufacture (CAD/CAM) integration [3]. However, there is nearly none ongoing research study specially focused on utilizing the innovative VR technology as a promising solution in purpose to enhance CAFD systems' capability and functionality.

## 2 Literature Study of the State of the Art of CAFDs

A great amount of research has been conducted in the computer-aided fixture design (CAFD) area. In order to demonstrate the state of the art of CAFD research, three latest typical CAFD systems are selected and reported as follows.

Shokri et al [4] recently reported a new software in 2008 that can be used to plan fixture configuration and assembly procedure for modular CMM measuring fixtures. The system is implemented applying VC++ in a Solidworks platform. An assembly algorithm is developed to design a suitable fixture configuration and assembly procedure by using 'minimum fixture elements' and 'simple fixture configuration' criteria. Using Set Theory, a fixture structure is considered as an assembly unit divided into several sub-assemblies and elements. Each sub-assembly is defined as a functional unit that comprises one contacting element, one connecting element and some other assembly elements.

Kang et al [5] addressed another fixture design system for networked manufacturing in 2007, which includes characteristics of rapid configuration designing, three-dimensional modeling, a standardized elements database. This CAFD system may also transfer information with various other systems. For the requirement of rapid configuration design, based on the analysis of the contents and characteristics of fixture design, a hybrid CBR/KBR (case-based reasoning /knowledge-based reasoning) fixture design method was applied to develop this fixture design system.

In comparison with the two CAFD systems, the interactive VR system may have the advantages of making the fixture design in a natural and instructive manner, providing better match to the working conditions, reducing lead-time, and improving fixture productivity and economy. VR system can also support the visualization or planning of a whole assembly process but the conventional CAFD systems can not achieve this.

Peng et al [6] reported a novel modular fixture design and assembly system based on VR in 2006, which comprises the Graphic Interface (GUI) module, Virtual Environment (VE) module and Database module. The GUI is basically a graphic interface that is used to integrate the virtual environment and modular fixture design actions. The VE provides the users with a 3D display for navigating and manipulating the models of modular fixture system and its components in the virtual environment. The database deposits all the models of environment and modular fixture elements, as well as the domain knowledge and useful cases.

Peng et al [7] addressed a precise manipulation approach to support the interactive design and assembly of modular fixture configuration in a virtual environment in 2008 in order to develop a VR-based modular fixture assembly design system.

Although Peng et al [6] [7] claim that their modular fixture design systems are based on VR technology, the essential physical features of VR actually can not be found in these two systems such as: mass, gravity, friction, applied force, elasticity and toppling. To some extent, these two CAFD systems should not be regarded as the sound VR simulations for fixture design. Moreover, all these four CAFD systems presented above lack the capability of the real-time 3D collision detection which is required to conduct the design interference check between fixturing elements, workpiece and machine tools. To fill the research gap, the intention of this paper is to develop the VFDAS system that can simulate these physical features of VR and the real-time 3D collision detection in accuracy so that the fixture design and assembly process is realistically carried out as if in the real physics world. The VFDAS system can also demonstrate the fixture physical behaviour in use and restore the real fixture design conditions in realism.

### **3 Aims and Objectives of VFDAS**

Computer-aided design (CAD) and computer-aided manufacturing (CAM) systems are usually referred to standard engineering utility utilized to reduce the time and cost of product design and manufacturing. Despite of well-known fact that there is an executive gap between CAD and CAM, Computer-aided process planning (CAPP) has been suggested as the link to achieve thorough CAD/CAM integration [8].

The function of a fixture is to hold a workpart firmly in position during a manufacturing process. Fixture design is basically regarded as a concurrent activity of process planning. Due to the trend toward high-precision and entire automation production, computerization of fixture design is urgently demanded to decrease the lead time and cost of product development. Therefore, computer-aided fixture design (CAFD) has been developed and employed as a part of computer aided design and manufacture (CAD/CAM) integration [9].

On the other hand, Virtual Reality (VR) has great potentials and promises regarding innovation applications and practical uses in real industry. Although a vast number of potential research for VR applications and systems distributed in various industry uses, have been investigated and implemented prosperously for around last thirty years e.g. product design, rapid virtual prototyping, virtual assembly (VA), manufacturing simulation and plant layout etc., however there is nearly none ongoing research exploration specially focus on “Virtual Reality for Fixture Design and Assembly”. As a result, the research gap which is urgently required for industrial development can be apparently identified and determined and further investigation appears very emergent and worthwhile for undertaking.

Consequently, the key objectives and aims we endeavor to achieve in this research are to establish a novel and innovative interactive VR system entitled Virtual Reality Fixture Design & Assembly System (VFDAS), which is employed to support modular fixture and assembly process by fixture designers who could be classified two categories of users. One category is the novice fixture designers such as: engineering students and academic technicians. Another category is the expert users from industry. Furthermore, the potential preliminary functionalities and usages of VFDAS we strive to invent can be summarized in several key points below:

- Implementation of interactive VR fixture design and assembly
- Generate Interactive VR animation simulation for assembly
- Analyze the machining strategies before conduct fixture design to avoid design interference
- Conduct kinematics analysis for fixture design evaluation
- Establish Interactive VR fixture parts library to facilitate fast virtual fixture design
- Visualization and optimisation of fixture design
- Problems detection and assembly evaluation prior to real manufacture
- A common role of communication between engineering designers and people from different domains
- Plant layout and production flow line simulation
- Conduct a analysis and evaluation regarding safety and ergonomics issues
- Get involved with academic values to support undergraduate engineering student
- Reverse 3D transfer data of fixture design models back to the CAD system after the overall design process.

#### **4 The Advancements of the VFDAS System**

The VFDAS system provides an interactive design platform in which the fixture design and assembly process for modular fixtures can be fulfilled by the fixture

designers in a VE. The comprehensive VFDAS system is employed to carry out the functionality of fixture element selection, fixture layout design, assembly planning and essential analysis etc. In VFDAS, the combination of many micro-focused research activities into a complete fixture design system was explored and

accomplished. More specifically, the VFDAS system allows the users to select the fixture elements from VR libraries, spatially manipulate each element to assemble on the arbitrary workpiece users provide, detect assembly interference and perform design verification.

Even though ordinary CAD and CAFD systems allow configuration evaluation by letting engineers visualize the final fixture assembly, these systems can not support the visualization or planning of the fixture assembly process [10]. In contrast, the VFDAS system provides the function for fixture designers to visualize, observe, support and evaluate the fixture planning and assembly process. The VFDAS system is also considered as a particular type of CAFD or CAMFD system in which the advantages of VR are used to improve modular fixture designs within a VR environment. In addition, two types of fixture designers are regarded as the users of the VFDAS system. One type is the novice fixture designers such as: engineering students and academic technicians. Another type is the professional fixture designers from the industry.

According to a great deal of literature, the implementation of accurate collision detection is a critical aspect of VR simulation system during a virtual manufacturing process [11]. For the development of the VFDAS system, a VR behaviour programming methodology is used to develop the physics behaviour simulation and the real-time 3D collision detection in accuracy. This programming methodology improves the research bottleneck of collision detection related to concave 3D objects in VEs. The framework of the VFDAS system was developed using this methodology.

Moreover, the VFDAS not only provides real-time 3D collision detection in accuracy but also achieves the goal of good physics behaviour simulation. The VFDAS system also demonstrates an interactive design platform that is based on a physically realistic VR environment. The physical properties associated with each fixture element that were developed in the VFDAS system refer to mass, gravity, friction, elasticity, collision detection, applied force, toppling and so on [12]. These physical properties are normally taken into account during the actual fixture design and evaluation process. These physical properties obey Newton's laws of physics.

## References

- [1] Dai, F. (ed.): *Virtual Reality for Industrial Applications*. Springer, Berlin (1998)
- [2] Ji, P., Choi, A.C.K., Tu, L.: VDAS: a virtual design and assembly system in a virtual reality environment. *Assembly Automation* 22(4), 337–342 (2002)
- [3] Yu, K.M., Lam, T.W., Lee, A.H.C.: Immobilization check for fixture design. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 217(4), 499–512 (2003)
- [4] Shokri, M., Arezoo, B.: Computer-aided CMM modular fixture configuration design. *International Journal of Manufacturing Technology and Management* 14(1-2), 174–188 (2008)

- [5] Kang, Y.G., Wang, Z., Li, R., Jiang, C.: A fixture design system for networked manufacturing. *International Journal of Computer Integrated Manufacturing* 20(2-3), 143–159 (2007)
- [6] Peng, G.L., Liu, W.J.: A novel modular fixture design and assembly system based on VR. In: *Proceedings of 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, pp. 2650–2655 (2006)
- [7] Peng, G.L., He, X., Yu, H.Q., Hou, X., Khalil, A.: Precise manipulation approach to facilitate interactive modular fixture assembly design in a virtual environment. *International Journal of Assembly Automation* 28(3), 216–224 (2008)
- [8] Surendra Babu, B., Madar Valli, P., et al.: Automatic modular fixture generation in computer-aided process planning systems. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 219(10), 1147–1152 (2005)
- [9] Yu, K.M., Lam, T.W., Lee, A.H.C.: Immobilization check for fixture design. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 217(4), 499–512 (2003)
- [10] Jayaram, S., Jayaram, U., Wang, Y., Tirumali, H., Lyons, K., Hart, P.: VADE: A Virtual Assembly Design Environment. *IEEE Computer Graphics and Applications* 19(6), 44–50 (1999)
- [11] Tesic, R., Banerjee, P.: Exact collision detection for a virtual manufacturing simulator. *IIE Transactions (Institute of Industrial Engineers)* 33(1), 43–54 (2001)
- [12] Li, Q., Chen, X., Cobb, S., Eastgate, R.: Physical behaviour simulation and exact collision detection within VFDAS. In: *Proceedings of the 12th Annual Conference of the Chinese Automation and Computing Society in the UK (CACSUUK 2006)*, Loughborough, UK, September 16, pp. 33–37 (2006)

# Multi-phenomenon Oriented Embedded Software Fault Diagnosis

Shunkun Yang, Ge Lin, and Lu Minyan

**Abstract.** In order to locate the software fault more quickly and accurately, the positioning statement or software module involved in different test cases were sorted by the possibility of attributing to the software failure, which was calculated by using the method of correlation coefficient. Then, method of range analysis and variance analysis were introduced to address the issue of the volatility in some condition, which can also be extended to solve the problem of different fault phenomenon with different importance factor. Finally an example of the method was given to verify its effectiveness.

**Keywords:** Range analysis, Variance analysis, Fault diagnosis, Fault localization, Software diagnosis.

## 1 Introduction

As software systems become more and more complex, software has been expanded and complexity keeps on increasing. The potential software defects are also increasing. Despite people have already done a lot of software testing, software failure is still inevitable. However, if these errors are triggered, they often have catastrophic consequences. Therefore, how to locate the fault quickly and efficiently has attracted many researchers' attention.

Currently the major software fault location methods including: Fault location based on static analysis[3]; Fault location based on dynamic analysis [1][2]; The cumulative analysis based on program behavior[4]; Fault Location based on the experiment (Delta debugging[5]). For the method based on static analysis, the scale is often subject to certain restrictions; Method based on dynamic analysis can only reveal the procedures behavior in the specific corresponding input, when the program logic complexity increases, the effectiveness of this approach will be

---

Shunkun Yang · Ge Lin · Lu Minyan

School of Reliability and System Engineering, Beihang University, Beijing, China  
e-mail: {ysk, lmy}@buaa.edu.cn, Gelingelin@126.com

limited; Program behavior analysis based on the cumulative fault location studies, mostly based on the differences comparison technology, for the success sequence be selected is the closest sequence of the failure sequence, sequence requirement is high, it may lead to some of the data sparseness; Rui Abreu proposed the fault location method[7] using the software program spectra, the main idea is that different statements will be run under different input, and there are two cases of operation results, one normal, and second, failure occurs. Then the failure probability of all the statements which are involved under these inputs will be assessed. The assessment result includes three correlation coefficient: Jaccard correlation coefficient, the correlation coefficient used in Tarantula fault location tool, and Ochiai correlation coefficient; But the author did not take the multiple-phenomenon into account.

## 2 Fault Location under Multiple - Phenomenon

When software fails under a specific input, its symptoms may be more than one. we will discuss how the software fault location performs when multiple fault phenomenon occur. Consider different inputs:  $t_1, t_2...t_m$ , all possible failure-contain statements involved:  $c_1,c_2...c_n$ , and different fault phenomenon appears under these inputs:  $f_1...f_k$ . We can use the information that whether the statements involved or not and fault phenomenon appears or not to locate the software’s fault.

**Table 1** Statements for Multi Fault Phenomenon

Input	Statements				Fault Phenomenon			
	C <sub>1</sub>	C <sub>2</sub>	...	C <sub>n</sub>	f <sub>1</sub>	...	f <sub>k</sub>	
t <sub>1</sub>	I <sub>11</sub>	I <sub>12</sub>	...	I <sub>1n</sub>	F <sub>11</sub>	F <sub>12</sub>	...	F <sub>1k</sub>
t <sub>2</sub>	I <sub>21</sub>	I <sub>22</sub>	...	I <sub>2n</sub>	F <sub>21</sub>	F <sub>22</sub>	...	F <sub>2k</sub>
...	...	...	I <sub>pn</sub>	...	...	...	F <sub>pk</sub>	...
t <sub>m</sub>	I <sub>m1</sub>	I <sub>m2</sub>	...	I <sub>mn</sub>	F <sub>m1</sub>	F <sub>m2</sub>	...	F <sub>mk</sub>

Where the value of  $I_{pq}$  is 0 or 1, respectively in the test case p, the component (or statement) q is not involved or involved; the value of  $F_{pq}$  is also 0 or 1, respectively in the test case (or input) p, the fault phenomenon of f does not appear or appear. We use  $a_{pq}(j)_i$  to stand for the situation that statements involved and fault phenomenon appears.

- $a_{00}(j)_i$ : The statement is not involved and the phenomenon doesn’t appear
- $a_{01}(j)_i$ : The statement is not involved but the fault phenomenon appears
- $a_{10}(j)_i$ : The statement is involved but the fault phenomenon doesn’t appear
- $a_{11}(j)_i$ : The statement is involved and the fault phenomenon appears.

When multiple fault phenomena appear, we can use the following correlation coefficient to measure the failure probability of every statement or module:

$$S_j(j) = \frac{\sum_{i=1}^k a_{11}(j)_i}{\sum_{i=1}^k a_{11}(j)_i + \sum_{i=1}^k a_{01}(j)_i + \sum_{i=1}^k a_{10}(j)_i} \tag{1}$$

$$S_r(j) = \frac{\sum_{i=1}^k \frac{a_{11}(j)_i}{a_{11}(j)_i + a_{01}(j)_i}}{\sum_{i=1}^k \frac{a_{11}(j)_i}{a_{11}(j)_i + a_{01}(j)_i} + \sum_{i=1}^k \frac{a_{00}(j)_i}{a_{10}(j)_i + a_{00}(j)_i}} \tag{2}$$

$$S_o(j) = \frac{\sum_{i=1}^k a_{11}(j)_i}{\sqrt{\sum_{i=1}^k (a_{11}(j)_i + a_{01}(j)_i) \cdot \sum_{i=1}^k (a_{10}(j)_i + a_{00}(j)_i)}} \tag{3}$$

Where  $\sum_{i=1}^k a_{11}(j)_i$  represents the sum of the situation that statement j is involved and the fault phenomenon appears under different inputs;  $\sum_{i=1}^k a_{01}(j)_i$  represents the sum of the situation that statement j is not involved but the fault phenomenon appears under different inputs;  $\sum_{i=1}^k a_{10}(j)_i$  represents the sum of the situation that statement j is involved but the fault phenomenon doesn't appear under different inputs.

When we use the three correlation coefficients to assess the failure possibility of the statement, we may find that the size trends of the three correlation coefficients are not same, and then we just need to take the more consistent results. When the results assessed by the three correlation coefficients aren't consistent, we treat the situation that statement involved or not as condition level and the situation that fault phenomenon appears as results level. When the conditions are at different levels, the results level changes under different condition level. If under different condition level, the volatility of the results level is big, that indicates when the statement is involved; its failure possibility is big. According to this idea, we introduce the range analysis and variance analysis methods to assess the failure possibility of each statement.

### 3 Range Analysis Method

The idea of range analysis method is that: when the statement is involved or not involved, if the differences between the situations that fault phenomenon appears and does not appear is big, that means the statement has great impact on the fault phenomenon, and it is more likely to be the cause of failure. We use R(j) to represent the statement or the module j's range, the size of range represents size of the failure possibility of the module or statement:

$$R(j)_i = \frac{a_{11}(j)_i - a_{01}(j)_i}{|\sum_{p=1}^m I_{pj}| \quad m - |\sum_{p=1}^m I_{pj}|} \tag{4}$$

Where  $a_{11}(j)_i$  represents the sum of the situation that statement j is involved and the fault phenomenon appears;  $a_{01}(j)_i$  represents the sum of the situation that statement j is not involved but the fault phenomenon appears;  $|\sum_{p=1}^m I_{pj}|$  represents





$$\begin{aligned}
 R(j) &= \frac{a_{11}(j)}{|\sum_{p=1}^m I_{pj}|} - \frac{a_{01}(j)}{m - |\sum_{p=1}^m I_{pj}|} \\
 &= \frac{a_{11}(j)}{a_{11}(j) + a_{10}(j)} - \frac{a_{01}(j)}{a_{01}(j) + a_{00}(j)} \\
 &= \frac{1}{1 + \frac{a_{10}(j)}{a_{11}(j)}} - \frac{1}{1 + \frac{a_{00}(j)}{a_{01}(j)}}
 \end{aligned} \tag{9}$$

We can find that when the value of  $a_{10}(j)/a_{11}(j)$  and  $a_{10}(j)/a_{11}(j)$  are small, the value of  $a_{00}(j)/a_{01}(j)$  and  $a_{00}(j)/a_{10}(j)$  are big, the assessment of these formulas are of the same size trend. When we use  $S_0(j)$   $S_T(j)$   $S_j(j)$ , the main idea is the proportion of the situation that the statement is involved and the fault phenomenon appears in all situations, the statement is more likely to be the cause of the software failure. However, the idea of range analysis and variance analysis method is that if the volatility of the situation that the statement is involved and not involved is bigger, the statement is more likely to be the cause of failure.

In the previous section, we only consider the situation that there is only one fault phenomenon or the importance of each fault phenomenon is same. When the importance of each fault phenomenon is different, we assume that the score of each fault phenomenon is  $W_i$ :

$$S_j(j)' = \frac{\sum_{i=1}^k w_i \cdot a_{11}(j)_i}{\sum_{i=1}^k w_i \cdot a_{11}(j)_i + \sum_{i=1}^k w_i \cdot a_{01}(j)_i + \sum_{i=1}^k w_i \cdot a_{10}(j)_i} \tag{10}$$

$$S_T(j)' = \frac{\sum_{i=1}^k w_i \cdot \frac{a_{11}(j)_i}{a_{11}(j)_i + a_{01}(j)_i}}{\sum_{i=1}^k w_i \cdot \frac{a_{11}(j)_i}{a_{11}(j)_i + a_{01}(j)_i} + \sum_{i=1}^k w_i \cdot \frac{a_{10}(j)_i}{a_{10}(j)_i + a_{00}(j)_i}} \tag{11}$$

$$S_0(j)' = \frac{\sum_{i=1}^k w_i \cdot a_{11}(j)_i}{\sqrt{\sum_{i=1}^k w_i \cdot (a_{11}(j)_i + a_{01}(j)_i) \cdot \sum_{i=1}^k w_i \cdot (a_{11}(j)_i + a_{10}(j)_i)}} \tag{12}$$

The equation deformation of range analysis is:

$$R(j)' = \frac{\sum_{i=1}^k w_i \cdot a_{1q}}{|\sum_{p=1}^m I_{pj}|} - \frac{\sum_{i=1}^k w_i \cdot a_{0q}}{m - |\sum_{p=1}^m I_{pj}|} \tag{13}$$

The equation deformation of variance analysis is:

$$S_j^2 = \frac{(\sum_{i=1}^k w_i \cdot a_{01})^2}{|\sum_{p=1}^m I_{p1}|} + \frac{(\sum_{i=1}^k w_i \cdot a_{10})^2}{m - |\sum_{p=1}^m I_{p1}|} - \frac{(\sum_{i=1}^k w_i \cdot a_{11} + \sum_{i=1}^k w_i \cdot a_{01})^2}{m} \tag{14}$$

## 5 Examples

We assume that there are 5 statements or modules:  $c_1$ ,  $c_2$ ,  $c_3$ ,  $c_4$ ,  $c_5$ , under the test cases  $t_1$ ,  $t_2$ ,  $t_3$ ,  $t_4$ ,  $t_5$ ,  $t_6$ , the situation that each is involved or not and the fault phenomenon  $f_1$ ,  $f_2$  appear or not, shown in Table 2.

**Table 2** Statements for Multi Fault Phenomenon

	$c_1$	$c_2$	$c_3$	$c_4$	$c_5$	$f_1$	$f_2$
$t_1$	1	0	0	0	0	0	0
$t_2$	1	1	0	0	0	0	0
$t_3$	1	1	1	1	0	0	1
$t_4$	1	1	1	0	0	0	0
$t_5$	1	1	1	1	1	1	0
$t_6$	1	1	1	0	1	0	1

The failure probability of each statement assessed by formula 1, formula 2, formula 3 and formula 4 is shown in Table 3. We can see that when we use  $S_j(j)$ ,  $S_T(j)$ ,  $S_0(j)$  to assess the failure probability,  $c_4$ ,  $c_5$  are more likely to cause the failure. And the size trends of  $S_j(j)$  and  $S_T(j)$  are same, but the size trend of  $S_0(j)$  is different from the others'. The results assessed by range analysis and variance analysis are also shown in Table 3. We can find that: when the phenomenon  $f_1$  appears, statement  $c_4$  and  $c_5$  are more likely to cause of the software failure; When the phenomenon  $f_2$  appears, statement  $c_3$  is more likely to be the cause of the failure. In Table 4, we consider the situation that different phenomenon occurs with different importance factor. We assume that the importance score of phenomenon  $f_1$  is 1, and  $f_2$ 's is 2, we can find that the failure probability of  $c_3$  and  $c_5$  is big.

**Table 3** Assessment Result of the Failure Probability of Each Statement

	C1		C2		C3		C4		C5	
$a_{11}(j)_i$	1	2	1	2	1	2	1	1	1	1
$a_{10}(j)_i$	5	4	4	3	3	2	1	1	1	1
$a_{01}(j)_i$	0	0	0	0	0	0	0	1	0	1
$1 \sum_{r=1}^n I_{rj} \cdot 1$	6		5		4		2		2	
$m-1 \sum_{r=1}^m I_{rj} \cdot 1$	0		1		2		4		4	
$S_j(j)$	0.25		0.3		0.375		0.4		0.4	
$S_T(j)$	0.5		0.66		0.73		0.77		0.77	
$S_0(j)$	0.5		0.59		0.61		0.52		0.52	
$R(j)_i$	0.17	0.33	0.2	0.4	0.25	0.5	0.5	0.25	0.5	0.25
$S_j^2$	0	0	0.033	0.133	0.083	0.33	0.33	0.083	0.33	0.083

**Table 4** Importance Scores  $w_1=1, w_2=4$

	C1	C2	C3	C4	C5
$S_j(j) \cdot$	0.3	0.36	0.45	0.38	0.38
$S_T(j) \cdot$	0.5	0.70	0.78	0.71	0.71
$S_0(j) \cdot$	0.548	0.6	0.67	0.53	0.53
$R(j) \cdot$	1.5	1.8	2.25	1.5	2.25
$S_j^2 \cdot$	0	2.7	6.75	28.3	28.3

## 6 Conclusions

We propose several methods to assess the failure probability of the statements. We extend Rui Abreu' spectrum-based fault location method [7] and apply it to the situation that multiple fault phenomenon occurrence. The range analysis and the variance analysis are used to assess the failure probability of each statement, explaining which measure we should take when the assessment results fluctuate.

## References

1. Weeratunge, D., Zhang, X., Sumner, W.N., Jagannathan, S.: Analyzing Concurrency Bugs Using Dual Slicing. CSD TR #10-004 (2010)
2. Jia, X., Wu, J., Jin, M., Liu, Y.: Novel scheme to locate software fault by aggregate analysis of program behaviors. *Journal of Beijing University of Aeronautics and Astronautics* 32(05), 607–611 (2006)
3. Yi, Z., Mu, X., Zhao, P., Zhang, X.: Software Fault Location Based on Checking Codes. *Computer Engineering* 33(12), 82–83 (2007) (in Chinese)
4. Agrawal, H., Demillo, R.A., Spafford, E.H.: Debugging with dynamic slicing and backtracking. *Softw. Pract. Exper.* 23(6), 589–616 (1993)
5. Zeller, A.: Isolating cause-effect chains from computer programs. *ACM SIGSOFT Software Engineering Notes* 27(6), 1–10 (2002)
6. Lei, Z.: An improved Similar Path Generation Algorithm and its Application for Fault Location. HuaZhong Normal University (2008)
7. Abreu, R., Mayer, W., Stumptner, M., et al.: Refining spectrum-based fault localization rankings. In: *Proceedings of the 2009 ACM Symposium on Applied Computing*, pp. 409–414. ACM, Honolulu (2009)

# Asymptotic Behavior of Random Walks with Resting State in Ergodic Environments\*

Xiang-dong Liu and Li-ye Zhu

**Abstract.** In this paper we consider asymptotic theorems for random walks with resting state in ergodic environments. Suppose that  $(\alpha_i, \beta_i), \alpha_i, \beta_i \in (0, \frac{1}{2}), i \in \mathbb{Z}$  are stationary and ergodic random variables (therefor non-independent),  $\gamma_i = 1 - \alpha_i - \beta_i \geq 0$ .  $\{(\alpha_i, \gamma_i, \beta_i), i \in \mathbb{Z}\}$  serves as an environment. This environment defines a random walk  $\{X_k\}$  (Called random walk with resting state on the line in random environment) which, when at  $x$ , moves one step to the right with probability  $\alpha_x$ , and one step to the left with probability  $\beta_x$ , or rests with probability  $\gamma_x$ . Recurrence-transience criteria and law of large numbers of the model in stationary and ergodic environments are obtained, which are extensions of the work of independent random environments.

## 1 Introduction

Let  $(E, \mathcal{E}, \theta, \pi)$  be an ergodic dynamical system and  $(\alpha, \beta)$  measurable function with values in  $(0, 1/2) \times (0, 1/2)$ . We denote  $\alpha(\theta^t \cdot), \beta(\theta^t \cdot)$  by  $\alpha_i(\cdot), \beta_i(\cdot)$  and we consider the following model. For a fixed  $e$  in  $E$ , let  $\{X_n; n \in \mathbb{Z}^+\}$  be a Markov chain on  $\mathbb{Z}$ , starting at 0, whose transition probabilities are given by

$$\begin{aligned} P_e^\omega(X_0 = 0) &= 1, \\ P_e^\omega(X_{n+1} = j | X_0 = 0, X_1 = i_1, \dots, X_n = i; e) &= \begin{cases} \alpha_i, & j = i + 1, \\ \beta_i, & j = i - 1, \\ \gamma_i, & j = i, \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

---

Xiang-dong Liu · Li-ye Zhu

Department of Statistics, Jinan University, Guangzhou Guangdong P.R.C

e-mail: [tliuxd@jnu.edu.cn](mailto:tliuxd@jnu.edu.cn), [zhuliy361@163.com](mailto:zhuliy361@163.com)

\* Supported by the Fundamental Research Funds for the Central Universities(10JYB2019).

where  $\gamma_i = 1 - \alpha_i - \beta_i, i \in Z, \{(\alpha_i, \beta_i, \gamma_i), i \in Z\}$  is usually called environment, then  $X_n$  is referred as a *Random Walk with resting state on the line In Random Environment*(RWIRE). The process  $(X_n)$  is not Markovian and has numerous applications in physical models.

Simple random walk in a random environment is deeply discussed, Solomon[1] and ALili [2] obtain recurrence-transience criteria and get the law of large numbers. Other results concerning the large deviation principle have been obtained by Greven and Hollander( $P_e$ )[11], Comets, Gantert and Zeitouni( $P$ )[12]. In discussing random walk with resting state on the line in random environment, when  $(\alpha_i, \beta_i)$  is a sequence of i.i.d random vectors, Liu and Dai[4,5], proved that  $\lim_{n \rightarrow \infty} X_n = +\infty, \lim_{n \rightarrow -\infty} X_n = -\infty$  with probability 1 according as  $E(\log(\beta_0/\alpha_0)) < 0, E(\log(\beta_0/\alpha_0)) > 0$ . Otherwise  $\liminf_{n \rightarrow +\infty} X_n = -\infty$  and  $\limsup_{n \rightarrow +\infty} X_n = +\infty$  with probability 1. The processes in ergodic environments is more difficult to deal with than in i.i.d environments and the method of applying in this paper is different to [1,2]. In this paper we want to make it clear what happens when  $(\alpha_i, \beta_i)$  are non-independent, and to check if the slow diffusion phenomenon appears in this case. Section 2 contains recurrence-transience criteria which is natural extensions of [1,2] and contains the law of large numbers. Let us consider the space of trajectories,  $Z^N$ , of the class (while the environment  $e$  is running over the set  $E$ ) of Markov chains  $(X_n)$ , equipped with the  $\sigma$ -algebra  $\mathcal{F}$  generated by the cylinder sets. The RWIRE evolves on the space  $(\Omega, \mathcal{F}, P)$  where  $\Omega = E \times Z^N, \mathcal{F} = \mathcal{E} \otimes \mathcal{J}$  and  $P$  (sometimes denoted by  $P_\pi$ ) is the probability measure defined by:  $P(A \times B) = \int_A P_e(B) d\pi(e)$ . for  $A \in \mathcal{E}$  and  $B \in \mathcal{J}$ .

**THEOREM.** Let  $B \subset Z^N$  be measurable. Suppose  $P_e(\{X_n\} \in B) = 1$  for a.e. environment  $e$ . Then  $P(\{X_n\} \in B) = 1$ .

*Proof.*  $P(\{X_n\} \in B) = \int P_e(\{X_n\} \in B) d\pi(e) = 1$ .

This theorem provides the basis for finding the limit behavior of  $\{X_n\}$  in the next section. The proofs of limit theorems for RWIRE will use the fact that the asymptotic behavior of  $X_n$  can be derived of the hitting times  $T_n$  where  $T_n = \inf\{k : X_k = n\}, n \in Z$  ( $\inf \emptyset = \infty$ ). The following Basic result and lemma play an essential role to obtain limit theorems.

**Basic result 1.1.** When an environment  $e \in E$  is fixed, i.e. under probability  $P_e$ , the random variables  $\tau_n = T_n - Y_{n-1}, n \geq 1$ , are independent and have respectively the laws  $L^{\theta^{n-1}e}$  where  $L^e$  stands for the law of  $\tau_1$ .

Throughout this paper we will use the following notation  $\sigma_i = \sigma_i(e) = \frac{\beta_i(e)}{\alpha_i(e)}$ .  $S(e) = \prod_{n=0}^{+\infty} \sigma_1 \sigma_2 \cdots \sigma_n$ .  $F(e) = \prod_{n=0}^{+\infty} \sigma_{-n} \sigma_{-n+1} \cdots \sigma_{-1}$ . where we agree that an empty product equals 1.  $E = E_\pi$  (respectively  $E_e$ ) will denote the expectation under probability  $P = P_\pi$  (respectively  $P_e$ ).

**Lemma 1.** When an environment  $e \in E$  is fixed, i.e. under  $P_e$ . Then

- (i)  $\sum_{n=1}^{\infty} (\sigma_{-1} \cdots \sigma_{-n})^{-1} = \infty, \sum_{n=1}^{\infty} \sigma_1 \cdots \sigma_n < \infty$ , implies  $\lim_{n \rightarrow \infty} X_n = \infty$  a.e.
- (ii)  $\sum_{n=1}^{\infty} (\sigma_{-1} \cdots \sigma_{-n})^{-1} < \infty, \sum_{n=1}^{\infty} \sigma_1 \cdots \sigma_n = \infty$ , implies  $\lim_{n \rightarrow \infty} X_n = -\infty$  a.e.
- (iii)  $\sum_{n=1}^{\infty} (\sigma_{-1} \cdots \sigma_{-n})^{-1} = \infty, \sum_{n=1}^{\infty} \sigma_1 \cdots \sigma_n = \infty$ , implies  $\{X_n\}$  is recurrent. In fact  $-\infty = \liminf_{n \rightarrow \infty} X_n < \limsup_{n \rightarrow \infty} X_n = +\infty$  a.e.

*Proof.* See [10].

## 2 Main Results and Proofs

**Theorem 1.** suppose that  $E(\log \sigma_0)$  is well defined (with possible values  $+\infty, -\infty$ ). let  $\sigma = \beta_0/\alpha_0$ . (i) if  $E \log \sigma < 0$ , then  $\lim_{n \rightarrow \infty} X_n = +\infty$  a.e. (ii) if  $E \log \sigma = 0$ , then  $-\infty = \liminf_{n \rightarrow \infty} X_n$  and  $\limsup_{n \rightarrow \infty} X_n = \infty$  a.e. (iii) if  $E \log \sigma > 0$ , then  $\lim_{n \rightarrow \infty} X_n = -\infty$  a.e.

*Proof.* when an environment  $e$  is fixed, according to lemma 1, one has  $P_e$ -almost surely:

$$\lim_{n \rightarrow +\infty} X_n = \begin{cases} +\infty & \text{if } S(e) < +\infty \text{ and } F(e) = +\infty. \\ -\infty & \text{if } S(e) = +\infty \text{ and } F(e) < +\infty. \end{cases}$$

in which cases the random walk is transient; and  $-\infty = \liminf_{n \rightarrow \infty} X_n < \limsup_{n \rightarrow \infty} X_n = +\infty$  if  $S(e) = F(e) = +\infty$ , in which cases the random walk is recurrent; Now, if  $e$  is chosen at random with respect to  $\pi$ , the sets  $D = \{e \in E : S(e) < +\infty\}$  and  $D' = \{e \in E : F(e) < \infty\}$  have probability 0 or 1 since they are  $\theta$ -invariant; and by invariant of  $\pi$ , one has that  $\pi(D) = 1$  implies  $\pi(D') = 0$ . Thus, for symmetry reasons, we have only to prove that  $\pi(D) = 1$  if and only if  $E \log(\sigma) < 0$ . If  $E \log(\sigma) < 0$ , then almost surely  $\sigma_1 \sigma_2 \cdots \sigma_n$  converges exponentially to 0 by Birkhoff's ergodic theorem. In fact,  $\lim_{n \rightarrow +\infty} \sum_{k=1}^n \log(\sigma_k) = E \log(\sigma_0)$ . a.s. So that the series  $S(e)$  is convergent with probability 1. Conversely if  $\pi(D) = 1$ , then almost surely  $\sigma_1 \sigma_2 \cdots \sigma_n$  convergent to 0 and  $\lim_{n \rightarrow +\infty} \sum_{k=1}^n \log(\sigma_k) = -\infty$ . But, according to [7] this sum cannot 'grow' slower than linearly. Thus  $\lim_{n \rightarrow +\infty} \frac{1}{n} \sum_{k=1}^n \log \sigma_k = E(\log \sigma_0) < 0$ . Symmetry result is  $\pi(D') = 1$  if and only if  $E \log(\sigma) > 0$ .  $\pi(D) = \pi(D') = 0$  if and only if  $E \log(\sigma) = 0$ . In fact, if  $E \log(\sigma) = 0$ , then  $\pi(D) = \pi(D') = 0$ . Conversely if  $\pi(D) = 0$ , then  $E \log \sigma \geq 0$ . if  $\pi(D') = 0$ , then  $E \log \sigma \leq 0$ . Thus the theorem is proved.  $\square$

**Theorem 2.** (The law of large numbers) (i) if  $E(S) < +\infty$ , then  $\lim_{n \rightarrow +\infty} (X_n/n) = v$  and  $\lim_{n \rightarrow +\infty} (T_n/n) = v^{-1}$  a.s., where  $v^{-1} = E((1 + \frac{\gamma_0}{\alpha_0} + \sigma)S)$ .

(ii) if  $E(F) < +\infty$ , then  $\lim_{n \rightarrow +\infty} (X_n/n) = -v'$  and  $\lim_{n \rightarrow +\infty} (T_{-n}/n) = v'^{-1}$  a.s., where  $v^{-1} = E((1 + \frac{\gamma_0}{\beta_0} + \sigma^{-1})F)$ .

(iii) if  $E(S) = +\infty$  and  $E(F) = +\infty$ , then  $\lim_{n \rightarrow +\infty} (X_n/n) = 0$  and  $\lim_{n \rightarrow +\infty} (T_n/n) = \lim_{n \rightarrow +\infty} (T_{-n}/n) = +\infty$  a.s. .

To obtain the law of large numbers of  $X_n/n$ , we need the following lemma. We can show that in a suitable probability space,  $X_n$  has stationary ergodic increments. This result plays an key role in providing that large number law.

Let us consider the following auxiliary process with values in  $E' = E \times \{-1, 0, 1\}$ : for  $(e, \omega) \in \Omega$ ,  $V_n = V_n(e, \omega) = (\theta^{X_n(e, \omega)}e, Y_n(e, \omega)) \quad n \geq 0$ . where  $Y_n = X_{n+1} - X_n$  is the  $n$ th increment of the RWIRE( $X_n$ ).

One can easily see that  $(V_n)$  is Markov chain. Following [8], it suffices to study the first component  $x_n = x_n(e, \omega) = \theta^{X_n(e, \omega)}$ , which describes the 'environment seen from the position of the random walk'. This is a Markov chain, with state space  $E$ , initial distribution  $\pi$  and transition operator  $Q$  given by:

$$Q\psi(e) = \psi(\theta e)\alpha(e) + \psi(\theta^{-1}e)\beta(e) + \psi(e)\gamma(e)$$

for a measurable bounded function  $\psi$ . By using a 'regenerative method', we construct an invariant probability measure, which is equivalent to  $\pi$  ( $\pi$  is not  $Q$ -invariant).

**Lemma 2.** Suppose  $E(S) < +\infty$  (respectively  $E(F) < +\infty$ ). Then there exists a  $Q$ -invariant probability  $\nu$ , which is equivalent to  $\pi$  with the explicit density  $h(e) = \nu(1 + \frac{\gamma_0}{\alpha_0} + \sigma)S(e)$  (respectively  $h'(e) = \nu'(1 + \frac{\gamma_0}{\beta_0} + \sigma^{-1})F(e)$ ), where  $\nu$  (respectively  $\nu'$ ) is a normalization constant.

*Proof.* Construction of the invariant probability measure. For symmetry reasons, we prove the result only when  $E(S) < +\infty$ . In this case, one can easily check by Theorem 2.1, that in particular,  $\limsup_{n \rightarrow +\infty} X_n = +\infty$  so that the hitting time  $T_1$  is finite with probability 1. set  $\nu(B) = E_\pi(\sum_{k=0}^{T_1-1} I_B(x_k))$ ,  $B \in \mathcal{E}$ . where  $I_B$  denotes the indicator of the set  $B$ .  $\nu(B)$  is the mean number of visit to the set  $B$ , by  $(x_n)$ , before time  $T_1$ . It is not difficult to see that  $\nu$  is a well-defined measure. We prove below that it is  $Q$ -invariant. For a measurable non-negative function  $f$ ,  $\nu(f) = \int_E f d\nu = \sum_{k=0}^{+\infty} E_\pi(f(x_k), T_1 > k)$ . Then one can check that

$$\begin{aligned} \nu(Qf) &= \sum_{k=0}^{+\infty} E_\pi(f(x_{k+1}), T_1 > k) \\ &= \sum_{k=0}^{+\infty} E_\pi(f(x_{k+1}), T_1 = k+1) + \sum_{k=0}^{+\infty} E_\pi(f(x_{k+1}), T_1 > k+1) \\ &= E_\pi(f(x_{T_1}), T_1 > 0) + \sum_{k=1}^{+\infty} E_\pi(f(x_k), T_1 > k) = \nu(f). \end{aligned}$$

In the last equality we have used

$$E_\pi(f(x_{T_1})) = \int f(\theta e) d\pi(e) = \int f(e) d\pi(e) = E_\pi f(x_0).$$

To obtain density of  $\nu$ . We introduce the random variables  $N_i = \#\{k, 0 \leq k \leq T_1 : X_k = i\}$ ,  $i \in Z^-$ . Then we have  $\nu(f) = E_\pi(\sum_{i \leq 0} f(\theta^i \cdot) N_i)$  After a few transformations and



the invariance of  $\pi$ , we obtain  $v(f) = \int_E f(e) [\sum_{i \leq 0} E_{\theta^{-i}e}(N_i)] d\pi(e)$ . Thus,  $v$  has density  $h(e) = \sum_{i \leq 0} E_{\theta^{-i}e}(N_i)$ , subject to proving the convergence of the random series. But  $v(E) = E(T_1)$ . To obtain  $ET_1$ , need the following lemma 3. let  $X^F(t)$  be the process obtain from  $X_t$  by reflecting  $X(t)$ . Given the environment  $e$ ,  $X^F(t)$  has transition function  $P_e(x, y)$  satisfying: If  $x \leq 0$ ,  $P_e(x, x+1) = \alpha_x$ ,  $P_e(x, x-1) = \beta_x$ ,  $P_e(x, x) = \gamma_x$ ,  $P_e(x, x+a) = 0$  otherwise.

**Lemma 3.** For a fixed environment  $e$ , the system of equations

$$u(1) = 1; \quad u(y) = \sum_{x \leq 1} u(x) P_e(x, y), y \leq 1;$$

has a nonnegative solution  $u(*)$ , for which  $\sum_{y \leq 1} u(y) < \infty$  if and only if  $E\{T_1|e\} < \infty$ . furthermore, if  $E\{T_1|e\} < \infty$  then the solution is unique and  $E\{T_1|e\} = \sum_{y \leq 1} u(y)$ .

*Proof.* See Breiman[9, 143-145] and recall that given the environment,  $X^F(t)$  is Markov chain. Define the random variable  $h(n), n \in \mathbb{Z}$  by  $h(0) = 1; h(a) = 0, a \geq 1; h(n-1) = h(n) \cdot \frac{\beta_n}{\alpha_n}, n \leq 0$ . for  $n \leq 0$ , let  $v(n) = h(n)/\alpha_n$ , and  $v(1) = 1$ . Then  $v(*)$  is the solution of Lemma 2. i.e.  $E(T_1|e) = (1 + \sigma_0 + \frac{\gamma_0}{\beta_0}) + \sum_{j=-\infty}^0 (1 + \sigma_{j-1} + (\frac{\gamma_{j-1}}{\beta_{j-1}})) \sigma_j \cdots \sigma_0$ . Taking the expectation with respect to the environment yields:  $E(T_1) = E((1 + \sigma + \gamma_0/\alpha_0)S)$  Now, when  $E(S) < +\infty$ , the density  $h(e) = v(1 + \sigma + \gamma_0/\alpha_0)S(e)$ , where  $V^{-1} = E(T_1) = E((1 + \sigma + \gamma_0/\alpha_0)S(e))$ . The equivalence between  $v$  and  $\pi$  holds since  $h$  is a positive function. Thus the lemma 2 is proved.

**Lemma 4.** Under probability  $P_v$ , the sequence  $(V_n)$  is stationary and ergodic.

*Proof.* The Markov chain  $(V_n)$  has state space  $E^V = E \times \{-1, 0, 1\}$  and transition operator  $K\varphi(e, y) = \varphi(\theta^y e, +1)\alpha(\theta^y e) + \varphi(\theta^y e, -1)\beta(\theta^y e) + \varphi(\theta^y e, 0)\gamma(\theta^y e)$ , for a measurable bounded function  $\varphi$ . The stationary of  $(V_n)$  comes from the fact that, under probability  $P_v$ , the distribution of  $V_0$  is  $\eta(de \times dy) = v(de) \otimes p_e(dy)$ , where  $P_e = \alpha(e)\delta_{+1} + \beta(e)\delta_{-1} + \gamma(e)\delta_0$  and where  $\delta_{+1}, \delta_{-1}$  and  $\delta_0$  stand for Dirac measures. Thus one can see that  $\eta$  is K-invariant, since  $v$  is Q-invariant. The remainder of the proof is essentially the same as in [8, Lemma 2]. Thus the lemma 4 is proved.

*Theorem 2.2 Proof:* We note that (i)-(iii) are mutually exclusive and take into account all possible situations. Under (i) or (ii) the RWIRE is transient while under (iii) it may be either recurrent or transient. The conditions in (i)-(ii) mean the fluctuations of  $\beta_0/\alpha_0$  around value 1 are 'small'. In the case of independent environments, (i)-(iii) correspond respectively to  $E\sigma < 1, E\sigma^{-1} < 1$  and  $E(\sigma)^{-1} \leq 1 \leq E(\sigma^{-1})$ . (ii) can be deduced from (i) by exchanging the roles of  $S$  and  $F$ . let us prove (i). we will initially calculate the limit of  $X_n/n$  under Probability  $P_v$ . Once this is done, the limit calculated remains valid under probability  $P = P_\pi$ , due to the equivalence of  $\pi$  and  $v$ , one can see  $P_\pi$  and  $P_v$  are equivalent.

We write  $X_n = \sum_{k=0}^{n-1} Y_k$  (where  $Y_k = X_{k+1} - X_k$ ) as the sum of its increment; then using lemma 2 and lemma 4, by the ergodic theorem,

$$\begin{aligned} \lim_{n \rightarrow +\infty} \frac{X_n}{n} &= \int_E E_e(Y_0) dv(e) \\ &= \int_E (\alpha_0(e) - \beta_0(e)) h(e) d\pi(e) = v \int_E (\alpha_0(e) - \beta_0(e)) \alpha(0)^{-1} S(e) d\pi(e) \\ &= v \int_E (1 + \sigma_1 S(\theta e)) d\pi(e) - v \int_E \sigma_0 S(e) d\pi(e) = v \quad P_V - a.s. \end{aligned}$$

The result for  $T_n$  can be shown that  $\lim_{n \rightarrow +\infty} \frac{T_n}{n} = \lim_{n \rightarrow +\infty} \frac{n}{X_n}$ . In fact, let  $k_n$  be unique nonnegative integers such that  $T_{k_n} \leq n < T_{k_n+1}$ . In case (i),  $X_n \rightarrow +\infty$  a.e. thus  $k_n \rightarrow \infty$  a.e. Now,  $X_n < k_n + 1$ ; also, by time  $n$  the random walk has already hit  $k_n$  and since it moves not more than one integer at each step  $k_n \leq X_n + (n - T_{k_n})$ .

Thus

$$\frac{k_n}{n} - (1 - \frac{T_{k_n}}{n}) \leq \frac{X_n}{n} < \frac{k_n + 1}{n}.$$

But the definition of  $k_n$  implies  $\lim_{n \rightarrow \infty} \frac{k_n}{n} = \lim_{n \rightarrow \infty} \frac{n}{T_n}$  and  $\lim_{n \rightarrow +\infty} \frac{T_{k_n}}{n} = 1$  a.e., it suffices to get. part(iii) follows by first considering the random times  $\tau_k = T_k - T_{k-1}, k \geq 1$  (this gives the result for  $T_n$ , and the result for  $T_{-n}$  can be derived similarly); Suppose that  $\limsup X_n = +\infty$  (if not,  $T_n = +\infty$  a.s. for large  $n$  and the result is obvious) then  $\tau_k < \infty$  with probability 1. By a classical truncation argument, one can apply the of large numbers for independent random variables (with arbitrary distributions) once an environment is fixed. This leads to  $\liminf_{n \rightarrow +\infty} (1/n) \sum_{k=1}^n \tau_k \geq E(\tau_1)$ . But  $E(\tau_1) = +\infty$ . So  $\lim_{n \rightarrow +\infty} T_n/n = +\infty$  a.s. and  $\lim_{n \rightarrow +\infty} \frac{X_n}{n} = \lim_{n \rightarrow +\infty} T_n/n = 0$  a.s. Thus theorem 2.2 is proved. It is easy to obtain the following corollary 1. ‘ □

**Corollary 1.** If  $(\alpha_i, \beta_i)$  are i.i.d,

- (i) if  $E\sigma < 1$  then  $\lim_{n \rightarrow \infty} \frac{X_n}{n} = \frac{1-E\sigma}{1+E\sigma+E(\gamma_0/\alpha_0)}$  a.s..
- (ii) if  $E(\sigma^{-1}) < 1$ , then  $\lim_{n \rightarrow \infty} \frac{X_{-n}}{n} = -\frac{1-E(\sigma^{-1})}{1+E(\sigma^{-1})+E(\gamma_0/\beta_0)}$  a.s..
- (iii) if  $(E\sigma)^{-1} \leq 1 \leq E(\sigma^{-1})$ , then  $\lim_{n \rightarrow \infty} \frac{X_n}{n} = 0$  a.s..

## References

1. Solomon, F.: Random walk in a random environment. Ann. Prob. 3, 1–31 (1975)
2. Alili, S.: Asymptotic Behaviour for random walks in random environments. J. Appl. Prob. 36, 334–349 (1999)
3. Hu, Y., Shi, Z.: The limits of Sinai’s simple random walk in random environment. Ann. Prob. 26, 1447–1521 (1998)
4. Liu, X.D., Dai, Y.L.: A class of random walks in random environments. J. Math. Study 35, 298–302 (2002)
5. Liu, X.D., Dai, Y.L.: A class of random walks on half-line in random environments. Acta. Mathematica. Scientia. 25, 98–102 (2005)

6. Key, E.S.: Recurrence and transience criteria for random walk in a random environment. *Ann. Prob.* 12, 529–560 (1984)
7. Kesten, H.: Sum of stationary sequences cannot grow slower than linearly. *Proc. Amer. Math. Soc.* 49, 205–211 (1975)
8. Kozlov, S.M.: The method of averaging and walks in inhomogeneous environments. *Russian Math. Surveys* 40, 73–145 (1985)
9. Breiman, L.: *Probability*. Addison-Wesley Reading, Massachusetts (1968)
10. Chung, K.L.: *A course in Probability Theory*. Academic Press, New York (1974)
11. Greven, A., den Hollander, F.: Large deviations for a random walk in random environment. *Ann. Prob.* 22, 1381–1428 (1994)
12. Comets, F., Gantert, N., Zeitouni, O.: Annealed and functional large deviations for one-dimensional random walk in random environment. *Prob. Theory. Relat. Fields* 118, 65–114 (2000)

# Research on the Spatio-temporal Pass Index Model in Emergency Evacuation\*

Liu Yi, Zhu Haiguo, Huang Quanyi, Zhao Yunxiu, Yuan Shengcheng, and Zhang Hui

**Abstract.** Based on the analysis of the mixed traffic evacuation in china, this paper presented the behavior characteristics and inherent laws of the mixed traffic evacuation, designed the quantification model of evacuation behaviors and road availability model. By using the BPR function and UE model, the mixed traffic Emergency Evacuation Spatio-temporal Pass Index Model (EESPIM) which possessed the dual-layer programming structure was designed by considering the maximum of road availability and the minimum of evacuation time as the index of optimal evacuation path choice. Finally, the optimal evacuation path choice by virtue of the improved heuristic algorithm, which provided the scientific evidence for decision-making in the emergency evacuation.

## 1 Introduction

In recent years, a variety of unexpected emergencies have posed a major threat to many people, meanwhile, accompanied by a regional emergency evacuation. However, traditional evacuation model ignored the reality of the situation for lack of considering the evacuation behaviors and irrational actions in emergency.

On the whole, some experts paid more attention to emergency evacuation mainly focused on two aspects at home and abroad, on the one hand, the evacuation model, evacuation time calculation and simulation in small regions or

---

Liu Yi

Institution of Public Safety Research, Tsinghua University, Beijing, China

e-mail: liuyi@tsinghua.edu.cn

Zhu Haiguo · Huang Quanyi · Zhao Yunxiu · Yuan Shengcheng · Zhang Hui

Institution of Public Safety Research, Tsinghua University, Beijing, China

\* Supported by National Natural Science Foundation of China (Grant No. 90924001; 70833003).

buildings [1-9, 11, 12-14] were presented. Cova and Church designed the Critical Cluster Model (CCM) for evacuation region risk assessment [9, 10], CHEN Xiang and LI Qiang used the Critical Cluster Model to give the traffic evacuation risk assessment model in cities by considering the hazardous materials [10]. On the other hand, In view of accessibility, some other experts analyzed the evacuation theory. However, accessibility measurement approaches were only illustrated by Euclidean distance, cost and time in practical applications [14, 15], which cannot satisfy the emergency evacuation.

Based on the analysis above, this paper presented the behavior characteristics and inherent laws of the mixed traffic emergency evacuation, designed the quantification model of evacuation behaviors and road availability model, used the BPR function and UE model, the mixed traffic Emergency Evacuation Spatio-temporal Pass Index Model(*EESPIM*) In this paper I first review recent progress in emergency evacuation technologies. The second major section then presents the mixed traffic Emergency Evacuation Spatio-temporal Pass Index Model (*EESPIM*) which possessed the dual-layer programming structure and the detailed calculation was also addressed. Because of some practical researches, the paper ends with some brief concluding points, and the case study will be presented in the following research paper. The flow chart of *EESPIM* calculation was illustrated in Fig. 1.

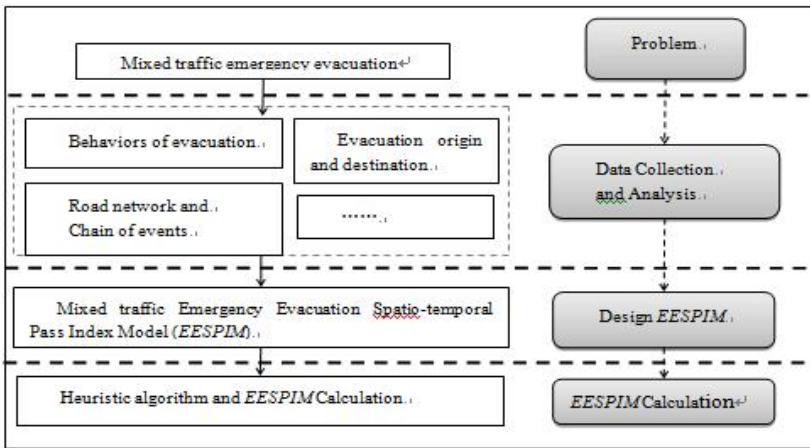


Fig. 1 Flow chart of *EESPIM* calculation

## 2 Emergency Evacuation Spatio-temporal Pass Index Model (*EESPIM*) DESIGN

### 2.1 Objective Function

The mixed traffic Emergency Evacuation Spatio-temporal Pass Index Model (*EESPIM*) was used for analyzing emergency evacuation quantitative indices,

optimizing the evacuation which referred to reliability and efficiency of emergency evacuation. The objective model can be defined as:

$$\max G = aR_r + bE_r \tag{1}$$

Where  $R_r$  is the reliability of emergency evacuation path  $r$ ,  $E_r$  is the efficiency of emergency evacuation path  $r$ ,  $a$ ,  $b$  are the dynamic parameters which can be adjusted by evacuation directors, calculated by means of experience, research results and historical data statistical analysis.

Because of the complexity of evacuation road, irrational behaviors of evacuee, and the chain of events in mixed traffic emergency evacuation, this paper represented the efficiency of emergency evacuation by using evacuation utility.  $E_r$  in (1) was transformed into evacuation utility, which can be defined as:

$$\max G = mR_r + nU_r \tag{2}$$

Where  $U_r$  is the utility of emergency evacuation path  $r$ , the greater its value, the shorter evacuation time,  $m$ ,  $n$  are the dynamic parameters which will affect *EESPIM*, can be calculated by means of experience, research results and historical data statistical analysis.

In order to extract the appropriate evacuation path, this paper used BPR function to analyze the evacuation impedance in each path, which can put the evacuee into the corresponding evacuation path. The internal relation between evacuation time and evacuation flux can be defined as:

$$t = t_0 \{ 1 + \alpha [\frac{x(a)}{c}]^\beta \} \tag{3}$$

Where  $t$  is the evacuation time,  $t_0$  is the mixed traffic time in normal state,  $x(a)$  is the actual flux in evacuation path,  $c$  is the pass capacity of evacuation path,  $\alpha=0.15, \beta=4$ (the value were recommended by U.S. Bureau of public Roads)

The reliability of evacuation path was represented by the probability of evacuation origin to destination. This paper improved the reliability of evacuation by import the quantification of evacuation behaviors and evacuation path availability, which can be defined as:

$$R_r = (1 - \delta) \cdot (1 - \prod_{m=1}^n P_m) \tag{4}$$

The expression can be explained as following:

1)  $P_m$  is the quantification of evacuation behaviors. The quantification of evacuation behaviors can be expressed by logit model, which can be defined as:

$$P_m = \frac{\exp(U_m)}{\sum_{n \in A_m} \exp(U_n)} \tag{5}$$

Where  $P_m$  is the probability of evacuation behavior, is  $U_m$  is the utility of evacuation behavior which can be defined as  $U_m = \omega X$ ,  $\omega$  is parameter array,  $X$  is observable parameters (such as age, evacuation experience and choices, et al.).

2)  $\delta$  is quantification of evacuation path availability, which can be expressed by fuzzy set theory. Based on studying the relation between evacuation path and “temporary obstacles (such as hill slide, flood, chains of event, et al.)”, the quantification set can be calculated as:

$$\mu_{R^*}(x, y) = k \left[ 1 - \frac{d^{02}((x, y), R)}{\varepsilon} \right] \tag{6}$$

Where  $R^*$  is composed of the boundary points of “temporary obstacles”,  $d^{02}$  is the function of distance between point(x, y) and “temporary obstacles”.

Because of the inverse relationship between value  $\delta$  of evacuation path availability and quantification of “temporary obstacles”, the  $\delta$  can be defined as:

$$\delta = 1 - \mu_{R^*}(x, y) \cdot \omega \tag{7}$$

Where  $\omega$  is the relating coefficient between evacuation path characteristic and “temporary obstacles”,  $\omega = 0$  represents the evacuation path is unavailable,  $\omega = 1$  represents the evacuation path is well available.

### 2.2 Dual-Layer Programming Model

The dual-layer programming model was composed of upper objective model ( $UM_1$ ) and upper objective model ( $UM_2$ ). the mixed traffic evacuation time can be calculated by using the  $UE$  model.

We assumed path nodes and segments constitute an evacuation path,  $x(a)$  is the actual flux in evacuation path,  $a$  is the segment of evacuation path, and then the dual-layer programming model is designed.

1)  $UM_1$  can be defined as:

$$\max G = mR_r + nU_r \tag{8}$$

s.t.

$$\begin{cases} \sum_r R_r = 1 \\ t = t_0 \left[ 1 + \alpha \left[ \frac{x(a)}{c} \right]^\beta \right] \\ T_r = \sum_r t \end{cases}$$

2)  $UM_2$  can be defined as:

$$\min G(x) = \sum_a \int_0^{x(a)} t_a(\omega) d\omega \tag{9}$$

s.t.

$$\begin{cases} q_{r,s} = \sum_k f_k^r & \forall r, s \\ x(a) = \sum_{rs} \sum_k \sigma_{ak}^r f_k^r & \forall a \\ f_k^r \geq 0 \end{cases}$$

Where  $t_a$  is the impedance of evacuation path  $a$ ,  $r, s$  are the evacuation origin and destination,  $q_{rs}$  is the evacuee of path  $r - s$  in OD,  $f_k^{rs}$  is the evacuation flux of the path  $k$  which located on the evacuation segment  $r - s$  in OD,  $\sigma_{a,k}^{rs}$  is the relating variable between the evacuation path which can be defined as:

$$\sigma_{a,k}^{rs} = \begin{cases} 0 & \text{others} \\ 1 & \text{path } a \text{ which located on the evacuation segment } r - s \end{cases}$$

### 2.3 EESPIM Algorithm Calculation

TheEESPIM was composed of  $UM_1$  and  $UM_2$  in this paper,  $UM_1$  was used for the maximum of road availability and  $UM_2$  was used for the minimum of evacuation time. If  $UM_2$  changed,  $UM_1$  would change also, when the impedance of each evacuation path in a balance state, the optimal evacuation path can be founded. The procedure of EESPIM calculation by virtue of heuristic algorithm can be explained as following:

- Putting the evacuee into evacuation paths by virtue of improved UE model
  - 1) Constructing the evacuation paths network with path nodes and segments.
  - 2) Initializing model.

Let  $t_a^s = t_s(0)$ , this paper putting the OD matrix into the evacuation path by using the All or Nothing Method, evacuation flux  $\{t_a^n\}$  can be calculated, iterate at one time, and  $t_a^n = t_a(x_a^{n-1})$  can be calculated.

- 3) Loading the evacuation paths network

Based on the calculation of  $\{t_a^n\}$ , putting the OD matrix into the evacuation path by using the All or Nothing Method, and then the evacuation flux  $\{x_a^n\}$  can be represented.

- 4) Analysis of UE Calculation Convergence
- Analyzing the optimal evacuation paths
    - 1) Calculating the impedance in each evacuation path by virtue of BPR function;
    - 2) Designing the spatio-temporal pass index data structure;
    - 3) Improving the Dijkstra algorithm to analyze optimal evacuation path.
  - Calculating the optimal evacuation path which satisfied  $\max G = mR_r + nU_r$ .

### 3 Concluding Comments and Future Work

Through the analysis of the mixed traffic evacuation in china, this paper analyzed the behavior characteristics and inherent laws of the mixed traffic evacuation. The



mixed traffic Emergency Evacuation Spatio-temporal Pass Index Model (*EESPI*M) which possessed the dual-layer programming structure was designed by considering the maximum of road availability and the minimum of evacuation time as the index of optimal evacuation path choice by improving the BPR function and UE model. Finally, the procedure of *EESPI*M calculation was presented by combining with heuristic algorithm. However, because of lacking the indispensable data which used for the *EESPI*M, the case study was not included in this paper. The main works in the future are the data analysis and *EESPI*M calculation, which can provide the scientific evidence for decision-making in emergency evacuation.

## References

- [1] Ye, W., Yu, Y.: Path planning based on extension strategy in unknown environment. *Computer Engineering and Applications* 46(19), 10–13 (2010)
- [2] Ji, L., Zhang, F., Fu, Y., et al.: 3D Interaction Techniques Based on Semantics in Virtual Environments. *Journal of Software* 17(7), 1535–1543 (2006)
- [3] Yin, X., Pan, X., Wu, Y.: Decision making of personnel emergency protective actions for toxic gas leakage accident. *Journal of Nanjing University of Technology (Natural Science Edition)* 32(1), 64–68 (2010)
- [4] Church, R.L., Sexton, R.M.: Modeling Small Area Evacuation: Can Existing Transportation Infrastructure Impede Public Safety? (2002), <http://www.ncgia.ucsb.edu/vital>
- [5] Lu, J., Fang, Z., Lu, Z., et al.: Mathematical model of evacuation speed for personnel in buildings. *Engineering Journal of Wuhan University* 35(2), 66–70 (2002)
- [6] Cova, T.J., Church, R.L.: Modeling community evacuation vulnerability using GIS. *International Journal of Geographic Information Science* (11), 763–784 (1997)
- [7] Church, R.L., Cova, T.J.: Mapping evacuation risk on transportation networks using a spatial optimization model. *Transp. Res. Pt. C* 8, 321–336 (2000)
- [8] Chen, X., Li, Q., Wang, Y., et al.: The Critical Cluster Model and Its Application in Accessibility Assessment of Public Bus Network. *Acta Geographica Sinica* 64(6), 693–700 (2009)
- [9] Song, Y., Pan, X., Yu, Z., et al.: Research on Connectivity of Mountainous Road Network for Emergency Evacuation. *Highway Engineering* 34(5), 29–32 (2009)
- [10] Kobes, M., Helsloot, I., et al.: Way finding during fire evacuation; an analysis of un-announced fire drills in a hotel at night. *Building and Environment* 45, 537–548 (2010)
- [11] Yuan, Y., Wang, D.: Route Selection Model in Emergency Evacuation under Real Time Effect of Disaster Extension. *Journal of System Simulation* 20(6), 1563–1566 (2008)
- [12] Oven, V.A., Cakici, N.: Modeling the evacuation of a high-rise office building in Istanbul. *Fire Safety Journal* 44, 1–15 (2009)
- [13] Nilsson, D., et al.: Evacuation experiment in a road tunnel: A study of human behavior and technical installations. *Fire Safety Journal* 44, 458–468 (2009)
- [14] Chen, J., Lu, F., Cheng, C.: Advance in Accessibility Evaluation Approaches and Applications. *Progress in Geography* 26(5), 100–110 (2007)
- [15] Braun, A., Bodmann, B.E.J., Musse, S.R.: Simulating virtual crowds in emergency situations. In: *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pp. 244–252 (2005)

# On Issues of Multi-path Routing in Overlay - Networks Using Optimization Algorithms

Sameer Qazi and Tim Moors

**Abstract.** Routing policies used in the Internet can be restrictive, limiting communication between source-destination pairs to one path, when often better alternatives exist. To avoid route flapping, recovery mechanisms may be dampened, making adaptation slow. Overlays have been widely proposed to mitigate the issues of path and performance failures in the Internet by routing through an indirect-path via overlay peer(s). Choosing alternate-paths in overlay networks is a challenging issue. Guaranteeing both availability and performance guarantees on alternate paths requires aggressive active probing of all overlay paths, which limits scalability when the number of overlay-paths becomes large. If path correlations could be determined, multi-media applications can benefit greatly if their traffic could be sent over multiple uncorrelated paths. Statistical approaches have been previously proposed for establishing path correlation for multi-path routing; In this paper we test the efficacy of such approaches in Internet scale overlay networks using real-world datasets.

## 1 Introduction

End-to-end paths in the Internet sometimes fail to deliver the Quality of Service (QoS) required by some applications. For many user-perceived performance failures/faults there exists a redundant path available which can be used to actually prevent or “mask” the fault from the end user by using quick switch-over mechanisms. One study [1] shows that for almost 80% of the paths used in the

---

Sameer Qazi

National University of Sciences and Technology, Karachi, Pakistan

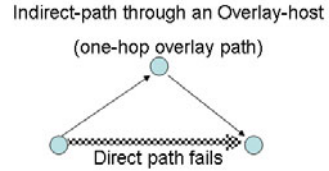
e-mail: sameer@pnec.edu.pk

Tim Moors

University of New South Wales, Kensington Campus, Sydney, Australia

e-mail: moors@ieee.org

**Fig. 1** Indirect-path through an overlay-host when direct-path fails.



Internet there is an alternate path with lower probability of packet loss. On detection of failure on primary direct-path, Internet routers switch to alternate direct-paths learnt through the Border Gateway Protocol (BGP). Despite being highly scalable, BGP only addresses reachability, without similar guarantees for QoS, which can take several minutes to recover from failures.

When the direct-path between two hosts fails, overlay networks can quickly establish an indirect-path through intermediate host/s. It has been found that majority of direct-path failures can be bypassed by an indirect path through a single intermediate overlay-host (*one-hop* overlay path) [2] (Figure 1).

Section 2 describes the Mean-Variance model theory which outlines how path correlations are established; Section 3 discusses the mathematics of the problem formulation. In Section 4 evaluates the performance of proposed model using real-world Internet data. Section 5 discusses related work and Section 6 summarizes the key findings of the paper.

## 2 Mean-Variance Model

The Mean-variance model solves the problem of optimal portfolio selection for an individual. Portfolio theory [4] explores how risk-averse participants construct portfolios in order to optimize expected returns for a given level of market risk e.g. profits from investment in stocks in stock-exchanges. In general given the mean, variance and co-variances of the returns from stocks, one can compute the maximum possible expected return for a given level of risk.

The mean-variance model discussed above could be adapted to perform path-selection in overlay-networks. Given paths with different levels of quality, one could use some combinations of the paths to minimize the fluctuations in quality. A selection process based on the mean-variance model would invariably use uncorrelated paths to minimize the variations in performance. Paths that do not share low-performance physical links are precisely those that exhibit low correlations in performance and would get picked by the selection process that minimizes variance.

## 3 Problem Formulation

Our formulations reflect the fact that real-time multimedia applications require low latency and/or low loss-rate. Given a collection of overlay-paths it is natural to model the throughput and latency of each path as a random variable with a given distribution. Then the problem of optimizing the end users experience

reduces that of trying to find the optimal allocation of packets to the given collection of paths so as to minimize the variation, subject to a fine choice of average latency or collective throughput. Once we compute a solution that determines a set of overlay paths with associated weights referring to fractions of the data flow to be conveyed through certain paths, we then assume that the transport process would transmit packets in a random and independent manner down the various paths with a probability that is proportional to their weights.

We formulate our problem as an optimization problem similar to that used by Antonova et al. [12]. Consider a collection of overlay paths  $i=1,2,3,\dots,n$  with latencies  $L_i$ . Let the covariance matrix be represented by  $C$ ;  $C_{ij} = covar(L_i, L_j)$ . Let  $w$  represent the vector of weights, where  $w_i$  is the fraction of data sent along overlay path  $i$ . Let  $\mu$  be the desired average latency. Here we assume that latency is a fungible commodity, in other words packets that arrive earlier than the average compensate for late packets.

Then we wish to solve the following optimization problem where  $w'$  stands for the transposed version of  $w$ .

#### *Optimization Problem*

$$\begin{aligned}
 & \text{Minimize} && w' C w \\
 & \text{subject to} && : \\
 & && w' L = \mu \\
 & && w' \mathbf{1} = 1 \\
 & && w \geq 0
 \end{aligned} \tag{1}$$

This optimization problem with convex objective function and convex constraints can be solved to the desired precision in polynomial time, since quadratic programming with a positive semi-definite Hessian matrix is in  $P$  [5]. A symmetric Hessian matrix  $M$  is said to be positive semi-definite if all its eigenvalues are non negative, or equivalently for all vectors  $x$  it is the case that  $x' M x \geq 0$ . It is easy to see that any covariance matrix  $C$  is positive semi definite since  $x' M x$  is just the variance of the collection of overlay paths weighted by  $x$  and hence non-negative.

## **4 Performance Evaluation**

To evaluate the performance of the proposed framework, we only require the requisite end to end path metric and topology information. Throughout the remainder of this paper, we analyze the performance of overlay networks using real Internet datasets, so it is important that the methodology of obtaining this datasets is explicitly described before proceeding any further. Our datasets include a deployed Internet wide US based experimental network, Active Measurement Project (AMP) [3], managed by National Laboratory for Applied Network Research (NLNR), which performs active measurements (delay measurements and trace routes) between 150 hosts in US and outside. From this set of AMP hosts we randomly selected 62 AMP hosts i.e. our overlay network consists of the paths between these 62 AMP hosts.

We use the same definition of a path anomaly as used by [7]. We define an anomaly as occurring when path metric (delay) exceeds its average value by a factor ( $k$ ) of the standard deviation ( $\sigma$ ) of the delay values in the previous 60 epochs, one hour for AMP network:

$$Path\ Delay > Path\ Delay_{average} + k\sigma \quad (2)$$

where  $k=1,2,3..$  is a tunable parameter to trigger an anomaly for small to large delay variations with increasing values of  $k$ , respectively. The one hour time window prevents oscillations [11] in overlay networks if paths are switched rapidly for minor and short lived performance gains. These values are typical of those used by Chua et al. [9] and Fei et al. [7]. We select  $k = 5$  to emulate path performance failures and the desired average latency  $\mu=100\text{msec}$ , typical for real time applications; e.g. VoIP. We run our trace based simulations in Matlab using the optimization problem formulation described before.

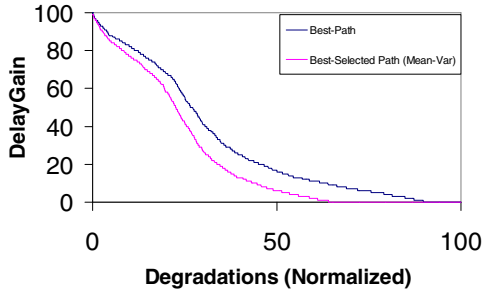
Figure 2 indicates that the multi-path routing algorithm is able to select the paths exhibiting the best performance benefits. However, two things warrant attention. What is the total number of paths selected by the algorithm and what percentage of traffic is sent along these paths?

Figure 3 indicates that in most cases (60%) more than 20 paths are selected to route packets. This is a very large number; the algorithm is roughly selecting a third of the paths. We determine the reasons for this in our analysis.

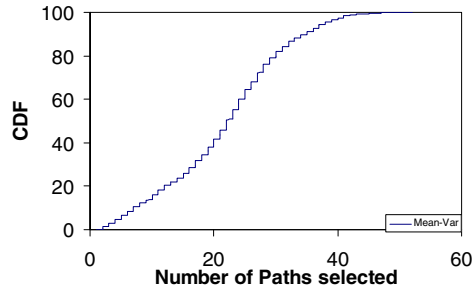
Next we look at what percentage of the traffic is split along these paths. The results are presented in Figure 4. Surprisingly, we do not see any strong pattern between weights assigned to paths (percentage of traffic routed) and their delay gain. Delay gain is:  $[\text{delay}_{\text{direct path}} - \text{delay}_{\text{1-hop overlay path}}] / \text{delay}_{\text{direct path}}$ . Figure 5 shows the correlation between weights assigned to paths i.e. percentage of traffic routed on them and the % delay gain (delay gain expressed as percentage) of the path. We even notice a negative correlation in a large number of cases. This is because:

1. Several paths are exhibiting similar levels of performance when the direct path undergoes performance degradation. This has been observed before by Anderson et al. [12] and path failures are often correlated, as significant portions of path segments are shared between one-hop overlay paths.
2. The direct paths on the AMP network often exhibit very small delays. Even when experiencing degradations their delay is still much lower than majority of average delay on the one hop overlay paths.
3. Path measurements are not GPS synchronized. We incorporate these differences by assigning measurements belonging to 60 second bins. Note that this is path delay measurement interval in the datasets considered. This cannot effectively capture the mutual path variations for short lived path degradations.

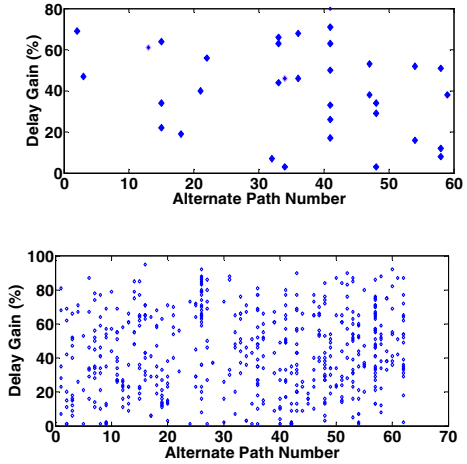
**Fig. 2** Performance comparison of path selected using Mean Variance Approach and best possible path; k=5



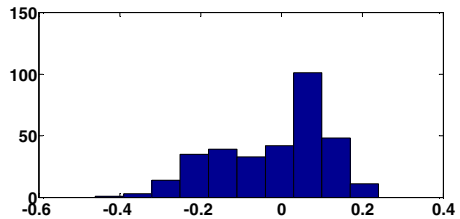
**Fig. 3** Number of paths selected out of 60 (max);k=5



**Fig. 4** Delay gain on alternate paths carrying (a) 10% or more; (b) 1% or more of the split traffic; (k=5,  $\mu=100\text{msec}$ ). x-axis indicates alternate path number, each has a choice of 62 overlay paths (including the two direct path formulations through the source and destination themselves acting as relays nodes)



**Fig. 5** Histogram of correlation. Correlation between weights assigned to paths and their delay gain merit



## 5 Related Work

Finding alternate-paths to act as a backup when primary routes fail in overlay networks is previously explored in [2]. Anderson et al, designed RON to be a resilient routing tool for the Internet by implementing a small overlay (50 nodes). Gummadi et al. [8] shows that in most cases alternate paths can be found using at most one overlay hop. Topology aware approaches have been extensively studied to optimize overlay path disjointness by Nakao et al. [6] and Fei et al. [7].

## 6 Conclusion

This paper presented the first practical analysis of multi-path routing for multi-media applications using the Mean-Variance heuristic using real world datasets. Disjoint path computation can be used as a heuristic to supplement measurement-based approaches [2] which are not scalable, or for alternate indirect-path computation when the direct path between two hosts is affected by a performance failure or an outage. Our results show that such heuristics can have only limited benefits if path information data is not collected aggressively and synchronously.

## References

- [1] Savage, S., et al.: The end-to-end effects of Internet path selection. In: SIGCOMM 1999: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, pp. 289–299 (1999)
- [2] Andersen, D., et al.: Resilient Overlay Networks. In: Proc. ACM Symp. on Operating Systems Principles, SOSP (2001)
- [3] Active Measurement Project (Amp), <http://watt.nlanr.net/>
- [4] Bailey, R.: Economics of Financial Markets (2003)
- [5] Nocedal, J., Wright, S.: Numerical Optimization. Springer; ISBN: 0-387-98793-2
- [6] Nakao, A., et al.: Scalable routing overlay networks. SIGOPS Oper. Syst. Rev. 40(1), 49–61 (2006)
- [7] Fei, T., et al.: How To Select A Good Alternate Path In Large Peer-To-Peer Systems? In: IEEE INFOCOM 2006 (2006)
- [8] Gummadi, K., et al.: Improving the Reliability of Internet Paths with One-hop Source Routing. In: USENIX Symp. Operating System Design and Implementation (OSDI), pp. 183–198 (2004)
- [9] Chua, D.B., et al.: Network Kriging. IEEE Journal on Selected Areas in Communications 24, 2263–2272 (2006)
- [10] Keralapura, R., Chuah, C.-N., Taft, N., Iannaccone, G., et al.: Race Conditions in Coexisting Overlay Networks. IEEE/ACM Transactions on Networking 16(1) (February 2008)
- [11] Andersen, D.G., Snoeren, A.C., Balakrishnan, H.: Best-Path vs. Multi-Path Overlay Routing. In: IMC 2003, Miami Beach, Florida, USA, October 27-29 (2003)
- [12] Antonova, D., Krishnamurthy, A., Ma, Z., Sundaram, R.: Managing a portfolio of overlay paths. In: NOSSDAV 2004, Ireland (2004)

# Reviews on Intelligent Building Systems in China

Feng Jiang, Min Gao<sup>\*</sup>, and Zhijun Wang

**Abstract.** This paper reviews existing research and applications in the area of IB. At first, Four-level architecture is proposed to categorize intelligent building (IB) systems in China. The common structure and functions of existing IB systems are then discussed. The paper also analyzes variety of techniques of current IB systems including fingerprint identification, automatic drainage, water supply, and premise distribution systems. Finally, directions and issues are identified for further research.

**Keywords:** component, intelligent buildings, systems, structure.

## 1 Introduction

Intelligent Buildings (IB) construction has been built up since the end of the last century in China. So far, China has no less than 4500 intelligent buildings

---

Feng Jiang

College of Civil Engineering, Chongqing University, Chongqing, China  
Construction Engineering Faculty, Chongqing Technology and Business,  
Institute Chongqing, China  
e-mail: jiangfeng@cqu.edu.cn

Min Gao

College of Software Engineering, Chongqing University, Chongqing, China  
e-mail: gaomin@cqu.edu.cn

Zhijun Wang

College of Civil Engineering, Chongqing University, Chongqing, China  
e-mail: wzj@cqu.edu.cn

<sup>\*</sup> Corresponding author.



according to statistics. The areas of most IB are 25-50 thousand square meters, e.g. Beijing Development Building, Shanghai Jinmao Tower, and so on.

With the rapid development of economic in China, a number of requirements and aspirations for IB systems have been come into being especially since the beginning of 21 century. What people concern about IB systems are comfort, convenience, and personalization in the context of working or living conditions. Furthermore, with the development of Information Technology (IT), more and IB systems emerged, and need to be researched and innovated. The contribution of the paper includes three aspects. Firstly, a novel classification has been listed according to different service condition. Secondly, a general structure and functions of IB systems are summarized. Thirdly, it identifies several issues and possible directions for further research on IB systems. The paper provides a strong foundation for researchers to construct IB systems.

## 2 Existing IB Overview

According to the different target users, IB can be divided into two categories. One is public buildings, such as hotels, airport terminals, and sports stadiums. Another is residential buildings, such as single or double villas, and apartments. Intelligent residential buildings are the basic units of digital cities and information nodes. Since the construction of digital cities promotes the development of the intelligent buildings and smart residential area, the concept of intelligent buildings has been taken great changes these years in China. At the same time, growing investment of IB in recent years has bring about a great demand for better methods and techniques for evaluating investments to maintain investment profitability[1]. The basic types of IB can be categorized to the following four levels according to the usage and service condition of intelligent systems [2-5]:

A-Level, occupies 5% of the total. The IB systems of this level are operated in good conditions with a certain degree of system integration, system expansion, and upgrade ability. For this reason, this level is called advanced level.

B-Level, accounts for about 15%. The IB systems of the level are difficult to meet the basic requirements of users. There are necessary IB subsystems in these buildings in normal state, but the ability to integration, update, and processing is weak.

C-Level, is about 30% among four types. Even if the IB systems were opened, only few parts of them were running under normal situation. The feature of this level is handicapped.

D-Level, is nearly equal to the sum of A-Level, B-Level and C-Level. IB of this level are hardly used and in a paralyzed state. In this level, despite IB have been invested millions or tens of millions of dollars, IB systems are operated and handled entirely by hand or traditional methods.

### 3 Structure and Function

With the construction of digital cities, the Chinese government not only keep achieving the integration of IT industry and Intelligent Building (IB), but steps up the pace of the usage of information technology, especially for the construction of e-government systems. In the "11th 5-year Plan" period, the e-government goal of construction and application plans is to develop and perform under the "Three Nets and One Library" ("3+1 system" for short), which refers to Internal Office Network, Official Resources Network, Public Management and Service Network and E-Government Information Resource Library. The all levels' government departments have established a basic information technology framework, and then gradually expand to various businesses. For instance, some government offices have achieved IT and IB framework, which consists of three parts: Information Systems, Intelligent Systems, and Infrastructure Systems.

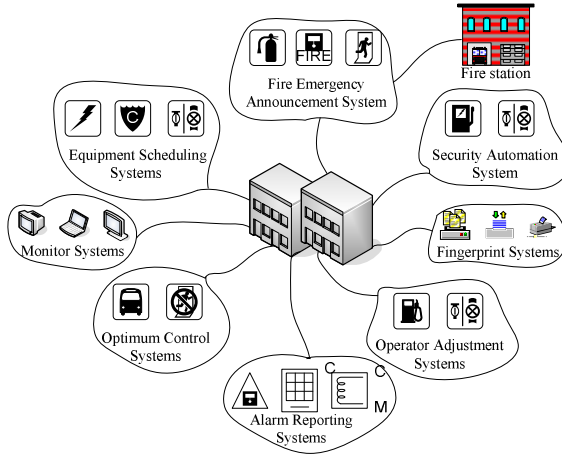
Information Systems include Network Systems, Information Application Systems, and Other Information Systems [6].

- Network Systems contain network switch subsystems, server subsystems, storage subsystems, network management subsystems, data centers, internet centers, telephone switches, and software systems.
- Information Application Systems include e-government, e-commerce, telemedicine, property management, office automation, information management, information security, and database systems.
- Other Information Systems consist of digital conference systems, video conferencing systems, multimedia display systems, wireless coverage systems, wireless access systems, and satellite TV systems. Intelligent systems include home intelligent, electrical and mechanical equipment, remote meter reading, security systems, fire alarm and linkage, parking, and environmental monitoring. Infrastructure systems include integrated wiring, piping, cable, lightning protection, and computer room.

### 4 IB Systems

IB systems traditionally refer to the Building Automation Systems, which optimize the start-up and performance of the lighting, heating, ventilating and air conditioning equipment, and such alarm systems.

The interaction of mechanical subsystems with IB systems is greatly increasing to provide comfortable, low energy use, and convenient living and working environment. At present, IB systems generally use computer-based monitor to coordinate, organize, and optimize building control the subsystems as follows (See Fig.1).



**Fig. 1** An Example of IB Systems

- **Equipment Scheduling Systems:** turning equipment off and on according to a schedule.
- **Optimum Control Systems:** turning heating and cooling equipment on in advance to ensure the building is at the required temperature during occupancy.
- **Operator Adjustment Systems:** accessing operator set-points that tune systems to changing conditions.
- **Monitor Systems:** logging of temperature, energy use, equipment start times, operator logon, etc.
- **Alarm Reporting Systems:** notifying the operator of failed equipment, out of limit temperature or pressure conditions.

Besides, some other modern and advanced systems are applied in the field of IB systems these years, e.g., fingerprint identification, automatic drainage, automatic water supply, premise distribution, and premise distribution systems.

#### ***4.1 Fingerprint Identification Systems***

The fingerprint identification system has gotten a great development as a reliable and highly feasible biometric identification technology [7], which has been applied in many fields such as justice, public security and security systems. Everyone is known to have a unique and immutable fingerprint which fingerprint is a series of ridges and furrows on the surface of the finger. When fingerprint systems have been more quick and convenient than before in today's modern society as well, people have to face the challenge of endless and boring information security risks. So that, the research of the fingerprint identification technology is very important in practice and theory [8].

## 4.2 Automatic Drainage Systems

The design concept of residential buildings has been using the traditional gravity drainage system without considering high demand in the past [9, 10]. Combining drainage control systems and draining water technology in the same layer will be the dominant direction of future design. With the use of information systems, the state of drainage systems can be observed, such as measuring of displacement, testing the amount of dirt, and water level abnormal monitoring in septic tanks, so as to achieve the living applicability, safety, and convenience.

## 4.3 Automatic Water Supply Systems

The automatic meter reading systems have been studied for more than 10 years in China. To borrow this technology to the residential building is the development trend of IB [11]. The goal of automatic water supply systems is to achieve the following main functions through information systems:

- To meet indoor humidity requirements through electronic information systems;
- To indicate water quality and to alarm;
- To integrate with automatic fire extinguishing systems;
- To combine solar energy with water temperature control.

## 4.4 Premise Distribution Systems

Premise Distribution Systems (PDS), as the basis of IB, transmit data, audio, graphics, and other information technologies within a building or between buildings, with the use of a series of high-quality standard materials and the combination of standard modules. PDS in a modern architecture, alike human body's nervous systems, realize the communications between voice, data equipments, switching equipments, and other facilities for the exchange of information. As an open network structure, PDS can be divided into five separate subsystems: workspace, horizontal, trunk, management, and equipment room subsystems. Each subsystem can be considered as an independent unit. A subsystem will not have to be modified when any subsystem changes.

- 1) The **workspace subsystem** includes a connection with a multi-core cable and plug connectors, playing a role of connecting terminal equipments and information sockets.
- 2) The **horizontal subsystem** is the information system from each work area outlet to cable distribution frame in the region of its management.
- 3) The **trunk subsystem** consists of all the vertical lines, providing equipments, and the wiring linkage between main distribution frame and floors.

- 4) The **management subsystem**, mainly connected by hardware equipments in crossway and linear-way, provides cable line connection, line orientation, and displacement management between various floors and patch panel on the overall levels of communication between the cables and trunks.
- 5) The **equipment room subsystems** [12] associate various devices and supporting hardware components. Meanwhile, it is the public system to interconnect each other.

## 4.5 Building Integrated Management Systems

Building integrated management system (BIMS) [13], unlike what mentioned upper from 4.1 to 4.4, is an advanced integrated system [14], relating to various subsystems integration and information sharing. BIMS should have functions to achieve an organic altogether with data, image sharing, and other systems. A BIMS deals with various system information and heterogeneous data. Point to point application integration is a kind of BIMS. When there is the need for cooperation between the two systems, the BIMS will coordinate for both the corresponding interface and interconnection.

## 5 Conclusion

This paper reviews existing research and application in the area of intelligent building at first. Four different levels are then proposed according to the usage and service condition of intelligent systems. The paper also analyzes and researches the structure and function of IB. Then five kinds of systems with advanced and innovative intelligent technologies are discussed in detail. Finally, some research trends of IB are identified according to existing research and applications. We hope that the analysis and discussion in this paper will provide a starting point for researchers to construct more effective and useful IB systems.

**Acknowledgment.** This work was supported by the Social Science Planning Project of Chongqing (Grant No.2010QNRW54), the Higher Education Teaching Reform Project of Chongqing (113281) and of Chongqing Technology and Business Institute (10313).

## References

1. Jiang, F.: Energy-saving technology of existing residential building in Chongqing area. *Journal of Central South University of Technology* 14(suppl.3) (2007)
2. Clements-Croome, D., Croome, D.J.: *Intelligent buildings: design, management and operation*. Thomas Telford Services Ltd. (2004)
3. Harrison, A., Loe, E., Read, J.: *Intelligent buildings in south East Asia*. Taylor & Francis Group (1998)
4. Wang, S., Xu, Z., Li, H., Hong, J., Shi, W.: Investigation on intelligent building standard communication protocols and application of IT technologies. *Automation in Construction* 13, 607–619 (2004)

5. Chen, Z., Clements-Croome, D., Hong, J., Li, H., Xu, Q.: A multicriteria lifespan energy efficiency approach to intelligent building assessment. *Energy & Buildings* 38, 393–409 (2006)
6. Ivanovich, M., Gustavson, D.: Future of intelligent buildings is now. *Hpac Heat Piping Air Cond.* 71, 73–79 (1999)
7. Matsumoto, T., Matsumoto, H., Yamada, K., Hoshino, S.: Impact of artificial gummy fingers on fingerprint systems, vol. 4677, pp. 275–289. Citeseer (2002)
8. Ratha, N.K., Bolle, R.: *Automatic fingerprint recognition systems*. Springer (2004)
9. Robinson, L.P.: *Automatic drain system*. Google Patents (2001)
10. Chow, V.T., Maidment, D.R., Mays, L.W.: *Applied hydrology*. McGraw-Hill, New York (1988)
11. Anaissie, E.J., Penzak, S.R., Dignani, M.C.: The hospital water supply as a source of nosocomial infections: a plea for action. *Archives of Internal Medicine* 162, 1483 (2002)
12. Hannemann, R., Marsala, J., Pitasi, M.: Pumped liquid multiphase cooling. ASME paper IMECE2004-60669 (2004)
13. Zutshi, A., Sohal, A.S.: Integrated management system. *Journal of Manufacturing Technology Management* 16, 211–232 (2005)
14. Boman, M., Davidsson, P., Skarmeeas, N., Clark, K., Gustavsson, R.: Energy saving and added customer value in intelligent buildings, pp. 505–517. Citeseer (1998)

# XBeGene: Scalable XML Documents Generator by Example Based on Real Data

Manami Harazaki, Joe Tekli\*, Shohei Yokoyama, Naoki Fukuta, Richard Chbeir, and Hiroshi Ishikawa

**Abstract.** XML datasets of various sizes and properties are needed to evaluate the correctness and efficiency of XML-based algorithms and applications. While several downloadable datasets can be found online, these are predefined by system experts and might not be suitable to evaluate every algorithm. Tools for generating synthetic XML documents underline an alternative solution, promoting flexibility and adaptability in generating synthetic document collections. Nonetheless, the usefulness of existing XML generators remains rather limited due to the restricted levels of expressiveness allowed to users. In this paper, we develop a novel XML By example Generator (XBeGene) for producing synthetic XML data which closely reflect the user's requirements. Inspired by the query-by-example paradigm in information retrieval, Our generator system i) allows the user to provide her own sample XML documents as input, ii) analyzes the structure, occurrence frequencies, and content distributions for each XML element in the user input documents, and iii) produces synthetic XML documents which closely concur, in both structural and content features, to the user's input data. The size of each synthetic document as well as that of the entire document collection are also specified by the user. Clustering experiments demonstrate high correlation levels between the specified user requirements and the characteristics of the

---

Manami Harazaki · Shohei Yokoyama · Naoki Fukuta · Hiroshi Ishikawa  
Department of Computer Science, Faculty of Informatics, Shizuoka University,  
Hamamatsu-shi, Shizuoka, Japan  
e-mail: [gs09046@s.inf.shizuoka.ac.jp](mailto:gs09046@s.inf.shizuoka.ac.jp),  [{yokoyama, fukuta, ishikawa}@inf.shizuoka.ac.jp](mailto:{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp)

Joe Tekli · Richard Chbeir  
LE2I Laboratory CNRS, University of Bourgogne, 21076 Dijon, France  
e-mail:  [{joe.tekli, richard.chbeir}@u-bourgogne.fr](mailto:{joe.tekli,richard.chbeir}@u-bourgogne.fr)

\* The author is supported in part by the Japan Society for the Promotion of Science, JSPS Post-doc Fellowship no: PE10006.

generated XML data, while timing results confirm our approach's scalability to large scale document collections.

## 1 Introduction

Artificial collections of XML documents have many applications in benchmarking, testing, and evaluation of several algorithms, tools, and systems. For instance, methods for XML similarity evaluation [1, 2, 3], version control and change management [4, 5], document classification [6, 7] and clustering [8, 9], all require large XML test collections in order to evaluate their effectiveness and/or efficiency levels. However, even though the need is overwhelming, it is often difficult to find or generate appropriate XML data. This is a significant stumbling block in system development. Furthermore, different applications require different kinds of XML documents with different characteristics.

On one hand, several downloadable XML datasets can be found on the web. Such datasets may be real documents of various types, e.g., *Plays of Shakespeare* [10], SIGMOD Record [11], DBLP [12], or XML benchmarks, e.g., [13, 14, 15]. Nonetheless, these are not always suitable to evaluate each and every application or algorithm. Most importantly, they are predefined by domain experts, and thus are not adaptable to each user's requirements and/or application constraints. On the other hand, a few tools to automatically generate XML data have been recently developed [16, 17, 18]. An XML generator is a program that produces synthetic XML documents according to given user constraints. While they seem more flexible than predefined XML datasets, the usefulness of existing generators remains rather limited due to the restricted levels of expressiveness allowed to users. Following [18], an XML dataset should satisfy three main properties to be useful for effective system testing and evaluation:

- (1) The data structure must conform to that of the target application.
- (2) The datasets should correspond to the expected workload and have expected characteristics. For instance, the user should be able to specify whether an optional element is to appear frequently, or less often, in the synthetic document, or whether its branching factor is large.
- (3) The data values must match the expected data distribution.

XMLGen [16] produces XML documents with random structures and contents. The user may only specify parameters that control the randomness of the result, e.g., the number of levels in the resulting tree, and the minimum and maximum number of children for a given level. In short, it does not seem to comply with any of the properties mentioned above. ToXGene [17] focuses on content distributions, such as constraints are specified locally, within an XML schema. Nonetheless, despite ToXGene's rich definition of the XML content that should be produced, the ability to control the generated structure of the XML data is very limited. Since all constraints are specified



locally in the schema, global structural properties of the document cannot be provided. Hence, ToXGen can generate datasets verifying properties 1 and 3. In [18], the authors introduce GxBE to generate XML document structures that satisfy a given DTD grammar and a sample input XML document. It uses a natural declarative syntax, which includes XPath expressions to allow users to express global and local characteristics for the desired target documents, such as the generated documents can require satisfaction of XPath expressions from a given workload. While GxBE is clearly more sophisticated than its predecessors in considering the structural properties of XML data, it only targets XML structure and does not deal with XML content. In other words, the datasets generated using GxBE verify properties 1 and 2.

**Table 1** Comparison of the existing approaches

	Property1	Property2	Property3	Property4
	Structure	Element/ Attribute characteristics	Element/ Attribute value distribution	XML Grammar
XMLGen [13]	x	x	x	Not required
ToXGene [17]	✓	x	✓	Required
GxBE [18]	✓	✓	x	Required
XBeGene(Our App.)	✓	✓	✓	Not required

Note that the ToXGene [17] and GxBE [18] generators require XML grammars as input to the data generation task, so as to produce synthetic documents. Nonetheless, we assume that input grammars are not always easy to come by, and that a user-friendly generator should provide the user with the possibility of using sample XML documents, necessary for the usefulness of automatic XML generators.

In this paper, we propose an XML By example Generator (XBeGene) capable of producing synthetic XML data which closely reflect the structural characteristics and content distributions required by the user. Our system: i) allows the user to provide her own sample XML documents as input, ii) analyzes the structure (i.e., parent/child relations, paths, hierarchical levels and ordering), occurrence frequencies (occurrence probability of each tag name), and content distributions (allowed values and their frequencies) for each XML element/attribute in the user input documents, and iii) produces synthetic XML documents which closely concur, in both structural and content features, to the user input data. The size of each synthetic document, as well as the size of the entire document collection are also specified by the user. In other words, our method provides the user with full control over the different aspects of the XML data generation process. In addition, our approach is designed to efficiently generate large collections of documents, and runs in average polynomial time.

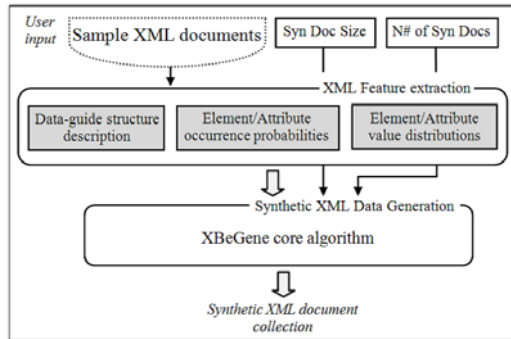


Fig. 1 Simplified activity diagram describing *XBeGene*.

## 2 Our XML Data Generator

We introduce an XML by example data generator, *XBeGene* that i) fulfills the *dataset usefulness* properties [18] discussed in the previous section, and ii) allows the user to provide her own sample XML documents as input to the generation process. *XBeGene* consists of two phases: i) feature extraction from real XML data, and ii) synthetic XML documents generation. The feature descriptions are consequently exploited as input to the XML data generator component, along with the number of synthetic documents and the maximum document size specified by the user. A simplified activity diagram describing the *XBeGene* prototype is depicted in Figure 1.

Technical details of *XBeGene* are described in Harazaki et. al. [22]

### 2.1 XML Feature Extraction

To generate synthetic XML data that closely resemble a real XML document collection, we start by extracting features that characterize the document collection, such as the XML structure, element/attribute occurrence probabilities and element/attribute value distributions.

#### 2.1.1 Data Structure

Efficient information extraction is difficult if the entire structure of the data cannot be understood [19]. First, the whole structure must be extracted from real data. To obtain a simple and useful description of the data structure of an XML document collection, we use data-guides [19], dynamically generated and maintained structural summaries for semi-structured databases.

The data-guide structure has been originally developed for the OEM data model [19]. Yet, transferring the notion of data-guide to the XML data model

is straightforward. A *data-guide*  $DG_C$  for a collection  $C$  of XML documents is a labeled directed graph structure fulfilling the following conditions:

- Each path in  $C$  exists in  $DG_C$ ,
- Each path in  $DG_C$  exists in  $C$ ,
- Paths in  $DG_C$  are unique (which is not the case in  $C$ ).

In case a common root for the document collection  $C$  does not exist, a virtual root node is added in constructing  $DG_C$ . Every leaf node of the tree is annotated with the IDs of the document leaves it represents. Note that a data-guide  $DG_C$  does not reflect the node ordering relations in the XML documents in  $C$ .

While data-guides are less expressive than XML grammars, they can be extracted and handled much more efficiently. Existing approaches to automatically generate an XML grammar from a collection of XML documents are typically NP-Complete [20], which is not practical in many applications, particularly in the context of our study which requires the fast processing of large collections of documents. Instead, we adopt XML data-guide extraction which is reasonably more efficient, i.e., of polynomial complexity, and thus better suited for the task at hand.

Our algorithm takes as input a collection  $C$  of parsed XML documents, and generates the corresponding dataguide structure  $DG_C$ . For every XML document, the algorithm goes through every element, and verifies whether the corresponding start tag event already exists in the main dataguide HashTable, otherwise it is added. Consequently, it verifies whether each outgoing edge from the current element node has already been accounted for in the data-guide, so as to otherwise add it.

### 2.1.2 Elements and Attributes Occurrence Probabilities

The occurrence rates of XML elements and attributes are typically biased in real XML document collections. This is a central piece of information required to better mirror the properties of real-XML data in synthetically generated documents. To reflect the bias in the occurrence count of each element/attribute in the generated data, we must obtain not only the number of distinct element/attribute tag names but also their occurrences probabilities, which can only be understood from analyzing the real data collection at hand.

In our approach, we estimate and record the occurrence probabilities of each distinct element and attribute in the real XML data collection  $C$ , and append probability scores to the corresponding element/attribute representatives in data-guide  $DG_C$  in order to provide a unified descriptive structure for the whole dataset. The weighted data-guide is thereafter denoted  $\overline{DG_C}$ . Note that attributes are processed as children of their encompassing nodes, their occurrence probabilities being estimated accordingly.

### 2.1.3 Elements and Attributes Value Distributions

In addition to mirroring the structural properties of the original XML documents and element/attribute occurrence probabilities, we expect our generated XML data to reflect the contents of the real data. In existing XML generators, users are required to specify element data-types, which are then exploited by the system to generate the values accordingly. Instead, in order to reflect real XML contents, we propose to select data values based on value lists extracted from the input documents, corresponding to each element in the real data collection. In other words, synthetic element values are defined by the vocabularies corresponding to their peers in real documents, such as incorrect values will not appear in the generated data.

We introduce a space-saving algorithm [21] to identify the top- $k$  element values in a data stream. It allows the user to specify the  $k$  parameter, and consequently identifies the most frequent values within the top- $k$  range. A number  $k$  of pairs of values and counters, i.e., (*value*, *count*), are stored, and initialized by the first  $k$  distinct values and their exact counts. Value counts are consequently incremented following their occurrences in the considered element/attribute value sequences.

## 2.2 Synthetic XML Document Generation

We use the information obtained through analysis of real XML data (feature extraction phase) for synthetic XML documents generation. Having evaluated the various structural and content properties of the input XML data, we consequently exploit these information for generating synthetic XML documents. Given a real data collection, the input to the data generation process consist of two main XML-based files: one to describe the structure and occurrence probabilities of XML elements/attributes, i.e., the weighted data-guide  $\overline{DG}_C$  corresponding to document collection  $C$ , the other to describe element/attribute value distributions. Furthermore, the user can choose between two methods for specifying data size: the total number of XML elements/attributes, or the number of data bytes in the resulting file. Our generator outputs synthetic XML documents that mirror the various structural and content features of real input documents.

Our algorithm for XML data generation, entitled XBeGene selects elements such as their occurrence probabilities that are most similar to those of real data. It compares the occurrence probability of each element in the generated XML dataset to that of its peer in the real data collection. Elements with the highest probability discrepancies, are gradually selected and added to the generated dataset. Note that the number of added elements is adjusted to maintain, not only the occurrence probability the added element itself, but also the probability of occurrence of its corresponding parent element. This process is repeated until the generated dataset reaches the specified data size.

### 2.3 Complexity Analysis

Let  $C$  be an sample XML document collection,  $N$  be the total number of nodes in  $C$ ,  $E$  the number of elements/attributes of distinct tag names in  $C$ ,  $depth$  the maximum hierarchical depth of synthetic documents, and  $N_{Syn}$  the total number of nodes in the synthetic documents. The overall complexity of our XML synthetic document generator simplifies to  $O(N \times E \times N_{Syn} \times xdepth)$ .

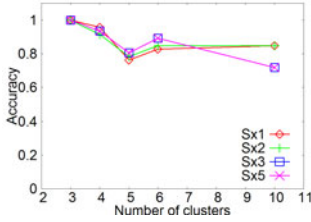
- The computational complexity of the XML\_DG\_Extract algorithm is of  $O(N)$  time. The  $\overline{DG_C}$  can be obtained within one scan of the XML documents in collection  $C$ .
- The computational complexity of the *Top-K\_SS* algorithm is of  $O(N)$  time. The value distribution can also be computed within one scan of the XML documents in collection  $C$ .
- The computational complexity of the *XBeGene* algorithm is of  $O(E \times N_{Syn} \times depth)$  time. The algorithm depends on the size of the documents to be created. In addition, dedicated processing is required to consider the many types of elements and attributes, so as to determine the synthetic nodes to be added. Moreover, the depth factor linearly increases the number of node processing loops in the data generation process.

## 3 Experimental Evaluation

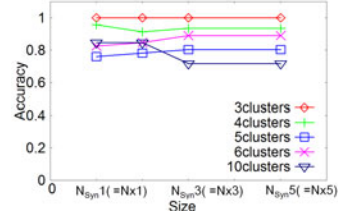
In this section, we conducted various experiments to test and validate our approach, in comparison with one of its most sophisticated alternatives: ToXGene [17]. First, we verify the system’s accuracy in generating synthetic documents which closely mirror the characteristics of real data. To do so, we estimate the similarity between real and synthetic XML documents using XML document clustering. Second, we evaluate the correlation in element occurrence probabilities and value distributions between the generated and real data, in order to verify the similarity between each pair of matching elements. Third, we assess the overall throughput of the system, considering different sizes of generated documents, in order to evaluate execution time. Our tests are based on the bibliography of SIGMOD Record [11], DBLP [12] and XMLGen [16]. Experimental results confirm our algorithm’s effectiveness and efficiency in generating synthetic XML documents of arbitrary sizes, in comparison with ToXGene, while preserving the structural and content features of user data.

### 3.1 Accuracy

Our first experiment measures the similarity between real XML data and the generated synthetic data, in order to evaluate XBeGene’s efficiency in mirroring the characteristics of user input XML documents. We provide as input



**Fig. 2** Accuracy of clustering (x-axis=number of clusters).

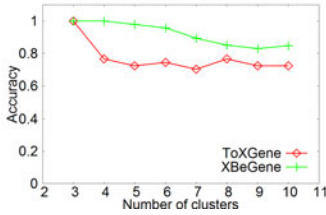


**Fig. 3** Accuracy of clustering (x-axis=size).

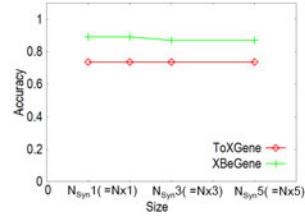
a number of heterogeneous documents, generate corresponding output documents, and consequently compare both input and output document sets. We exploit document clustering to assess the similarity between input and output document sets. The main idea is to verify whether the clustering result of input documents correlates with the clustering result of output documents. In our experiment, we utilize an agglomerative hierarchical clustering algorithm, built upon an XML path similarity model to evaluate the similarity between XML documents [23]. The approach developed in [23] provides an improved path similarity approach taking into account the main properties of XML documents in the comparison process, while preserving the same complexity levels as existing path-based methods. Similarity between real and synthetic XML data is evaluated based on clustering accuracy:

$$accuracy = \frac{\text{number of documents correctly clustered}}{\text{total number of documents}}$$

In other words, the closer input and output clusters are, the higher the similarity between real and generated XML documents, and thus the higher the accuracy of the XML data generation process in preserving the characteristics of the user input XML data. We utilize as input data real XML documents from the SIGMOD Record [11], and compare our system's results to those generated by ToXGene [17]. First, we perform clustering on the real document collection and identify the input clusters  $C_1; C_2; \dots; C_n$ . Second, we perform clustering on the synthetic documents generated by XBeGene (based on the SIGMOD collection) and by ToXGene, and identify the output clusters  $S_1; S_2; \dots; S_n$ . The size of the real document collection, i.e., sum of the sizes of all input documents, is denoted  $N$ , such as the size of the generated data varies w.r.t.  $N$ ;  $N_{Syn1} = 1 \times N$ ,  $N_{Syn2} = 2 \times N$ ,  $N_{Syn3} = 3 \times N$ ,  $N_{Syn5} = 5 \times N$ . Consequently, we verify how closely clusters  $C_i$  correspond to clusters  $S_j$ , correlating clustering results by evaluating accuracy. Fig. 2 and Fig. 3 depict the accuracy of XBeGene w.r.t. the number of clusters



**Fig. 4** Comparing average accuracy between XBeGene and ToXGene (x-axis=number of clusters).



**Fig. 5** Comparing average accuracy between XBeGene and ToXGene (x-axis=size).

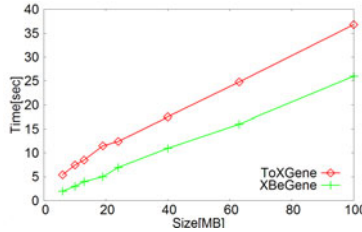
and document collection size respectively. Fig. 2 depicts clustering accuracy (y-axis) w.r.t. the number of real clusters (x-axis). For instance, when the number of real clusters is equal to 3, synthetic documents are distinguished by 3 different XML grammars corresponding to 3 real XML data sets. Experimental results show that accuracy levels remain almost steady (varying from 0.7 to 1.0) when increasing the number of real clusters. Similarly, accuracy remains steady when varying the size of the synthetic document collection. Therefore, results show that XBeGene generates synthetic XML documents of varied sizes which are highly similar to the input XML documents. In addition, Fig. 4 depicts the average accuracy levels attained using our method and ToXGene. Fig. 5 depicts clustering average accuracy (y-axis) of our method and ToXGene. Our approach steadily outperforms its predecessor, regardless of the number of clusters considered and document collection size. Recall that ToXGene underlines limited capabilities in controlling the structure of the generated XML data, which explains the poor accuracy results.

### 3.2 Occurrence Probabilities and Value Distributions

In addition, we compare element occurrence probabilities and value distributions between synthetic and real XML documents in order to verify the similarity between each pair of matching elements. We generate synthetic documents of different sizes (60MB, 270MB) based on the XMark document (115MB) and calculate the respective element occurrence probabilities for synthetic documents. In same way, we generate a synthetic document using ToXGene and calculate the each element occurrence probabilities. Then, we compare the element occurrence probabilities of XMark, ToXGene and XBeGene. Table 2 contrasts the average scores in the XMark document, and the synthetic documents collection generated based on XMark by using ToXGene and XBeGene. Results in Table 2 show high correlation levels for all elements in XMark and the documents generated using XBeGene, regardless of the synthetic document collection size. While ToXGene provides advanced settings to capture element contents, however it provides limited control over

**Table 2** Average occurrence probabilities for each element in the data collection[%]

element	XMark 115MB	XBeGene 60MB	XBeGene 270MB	ToXGene 60MB
/asia/item/mailbox/mail	60.250	60.250	60.500	45.000
/australia/item/mailbox/mail	60.455	60.455	58.778	49.545
/watches/watch	88.350	88.350	88.350	78.500
/open_auction/bidder	90.300	90.300	100.000	82.670
/categories/description/text/bold	5.400	6.000	5.9000	36.240
/close_auction/description/emph	5.700	5.700	5.700	30.000

**Fig. 6** Execution time.

the XML data structure, which explains the lower correlation levels in Table 2. In other words, results concur and confirm the high accuracy levels obtained in our clustering experiments described in the previous paragraph.

### 3.3 Throughput

In addition to validating the effectiveness of XBeGene in generating documents which mirror the characteristics of real data, we assess XBeGene’s efficiency levels, and assess its execution time.

Our method is of  $O(N \times E \times N_{Syn} \times depth)$ . We verified our method’s linear dependency on each of the latter mentioned complexity parameters. Hereunder, we present timing results corresponding to  $N_{Syn}$ , i.e., size of the generated document collection, as the most significant parameter in practical applications. We generate synthetic documents of different sizes (6MB–100MB) based on the DBLP document collection, and store the average time elapsed to generate each synthetic document set. In the same way, we generate documents of different sizes using ToXGene, and measure the average time to generate documents. Timing results are reported in Table 3 and depicted in Fig. 6. Results show that XBeGene’s execution time is lower, on average, than ToXGene and grows almost linearly w.r.t. synthetic document size. In short, results confirm our approach’s scalability in generating large document collections.



**Table 3** Size and generation time of synthetic data

size[MB]	XBeGene		ToXGene	
	time[sec]	throughput [MB/sec]	time[sec]	throughput [MB/sec]
6	2.0	3.000	5.43	1.105
10	3.0	3.333	7.5	1.333
20	5.0	4.000	11.5	1.740
40	11.0	3.636	17.5	2.286
100	26.0	3.846	36.8	2.717

## 4 Conclusions

In this paper, we introduced an XML-by-example generator for producing synthetic XML documents which closely satisfy the structural characteristics and content distributions of user-defined input data. Our method uses weighted data-guides as an efficient means to describe the structural features of the input document collection, and a dedicated space-saving algorithm for assessing element/attribute value distributions. Our theoretical study as well as our experimental evaluation showed that our approach is efficient and scalable in generating XML documents of arbitrary sizes which mirror the structural and content features of user provided real document collections.

**Acknowledgements.** The XML clustering method is provided by Cheikh Bedy Mohamed (Graduate School of Informatics, Shizuoka University, Japan). This research was partially supported by Scientific Research(B)(19300026), and *Japan Society for the Promotion of Science (JSPS)* 2010 research fellowship n.PE10006.

## References

- [1] Tekli, J., Chbeir, R., Yétongnon, K.: A hybrid approach for XML similarity. In: van Leeuwen, J., Italiano, G.F., van der Hoek, W., Meinel, C., Sack, H., Plášil, F. (eds.) SOFSEM 2007. LNCS, vol. 4362, pp. 783–795. Springer, Heidelberg (2007)
- [2] Tekli, J., Chbeir, R., Yetongnon, K.: Extensible User-Based XML Grammar Matching. In: Laender, A.H.F., Castano, S., Dayal, U., Casati, F., de Oliveira, J.P.M. (eds.) ER 2009. LNCS, vol. 5829, pp. 294–314. Springer, Heidelberg (2009)
- [3] Helmer, S.: Measuring the Structural Similarity of Semistructured Documents Using Entropy. In: Proceedings of the International Conference on Very Large Databases (2007)
- [4] Cobena, G., Abiteboul, S., Marian, A.: Xydiff, tools for detecting changes in XML documents (2001), <http://www.rocq.inria.fr/?cobena/XyDiffWeb/>
- [5] Cobena, G., Abiteboul, S., Marian, A.: Detecting Changes in XML Documents. In: Proceedings of the IEEE International Conference on Data Engineering, ICDE (2002)
- [6] Bertino, E., Guerrini, G., Mesiti, M.: A Matching Algorithm for Measuring the Structural Similarity between an XML Documents and a DTD and its Applications. Elsevier Computer Science (29), 23–46 (2004)

- [7] Candillier, L., Tellier, I., Torre, F.: Transforming XML Trees for Efficient Classification and Clustering. In: Proceedings of the Workshop of the Initiative for the Evaluation of XML Retrieval (INEX), pp. 469–480 (2005)
- [8] Dalamagas, T., Cheng, T., Winkel, K., Sellis, T.: A Methodology for Clustering XML Documents by Structure. *Information Systems* 31(3), 187–228 (2006)
- [9] Nierman, A., Jagadish, H.V.: Evaluating structural similarity in XML documents. In: Proceedings of the ACM SIGMOD International Workshop on the Web and Databases (WebDB), pp. 61–66 (2002)
- [10] Bosak, J.: The Plays of Shakespeare in XML (1999), <http://xml.coverpages.org/bosakShakespeare200.html>
- [11] SIGMOD Record Document Collection, <http://www.sigmod.org/record/xml/>
- [12] The DBLP Computer Science Bibliography, <http://www.informatik.uni-trier.de/ley/db/>
- [13] Schmidt, A., Waas, F., Kersten, M., Carey, M., Manolescu, I., Busse, R.: XMark: a Benchmark for XML Data Management. In: Proceedings of the International Conference on Very Large Data Bases (VLDB), pp. 974–985 (2002)
- [14] Yao, B., Ozsu, M., Khandelwal, N.: XBench: Benchmark and Performance Testing of XML DBMSs. In: Proceedings of the International Conference on Data Engineering (2004)
- [15] Runapongsa, K., Patel, J., Jagadish, H., Chen, Y., Al-Khalifa, S.: The Michigan Benchmark: towards XML Query Performance Diagnostics. *Information Systems* (2006)
- [16] Aboulnaga, A., Naughton, J., Zhang, C.: Generating Synthetic Complex-Structured XML Data. In: Proceedings of the ACM SIGMOD International Workshop on the Web and Databases (WebDB), pp. 79–84 (2001)
- [17] Barbosa, D., Mendelzon, A., Keenleyside, J., Lyons, K.: ToXgene: an Extensible Template-based Data Generator for XML. In: Proceedings of the ACM SIGMOD International Workshop on the Web and Databases (WebDB), pp. 49–54 (2002)
- [18] Cohen, S.: Generating XML Structure Using Examples and Constraints. In: Proceedings of the Very Large Data Bases Endowment (PVLBD), vol. 1(1), pp. 490–501 (2008)
- [19] Goldman, R., Widom, J.: data-guides: Query Formulation and Optimization in Semistructured Databases. In: Proceedings of the International Conference on Very Large Databases (VLDB), pp. 436–445 (1997)
- [20] Garofalakis, M., Gionis, A., Rastogi, R., Seshadri, S., Shim, K.: Xtract: A system for extracting document type descriptors from XML documents. In: Proceedings of the ACM International Conference on Management of Data (SIGMOD), Dallas, Texas, USA (2000)
- [21] Metwally, A., Agrawal, D., El Abbadi, A.: Efficient computation of frequent and top-k elements in data streams. In: Eiter, T., Libkin, L. (eds.) ICDT 2005. LNCS, vol. 3363, pp. 398–412. Springer, Heidelberg (2005)
- [22] Harazaki, M., Tekli, J., Yokoyama, S., Fukuta, N., Chbeir, R., Ishikawa, H.: Technical Report of the XBeGene: Scalable XML Documents Generator, [http://db-lab.cs.inf.shizuoka.ac.jp/paper/tech\\_xbegene.pdf](http://db-lab.cs.inf.shizuoka.ac.jp/paper/tech_xbegene.pdf)
- [23] Mohamed, C.B., Yokoyama, S., Fukuta, N., Ishikawa, H., Chbeir, R.: New Approach for Computing Structural Similarity between XML Documents. Master's thesis, Shizuoka University, Japan (2010)

# A Preliminary Activity Recognition of WSN Data on Ubiquitous Health Care for Physical Therapy

S.-Y. Chiang, Y.-C. Kan, Y.-C. Tu and H.-C. Lin\*

**Abstract.** The physical therapy with ubiquitous health care (UHC) for geriatrics training or stroke patients requires continuous and routine rehabilitation during the cure period. The physiatrists are hereby the feedback clinical record to design necessary assistant programs. The successful treatment usually concerns whether the patients follow the therapeutic assignment without interruption. This study hence developed a set of wireless sensor network (WSN) devices including the accelerometer and gyroscope to measure the essential movement of human body. At this initial stage, the sensor data of static and dynamic postures for lying, sitting, standing, walking, and running were calibrated by the fuzzy algorithm with an overall accuracy rate at best to 99.33%. The approach may support for monitoring patient's remedy process at home for ubiquitous health care of physical therapy.

## 1 Introduction

The motion detection is one of the most important issues of the health care system, particular for monitoring daily activity. Based on the physical therapy and

---

S.-Y. Chiang · Y.-C. Tu

Department of Information and Telecommunications Engineering,  
Ming Chuan University, Gui-Shan, Taoyuan 333, Taiwan  
e-mail: sychiang@mail.mcu.edu.tw

Y.-C. Kan

Department of Communications Engineering, Yuan Ze University, Chung-Li,  
Tao-Yuan 32003, Taiwan  
e-mail: yckan@saturn.yzu.edu.tw

H.-C. Lin

Department of Health Risk Management, China Medical University, Taichung 404, Taiwan  
e-mail: snowlin@mail.cmu.edu.tw

\* Corresponding author.

medicine, the stroke patients usually need to continuously and routinely repeat specific motions during the remedy period of the rehabilitation treatment. If the physiatrists can acquire the rehabilitation records of patients during the cure process through the ubiquitous health care (or u-healthcare, UHC) with ambulatory measurement, it is helpful to design necessary assistant programs. The activity recognition hence can be a potential technique to achieve this scope.

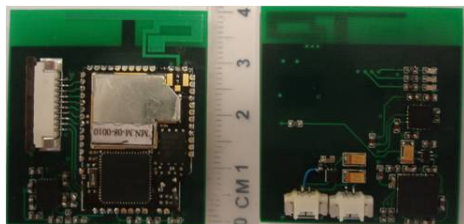
The regular postures of human beings essentially include lying, sitting, standing, and so on. According to different circumstance, these activities can be caught and recognized by appropriate algorithms [1]. In general, the methods of image [2] and non-image [3][4] processing are the most popular categories to monitor the postures and motions of human body. However, as considering the privacy of people, the cost of equipments, and convenience in an open environment, it is more reasonable to adopt the modern technology for the UHC. The concept of UHC is extended from the homecare personal area network that provides intelligent monitoring with ad hoc network. With the multi-tier telemedicine system due to the wireless body area network, the real-time analysis of sensors' data was performed by computer-assisted rehabilitation. Thus the ambulatory monitoring can help warnings on user's state, level of activity, and environmental conditions [5].

In the past decade, the wireless sensor network (WSN) has been widely spread out for different types of healthcare. It detects various vital signals such as blood pressure, impulse, motion as well as variation of home environment [6][7]. Many studies approved this technique on a variety of home healthcare systems since it provides wearable, portable, and mobile functionalities for usage to ubiquitously combine patients' healthcare information with clinical data in hospital [8]. For this approach to the physiotherapy and rehabilitation, it is helpful to measure patients' motions by creating a pattern library of WSN data for activity recognition.

In this study, we developed the wearable sensor by integrating the accelerometer and the gyroscope with the WSN devices to detect the signals of body movements. The preliminary measurement was designed by locating the sensors on the chest and the left thigh to transport data. The fuzzy algorithm was applied for advanced data recognition.

## 2 WSN Measurements

One of the primary tasks of the u-healthcare for physical therapy is assisting physiatrists to record patients' rehabilitation data continuously for tracking their essential motions during the cure period. It usually includes lying, sitting, standing, walking, and running, which imply the spatial movement of specific body part. Herein, we implemented two sets of WSN sensors for measuring desired body motions. The proposed WSN sensors are designed by remodeling MEMS modules of



**Fig. 1** (a) WSN mote, (b) Accelerometer and gyroscope modules.

the accelerometer and gyroscope with the WSN mote and antenna to transport the signals.

The practical devices include (1) a battery-powered mote shown in Fig 1(a), which supports embedded micro control unit (MCU) and radio frequency (RF) for processing and delivering the sensed signals; (2) a triaxial accelerometer and a bi-axial gyroscope shown in Fig 1(b), which can detect kinetic acceleration and angular velocity, respectively; and, in addition, a flash memory card which can temporarily store the sensed data. The sensor A and B are fixed at the chest and the left thigh, respectively, in which the forward or backward direction is defined as z axis while the up-down and right-left directions represent x and y axes, respectively. Herein, several features are extracted for each of the six data streams (acceleration and angular speed along x, y, and z axes).

## 2.1 Feature Extraction

- (1) Acceleration: denotes  $g_x$ ,  $g_y$ , and  $g_z$  in gravity of x-, y-, and z-axis.
- (2) Angular velocity: denotes  $\omega_x$ ,  $\omega_y$ , and  $\omega_z$  along x-, y-, and z-axis, respectively.
- (3) Tilt angle: represents the angle between x and z components of the acceleration vector ( $g_x$ ,  $g_y$ ,  $g_z$ ) in gravity can be calculated by  $\theta = \tan^{-1}(g_z/g_x)(180/\pi)$ .
- (4) Mean value: determines the average value ( $\mu$ ) of the number (n) of acceleration components ( $x_i$ ) in gravity during one second.
- (5) Standard Deviation: performs the probability density function (PDF) of probability versus ADC count of the sensed data. The data distribution returns random variables since a range of output voltage might be mapped to the same gravity component as a more-to-one relationship.

Therefore, this feature includes  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_z$  that represent the standard deviation of acceleration in gravity x, y, and z-axis, respectively.

## 2.2 Evaluation of Acceleration

The WSN data of acceleration components sent by the sensors must be calibrated before the device is worn for measuring individual body posture. We hence evaluated these components due to lying, sitting, standing, walking, and running through an initial test.

## 2.3 Application of Fuzzy Algorithm

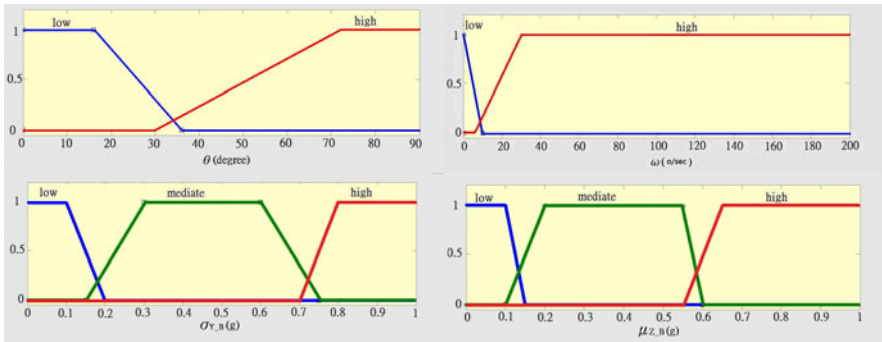
In this study, we apply the fuzzy algorithm to calibrate measured data. During the procedure, the acceleration per second was firstly classified into fuzzy sets and became the input feature of fuzzy system. Then, after fuzzification, we induced the fuzzy rules with the activity recognition pattern library. Consequently, with defuzzification, we could obtain the output of static postures or dynamic motions and store in the database.

### 3 WSN Data Calibration and Activity Recognition

According to the sensed data, we extract the reasonable input features like  $\mu_{z\_B}$ ,  $\sigma_{y\_B}$ ,  $\theta_A$ ,  $\theta_B$  and  $\omega_{x\_B}$  as well as the desired output postures (e.g., lying, sitting and standing) and motions (e.g., walking and running).

#### 3.1 Fuzzy Rules for Calibration

1)  $\theta_A$ : The feature  $\theta_A$  is defined by the angle between x and z axes and can be detected by the sensor A. Each set of the sensed  $\theta_A$  data was collected per second to yield Gaussian distribution in which the peak value can be adopted as the input features of the fuzzy system. As shown in Fig 2a, we define two membership functions ranged from  $0^\circ$  to  $36^\circ$  as the low-angle and above  $36^\circ$  as the high-angle.



**Fig. 2** Membership function of (a) tilt angle (left-up), (b) angular velocity (right-up), (c) standard variance of the probability density function due to measured acceleration components in gravity (left-down), (d) mean value of measured acceleration components in gravity (right-down).

2)  $\theta_B$ : The feature  $\theta_B$  can be detected by the sensor B and follows the same data processing as  $\theta_A$ . We can regulate the variation between  $\theta_A$  and  $\theta_B$  to determine the relative position of the chest and the left thigh and to recognize the postures of lying, sitting, and standing. For example, if the initial angles of  $\theta_A$  and  $\theta_B$  are identical to 0 as “standing,” then the condition of ( $\theta_A=0^\circ$  and  $\theta_B=90^\circ$ ) will be “sitting”.

3)  $\omega_{x\_B}$ : The feature  $\omega_{x\_B}$  is provided by the gyroscope of the sensor B to judge the static posture or the dynamic motion. According to the practical test, the angular velocity was approaching 0 as steady status, but its absolute value would pass a level ( $> 10^\circ/\text{sec}$  by referring Fig 2b) for the apparent motions. Hence, we can calibrate this rule with fuzzy variables as “the thigh is starting to move”.

4)  $\sigma_{y\_B}$ : The dynamic motions such as walking and running can be calibrated by the standard deviation of sway conditions along the y direction due to the sensor B. The membership function of  $\sigma_{y\_B}$  is shown in Fig 2c. The motion of changing

posture is defined as “in moving” since it could be either “walking” or “running”. We use  $\sigma_{Y_B}$  of gy data distribution as the level “in moving”. Herein, the motion is “in moving,” “walking,” or “running” when  $\sigma_{Y_B}$  is less than 0.2g, between 0.2g and 0.6g, or greater than 0.6g, respectively.

5)  $\mu_{Z_B}$ : Besides of  $\sigma_{Y_B}$ , we can also consider  $\mu_{Z_B}$ , which is the mean value of  $g_z$  in absolute integer value, to categorize different types of activity recognition. The membership function of  $\mu_{Z_B}$  is shown in Fig 2d that a variety of ranges present movement levels.

The input features above can calibrate sensed WSN data through the fuzzy system for the output movement. Fig 3 illustrates the membership function of output.

### 3.2 Fuzzification and Defuzzification

The first step of fuzzification is finding the membership function. In this test, as shown on Fig 2a, the  $\theta_A$  and  $\theta_B$  can be considered as the low angle if they are in the range between  $0^\circ$  and  $36^\circ$ ; otherwise they are classified as the high angle that is grater than  $36^\circ$ . Similarly, the membership function shown in Fig 2b implies a cut for  $\omega_{X_B}$  to detect the motion is starting. Furthermore, for “walking” and “running”, Fig 2c and 2d provide the required membership functions for fuzzification of  $\sigma_{Y_B}$  and  $\mu_{Z_B}$  to find the criteria. At once the membership functions of input and output features are obtained, the fuzzy rules can be calibrated for the activity recognition pattern. For defuzzification, the MATLAB™ toolbox was applied with the discretization technique to carry out the solution.

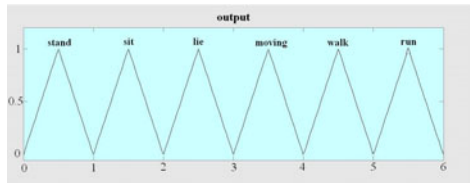


Fig. 3 Output of activity recognition pattern after defuzzification

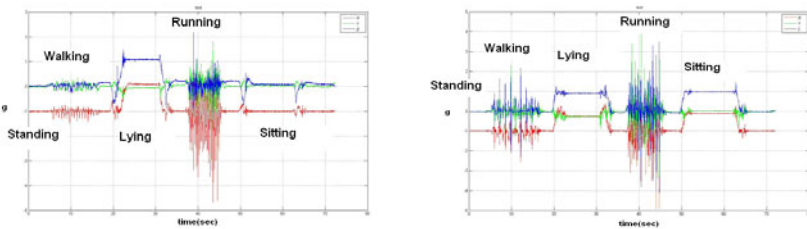


Fig. 4 Measured acceleration components by sensor A (left) and sensor B (right).

## 4 Results

In this study, we recruited three volunteers who wore the developed WSN sensor A and B on the chests and thighs, respectively, to collect sensed data. Then, the movements of static postures and dynamic motions were calibrated through fuzzy algorithm to yield the pattern library of activity recognition. Fig 4 shows the data distribution of acceleration components corresponding to movements sensed by the sensor A and B. The input features including  $\omega_{X\_B}$ ,  $\theta_A$ ,  $\theta_B$ ,  $\sigma_{Y\_B}$ , and  $\mu_{Z\_B}$  for the fuzzy system were extracted and the output movement for lying, sitting, standing, in moving, walking, and running were finally obtained.

**Table 1** The successful rate of activity recognition test

Sample	Standing	Sitting	Lying	In Moving	Walking	Running	Average
1	100	100	100	100	100	100	100
2	100	100	100	94	94	100	98
3	100	100	100	100	100	100	100
<b>Average</b>	100	100	100	98	98	100	99.33

By comparing the calibrated data and real motions, we found the good recognition rate as results show in Table 1. In the test, most of test movements were recognized with the best successful rates approaching 99.33% in average while that of the static postures and dynamic motions were approximately 100% and 98.67%, respectively. These results approved accuracy and stability of the developed devices in this study.

## 5 Conclusion Remarks

In this study, the wearable WSN sensor was integrated with the accelerometer and gyroscope to detect the signals of body movements including static postures and dynamic motions. The sensed data could be transported to the backend server via WSN by wearing the sensors on the chest and the left thigh. The fuzzy algorithm with the qualified features was applied for calibrating these data. With processes of fuzzification and defuzzification, the activity recognition pattern was created to obtain output movements. The proposed algorithm contains more flexible rules as recognizing the sensed data by comparing with traditional threshold criteria. The results performed the successful recognition rate at best to 99.33% in average and this preliminary study possessed the potential in advanced application on the ubiquitous health care for physical therapy.

**Acknowledgments.** This work is supported under Grant CMU 97-181, NSC 98-2221-E-130-016, NSC 99-2221-E-130-022, and NSC 100-2625-M-039 -001.



## References

- [1] Medjahed, H., Istrate, D., Boudy, J., Dorizzi, B.: Human Activities of Daily Living Recognition Using Fuzzy Logic for Elderly Home Monitoring. In: Proc. IEEE FUZZ, pp. 2001–2006 (2009)
- [2] Wren, C.R., Azarbajegani, A., Darrell, T., Pentland, A.P.: Real-Time Tracking of the Human Body. *IEEE Trans. Patt. Analys. Mach. Intellig.* 19(7), 780–785 (1997)
- [3] Yeoh, W.-S., Pek, I., Yong, Y.-H., Chen, X., Waluyo, A.B.: Ambulatory Monitoring of Human Posture and Walking Speed Using Wearable Accelerometer Sensors. In: Proc. IEEE on Eng. in Med. Biology Society, Canada, pp. 5184–5187 (2008)
- [4] Yeoh, W.-S., Wu, J.-K., Pek, I., Yong, Y.-H., Chen, X., Waluyo, A.B.: Real-Time Tracking of Flexion Angle by Using Wearable Accelerometer Sensors. In: Proc. 5th Int'l Workshop on Wearable and Implantable Body Sensor Networks, China, pp. 125–128 (2008)
- [5] Jovanov, E., Milenkovic, A., Otto, C., de Groen, P.C.: A wireless body area network of intelligent motion sensors for computer assisted physical rehabilitation. *J. NeuroEngineering Rehab.* 2(6) (2005); doi:10.1186/1743-0003-2-6
- [6] Virone, G., et al.: An Assisted Living Oriented Information System Based on a Residential Wireless Sensor Network. In: Proc. 1st Transdisciplinary Conf. Dist. Diagn., Home Healthcare, pp. 95–100 (2006)
- [7] Andréu, J., Viúdez, J., Holgado, J.A.: An Ambient Assisted-Living Architecture Based on Wireless Sensor Networks. In: Proc. 3rd Symp. Ubiqu. Comp. Ambient Intellig., vol. 51, pp. 239–248.
- [8] Steele, R., Secombe, C., Brookes, W.: Using Wireless Sensor Networks for Aged Care: The Patient's Perspective. In: Proc. Pervas. Health Conf. and Workshops, pp. 1–10 (2006)

# Development and Evaluation of Question Templates for Text Mining

Norio Ishii, Yuri Suzuki, Takashi Fujii, and Hironobu Fujiyoshi

**Abstract.** In this research, we developed a question template that makes it possible to efficiently gather text data for analysis, as the first step in building an environment that will enable persons who are inexperienced in text mining to analyze qualitative data easily. We created question templates using lists, conjunctive, and parts of speech formats. We conducted evaluation tests targeting two class questionnaires, and confirmed the effectiveness of those questionnaires.

## 1 Introduction

Data can be categorized into quantitative data, which comprises mainly figures, and qualitative data, which comprises free answers and other forms of prose. Testing methods have been established for analyzing quantitative data, and data mining technologies are being used with increasing frequency. The analysis of qualitative data, however, such as text mining of free text, is often criticized for its lack of scientific foundation. Nevertheless, in the case of qualitative data, it is possible to analyze aspects of emotion and sensation that cannot be analyzed easily in the case of quantitative data [1][3].

The goal of this research is to create a template and manual that will enable anyone to execute text mining easily, and to create an environment that will enable even those unfamiliar with text mining to easily analyze quantitative data. The first step in this process is to examine question formats and methods for

---

Norio Ishii

Aichi Kiwami College of Nursing, Jogan-dori 5-4-1, Ichinomiya, Aichi, 491-0063 Japan  
e-mail: n.ishii.t@aichi-kiwami.ac.jp

Yuri Suzuki · Takashi Fujii · Hironobu Fujiyoshi

College of Engineering, Chubu University, Matsumoto-cho 1200, Kasugai, Aichi,  
487-8501 Japan  
e-mail: {yuris, fujii, hf}@cs.chubu.ac.jp

effectively gathering the text data that will be subject to analysis. We will use the two experiments outlined below to examine the differences in response results derived from the various question formats.

## 2 Experiment 1

### 2.1 Method

The purpose of this experiment is to examine optimum question formats in a classroom questionnaire where subjects are expected to use mainly nouns in their responses.

The classroom questionnaires were conducted in the “Operations Research” class that is part of the curriculum in the Chubu University College of Engineering Department of Computer Science. The questionnaires were conducted in seven out of the 15 classes that comprised the entire course. The themes of the seven classes were AHP, Prediction, the Simplex Method, Transport Problems, Inventory Management, economic calculations, and waiting queues.

In this experiment, in addition to questionnaires using a free text format, we created question templates using lists, conjunctive, and parts of speech formats; for example: lists, in which the subjects respond in point form; conjunctive, in which the responses take the form of “A and B and C...”, and parts of speech, in which subjects are instructed to respond using nouns or adjectives. The templates comprised two columns: one for entering quantitative data, and another for entering qualitative data. The students were asked to provide quantitative data for four items: age, year of study, gender, and level of understanding. Responses for “level of understanding” were given using a four-point scale (“Understood clearly = 4” to “Didn’t understand at all = 1”). For the qualitative data, the students responded according to question formats that specified what was understood or not understood through each of the classes.

The above data was analyzed using a KH Coder [2]. Based on the results of prior surveys, it was confirmed that when the question format was “free text responses,” there were many superfluous words, and pre-processing of data was troublesome, so “free text responses” were excluded from this analysis. We conducted a comparison using three types of responses: lists, conjunctives, and parts of speech. We also examined the effectiveness of analyzing qualitative data in combination with quantitative data such as level of understanding of the classes.

### 2.2 Results

In this experiment, we asked questions about what was understood and what was not understood, so we expected that the key words in many of the responses would be nouns. We thus compared the noun appearance rate in the responses to each questionnaire.

**Table 1** Response results for students who responded with understanding level of 1

	Items that students could understand		Items that students could not understand		
	Nouns	Nouns converted to verbs	Nouns	Nouns converted to verbs	
plex	3 average	5 plex	7 calculation	6	
geometry	3 regression	2 stepping stones	3 evaluation	2	
northwest	3 creation	2 curve line	2 regression	1	
minus	2 consistency	2 economy	2 contribution	1	
stepping stones	2 auxiliary	2 method	2 lecture	1	
solution method	1 test	1 northwest	2 selection	1	
coefficient	1 solution	1 how to create	1 synthesis	1	
first	1 calculation	1 formula	1 orthogonal	1	

**Table 2** Response results for students who responded with understanding level of 4

	Items that students could understand		Items that students could not understand		
	Nouns	Nouns converted to verbs	Nouns	Nouns converted to verbs	
geometry	16 average	31 hierarchy	2 calculation	2	
plex	14 order	23 pair	1 alignment	1	
stepping stones	9 calculation	16 structure	1 explanation	1	
northwest	9 consistency	14 items	1 synthesis	1	
how to write	2 test	1			
straight line	2 exercise	1			
pair	1 confidence	1			
perfection	1 constraint	1			

The results show that in the case of both “Conjunctive” and “Parts of speech,” the noun appearance rate varied from one questionnaire to the next, but in the case of “Lists,” the ratio of noun appearance was relatively stable regardless of the questionnaire. Particularly in the fourth class, on “Transport problems,” the noun appearance rate for “Conjunctive” and “Parts of speech” decreased dramatically compared to the preceding questionnaire, to around 35%. In contrast, the appearance rate for “Lists” was similar to the preceding questionnaire, at nearly 50%.

Next, we investigated the effectiveness of analysis combining level of understanding, which is quantitative data, and the appearance of “nouns” and “nouns converted to verbs,” which is qualitative data. Table 1 shows the response results (“nouns” and “nouns converted to verbs”) for students who responded with an understanding level of 1, and Table 2 shows the response results (“nouns” and “nouns converted to verbs”) for students who responded with an understanding level of 4.

From Tables 1 and 2, if we compare the responses of students with understanding levels 1 and 4, we can see that there are 16 cases in which “calculation” appeared as an item that the students with understanding level 4 could understand, and two cases in which “calculation” appeared as an item that these students could not understand. In contrast, there is one case in which “calculation” appeared as an

item that the students with understanding level 1 could understand, and six cases in which “calculation” appeared as an item that these students could not understand. This means that students with understanding level 4 understood calculation problems, and suggests that subjects with understanding level 1 did not understand calculation problems very well.

Other words that appeared as items which could not be understood by students with understanding level 1 were “plex,” “stepping stones,” and “economy.” When we investigated the content of the questionnaire on the origin of these words, we confirmed that “plex” means “simplex method,” “stepping stone” means “stepping stone method,” and “economy” means “economic problems” like “final worth factor.” In other words, we can assume that while subjects with understanding level 4 understood these problems, there were many subjects with understanding level 1 who could not understand.

By combining the level of understanding with the nouns appearing, it is possible to make observations that could not be made only through an analysis of the nouns appearing. This type of knowledge is extremely effective in improving classes. Based on the above, we can say that we have confirmed the effectiveness of question templates that combine both quantitative and qualitative data when gathering questionnaire data.

## **3 Experiment 2**

### ***3.1 Method***

The purpose of this experiment is to test the validity of the knowledge obtained through Experiment 1; namely, that it would be best to use a “Lists” question template for questionnaires designed to gather data comprising nouns in the form of key words.

The subjects were students in “Creation B,” an open course in the Chubu University College of Engineering. The first half of this class comprised lectures in which the students learned about the fundamentals of programs, robots, etc. The second half comprised activities including the assembly of robots, the creation of programs to drive those robots, and a competition in which the robots raced. The students responded to questions on a computer using a question template in a “Lists” format regarding their student number, name, level of understanding, and impressions, as well as the items that they were and were not able to understand.

### ***3.2 Results***

In this experiment, like in Experiment 1, we expected that the key words appearing in many of the responses would be nouns indicating what was understood and

**Table 3** Results for number of noun types and appearances in 2008 and 2009 academic year

Year	2008	2009
Questionnaire format	Free text	Lists
Number of respondents	115	135
Total number of words extracted	6113	8967
Appearance ratio (%)	36.02	29.62
Number of different words	1015	1076
Appearance ratio (%)	46.54	46.38

**Table 4** Results for coding analysis in 2009 academic year

	Number of sentences	Lectures		Competition	
		Hardware	Software	Hardware	Software
Level 1	Comments	50 (10.00%)	10 (20.00%)	24 (8.33%)	4 (16.67%)
	Items that students could understand	38 (15.79%)	9 (23.68%)	9 (0.00%)	2 (22.22%)
	Items that students could not understand	42 (4.76%)	1 (2.38%)	25 (12.00%)	1 (4.00%)
Level 2	Comments	197 (10.66%)	49 (24.87%)	194 (13.92%)	36 (18.56%)
	Items that students could understand	119 (8.40%)	33 (27.73%)	78 (10.26%)	10 (12.82%)
	Items that students could not understand	137 (2.19%)	22 (16.06%)	136 (8.09%)	13 (9.56%)
Level 3	Comments	1018 (18.57%)	202 (19.84%)	821 (16.69%)	144 (17.54%)
	Items that students could understand	787 (23.00%)	215 (27.32%)	453 (9.27%)	38 (8.39%)
	Items that students could not understand	691 (2.46%)	29 (4.20%)	507 (3.94%)	17 (3.35%)
Level 4	Comments	396 (17.93%)	60 (15.15%)	545 (18.90%)	69 (12.66%)
	Items that students could understand	284 (20.77%)	64 (22.54%)	297 (8.42%)	18 (6.06%)
	Items that students could not understand	245 (3.27%)	5 (2.04%)	356 (1.12%)	3 (0.84%)

what was not understood. We thus analyzed the number of noun types and the number of noun appearances for each questionnaire. The analysis results for number of noun types and appearances are shown in Table 3.

From Table 3, we confirmed increases in both the number and types of nouns. The number of nouns appearing increased dramatically, from 6,113 in 2008 to 8,967 in 2009. This is most likely because the response format changed from “Free text” to “Lists,” which are easier to answer.

Next, we conducted analyses applying coding rules based on the themes of “Hardware,” “Software,” and “Both (LEGO)” for both “Lectures” in the first half of the classes and “Competition” in the second half. The results of this analysis are

shown in Table 4. The levels of understanding shown in this table were evaluated using a 4-point scale. As in Experiment 1, if a student responds with a “1” it indicates a low level of understanding, while a response of “4” indicates that the level of understanding is high.

From Table 4, we confirmed that the ratio of response content for subjects with an understanding level of 1-2 changed from “Software” to “Software,” the ratio of response content for subjects with an understanding level of 3 changed from “Software” to “Hardware,” and the ratio of response content for subjects with an understanding level of 4 changed from “Hardware” to “Hardware.” In other words, we confirmed that students with a low level of understanding provided many responses related to “Software,” regardless of whether those responses were referring to “Lectures” or “Competition,” and that students with a high level of understanding provided a greater number of responses related to “Hardware” than other students. This is new knowledge obtained through a combination of coding analysis and quantitative data, which could not be obtained through coding analysis alone.

From the results of this experiment, we confirmed that the appearance rate was especially stable for nouns in the responses to “Lists” questions. These results support the validity of the knowledge obtained through Experiment 1; namely, that it would be preferable to use “Lists” type questions in a questionnaire designed to gather responses in the form of key words. As in the case of Experiment 1, we confirmed that useful data was obtained by using question templates that combine both quantitative and qualitative data when gathering questionnaire data.

## 4 Conclusions

In this research, we developed a question template that makes it possible to efficiently gather text data for analysis, as the first step in building an environment that will enable persons who are inexperienced in text mining to analyze qualitative data easily. We conducted evaluation tests targeting two class questionnaires, and confirmed the effectiveness of those questionnaires. In the future, in addition to developing manuals and other tools, we plan to conduct data analysis trials using these templates, and to conduct more detailed analysis, for example to clarify the scope of applicability for each type of question format.

## References

1. Abe, H., Hirano, S., Tsumoto, S.: Mining a clinical course knowledge from summary text. In: The 19th Annual Conference of the Japanese Society for Artificial Intelligence, 1F4-04 (2005)
2. Coder, K.H.: <http://khc.sourceforge.net/>
3. Nasukawa, T.: Implementation of text mining in practice. *Journal of The Japanese Society for Artificial Intelligence* 24(2), 275–282 (2009)

# Segmentation of Fibro-Glandular Discs in Digital Mammograms Using Log-Normal Distribution

Y. A. Reyad, A. El-Zaart, H. Mathkour, M. Al-Zuair, and H. Al-Salman

**Abstract.** Segmentation is an important and challenging task in a computer-aided diagnosis (CAD) system. Accurate segmentation could improve the accuracy in lesion detection and characterization. In this paper, we present a new technique for Segmentation of Fibro-Glandular Discs based on Split-merge technique applied on histogram of the breast image and the Log-Normal distribution. The proposed new segmentation technique aims at improving the performance level of breast mass segmentation in mammography to provide accurate features for classification purposes. The proposed technique has been experimented with using various breast images. We observed better segmentation results of the proposed technique as compared to the existing methods.

**Keywords:** Breast mass segmentation, Log-Normal distribution, Digital mammograms, Computer-Aided Diagnosis, Split Merge technique.

## 1 Introduction

Image segmentation and classification have many applications including early detection of breast cancer. Image segmentation aims at partitioning the image into physically meaningful regions to identify the objects of interest in the image. There are four broad classes of segmentation methods: classification-based methods, edge-based methods, region-based methods and, hybrid methods [1], [3]. The principal approach of segmentation is based on thresholding that is related to the problem of the thresholds estimation.

Image thresholding is essentially a pixel classification problem [4]. It is based on the assumption that the objects can be distinguished by their gray levels. The

---

Y.A. Reyad  
College of Computer and Information Sciences-King Saud University Riyadh,  
Kingdom of Saudi Arabia  
e-mail: yasali@ksu.edu.sa



main objective is to classify the pixels of a given image into two or more classes [5]: those pertaining to the background and those pertaining to one or more objects. Each class will include pixels with gray values that falls between two certain threshold values,  $T_i$  and  $T_j$  where  $T_i \& T_j \in [t_0, t_1, t_2, \dots, t_n]$  and  $t_0 = 0$  and  $t_n = 255$ . Thresholding is a popular tool for image segmentation for its simplicity, especially when real time processing is required. Threshold selection can be categorized into two classes, local methods and global methods [5], [6]. The global thresholding methods segment an entire image with a single threshold using the gray level histogram of image. The local methods partition the given image into a number of sub-images and select a threshold for each of sub-images. The global thresholding techniques are easy to implement and less expensive in computation. Therefore, they are superior to local methods in terms of many real image processing applications [6], [2].

Due to the drawbacks of the classical split merge technique like compact region and the Rough edge [3], therefore, the split merge technique will be applied to the histogram of the image. Histogram modes are not always modeled as normal distribution, for this reason, the Log-Normal distribution will be used to model the modes of the histogram. In this paper, the split-merge technique is applied to the histogram of the breast image and the Log-Normal distribution is used model the image histogram data. The objective of this technique is to split the histogram (non-homogenous region) into distinct homogenous regions then merge the incorrect splitted classes if it apply the merge criteria.

Image histogram can be seen as a mixture of statistical distribution [7]. Modes of the histogram can be symmetric or non-symmetric. Log-normal distribution is a probability distribution of a random variable whose logarithm is normally distributed [9], [10]. It has certain similarities to the Gaussian and Gamma distribution and is also skewed to the right (positively skewed). Therefore the Log-Normal distribution is used to model symmetric and moderate positively skewed data. The rest of this paper is organized as follows. Section 2 describes the proposed method. Section 3 presents the experimental results. Section 4 concludes the paper.

## 2 Methodology

Mammography is a specific type of imaging that uses a low-dose x-ray system to examine breasts. In mammography image, the background of the breast affects the histogram of the breast region if the image is used as it is. Thus, the background interval should be removed from the histogram before applying segmentation for the breast. Below, the breast detection steps and the breast segmentation process are introduced.

### 2.1 Breast Detection Steps

In mammography image, the presence of higher gray values identifies the breast region from the non-breast region. Therefore, the breast region can be segmented

using a bimodal thresholding technique. The proposed steps for detecting the breast are as follows:

1. Calculate the mammogram image histogram.
2. Calculate the initial threshold value as the midway between the minimum and maximum gray levels of the image.
3. Calculate the statistical parameters of Log-Normal distributions.
4. Estimate the final threshold value based on the statistical parameters.

After applying a bimodal segmentation for the breast image the breast histogram is modified by changing the background interval in the histogram to zeros.

## 2.2 Breast Segmentation Process

The proposed segmentation process is applied to the detected breast only. The proposed method consists of two major operations: The split operation and the merge operation.

**The Split Operation:** It first considers the whole histogram as one mode and applies a homogenous test based on the variance of the mode, if it is not homogenous then split it into two modes  $M_1$  and  $M_2$  using threshold  $T_1$ . Then we apply the same operation to each mode. The histogram is modeled as a mixture of Log Normal distribution for estimating the threshold  $T_i$  between every two es  $M_i(p_i, \mu_i, \sigma_i)$ ,  $M_{i+1}(p_{i+1}, \mu_{i+1}, \sigma_{i+1})$ .

Thresholds are selected at the valleys of the mixture of Log Normal distribution. The optimal threshold  $T_i$  is the value at the abscissa where both Log Normal functions are equal.

$$w_i * G(T_i, \mu_i, \sigma_i) = w_{i+1} * G(T_i, \mu_{i+1}, \sigma_{i+1}) \quad (1)$$

Where  $w_i$  is the initial probability and  $G(T, \mu, \sigma) = \frac{1}{T\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln T - \mu}{\sigma}\right)^2}$

The threshold  $T_i$  between two Log-Normal distributions can be calculated by solving equation (1) in T so that T will equal:

$$\log(T_i) = \frac{-b\sqrt{b^2 - 4ac}}{2a} \quad (2)$$

Where,  $a = \frac{\sigma_i^2}{2} - \frac{\sigma_{i+1}^2}{2}$ ,  $b = (\sigma_{i+1}^2\mu_i - \sigma_i^2\mu_{i+1})$ ,  $c = \frac{1}{2}(\sigma_i^2\mu_{i+1}^2 - \sigma_{i+1}^2\mu_i^2) + \sigma_i^2\sigma_{i+1}^2 \log\left(\frac{w_i\sigma_{i+1}}{w_{i+1}\sigma_i}\right)$

Now the threshold between each two modes can be estimated by (2). The split operation works as a binary tree and stop when all the resulting subclasses are homogeneous.

**The Merge Operation:** The split operation may result in incorrect classes which should be merged. The merge operation defines two criteria to merge the incorrect classes. These two criteria are minimum class members (MCM) and the minimum class mean distance (MCD).

In MCM measure, if the number of pixels in class  $C_i$  is less than MCM value, then this class will be merged with the nearest class to it, i.e. the class who has the closest mean. In the MCD measure the difference in mean between a class and its adjacent class is computed. If it is less than MCD then the two classes are merged.

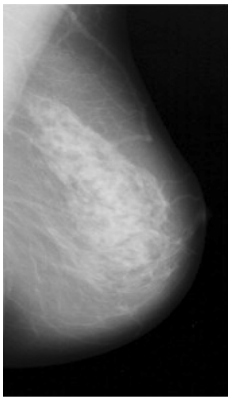
### 3 Experimental Results

The developed Fibro-glandular discs segmentation algorithm using Log Normal distribution is applied on different real mammography image. The obtained results are compared with those obtained by expectation maximization (EM) technique with Gaussian and Gamma distributions. A Fibro-glandular disc is the regions of interest in mammography images. It is Appears as a bright area in the original images or a dark area in the negative of mammography images. These areas are always represented by the first class in histograms of the negative image.

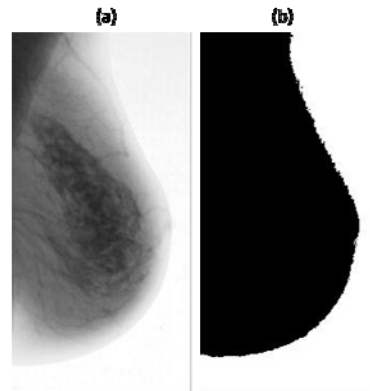
The proposed method will be applied on negative images. The fibro-glandular discs regions in mammography images are bright, see Fig. 1. The negative mammography image is used in order to have dark regions for fibro-glandular discs.

**Table 1** Numerical result of the proposed method compared to Gamma and Gaussian distributions

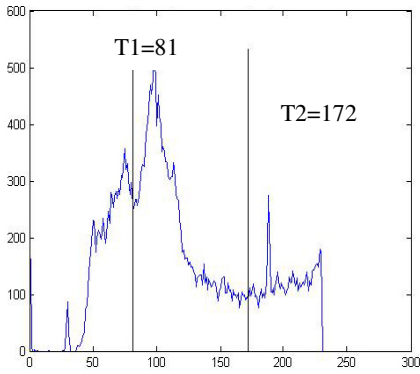
User input parameter		Thresholding values
<i>Split parameter</i>	<i>Merge parameter</i>	Log Normal :T1=81, T2=172
Sigma=30	MCM=5000 MCD=30	Gaussian: T1=90, T2=170 Gamma: T1=80, T2=146



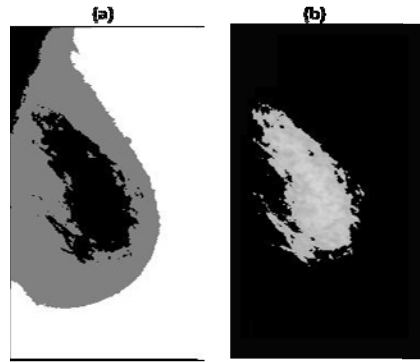
**Fig. 1** Original mammography image



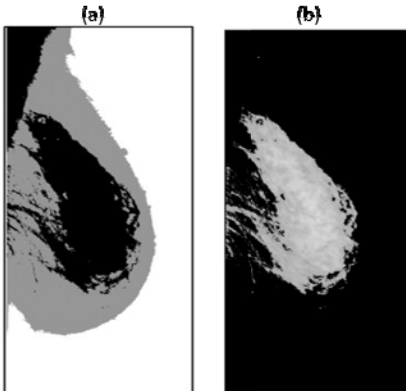
**Fig. 2** Breast detection: (a) negative mammography image, (b) detected breast



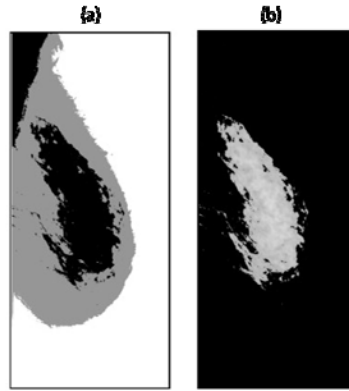
**Fig. 3** Histogram of the breast after applying the proposed method



**Fig. 4** Log-Normal distribution, (a) Segmented image, and (b) Fibro-glandular disc area



**Fig. 5** Gaussian distribution [8] (a) Segmented image, and (b) Fibro-glandular disc area



**Fig. 6** Gamma distribution: (a). segmented image, and (b) fibro-glandular disc area

An original mammography image is presented in Fig. 1. The negative image is shown in Fig. 2-a, and the detected breast is shown in Fig. 2-b. After applying the proposed method on the modified histogram of the detected breast we obtain the segmented breast result in Fig. 4-a. In Fig. 4-b, the fibro glandular disc area is defined separately. Fig 3 shows the histogram of the segmented image. The segmented result and the fibro glandular disc area using the Gaussian method are shown in Fig. 5 and table (1) show the numerical result. In Fig. 6 the segmented result and the fibro glandular disc area using the Gamma method are shown.

Consequently, it is concluded that the Log Normal distribution best model the breast data and the result of the proposed method give better results than those obtained through the Gaussian method and similar to or better than those obtained

by Gamma method. Comparing the result of the three methods visually we find that the proposed method segment the fibro glandular disc more accurately.

## 4 Conclusions

Breast cancer detection is a major health problem around the world. X-ray mammography remains the key screening tool for detecting breast abnormalities. Segmentation of the breast mammography image helps in the detection of fibro glandular disc accurately. The key point in this algorithm is modeling the histogram as a mixture of Log-Normal distributions and applying the split and merge operations on the histograms of the breast images.

From the evaluation of the resulting images, we conclude that the proposed thresholding method yields better results. The Log-Normal distribution can model symmetric and positively skewed data. This makes it possible to yield better segmented image than existing method that relies on Gaussian distribution. It is concluded that the Log Normal distribution best model the breast data and that the proposed method provided better results than those obtained through the Gaussian and Gamma methods. As reported earlier, comparing the result of these methods we find that the proposed method segment the fibro glandular disc more accurately.

**Acknowledgment.** This work was supported by the Research Center of the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.

## References

1. Dubeya, R.B., Hanmandlub, M., Gupta, S.K.: A comparison of two methods for the segmentation of masses in the digital Mammograms. *Computerized Medical Imaging and Graphics* 34, 185–191 (2010)
2. Oliver, A., Freixenet, J., Martí, J., Pérez, E., Pont, J., Denton, E.R.E., Zwiggelaar, R.: A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis* 14, 87–110 (2010)
3. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. Prentice-Hall (2008)
4. Weszka, J.S.: A survey of threshold selection techniques. *Comput. Graphics Image Process* 7, 259–265 (1978)
5. Feng, Wenkang, S., Liangzhou, C., Yong, D., Zhenfu, Z.: Infrared Image Segmentation with 2-D Maximum Entropy Method based on Particle Swarm Optimization (PSO). *Pattern Recognition Letters* 26(5), 597–603 (2005)
6. Yong, Y., Zheng, C., Lin, P.: Image thresholding based on spatially weighted fuzzy c-means clustering algorithm. *GVIP 05 (V3)* (2005)
7. El-Zaart, A.: Image Thresholding using ISODATA Technique with Gamma Distribution. *Journal of Pattern Recognition and Image Analysis* 20(1), 29–41 (2010)
8. El-Zaart, A.: Expectation-Maximization technique for fibro-glandular discs detection in mammography images. *Computers in Biology and Medicine* 40, 392–401 (2010)
9. Aitchison, J., Brown, J.A.C.: *The Lognormal Distribution* (1957)
10. Evans, M., Hastings, N., Peacock, B.: *Statistical distribution*, 2nd edn. (February 2000)

# An OOAD Model of Program Understanding System's Parser

Norazimah Rosidi, Nor Fazlida Mohd Sani, and Abdul Azim Abd Ghani

**Abstract.** This paper describes a model of parser for program understanding system. This model is developed by using Unified Modeling Language (UML). The UML is a common notation for structured modeling within Object-oriented analysis and design (OOAD) framework. It helps to specify, visualize and document models of software system. The objective of developing this model is to capture and document the details of this parser. The parser may build parse tree and abstract syntax tree (AST) which representing data structure of all elements of the source code. It also generate control flow graph (CFG) to show the flow of program. This paper describes the process of developing this model, which include creating use case diagram, class diagram, and behavior diagrams.

## 1 Introduction

Program understanding system is an activity that enables the programmer to know meaning for programming codes [2]. Process to understand the codes are difficult especially for novice. How the documentation of the system will assist for understanding the system. There are two broad categories of documentation that are used as an aid in program understanding: textual and graphical [3]. The Unified Modeling Language (UML), the most widely used object-oriented design representation and arguably an emerging standard format of graphical documentation.

In process of gathering the requirements of an application, we have to model the application in understandable view. The UML is the industry-standard language for specifying, visualizing, constructing and documenting the artifacts of software systems. It simplifies the complex process of software design, making a “blueprint” of construction [11]. It relatively detailed design diagrams used either for reverse engineering to visualize and better understanding existing code in UML diagram or code generation in forward engineering.

---

Norazimah Rosidi · Nor Fazlida Mohd Sani · Abdul Azim Abd Ghani

Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

e-mail: azie\_ros@yahoo.com, {fazlida, azim}@fsktm.upm.edu.my

In this paper we focus on UML design for parser in program understanding system. Since our research for object-oriented program, it is most appropriate that the model of this parser is described by using Unified Modeling Language. The UML is a common notation for structured modeling within OOAD framework.

The UML defines nine kinds of diagrams which one can mix and match to assemble each view. For static views these are use case, class, object, component, and deployment diagram, while for dynamic views these are sequence, collaboration, statechart, and activity diagram. For example, the static aspects of a system's database might be visualized using class diagrams; its dynamic aspects might be visualized using collaboration diagrams [11].

This paper contains of seven sections. In next section is related works for this research. In section 3 is explanation about Unified Approach methodology. Section 4 is about creating the use case for this research. Section 5 shows the class diagram. Behavior diagrams that include sequence diagram, collaboration diagram, statechart diagram and activity diagram are explained in section 6. The last section is the conclusion for this UML model for this research.

## **2 Related Works**

### ***2.1 Program Understanding and Parsing***

Program understanding plays an important role in many software engineering and reengineering activities such as code maintenance, debugging, restructuring and software reuse [6]. For example, in program understanding system CONCEIVER++, the intermediate representation that had been used is an OO-CFG which created for object-oriented programming code especially for java programming code [6].

Parsing in program understanding aims to produce representations of source code that can be used for program analysis and design recovery purpose [5]. Parser or syntax analyzer takes the tokens as input and check if the source codes match with the grammar. The parser may build parse tree and abstract syntax tree (AST) of the code. Parse tree and AST are data structure representing all the main elements of the input but in a compact way in AST. The source code also will analyze to build control flow graph (CFG) where the nodes represent the statements in the source code the edge represents as flow of control.

### ***2.2 Unified Modeling Language***

The Unified Modeling Language began in 1994 as an attempt to unify the Booch and OMT models but quickly developed into a broadly based effort to standardize object-oriented modeling concepts, terminology, and notation [10]. In 1997, the Object Management (OMG) [7] released the UML one of the purposes of UML was to provide the development community with a stable and common design

language that could be used to develop and build computer applications. The UML is not Object-oriented Analysis and Design or method, it is just diagramming notation. The OMG was improved their UML by released the UML 2.0 that was consists 13 diagrams with added composite structure diagram, package diagram, communication diagram and timing diagram.

The process of gathering and analyzing an application's requirements, and incorporating them into a program design is a complex one and regardless of the methodology that used to perform analysis and design, the UML can be used to express the result.

### ***2.3 Object-Oriented Analysis and Design***

In software there are several ways to approach a model. The two most common ways are from an algorithmic perspective and from an object-oriented perspective. As I mentioned above, the object-oriented perspective is seen outpacing the non-object oriented [11]. Object-oriented methodology is a set of methods, models, and rules for developing systems. In this research, some of the involved UA are object-oriented analysis and object-oriented design.

Object-oriented analysis and design (OOAD) is a software engineering approach to constructing software systems by building object-oriented models that abstract key aspects of the target system and by using the models to guide the development process [10].

OOA phase of the Unified Approach uses actors and use cases to describe the system from the users' perspective. The actors are external factors that interact with the system; use cases are scenarios that describe how actors use the system. During OOA there is an emphasis on finding and describing the objects or concepts in the problem domain [9].

OOD is a design method in which a system is modeled as a collection of cooperating objects are treated as instances of a class within a class hierarchy. During OOD there is an emphasis on defining software objects and how they collaborate to fulfill the requirements [9].

## **3 Unified Approach (UA)**

We are using UA as our methodology for this research to show the model of the program by using UML. This methodology has been chosen because of the research will be developed by using object-oriented language.

Unified approach (UA) is a methodology for software development which based on methodologies proposed by Booch, Rumbaugh, and Jacobson. UA establishes a unifying and unitary framework around their works by utilizing the unified modeling language (UML) [1]. The Unified Modeling Language (UML) is the de facto standard for modeling modern software applications [3].

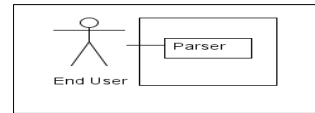


The UML is typically used as a part of a software development process, with the support of a suitable CASE tool, to define the requirements, the interactions and the elements of the proposed software system. UML helps specify, visualize, and document models of software system including their structure and design, in way that meets all of these requirements.

The UA to software development revolves around the following processes and concepts. The processes are: use-case driven development, OOA, OOD, incremental development and prototyping, and continuous testing [1].

## 4 Creating Use Case Diagram

Use case diagram is one of the UML models that describe the functionality of the system from the user's point of view. The use case model shows the relationship between actors (outside) and use case (inside). In this research, we have identified one actor as end-user and only one use case shown as in Figure 1. In Figure 2 shown the flow of event in this research for develop parse tree, AST and CFG.



**Fig. 1** Use Case Diagram for Parser

### 1. Flow of Events for the Parser in Program Understanding

#### 1.1 Preconditions

User must have source code that will be analyze

#### 1.2 Main Flow

This use case begins when the End-User start this system with insert the program code to analyze in parser generator {1}. The system will parse the code and create the parse tree and abstract syntax tree (AST) {2}. With the same code the user can build the control flow graph (CFG) by using CFG generator {3}. The CFG for the code will be calculating to get their complexity {4}.

#### 1.3 Subflows (if applicable)

#### 1.4 Alternative Flows

- {1}: Invalid type of source code inputted. User can re-insert the code or terminate the use case
- {2}: The system cannot be parsed. User can choose either the use case begin again or terminate the use case.
- {3}: The system cannot be build into CFG. User can choose either the use case begin again or terminate the use case.
- {4}: The system cannot be calculated. User can choose either the use case begin again or terminate the use case.

**Fig. 2** Flow of event for parser in program understanding system

### 5 Class Diagram

Class Diagram shows the static structure of the model that describes the structure of a system by showing the classes and their relationship, connected as a graph to each other and to their contents. In this research, class diagrams that included are parser generator, CFG generator, and complexity measurement as show in Figure 3.

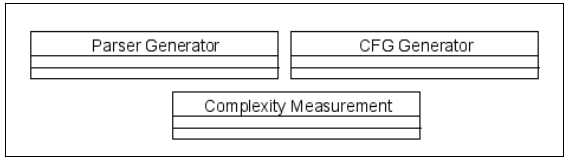


Fig. 3 Main Class Diagram

### 6 Behavior Diagram

Behavior diagram is a dynamic model in UML. Dynamic model can be viewed as a collection of procedures or behaviors that, taken together, reflect the behavior of the system over time [1]. Behavior diagrams included of interaction diagram, statechart diagram and activity diagram. Interaction diagram consist of sequence diagram and collaboration diagram.

Sequence diagrams are used to display the interaction between users, screens, objects and entities within the system. It provides a sequential map of message passing between objects over time. Figure 4 shows the sequence for parsing process which has user interface, parser generator, CFG generator, and complexity measurement as objects.

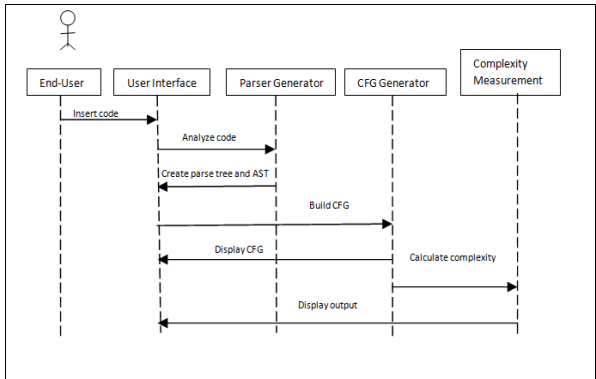


Fig. 4 Sequence diagram

Collaboration diagram is another type of interaction diagram. It represents a collaboration, which is a set of objects related in a particular context, and interaction, which is a set of messages exchanged among the objects within the collaboration to achieve desired outcomes [1]. Figure 5 shows the collaboration diagram for the parsing process.

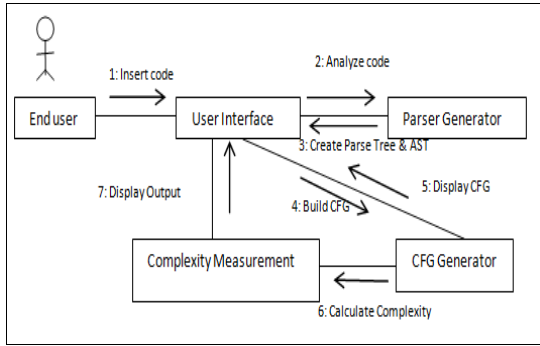


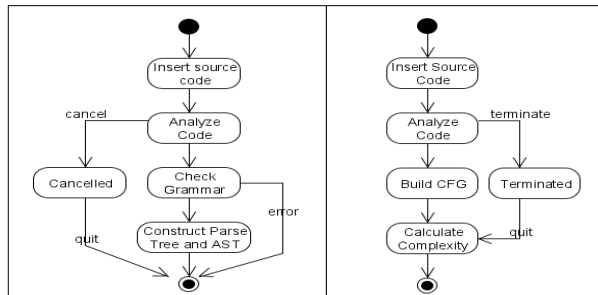
Fig. 5 Collaboration Diagram

Statechart diagrams and activity diagrams are another dynamic model in UML. State charts are used to detail the transitions or changes of state an object can go through in the system. They show how an object moves from one state to another and the rules that govern that change. State charts typically have a start and end condition.

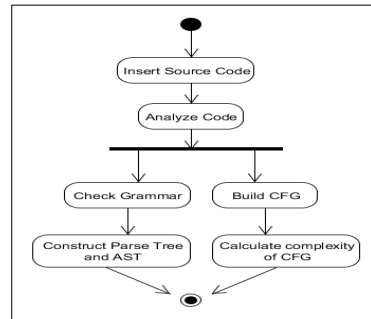
Activity diagrams are used to show how different workflows in the system are constructed, how they start and the possibly many decision paths that can be taken from start to finish. They may also illustrate the where parallel processing may occur in the execution of some activities.

Figure 6(a) shows the statechart for the parser state. Insert source code, parse the source code, canceled and construct parse tree and AST are the states in the parser state. Figure 6(b) shows the statechart diagram for the CFG builder and the last one Figure 7 shows activity diagram for this research.

Fig. 6 (a) Statechart Diagram for parser process and (b) statechart diagram for CFG builder



**Fig. 7** Activity Diagram parser show the workflow of the parser.



## 7 Conclusions

This paper has presented and discussed about the UML model for parsing process in program understanding system. The model uses UML for modeling, describing, analyzing and designing an application. The diagrams include use case diagram, flow of event, class diagram, sequence diagram, collaboration diagram, statechart diagram and activity diagram. The model developed is important for next phase of research which system development.

**Acknowledgements.** The authors acknowledge the financial support (RUGS) received from Ministry of High Education, (MOHE), Malaysia via Universiti Putra Malaysia.

## References

1. Bahrami, A.: Object-oriented Systems Development Using the Unified Modeling Language. McGraw-Hill, Boston (1999)
2. Sani, N.F.M., Zin, A.M., Idris, S.: Analysis and Design of Object-oriented Program Understanding System. *International Journal of Computer Science and Network Security* 9(1) (2009)
3. Tilley, S., Huang, S.: A Qualitative Assessment of the Efficacy of UML Diagrams as a Form of Graphical Documentation in Aiding Program Understanding. In: *Proc. of the 21st Annual International Conference on Documentation*, pp. 184–191 (2003)
4. Sani, N.F.M., Zin, A.M., Idris, S.: The Model of Program Understanding Systems using UML. In: *Proc. of Bengkel Sains Pengaturan ATUR 2003*, pp. 143–155 (2003)
5. Kontogiannis, K., Mylopoulos, J., Wu, S.: *Towards Environment Re-Target Parser Generators*. AISE, pp. 407–437. Springer, New York (2001)
6. Sani, N.F.M., Zin, A.M., Idris, S.: Object-oriented Code Representation of Program Understanding System. In: *International Symposium on Information Technology*, vol. 1, pp. 1–3 (2008)
7. Object Management Group, <http://www.omg.org>
8. Kozacynski, W., Ning, J.Q.: *Automated Program Understanding by Concept Recognition*. Automated Software Engineering (1994)

9. Larman, C.: *Applying UML and Patterns: An Introduction to Object-oriented Analysis and Design and Iterative Development*, 3rd edn. Prentice Hall PTR (2004)
10. Rumbaugh, J.: *Object-oriented Analysis and Design*, 4th edn., *Enclopedia of Computer Science*, pp. 1275–1279 (2003)
11. Andria, F.: *Object-oriented Analysis and Design using the Unified Modeling Language*. George Mason University (1999)

# The Unified Modeling Language Model of an Object Oriented Debugger System

Noor Afiza Mohd Ariffin, Nor Fazlida Mohd Sani, and Rodziah Atan

**Abstract.** In this paper will explain a model that being used to develop an Object Oriented Debugger (OOD). Unified Modeling Language (UML) is the best choice model and suitable to design the Object Oriented Debugger which will be developed in an object oriented programming environment. The model will provide an ability to capture the characteristics of a system by using notations in the process of designing and implementing the system. The objective of this paper is to capture the structure and behavior of the Object Oriented Debugger system by using the UML diagram. The model also can ease the readability of the documentation for the maintenance purposes. This paper starts with explanation about the concepts of Object Oriented Debugger and has divided into three sections which include creating use case, capturing process flow and finding objects interactions. Appropriate UML diagrams for Object Oriented Debugger system are presented in relevant sections of this paper.

## 1 Introduction

Debugger is a computer program that is used to reduce the errors in programming code. In development activity, the debugging process is very important. Its offer more sophisticated function such as single stepping, breakpoint and fix the errors depending on the programmer understanding or expertise. For novice programmers, it is a big challenge to understand the meaning of each error in program code especially for logic errors. The same problem also confront by the experienced. The process of finding and fixing the logic errors is more difficult rather than finding and fixing syntax errors. To helps and solve the problem that occurs among the novice programmers this research try to making the debugger that can be more understandable to the novice programmers. So, we present an automated

---

Noor Afiza Mohd Ariffin · Fazlida Mohd Sani · Rodziah Atan  
Faculty of Computer Science and Information, Universiti Putra Malaysia  
e-mail: afiz\_efy@yahoo.com, {fazlida,rodziah}@fsktm.edu.my

debugger named Object Oriented Debugger, a system for analyzing code written in Java language which can handle the problem of understanding on object oriented programming and debugging program among the novice programmers. It also to determine the difficulty of finding logic error among the beginning programming student by the analyze source code to localize, find logic error and provide more user-friendly error messages when an uncaught exception occurs.

The purpose of this research is to develop the Object Oriented Debugger that has been designed to understand a program written in a structured language, such as Java. This research has been carried out from the CONCEIVER++ (Sani et al., 2009) and Adil (Zin et al., 2000) system as well as extending the plan formalism to include the logical errors and designed as a new automated debugger which can debug on different style of written programming code. The objective of this paper is to capture the structure and behavior of an Object Oriented Debugger based on the diagram UML, also to make documentation more readable.

In this paper, we will describe an Object Oriented Debugger by using the Unified Modeling Language (UML). The UML has become the standard notation for object-oriented modeling system (Ali, 1999). The UML use the notations to express the design of a software system. The UML process model starts by seeking the requirement and ideas of the system via Use Case diagram, the identification the steps involve in each requirement in Activity diagram and define the external interfaces that need in the system. The identification of major function that need to provide in the system and also the identification of the relations among the elements in the system.

**Table 1** A capability of existing automated debugger

Name	Capabilities
Backstop tool	Is a tool to debug a runtime error in Java Application and also provide an error message to user.
Espresso	Is a debugger to detect much type of error such as syntax, semantic and logic error in Java program code also provide an error messages to user.
CMeRun	Program Logic Debugging is develop to detect logic error in C++ among student in course CS1/CS2, with allowing the user to see each of statement is executed.
CAP	Code Analyzer Pascal is an automated self assessment tool to detect syntax, logic and style errors in Pascal program. It also provides an error message to the user.
WPOL	Is a Web Plan Object Language that is incorporate with a plan paradigm. Explain about the plan management.
J.Bixbe	Use in Java application by showing the structure and functioning on the UML level. Can find bugs and also can show the weakness and insufficiencies in application.
DBG PHP	Is an open source debugging for PHP program. It is support GUI interface and also can debug in 18 different platforms whether as locally or remotely.
Jswat	Is a Java standalone debugger. Have a capability to detect bug by breakpoint and analyze the code with displaying the object relationship.
PyChecker	PyChecker is a software development tools used to find programming errors in Python language. It only finds the missing doc strings.
Adil	Use in subset of C language. Finding bugs by understand the code, localize and explain the bugs. It also uses the cliché that stored as a library plan in knowledge-base.

## 2 Related Work

### 2.1 Debugger

Debugger is a computer program that used to find bugs in program code during the execution process. Its find bugs by the analyze source code to localize a bugs and fixing this bugs. In Table 1 shown the existing debugger tool with different capabilities. All debugger has given different capabilities which will be take account into this research.

### 2.2 UML Notation

The UML is a modeling language for designing the software system [13]. All the requirement of the system will be describe and model in UML structure. The UML notation is very important in modeling. Use the appropriate notations to make the model more understandable. The standard UML notation that used to describe the Object Oriented Debugger is Class notation, Collaboration notation, Use Case notation and Interaction notation.

### 2.3 Object Oriented

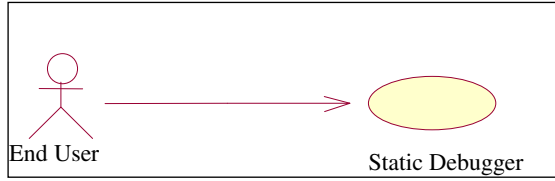
Object oriented programming is a new approach to programming which address these systems issues, moving the focus from programs to software (Anderson, 1988). With using the object oriented, it easier to build and understand the systems. So, in this research we present an automated debugger for object oriented programming.

### 2.4 Creating Use Case

Use case is a type of behavioral of the system that illustrates using the use case diagram. Use case is useful for analysts to more understand the flow of the system and it's also can help the analysts to partitioning the functionality of a system. Its present a functionality of a system with showing the each of function are performed by actor. From the use case diagram, we can see the relationship between the use case and actors in the system. In Object Oriented Debugger, we have one actor as a user. In Object Oriented Debugger use case diagram, we have identified five use cases which can access by user. It's consists of User Interface, Selection Program, View Program, Run Program and Modify Program. This use case begins when user enter to the system and select the program. The system will parse each of line codes in the cod program. The codes that already parse will stored into database. After that, the system will check the code refer to the plan base in database. Finally, the system produce the output that is description about the error appear in the program code. Fig. 1 shows the use case diagram for Object Oriented Debugger.

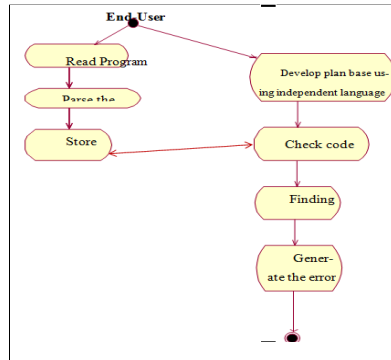


**Fig. 1** Use case diagram for Object Oriented Debugger



### 2.5 Capturing Process Flow

Activity diagram can be used to describe the stepwise activities of each component in the system. There are 7 components that involve in the Object Oriented Debugger, it's begin with the user enter to the system and select the program. The system will parse each of line codes in the program code. The codes that already parse will stored into database. After that, the system will check the code refer to the plan base in database. Finally, the system produce the output that is description about the error occur in the program code. The user The Activity Diagram for Object Oriented Debugger is show Figure 2.



**Fig. 2** Activity diagram for Object Oriented Debugger

### 2.7 Finding Object Interactions

There are two types of object interaction diagram that are sequence diagram and collaboration diagram. We use the sequence diagram to represent the flow of messages, events and actions between the objects of a system in a time sequence [14]. It's also used to describe the sequence of actions that need to complete a scenario of the system. The collaboration or interaction diagram shows the relationship and interactions among the objects in the system. Class diagram is one of the Unified Modeling Language. It's used to describe the structure of a system by showing the system classes, attributes and relationship between the classes. In Object Oriented Debugger we classified seven classes that are User Interface, Select Program, View Program, Run Program, Code Parser, Database and Plan Base. The object in this system consists of End-User. This system begins with User Interface

class which user can enters to the system and the Select Program class allowed the user to select the program. The View Program class allow user to view the program that their selected. When Run Program is function, the program will be processed, the system will parse each of line codes in the program code is done by Code Parser class. The codes that already parse will stored into Database. After that, the system will check the code refer to the Plan Base in database. Finally, the system produce the output that is description about the error occur in the program code. The several main classes in Object Oriented Debugger are shown in Fig. 3 below. The sequence diagram and collaboration diagram are shown in Fig. 4 and Fig. 5.

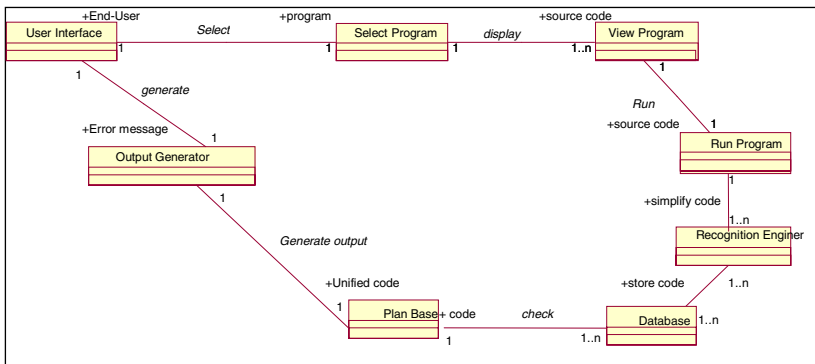


Fig. 3 Class diagram for the Object Oriented Debugger.

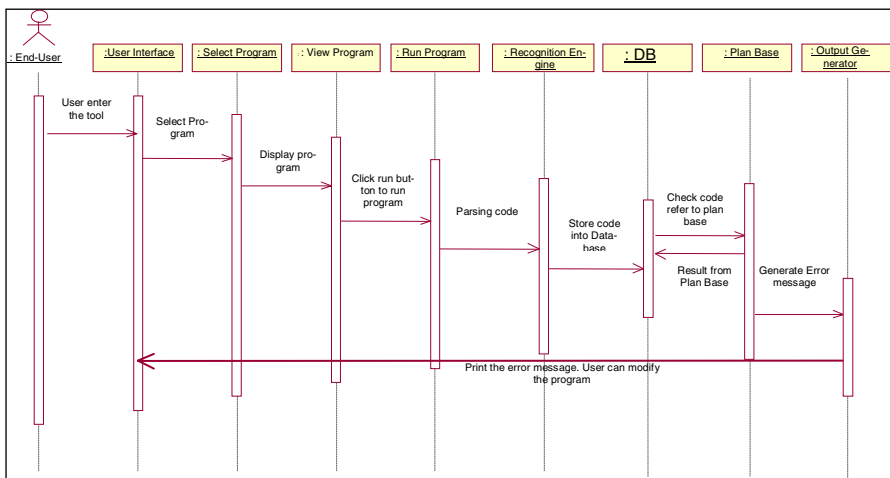


Fig. 4 Sequence diagram for Object Oriented Debugger.

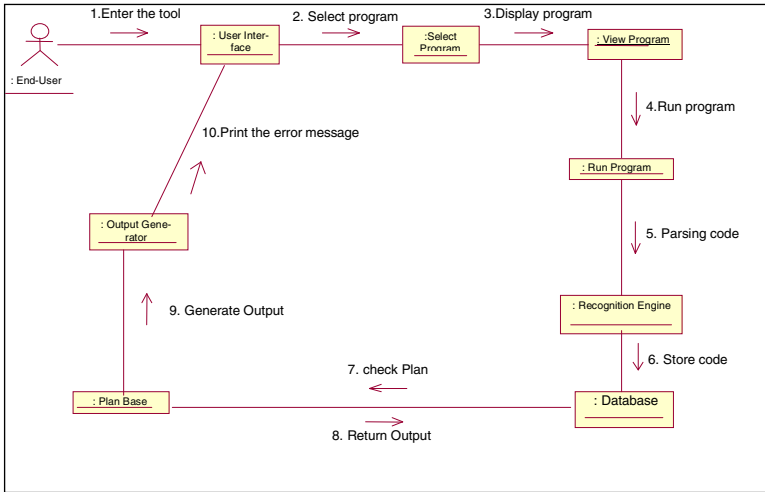


Fig. 5 Collaboration diagram for Object Oriented Debugger.

### 3 Conclusion

This paper has presented a process model of the Object Oriented Debugger by using the Unified Modeling Language (UML). It's consisting of two parts that are object oriented analysis and object oriented design. All the developing and designing are based on the model in UML. This paper has explained on the concepts of Object Oriented Debugger and has divided into three sections which include creating use case, capturing process flow and finding objects interactions. Appropriate UML diagrams for Object Oriented Debugger system also has presented. This model is accurate as a major part in order to develop the object-oriented Debugger which is the next phase of this research.

**Acknowledgement.** Special thanks from author to financial support (Science Fund) received from the Ministry of Science, Technology and Innovation (MOSTI), Malaysia via University Putra Malaysia. The supervisor of the research project, Dr. Nor Fazlida Mohd Sani, the committee member, Dr. Rodziah Atan.

### References

1. Ali, B.: Object Oriented System Development using the Unified Modeling Language. McGraw-Hill, Boston (1999)
2. Sani, N., Zin, A., Idris, S.: Implementation of Conceiver++: An Object-Oriented Program Understanding System. Journal of Computer Science 5(12), 1009–1019 (2009)

3. Zin, A., Aljunid, S., Shukur, Z., Nordin, M.: A Knowledge-Based Automated Debugger in Learning System. Research Notes in Artificial Intelligence, University Kebangsaan Malaysia (2000)
4. Al-Omari, H. M. A.: CONCEIVER: Not Just Another Program Understanding System, Ph.D Thesis, UKM (1999)
5. Al-Omari, H.M.A.: CONCEIVER: A program understanding system. Ph.D. Thesis, University Kebangsaan Malaysia (1999)
6. Swat, J.: JSwat Jaba DEBUGGER (2009), <http://code.google.com/p/jswat/>
7. Murphy, C., Kim, E., Kaiser, G., Cannon, A.: Backstop: A Tool For Debugging Runtime Errors. In: Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education, Portland, OR, USA, pp. 73–177 (2004); ISBN: 978- 1-59593-799-5
8. Bixbe, J.: Another Way to Debug (2006), <http://www.jbixbe.com/>
9. Schorsch, T.: Cap: An Automated Self-Assessment Tool to Check Pascal Programs for Syntax, Logic and Style Errors. In: Proceedings of the Twenty-sixth SIGCSE Technical Symposium on Computer Science Education, Nashville, Tennessee, United States, pp. 168–172 (2006); ISBN:0-89791-693-X
10. Etheredge, J.: CMeRun: Program Logic Debugging Courseware for CS1/CS Students. In: Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education, Norfolk, Virginia, USA, pp. 22–25 (2004); ISBN:1-58113-798-2
11. Ebrahimi, A., Schweikert, C.: Empirical Study of Novice Programming with Plans and Objects. SIGCSE Bulletin (38), 4 (2006)
12. Neal, N., Eric, C.N., Michele, M.: PyChecker: Finding Bugs in Other People's Programs (2005)
13. UML Tutorial - What is UML, [http://atlas.kennesaw.edu/~dbraun/csis4650/A&D/UML\\_tutorial/what\\_is\\_uml.htm](http://atlas.kennesaw.edu/~dbraun/csis4650/A&D/UML_tutorial/what_is_uml.htm)
14. UML Sequence Diagram Tutorial, <http://www.sequencediagrameditor.com/uml/sequence-diagram.htm>

# A FPGA-Based Real Time QRS Complex Detection System Using Adaptive Lifting Scheme

Hang Yu, Lixiao Ma, Ru Wang, Lai Jiang, Yan Li, Zhen Ji, Yan Pingkun, and Wang Fei

**Abstract.** This paper presents a real time QRS wave detection system implemented using the FPGA. Based on the Adaptive Lifting Scheme, the digitized ECG sequence is directly processed in the spatial domain, thus greatly simplifies the system design and reduces the memory usage. In order to synchronize the input ECG sampling data and the FPGA based ALS system, a pipeline architecture interface is proposed. It works as a buffer and the ECG data is processed in a real time manner. The system is implemented on the XUPV5-LX110T evaluation platform, and validated by using ECG samples from the MIT-BTH Arrhythmia Database. Experimental results show that the system achieves 98.688% detection accuracy, while the dynamic power consumption is only ~20mW.

**Keywords:** Real time QRS Detection; Adaptive Lifting Scheme; FPGA.

## 1 Introduction

In recent years, the Field Programmable Gate Array (FPGA) technology develops rapidly, and has been widely used in medical related applications. Because of its low power consumption, flexible programming, short development cycle and easy

---

Hang Yu · Lixiao Ma · Ru Wang · Lai Jiang · Yan Li  
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen

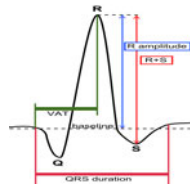
Shenzhen City Key Laboratory of Embedded System Design, Shenzhen University

Yan Pingkun  
OPTIMAL, Xi'an Institute of Optics and Precision Mechanics, Xi'an

Wang Fei  
IBM Almaden Research Center, San Jose, CA 95120, U.S.A.

for transplant, the FPGA is advantageous to be used for implementing the real time processing systems. For example, Yang and Huang presented a real-time electrocardiogram (ECG) monitoring system with the digital filtering and data compression arithmetic integrated on the FPGA [1]. A FPGA-based Fetal QRS complex detection system developed by M. I. Ibrahimy can be used to assess the fetal condition of an unborn baby [2].

ECG signal represents the electrical activity of heart. As shown in Fig. 1, a typical ECG trace, also named as QRS complex, consists of three parts: a R-wave is the only upward deflection, and the downward deflection before and after the R-wave are named Q- and R- wave, respectively [3]. The information obtained from analyzing the signal can be used to discover different types of heart disease [4]. ECG signal is a quasi-periodic signal, and it contains redundant information. Therefore, when analyzing the ECG data, data compression method is generally applied. Up to date, the most popular method to compress the QRS complex in real time is based on the discrete wavelet transform (DWT) [5]. However, because signal convolution is involved in the data processing, this method needs large amount of calculation, and requires high storage space when implemented using the FPGA. Targeted to reduce the processing complexity, thus improves the overall system performance, Daubechies and Sweldens proposed a lifting scheme based on the wavelet transform [6]. This method is flexible and the required double-orthogonal wavelet can be easily established. However, the ECG signal is not always smooth and simply applying the lifting scheme will cause signal distortion. To solve this problem, Claypoole.R and Baraniuk.R proposed the Adaptive Lifting Scheme (ALS) [7]. Without converting data into frequency domain through time- and memory- consuming Fourier transform, the ALS processes data directly in the spatial domain. The wavelet coefficients are updated locally, thus no extra memory are needed for saving the intermediate values. Because of its effectiveness, the ALS is more suitable for hardware implementation.



**Fig. 1** Schematic representation of the QRS complex

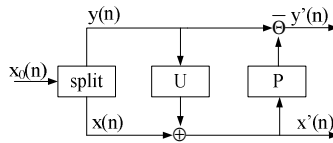
In this paper, the FPGA implementation of a real time QRS detection system based on the ALS method is presented. This system takes the digitized ECG signal as input, and detects the presence of the QRS complex. The implemented system is tested using ECG samples from the MIT-BTH database [8], and the effectiveness of the system is validated by showing a near-perfect detection rate. With continued research efforts, other processing modules can be added to the system for expanded functionalities.

This paper is divided into five sections. In section 2, the ALS algorithm used in this system for ECG data compression is briefly discussed. Section 3 presents the FPGA implementation of the system. The system is validated by directly feeding the digitized ECG samples, and the experimental results are summarized in section 4. The concluding remarks are given in section 5.

## 2 ALS for ECG Signal

For the digitized ECG sequence, the correlation structure is typically local, thus data compression method is generally applied to reduce the required memory cells in hardware implementation. In this study, the ALS is utilized to compress the ECG signal.

The ALS is an adaptive signal processing scheme, and it adaptively exploits correlation among neighboring samples to build a sparse approximation. Unlike traditional wavelet constructions, the ALS is entirely spatial and requires much less number of calculations [9]. It is also able to find the localized signal transitions, thus is advantageous in processing non-stationary time-varying signals, such as the ECG signals.



**Fig. 2** Illustration of processing ECG sequence using the ALS

Fig.2 illustrates the data processing steps of the ALS, and symbol  $\oplus$  and  $\ominus$  in the figure are the generalized addition and subtraction operators. Data compression using the ALS can be divided into three steps: split, update, and predict.

The original data sequence  $x_0(n)$  is first split into the two signal set  $x(n)$  and  $y(n)$ .  $x(n)$  includes the even indexed samples of  $x_0(n)$ , and  $y(n)$  includes the odd indexed samples of  $x_0(n)$ .  $x(n)$  and  $y(n)$  can then be written as

$$x(n) = x_0(2n), \quad y(n) = x_0(2n + 1) \tag{1}$$

Contrary to an ordinary lifting scheme, in the ALS, the wavelet transform is performed by first updating and then predicting. The updating process generates  $x'(n)$  from  $y(n)$  through an updating operation  $U$ .  $x'(n)$  is effectively an low-pass filtered approximation of  $x(n)$ , and therefore contains less noise power. Using  $x'(n)$ , instead of  $x(n)$ , in the predicting process thus yields better accuracy. The data set  $y'(n)$  generated from the predicting operation  $P$  contains the detailed signal information, and is used directly for QRS detection.

Following the procedure, the input ECG samples are essentially compressed by a factor 2.

In this work, the predictor P and updater U are defined as in (2) and (3),

$$P(n) = \frac{1}{2}[x'(n-1) + x'(n)] \tag{2}$$

$$U(n) = \begin{cases} y'(n) & del1 > del2 \\ \frac{1}{2}[y'(n-1) + y'(n)] & del1 = del2 \\ y'(n-1) & del1 < del2 \end{cases} \tag{3}$$

$$del1 = |x'(n) - y'(n-1)| \quad del2 = |x'(n) - y'(n)|$$

### 3 Implementation of the Real-Time Detection System

A real time QRS complex detection system is implemented using the FPGA. The digitalized ECG input samples are first compressed using the ALS, and only the predicted data sequence  $y'(n)$  is used for signal analysis. To detect the QRS complex, a threshold value is directly subtracted from  $y'(n)$ , and a QRS complex is detected if the subtraction returns a positive result. The threshold value is adaptively generated from  $y'(n)$  to remove the DC drift within the samples.

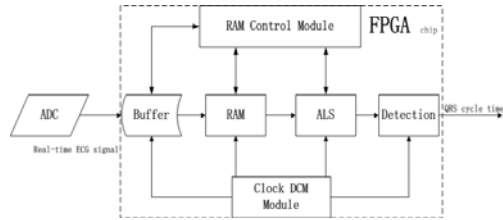


Fig. 3 Systematic diagram of the real time QRS detection system

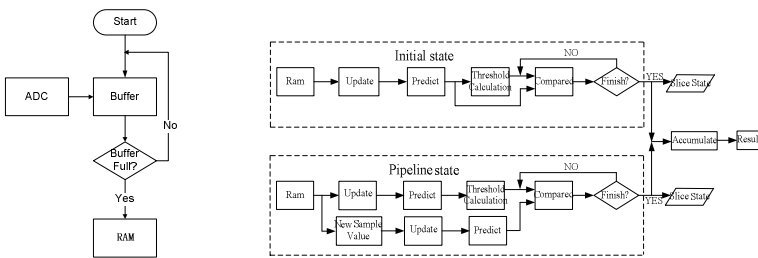


Fig. 4 Data flow: (a) Data receiving flow in the Buffer, (b) processing and analysis flow

The systematic level diagram of the FPGA based QRS detection system is shown in Fig.3. The blocks implemented on the FPGA include a Buffer stage, the RAM for data storage, the ALS module to compress the stored data, and the QRS detection module.

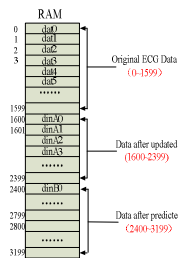


The Buffer stage, which is able to store 1600 digitized samples, acts as the interface between the analog-to-digital converter (ADC) and the data processing unit. This interface is necessary because generally the sampling frequency for the ECG signal is much lower compared with the FPGA clock rate. In its initial state, the Buffer is empty, and the digitized samples are saved to the Buffer in a sequential manner. Once the Buffer is full (buffered samples reach 1600), data are replicated to the RAM for processing. As the data sequence is continuously pushed into the Buffer, the oldest data are pushed out, and the Buffer is updated constantly. The system monitors the input data, and every time the new 100 samples are pushed in, the data in the Buffer, including the 1500 old samples and 100 new samples, are pipelined to the RAM. The data receiving flow is summarized in Fig.4(a).

Fig. 4(b) shows the data processing and analysis flow. The initial state represents the operating phase that the first 1600 digitized samples are copied into the RAM. In both the initial state and pipeline state, data are first split into even-indexed and odd indexed sets based on their memory address. The predictor and updater (P and U) are then adaptively generated in the ALS module. Within each operating phase, only the predicted result, which is the compressed version of the original input sequence, is used for QRS detection.

The required threshold value is also adaptively generated from the compressed data. The compressed data is divided into eight sections, with each containing 50 samples. To calculate the threshold, the maximum values in all sections are averaged, and the threshold value is set as one half of the averaging result. The threshold is calculated every time the data sequence is pipelined to the RAM, and therefore the DC drift in the ECG signal can be removed.

The compressed data from the ALS module is compared with the generated threshold to detect the QRS complex. An accumulator monitors the comparison result, and if the data is larger than the threshold, the accumulator output is increased by 1. The comparison continuously takes place, and it only stops when all compressed data in the current operating phase has been processed.



**Fig. 5** RAM address allocation in the implemented system

Totally 3200 RAM cells are required for the data processing, and the RAM address map is shown in Fig. 5. RAM cells with address from 0 to 1599 are used

to store the original data copied from the Buffer stage, and RAM cell 1600-2399 and 2400-3199 are allocated for the updated and the predicted data, respectively. Without allocating extra memory cells for intermediate calculation values, data stored in the RAM cells are read by the ALS module, processed in real time, and saved back to their original addresses. Only the predicted data stored in RAM cell 2400-3199 are used by the data comparison module for QRS detection.

## 4 Experimental Results

A real time QRS detection system is implemented using the FPGA device XUPV5-LX110T (Package FF1136) with speed grade -3. The Virtex5-LX110T chip is fabricated using CMOS 65 nm technology, and it includes 69120 logic cells, 640 I/O channels and 16 clock regions [10].

Behavioral models of the processing modules are first written in the Verilog HDL, and then downloaded to the FPGA chip to realize the detection system. The platform uses a XCF32P which includes 32Mbyte of flash PROMs. The hardware resource occupation of the implemented system on the FPGA platform is summarized in Table 1. It shows that the system only utilizes 1.3% of the hardware resource that the FPGA can provide, and therefore the functionality of the implemented system can be easily expanded by integrating other ECG analysis modules. The dynamic power consumption of the implemented system can also be estimated. With the FPGA clock rate set to 50 MHz, the system only consumes ~20mW of power.

**Table 1** Summary of the FPGA Hardware utilization

<i>Resource Name</i>	<i>Occupied</i>	<i>Total Number</i>	<i>Proportion</i>
Slices	877	69120	1.3%
Slice Flip Flops	856	69120	1.2%
LUTs	1705	69120	2.5%
I/O	64	640	10.0%
GCLK	1	32	3.1%

**Table 2** Summary of the Test Result

<b>Sample</b>	<b>BN</b>	<b>FP</b>	<b>FN</b>	<b>Accuracy (%)</b>
100	2272	0	7	99.692
101	1865	0	7	99.623
103	2084	0	7	99.663
104	2229	23	0	98.968
105	2572	0	48	98.134
107	2137	0	0	100
108	1774	0	163	90.812
109	2534	0	17	99.329
111	2127	0	17	99.201
112	2539	0	8	99.685
114	1879	0	18	99.042
Total	24012	23	292	98.688

To test the efficacy of this FPGA-based system, ECG samples from the MIT-BTH Arrhythmia Database are utilized. Totally 11 samples with identification numbers ranging from 100 to 114 are utilized for the test, with each lasting about 30 minutes. The test results are summarized in Table 2. In the table, BN represents the actual number of the QRS complexes, and FP and FN are the number of falsely detected and miss-detected waves. The detection accuracy is defined as the ratio between the numbers of the detected waves to the beat number, BN. It can be seen from the table that the detection accuracy varies with different samples, but in all the tested cases, the detection rate is higher than 90%, and the overall detection accuracy reaches 98.688%. In the experimental results, the slightly lower detection accuracy for sample 108 might be caused by the large noise embedded in the original signal (A small portion of the sample is plotted in Fig. 6). The detection rate for this sample might be able to be improved if additional filtering devices are included in the system.

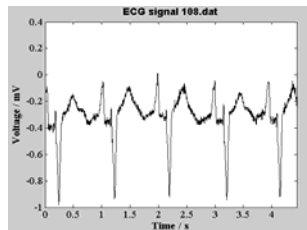


Fig. 6 Sample 108 with baseline wandering noise

## 5 Conclusion

This paper presents a real-time data processing system implemented on the FPGA for QRS complex detection. The system utilizes the ALS to compress the digitized ECG data, and therefore minimizes the required memory cells in hardware implementation. Novel pipeline architecture is used as the interface in order to synchronize the data flow between the ADC and the FPGA system. Implemented on the Virtex-5 FPGA platform, the system only utilizes 1.3% of the overall hardware resource, thus allows system functionality to be expanded in the future versions. The effectiveness of the system is validated using the ECG samples from the MIT-BTH arrhythmia database. The test results show that the overall system detection accuracy reaches 98.688% for ordinary QRS complexes.

**Acknowledgment.** This work is partially supported by NSFC 60901016, GDSF S2011040001460, SRF for ROCS, SEM, and the SZU R/D Fund 801000172.

## References

1. Yang, Y., Huang, X., Yu, X.: Real-Time ECG Monitoring System Based on FPGA. In: 33rd Annual Conference of the IEEE Industrial Electronics Society, pp. 2136–2140 (2007)

2. Ibrahimy, M.I., Reaz, M.B.I., Mohd-Yasin, F., Khoon, T.H., Ismail, A.F.: Fetal QRS Complex Detection Algorithm for FPGA Implementation. In: CIMCA-IAWTIC, pp. 846–850 (2005)
3. [http://en.wikipedia.org/wiki/qrs\\_complex](http://en.wikipedia.org/wiki/qrs_complex)
4. Hsieh, J.C., Yu, K.C., Chuang, H.C., Lo, H.C.: The Clinical Application of an XML-Based 12 Lead ECG Structure Report System. *Computers in Cardiology*, 533–536 (2009)
5. Wang, X., Zhao, H.: A Novel Synchronization Invariant Audio Watermarking Scheme Based on DWT and DCT. *IEEE Transactions on Signal Processing* 54(12), 4835–4840 (2006)
6. Daubechies, I., Sweldens, W.: Factoring wavelet transform into lifting steps. *Journal of Fourier Analysis and Applications*, 245–267 (1998)
7. Claypoole, R., Baraniuk, R., Nowak, R.: Adaptive Wavelet Transforms via Lifting. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1513–1516 (May 1998)
8. MIT-BIH Arrhythmia database, <http://www.physionet.org/physiobank/database/mitdb/>
9. Sweldens, W.: The lifting scheme: A construction of second generation wavelets. *Siam Journal on Mathematical Analysis*, 511–546 (1998)
10. Xilinx Company, Virtex-5 FPGA User Guide

# A Novel Temporal-Spatial Color Descriptor Representation for Video Analysis

Xiang-wei Li, Gang Zheng, and Kai Zhao

**Abstract.** A novel temporal-spatial color descriptor representation is developed based on Rough Sets (RS) in compressed domain. Firstly, DCT coefficients and DC coefficients are extracted from original video sequences. Secondly, an Information System Table is constructed using DC coefficients. Thirdly, a concise Information System Table is achieved by using the reduction theory of RS. The Core contains important visual color information and eliminates the redundant video information. At the same time, DC coefficients also contain spatial information of each frame, so the Core of Information System can regard as effective temporal-spatial color descriptor.

## 1 Introduction

Developed with the international standard of Mpeg-1 and Mpeg-2, There has been a rapid increase of video volume in past few years. Mpeg-4 introduced the concept of Audio-Video (AV) object, which enhanced the storage efficiency of video data, and video information were efficiently indexed and retrieved by various content based indexing and retrieval algorithm [2]. The major objective of MPEG-7 standard is to allow interoperable searching, indexing, filtering, and access of AV content by enabling interoperability among devices and applications that deal with AV content description. Mpeg-7 describes specific features of AV content, as well as information related to AV content management [3].

For the convenience of video analysis, many effective visual descriptors have defined by international standards Mpeg-7 [4]. Among the low level features, color is the most effective visual content description, especially, the use of low-level

---

Xiang-wei Li · Gang Zheng · Kai Zhao  
Lanzhou Polytechnic College, Lanzhou 730050, China

Xiang-wei Li  
Key Laboratory of Gansu Advanced Control for Industrial Processes, Lanzhou 730050

visual color features to find interesting information from video database or data warehouse has drawn much attention in recent years. More detailed information regarding the color descriptors in Mpeg-7 may be found in the references and other related Mpeg document. Color histogram is also the most widely used color descriptor in video content. A color histogram captures global color distribution in a frame of video sequences, but the performance is quit poor due to the lack of spatial color distribution information. Among various Mpeg-7 color visual descriptors, Color Structure Descriptor (CSD) always gives the good retrieval rate. The main merits are it can use spatial color distribution information in CSD. However, CSD has redundancy information as it uses a fixed color space for the histogram representation. To solve this limitation, a novel data analysis tool is introduced, i.e., attributes reduction theory of RS, which has successfully been used in many application domains, such as machine learning, expert system and pattern classification [5, 6]. The main advantage of rough sets is that it does not need any preliminary or additional information about data, like probability in statistics, or basic probability assignment in Dempster-Shafer theory and grade of membership or the value of possibility in fuzzy sets [7, 8]. It can effectively overcome the redundant information and preserve the temporal spatial information of video sequences.

## 2 Basic Theory of Rough Sets

### 2.1 Indiscernibility Relation

Let  $U$  be a universe of discourse and  $X$  be a subset of  $U$ . An equivalence relation,  $R$ , classifies  $U$  into a set of subsets  $X_i$  in which the following conditions are satisfied:

- (1)  $X_i \subseteq U, X_i \neq \emptyset$  For any  $i$ .
- (2)  $X_i \cap X_j = \emptyset$  For any  $i, j$ .
- (3)  $\bigcup_{i=1,2,\dots,n} X_i = U$

Any subset, which called a category, class or granule, represents an equivalence class of  $R$ . A category in  $R$  containing an object  $x$  is denoted by  $[x]_R$ . For a family of equivalence relations, an indiscernibility relation  $P \subseteq R$  over  $P$  is denoted by  $IND(P)$  and is defined by (1).

$$IND(P) = \bigcap_{R \in P} IND(R) \quad (1)$$

### 2.2 Lower and Upper Approximations

The set can be divided according to the basic sets of, namely a lower approximation set and upper approximation set. Approximation is used to represent the roughness of the knowledge. Suppose a set represents a vague concept, then the  $R$ -lower and  $R$ -upper approximations of  $X$  are defined by equation (2) and equation (3).

$$\underline{R}X = \{x \in U : [x]_R \subseteq X\} \tag{2}$$

Equation (4) is the subset of X, such that X belongs to X in R, is the lower approximation of X.

$$\overline{R}X = \{x \in U : [x]_R \cap X \neq \emptyset\} \tag{3}$$

Equation(5) is the subsets of all X that possibly belong to X in R, thereby meaning that may or may not belong to X in , and the upper approximation contains sets that are possibly included in X. R-positive, R-negative, and R-boundary regions of X are defined respectively by equation(4), equation(5) and equation(6).

$$POS_R(X) = \underline{R}X \tag{4}$$

$$NEG_R(X) = U - \overline{R}X \tag{5}$$

$$BNR(X) = \overline{R}X - \underline{R}X \tag{6}$$

### 2.3 Attributes Reduction and Core

In RS theory, an Information Table is used for describing the object of universe, it consists of two dimensions, each row is an object, and each column is an attribute. RS classifies the attributes into two types according to their roles for Information Table: Core attributes and redundant attributes. Consequence, the minimum condition attribute set can be obtained, which is called reduction. One Information Table might have several different reductions simultaneously. The intersection of the reductions is the Core of the Information Table and the Core attribute are the important attribute that influences attribute classification.

A subset B is a reduction of C with respect to R if and only if

$$(1) POS_B(R) = POS_C(R), \text{ and}$$

$$(2) POS_{B-\{a\}}(R) \neq POS_C(R), \text{ For any } a \in B$$

And, the Core can be defined by equation (7)

$$CORE_C(R) = \{c \in C \mid \forall c \in C, POS_{C-\{c\}}(R) \neq POS_C(R)\} \tag{7}$$

## 3 The Developed Method

### 3.1 Extracting DCT Coefficients

In term of Mpeg international video standard, the video sequences in compressed domain consist of I, P and B frame. The I frame is the base of video information in sequence, which uses DCT to compress in spatial, so the DCT coefficients can represent the full video information and be easily extracted from video sequences directly. We can represent this process as following:

$$\psi(P(x),t) \xrightarrow{\text{extract}} \text{DCT coefficients}$$

Where  $\psi(P(x),t)$  denotes the video sequences.

### 3.2 Extracting DC Coefficients

The DCT coefficients are made of DC coefficients and AC coefficients, DC coefficients denote the average and most important visual information in video frame. So we can utilize the DC coefficients to represent the video frame. This process can be described as following:

$$\text{DCT coefficients} \xrightarrow{\text{reprecasing}} \text{DC coefficients}$$

### 3.3 Constructing Information System

We have got the DC coefficients of each frame, so we can construct an Information System Table using it. Each row is a DC coefficient, and each column is a frame. The process can be described as following.

$$\text{DC} \xrightarrow{\text{construct}} \text{Information System Table } S = \{U, A, V, f\}$$

Where U is sets, denotes all the object of Information System, A is also a sets, denotes all attributes in Information System, V is the sets of attributes value, f is a function denotes the relations between objects and attributes.

3.4 Reduction of Information System.  
According to 2.3, the attributes in the Information Table can be divided into two types according to their roles: Core attributes and redundant attributes. The most important information is the Corsets of Information System. The Corsets represent Core of Information System. Because the Corsets are the frames that can not be reduced, so it is the salient and interesting content in video sequences. Table 1 is the classical reduced Information System using above method.

**Table 1** The reduced Information System

frame	frame1	frame3	frame4	frame5	frame6	frame7
DC1	0.35355	0.27771	0.35355	0.41572	0.35355	0.35355
DC2	0.35355	0.27779	0.35355	0.41573	0.35354	0.35355
DC3	0.35355	0.27779	0.35356	0.19134	0.35355	0.35355
DC4	0.35356	0.27779	0.35355	0.19134	0.35351	0.35355
DC5	0.35355	0.27779	0.35355	0.19134	0.35336	0.35355
DC6	0.27779	0.27779	0.35356	0.19134	0.35345	0.35355
DC7	0.27779	0.27779	0.35355	0.19134	0.35353	0.35355
DC8	0.23761	0.27779	0.35355	0.35355	0.35355	0.35355



### 3.4 Generating Temporal-Spatial Color Descriptor

Core is the most important attributes in Information System that can not be reduced, which represents the most important and interesting information in video sequences. Since Core eliminate many redundant frames, the volume of video shortened dramatically. At the same time, the DC coefficients of row in reduced Information System contain not only the color feature, but also the spatial information of each frame and the temporal information. So the achieved Information System Table can regard as good descriptor for video retrieval.

## 4 Experimentation

A serial of Mpeg video sequences are selected to examine the performance of proposed algorithm. Table 2 shows the detailed evaluation results of five different video sequences.

**Table 2** Detailed evaluation results of four sports video

Video shot	Frames in original shot	Frame in core	evaluation
No.1	112	17	good
No.2	174	33	good
No.3	317	42	excellent
No.4	126	13	good

The experiment results demonstrate that proposed algorithm can effective and scientific generates temporal-spatial color descriptor in compressed domain. It considered not only temporal information by using a series of frame, but also spatial information by using DC coefficients in each frame. In addition, all operation accomplished in the compressed domain, the amount of data reduced dramatically, so the algorithm is very efficient and scientific.

## 5 Conclusions

In this paper, a novel temporal-spatial color descriptor extraction algorithm is developed. DCT coefficients and DC coefficients are extracted directly from raw video sequences, and an Information System constructed to represent video shot, then attributes reduction theory of RS is introduced for redundant data processing. At last, a concise and important Core of Information System generated to describe temporal-spatial color descriptor. Our experiments have shown that better results are achieved compared to previous technologies.

**Acknowledgments.** The work is supported by the Gansu Natural Science Foundation of china under grant (No. 1107RJZA170).

## References

- [1] Dorairaj, R., Namuduri, K.R.: Compact combination of MPEG-7 color and texture descriptors for image retrieval. In: Conference record of the Thirty-Eighth Asilomar Conference on Signals, Systems & Computers, vol. 1, pp. 387–391 (2004)
- [2] Sikora, T.: The MPEG-7 visual standard for content description-an overview. *IEEE Transactions on Circuits and Systems for Video Technology* 11(6), 696–702 (2001)
- [3] Salembier, P., Smith, J.R.: MPEG-7 multimedia description schemes. *IEEE Transactions on Circuits and Systems for Video Technology* 11(6), 748–759 (2001)
- [4] Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V.: Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11(6), 703–715 (2001)
- [5] Kotlowski, W., Dembczynski, K., Greco, S.: Stochastic dominance-based rough set model for ordinal classification. *Information Sciences* 178(21), 4019–4037 (2008)
- [6] Parthalaïn, N.M., Shen, Q.: Exploring the boundary region of tolerance rough sets for feature selection. *Pattern Recognition* (2008) (in Press) (accepted manuscript)
- [7] Wang, C., Wu, C., Chen, D.: A systematic study on attribute reduction with rough sets based general binary relations. *Information Sciences* 178(9), 2237–2261 (2008)
- [8] Yao, Y.Y., Zhao, Y.: Attribute reduction in decision-theoretic rough set models. *Information Sciences* 178(17), 3356–3373 (2008)

# Author Index

- Abdullah, Saad 205  
Ahmed, Faiyaz 219  
Aleksovska-Stojkovska, Liljana 287  
Al-Salman, H. 475  
Al-Zuair, M. 475  
Anh, Duong Tuan 93  
Anh, Truong Ngoc 321  
Ariffin, Noor Afiza Mohd 489  
Atan, Rodziah 489
- Bai, Yong 255  
Bergelt, Rene 193  
Buailes, Jovani Alberto Jiménez 359
- Cai, Shi-Min 121  
Cao, Qian 407  
Chakraborty, Rudraneel 219  
Chan, Chien-Hui 315  
Chan, Chin-Hong 337  
Chang, Chung-Yi 315  
Chang, Hsien-Tsung 115  
Chbeir, Richard 449  
Chen, Gang 225  
Chen, Ja-Hao 337  
Chen, Xue 269  
Cheng, Chen-Yang 337  
Chiang, S.-Y. 461  
Choi, Bumkyoo 371  
Choi, Jeongheon 379  
Colace, Francesco 181
- Dali, Guo 175  
De Santo, Massimo 181  
Di, Wu 1
- Dong, Chunjiao 15  
Dy, Ed Darcy 401
- Ebecken, Nelson F.F. 295  
Ekštejn, Kamil 79  
El-Zaart, A. 475  
Eom, Jeongho 371
- Faisal, Zaman Md. 47  
Fan, Yang-Yang 269  
Fang, Miao 231  
Fei, Wang 497  
Feng, Deying 55  
Feng, Min 261  
Fu, Lidong 23  
Fu, Zhong-Qian 121  
Fujii, Takashi 469  
Fujiyoshi, Hironobu 469  
Fukuta, Naoki 449
- Gao, Lin 23  
Gao, Min 441  
Ghani, Abdul Azim Abd 481  
Glockner, Matthias 193
- Haiguo, Zhu 429  
Hamuro, Yukinobu 187  
Harazaki, Manami 449  
Hardt, Wolfram 193  
He, Huaidong 129  
Hirose, Hideo 47  
Hong, Lei 121  
Hori, Yukio 163  
Hosain, Shazzad 205, 219  
Hsieh, Nan-Chen 315

- Hu, Xiaojun 275  
 Huang, Cheng-Ta 135  
 Huang, Po-Han 41  
 Huang, Xiao 225  
 Huang, Yu-Hsun 337  
 Huda, Md. Nurul 249  
 Hui, Zhang 429  
 Hussain, Aini 281
- Imai, Yoshiro 163  
 Ishii, Norio 469  
 Ishikawa, Hiroshi 449
- Ji, Zhen 497  
 Jiang, Feng 441  
 Jiang, Lai 497  
 Jinyun, Yao 35  
 Jiuyang, Hou 1
- Kaiser, Gail 349  
 Kan, Y.-C. 461  
 Kantola, Jussi 329  
 Karlström, Per 199  
 Keh, Huan-Chao 315  
 Khalilian, Madjid 73  
 Koo, Jajin 379  
 Kosaka, Manabu 393  
 Krčmář, Lubomír 79  
 Kuo, Su-Hui 321
- Lai, Yingxu 211  
 Lan, Tsuo-Hung 337  
 Lan, Zhangli 255  
 Lang, Bo 239  
 Lee, Moon Kyu 371  
 Lee, Seongjin 379  
 Lee, Wen-Tin 41  
 Li, Li 321  
 Li, Qiang 407  
 Li, Wei 239  
 Li, Xiang-wei 505  
 Li, Yan 497  
 Li, Zhi-Yu 269  
 Lin, Ge 413  
 Lin, H.-C. 461  
 Liu, Dake 199  
 Liu, Hongnan 211  
 Liu, Jonathan C.L. 135  
 Liu, Shu-Wei 115  
 Liu, Xiang-dong 421
- Liu, Xianglong 239  
 Liu, Yinhong 15  
 Liu, Yufeng 225  
 Loskovska, Suzana 287  
 Luna, Jaime Alberto Guzmán 359  
 Luo, Jun 275
- Ma, Lixiao 497  
 Maceren, Silvin Federic 401  
 Mahmud, M. Sultan 205  
 Marcial, Dave E. 401  
 Mathkour, H. 475  
 Matsumoto, Kazunori 143, 155  
 Meng, Xiangzhong 87  
 Meng, Xiao-Bo 269  
 Minyan, Lu 413  
 Mohd Zainal, Mohd Ridzuwary 281  
 Monira, Sumi S. 47  
 Moors, Tim 435  
 Morita, Hiroyuki 187
- Nakayama, Takashi 163  
 Napoletano, Paolo 181  
 Nguyen, Man 61  
 Nie, Rongmei 275
- Okamoto, Kouki 155
- Pardo, Ingrid Durley Torres 359  
 Phan, Doan 61  
 Pingkun, Yan 497
- Qazi, Sameer 435  
 Qiang, Li 1  
 Qiu, Ming-yi 307  
 Quanyi, Huang 429
- Reyad, Y.A. 475  
 Rosidi, Norazimah 481
- Safarinejadian, Behrouz 29  
 Saga, Ryosuke 155  
 Samad, Salina Abdul 281  
 Sani, Nor Fazlida Mohd 481, 489  
 Santos, Fatima F. 295  
 Sarno, Erivn Rygl 401  
 Segev, Aviv 329  
 Shao, Chunfu 15  
 Shen, Naiqi 225  
 Shengcheng, Yuan 429  
 Sheth, Swapneel 349

- Son, Nguyen Thanh 93  
Sonehara, Noboru 249  
Song, Jia 261  
Sugimura, Hiroshi 143  
Sun, Hung-Min 385  
Sun, Wen-wen 107  
Suzuki, Yuri 469
- Tan, Arie H. 343  
Tan, Fabian H. 343  
Tekli, Joe 449  
Tiejun, Li 175  
Tran, Tan 61  
Tsuji, Hiroshi 155  
Tu, Y.-C. 461
- Vodel, Matthias 193
- Wang, Jianghong 87  
Wang, Lei 129  
Wang, Ru 497  
Wang, Shih-Jeng 135, 385  
Wang, Tien-Chin 321  
Wang, Wei-Jen 135  
Wang, Yingshuang 225  
Wang, Zhe 107  
Wang, Zhijun 441  
Watanabe, Toshinori 365  
Weng, Chi-Yao 385  
Wenhui, Yang 231  
Won, Youjip 379
- Xiaoyan, Lu 35  
Xijun, Ke 175  
Xu, Xi-dong 307
- Yamada, Shigeki 249  
Yan, Xie 231  
Yang, Cheng 55  
Yang, Cheng-Hsing 385  
Yang, Erlong 129  
Yang, Jie 55  
Yang, Shunkun 413  
Yang, Zhen 211  
Yeh, Dowming 41  
Yi, Liu 429  
Yin, Fang 261  
Yokoyama, Shohei 449  
Yoo, Keedong 9  
Yu, Hang 497  
Yunxiu, Zhao 429
- Zhang, Nuo 365  
Zhao, Bo 275  
Zhao, Dan 15  
Zhao, Guanwei 101  
Zhao, Hui 275  
Zhao, Kai 505  
Zheng, Gang 505  
Zhihao, Tang 175  
Zhou, Chun-guang 107  
Zhou, Pei-Ling 121  
Zhou, Tong 107  
Zhou, Wenbiao 199  
Zhou, Xian 269  
Zhou, Yi 255  
Zhu, Hai 269  
Zhu, Li-ye 421