Hua Wang Lei Zou
Guangyan Huang Jing He
Chaoyi Pang Haolan Zhang
Dongyan Zhao Yi Zhuang (Eds.)

LNCS 7234

# Web Technologies and Applications

APWeb 2012 International Workshops:
SenDe, IDP, IEKB, MBC
Kunming, China, April 2012, Proceedings

Springer

Lecture Notes in Computer Science 7234

*Commenced Publication in 1973*
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Hua Wang Lei Zou Guangyan Huang
Jing He Chaoyi Pang Haolan Zhang
Dongyan Zhao Yi Zhuang (Eds.)

# Web Technologies and Applications

APWeb 2012 International Workshops:
SenDe, IDP, IEKB, MBC
Kunming, China, April 11-13, 2012
Proceedings

Springer

Volume Editors

Hua Wang
University of Southern Queensland, Toowoomba, QLD, Australia
E-mail: wang@usq.edu.au
Lei Zou
Peking University, Beijing, China
E-mail: zoulei@icst.pku.edu.cn
Guangyan Huang
Victoria University, Melbourne, Australia
E-mail: guangyan.huang@vu.edu.au
Jing He
Victoria University, Melbourne, Australia
E-mail: jing.he@vu.edu.au
Chaoyi Pang
Australian eHealth Research Centre, Herston, QLD, Australia
E-mail: chaoyi.pang@csiro.au
Haolan Zhang
NIT, Zhejiang University, China
E-mail: haolan.zhang@nit.zju.edu.cn
Dongyan Zhao
Peking University, Beijing, China
E-mail: zhaody@pku.edu.cn
Zhuang Yi
Zhejiang Gongshang University, Hangzhou, China
E-mail: zhuangyi@zjuem.zju.edu.cn

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-29425-9 e-ISBN 978-3-642-29426-6
DOI 10.1007/978-3-642-29426-6
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2012934825

CR Subject Classification (1998): H.3, I.2, H.4, C.2, J.1, H.5

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI



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

Printed on acid-free paper

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

# General Chairs' Welcome Message

Welcome to APWeb 2012, the 14th Edition of Asia Pacific Web Technology Conference. APWeb is a leading international conference on research, development, and applications of Web technologies, database systems, and information management and software engineering, with a focus on the Asia-Pacific region. Previous APWeb conferences were held in Beijing (1998), Hong Kong (1999), Xian (2000), Changsha (2001), Xian (2003), Hangzhou (2004), Shanghai (2005), Harbin (2006) Huangshan (2007), Shenyang (2008), Suzhou (2009), Busan (2010), and Beijing (2011).

APWeb conferences cover contemporary topics in the fields of Web management and WWW-related research and applications, such as advanced application of databases, cloud computing, content management, data mining and knowledge discovery, distributed and parallel processing, grid computing, Internet of things, Semantic Web and Web ontology, security, privacy and trust, sensor networks, service-oriented computing, Web community analysis, Web mining and social networks.

The APWeb12 program featured a main conference and five satellite workshops. The main conference had two eminent keynote speakers—Patrick McDaniel from Pennsylvania State University, USA, and Vijay Varadharajan from Macquarie University, Australia—39 full papers, 34 short papers, and 5 demo papers.

The five workshops reported on some recent developments and advances in contemporary topics in the related fields of: Information Extraction and Knowledge Base Building (IEKB) Data Mining for Information Security and Privacy Intelligent Data Processing (IDP 2012) Sensor Networks and Data Engineering (SenDe 2012), and Mobile Business Collaboration (MBC 2012).

Both main conference program and workshop sessions were of high quality owing to the strong support and commitment from their international Program Committee. We wish to take this opportunity to thank Program Co-chairs Michael Sheng, Guoren Wang and Christine S. Jensen for their dedication and effort in ensuring a high-quality program. I would also like to thank Workshop Chairs Hua Wang and Lei Zou, and each workshop organizer, for their contribution to developing an interesting and attractive workshop program.

Many colleagues helped toward the success of APWeb 2012. They are Local Arrangements Co-chairs: Jianfeng He and Guangyan Huang; Financial Chair: Jing He; Publication Chair: Guandong Xu; Industry Chairs: Gary Morgan and Qingzhong Li; Demonstration Chair: Chaoyi Pang ; Publicity Co-chairs: Haolan Zhang and Jiangang Ma; and Webmasters: Zhi Qiao and Zhangwei Jiang.

We would like to sincerely thank our financial supporters and sponsors. The following organizations generously supported and sponsored APWeb 2012: Hebei University of Engineering, Nanjing University of Finance and Economics, National Science Foundation of China, Kunmin University of Technology, Graduate University of Chinese Academy of Science, and Victoria University.

We wish also to thank the APWeb Steering committee led by Xuemin Lin for offering the opportunity to organize APWeb 2012 in Kunming.

Finally, we wish to thank the host Kunming University of Technology and the local Arrangements Committee and volunteers for the assistance in organizing this conference. The following members helped with the registration, accommodation, and various logistics: Jing Yang, Xun Zhou, and Shang Hu.

April 2012

Xaoxue Zhang
Yanchun Zhang
Masaru Kitsuregawa

# Preface

Following the tradition of the APWeb conference, a leading international conference on research, development and applications of Web technologies, database systems, information management and software engineering, there were several satellite workshops of the 14th Asia Pacific Web Conference Workshops held in Kunming on April 11, 2012. This year, four workshops were organized in conjunction with APWeb 2012, coving a wide range of topics, such as information extraction, information security, intelligent data processing, sensor network and business intelligence. These topics play very important roles in creating the next-generation information technology architectures and solutions.

The high-quality program would not have been possible without the authors who chose APWeb 2012 workshops as a venue for their publications. This proceedings volume compiles the papers selected for presentation at the following four workshops:

- The First Workshop on Sensor Networks and Data Engineering(SenDe 2012)
- The First International Workshop on Intelligent Data Processing(IDP 2012)
- Workshop on Information Extraction and Knowledge Base Building(IEKB 2012)
- The Second International Workshop on Mobile Business Collaboration(MBC 2012)

We are very grateful to the workshop organizers and Program Committee members who put a tremendous amount of effort into soliciting and selecting research papers with a balance of high quality, new ideas, and novel applications. Furthermore, we would like to take this opportunity to thank the main conference organizers for their great effort in supporting the APWeb 2012 workshops.

We hope that you enjoy reading the proceedings of the APWeb 2012 workshops.

April 2012

Hua Wang
Lei Zou

# Message from the SenDe 2012 Co-chairs

Welcome to SenDe 2012, the first Sensor Networks and Data Engineering workshop. Data become more and more critical in our daily life, especially the data gathered by sensors from the physical world and data generated by people on the Web (virtual world). SenDe 2012 focused on discussing research problems arising from the following three aspects:

**(1) Wireless sensor networks and sensor data analysis for risk monitoring in the physical world.** Paulo de Souza was invited to give a talk—"Technological Breakthroughs in Environmental Sensor Networks"—which focused on the development of better sensors and more efficient sensor networks for monitoring environments. Wei Guo, Lidong Zhai, Li Guo, and Jinqiao Shi proposed a worm propagation control based on spatial correlation to resolve the security problems in wireless sensor network. Li Yang, Xiedong Cao, Jie Li, Cundang Wei, Shiyong Cao, Dan Zhang, Zhidi Chen, and Gang Tang modelled the security problem in the SCADA system for oil and gas fields into an online digital intelligent defense process. Xiedong Cao, Cundang Wei, Jie Li, Li Yang, Dan Zhang, and Gang Tang designed an expert system to forecast the potential risks of geological disasters through analyzing the real-time data of the large-scale network in the SCADA system.

**(2) Efficient data processing for query and knowledge discovery.** Ziqiang Deng, Husheng Liao, and Hongyu Gao presented a hybrid approach combining finite state machines and a holistic twig matching algorithm for memory-efficient twig pattern matching in XML stream query. Hongchen Wu, Xinjun Wang, Zhaohui Peng, Qingzhong Li, and Lin Lin provided a PointBurst algorithm for improving social recommendations when there are no, or too few, available trust-relationships. Yang Li, Kefu Xu, Jianlong Tan, and Li Guo provided a method of group detection and relation analysis for Web social networks. Xingsen Li, Zhongbiao Xiang, Haolan Zhang, and Zhengxiang Zhu develop a method for extension transformation knowledge discovery through further digging on the basis of traditional decision trees. Zhuoluo Yang, Jinguo You, and Min Zhou provided a two-phase duplicate string compression algorithm for real-time data compression.

**(3) Distributing storage and balancing workload in the cloud.** Keyan Liu, Shaohua Song, and Ningnan Zhou provided a flexible parallel runtime mechanism for large-scale block-based matrix multiplication for efficient use of storage. Xiling Sun, Jiajie Xu, Zhiming Ding, Xu Gao, and Kuien Liu propose an efficient overload control strategy for handling workload fluctuations to improve service performances in the cloud.

We accepted 10 from 27 submissions, with an acceptance rate of 37%. Each submission was carefully reviewed by at least three Program Committee members. We wish to thank the invited speaker: Paulo de Souza (ICT Centre, CSIRO, Australia), all of the authors, participants, and Program Committee members of the SenDe 2012 workshop.

April 2012

Jing He
Guangyan Huang
Xiedong Cao

# Message from the IDP 2012 Co-chairs

Intelligent data processing (IDP) has attracted much attention from various research communities. Researchers in related fields are facing the challenges of data explosion, which demands enormous manpower for data processing. Artificial intelligence and intelligent systems offer efficient mechanisms that can significantly reduce the costs of processing large-volume data and improve data-processing quality. Practical applications have been developed in different areas including health informatics, financial data analysis, geographic systems, automated manufacturing processes, etc.

Organized in conjunction with the 14th Asia Pacific Web Conference (APWeb 2012), the purpose of IDP 2012 was to provide a forum for discussion and interaction among researchers with interest in cutting-edge issues in intelligent data processing. IDP 2012 attracted 21 submissions from Japan, Korea, Hong Kong, Taiwan, China, and India with a 33.3% acceptance rate. Each paper was carefully reviewed by two or three members of an international Program Committee (PC). Seven papers were selected to be included in this volume. These papers cover a wide range of both theoretical and pragmatic issues related to data processing.

IDP 2012 was sponsored by APWeb 2012. We would like to express our appreciation to all members of the APWeb 2012 Conference Committee for their instrumental and unfailing support. IDP 2012 had an exciting program with a number of features, ranging from keynote speeches, presentations, and social programs. This would not have been possible without the generous dedication of the PC members and external reviewers, and of the keynote speaker, Clement H.C. Leung of Hong Kong Baptist University. We especially would like to thank Jiming Liu of Hong Kong Baptist University, Yanchun Zhang of Victoria University, and Xinghuo Yu of RMIT University for their thoughtful advice and help in organizing the workshop.

April 2012

Chaoyi Pang
Junhu Wang
Hao Lan Zhang

# Message from the IEKB 2012 Chair

Information extraction (IE) techniques aim at extracting structured information from unstructured data sources. Some low-level information extraction techniques have been well studied. Now, more IE research focuses on knowledge acquisition and knowledge base building. This workshop focused on issues related to information extraction and building knowledge base. IEKB 2012 was held in conjunction with the APWeb 2012 conference in Kunming, China. IEKB 2012 aimed at bringing together researchers in different fields related to information extraction who have common interests in interdisciplinary research. There were 11 submissions to IEKB 2012, and only five papers were accepted. The workshop provided a forum where researchers and practitioners can share and exchange their knowledge and experience.

April 2012 Dongyan Zhao

# Message from the MBC 2012 Co-chairs

The recent advancement of workflow technologies and adoption of service-oriented architecture (SOA) have greatly facilitated the automation of business collaboration within and across organizations to increase their competiveness and responsiveness to the fast evolving global economic environment. The widespread use of mobile technologies has further resulted in an increasing demand for the support of mobile business collaboration (MBC) across multiple platforms anytime and anywhere. Examples include supply-chain logistics, group calendars, and dynamic human resources planning. As mobile devices become more powerful, the adoption of mobile computing is imminent. However, mobile business collaboration is not merely porting the software with an alternative user interface, but rather involves a wide range of new requirements, constraints, and technical challenges.

The Second International Workshop on Mobile Business Collaboration (MBC 2012) was held on April 11, 2012, in Kunming in conjunction with APWeb 2012. The overall goal of the workshop was to bring together researchers who are interested in the optimization of mobile business collaboration.

The workshop attracted nine submissions from Germany, USA, Korea, and China. All submissions were peer reviewed by at least three Program Committee members to ensure that high-quality papers were selected. On the basis of the reviews, the Program Committee selected six papers for inclusion in the workshop proceedings (acceptance rate 66%).

The Program Committee of the workshop consisted of 14 experienced researchers and experts. We would like to thank the valuable contribution of all the Program Committee members during the peer-review process. Also, we would like to acknowledge the APWeb 2012 Workshop Chairs for their great support in ensuring the success of MBC 2012, and the support from the Natural Science Foundation of China (No. 60833005).

April 2012

Dickson W. Chiu
Jie Cao
Yi Zhuang
Zhiang Wu

# Organization

## Executive Committee

### Workshops Co-chairs

| | |
|---|---|
| Hua Wang | University of Southern Queensland, Australia |
| Lei Zou | Peking University, China |

The First International Workshop on Sensor Networks and Data Engineering (SenDe 2012)

### Workshop Chairs

| | |
|---|---|
| Jing He | Victoria University, Australia |
| Guanyan Huang | Victoria University, Australia |
| Xiedong Cao | Southwest Petroleum University, China |

The First International Workshop on Intelligent Data Processing (IDP 2012)

### Workshop Chairs

| | |
|---|---|
| Chaoyi Pang | CSIRO, Australia |
| Junhu Wang | Griffith University, Australia |
| Hao Lan Zhang | NIT, Zhejiang University, China |

Workshop on Information Extraction and Knowledge Base Building (IEKB 2012)

### Workshop Chairs

| | |
|---|---|
| Dongyan Zhao | Peking University, China |

The Second International Workshop on Mobile Business Collaboration (MBC 2012)

### Worksop Chair

| | |
|---|---|
| Jie Cao | Nanjing University of Finance and Economics, China |

### Program Chairs

| | |
|---|---|
| Dickson Chiug | Dickson Computer Systems, HK, China |
| Yi Zhuang | Zhejiang Gongshang University, China |
| Zhiang Wu | Nanjing University of Finance and Economics, China |

## Program Committee

### SenDe 2012

| | |
|---|---|
| Yong Shi | The University of Nebraska Omaha, USA |
| Ying Liu | Graduate University, Chinese Academy of Sciences, China |
| Chen Jiang | IBM, China |
| Xun Yi | Victoria University, Australia |
| Yingjie Tian | Graduate University, Chinese Academy of Sciences, China |
| Peng Yi | University of Electronic Science and Technology of China |
| Jianping Li | Institute of Management Science, Chinese Academy of Sciences, China |
| Xiuli Liu | Academy of Mathematics and System Sciences, Chinese Academy of Sciences, China |
| Haizhen Yang | Graduate University, Chinese Academy of Sciences, China |
| Peng Zhang | Institute of Computing Technology, Chinese Academy of Sciences, China |
| Aihua Li | Central University of Finance and Economics, China |
| Xiantao Liu | Southwest Petroleum University, China |
| Kou Gang | University of Electronic Science and Technology of China |
| Yanchun Zhang | Victoria University, Australia |
| Chaoyi Pang | Australia e-health Research Centre |
| Mehmet Yildz | IBM, Australia |
| Tianshu Wang | Lenovo Ltd., China |
| Jing Gao | Illinois University at Champion, USA |
| Jing Yang | Graduate University, Chinese Academy of Sciences, China |
| Jinghua Li | China University of Political Science and Law |
| Jinjun Chen | University of Technology, Sydney |
| Gengfa Fang | Macquarie University, Australia |
| Hao Lan Zhang | NIT, Zhejiang University, China |

### IDP 2012

| | |
|---|---|
| Akinori Abe | NTT, Japan |
| Clement Leung | Hong Kong Baptist University, Hong Kong |
| David Taniar | Monash University, Australia |
| D. Frank Hsu | Fordham University, USA |
| Feng Xia | Dalian University of Technology, China |
| Gansen Zhao | South China Normal University, China |

| | |
|---|---|
| Jiming Liu | Hong Kong Baptist University, Hong Kong |
| Jinli Cao | Latrobe University, Australia |
| Jing He | Victoria University, Australia |
| Ke Deng | University of Queensland, Australia |
| Wenhua Zeng | Xiamen University, China |
| Wei Peng | AUSTRAC, Australia |
| Xinghuo Yu | RMIT University, Australia |
| Xingquan Zhu | University of Technology, Sydney, Australia |
| Xingsen Li | NIT, Zhejiang University, China |
| Xiaohui Tao | University of Southern Queensland, Australia |

**IEKB 2012**

| | |
|---|---|
| Degen Huang | Dalian University of Technology, China |
| Wei Jin | North Dakota State University, USA |
| Kang Liu | Automation Institute of CAS, China |
| Xiaohui Liu | MSRA, China |
| Liyun Ru | Sohu Inc., China |
| Gordon Sun | Tencent Inc., China |
| Xiaojun Wan | Peking University, China |
| Furu Wei | MSRA, China |
| Lei Zou | Peking University, China |

**MBC 2012**

| | |
|---|---|
| Patrick C.K. Hung | University of Ontario Institute of Technology, Canada |
| Samuel P.M. Choi | The Open University of Hong Kong, China |
| Eleanna Kafeza | Athens University of Economics and Commerce, Greece |
| Baihua Zheng | Singapore Management University, Singapore |
| Edward Hung | Hong Kong Polytechnic University, China |
| Ho-fung Leung | Chinese University of Hong, China |
| Zakaria Maamar | Zayed University, UAE |
| Stefan Voss | University of Hamburg, Germany |
| Cuiping Li | Renmin University, China |
| Chi-hung Chi | National Tsing Hua University, Taiwan, China |
| Stephen Yang | National Central University, Taiwan, China |
| Ibrahim Kushchu | Mobile Government Consortium International, UK |
| Jidong Ge | Nanjing University, China |
| Huiye Ma | CWI, The Netherlands |
| Pirkko Walden | Abo Akademi University, Finland |
| Raymond Wong | National ICT, Australia |

| | |
|---|---|
| Haiyang Hu | Hangzhou Dianzi University, China |
| Matti Rossi | Helsinki School of Economics, Finland |
| Achim Karduck | Furtwangen University, Germany |
| Xiangmin Zhou | CSIRO Canberra ICT Center, Australia |
| Hoyoung Jeung | EPFL, Switzerland |
| Zaiben Chen | The University of Queensland, Australia |
| Ruopeng Lu | SAP Research CEC Brisbane, Australia |
| Quanqing Xu | National University of Singapore, Singapore |
| Mohammed Eunus Ali | The University of Melbourne, Australia |
| Zhenjiang Lin | The Chinese University of Hong Kong, China |

## Sponsoring Institutions

Chinese Academic of Science, China
Hebei University of Engineering, China
Kunming University of Science and Technology, China
Northeast University, China
Nanjing University of Finance and Economics, China
Victoria University, Australia

# Table of Contents

## The 1st International Workshop on Sensor Networks and Data Engineering (SenDe 2012)

## The 1st International Workshop on Intelligent Data Processing (IDP 2012)



## Workshop on Information Extraction and Knowledge Base Building (IEKB 2012)

## The 2nd International Workshop on Mobile Business Collaboration (MBC 2012)

# Keynote Speech: Technological Breakthroughs in Environmental Sensor Networks

Paulo de Souza

ICT Centre, CSIRO, Hobart TAS 7000 Australia

Sensor networks have significantly reduced the costs of data gathering in the environment. Real-time sensor data are being used for scientific and industrial purposes such as calibration of forecast modeling, environmental monitoring and decision-making processes in business impacted by environmental changes.

The development of better sensors (e.g., resistant to biofouling, more accurate or fast) and more efficient sensor networks (e.g., design, power efficiency and communication) is making sensor data cheaper and sensor networks more attractive. It is also noticeable the development of new information architectures addressing issues varying from simple interoperability to provenance, knowledge discovery, and data harvesting using ontologies and semantic principles. But all these efforts are incremental and one could claim that they are not real technological breakthroughs.

Informed decision-making in real-time supported by an environmental sensor network with unprecedented space coverage and time representation is still not a common reality.

This work introduces a number of technological breakthroughs that are being pursuit and the challenges they represent. Once achieved, these breakthroughs will enable the deployment of real ubiquitous sensor networks. These networks will provide the basis of situation awareness that will support real-time decision-making processes across domains with impact to different users and transform the way we perceive and understand nature.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, p. 1, 2012.

# A New Formal Description Model of Network Attacking and Defence Knowledge of Oil and Gas Field SCADA System*

Li Yang[1], Xiedong Cao[1], Jie Li, Cundang Wei[1], Shiyong Cao[1], Dan Zhang[1], Zhidi Chen[1], and Gang Tang[2]

[1] Southwest Petroleum University, Chengdu, 610500, P.R. China
[2] Sinopec Southwest Petroleum Branch, P.R. China
scncyl@126.com

**Abstract.** In this paper, we analyse the factors affecting the network security of a gas SCADA system. We model the security problem in the SCADA system into an online digital intelligent defensing process, including all reasoning judgment, thinking and expression in attacking and defence. This model abstracts and establishes a corresponding and equivalent network attacking and defence knowledge system. Also, we study the formal knowledge theory of SCADA network for oil and gas fields though exploring the factors state space, factors express, equivalence partitioning etc, and then put forward a network attack effect fuzzy evaluation model using factor neural network theory. The experimental results verify the effectiveness of the model and the algorithm, which lays the foundation for the research of the simulation method.

**Keywords:** SCADA, Factors Knowledge, Knowledge Description, Factors Express.

## 1 Introduction

This article analyzes the composition and network topology structure of the oil and gas SCADA system. We model the attacking and defense of SCADA system into an online digital intelligent defensing process, including all reasoning judgment, thinking and expression in attacking and defense. This model abstracts and establishes a corresponding and equivalent network attacking and defense knowledge system. By the introduction of factor neural network theory, the comprehensive consideration and analysis of various factors including the goals and related conditions, against attack cognitive function relation, the system state and support personnel ability level, and the relationship between the attack factors, a formal description model of network attacking and defense knowledge for the oil and gas SCADA system is proposed.

The remainder of the paper is organized as follows. We review the relevant literature on the Network Attacking and Defence in Section 2. Section 3 then describes the formal description model of SCADA network attacking and defence knowledge. The attack and defense task model is presented and experimental result is discussed in

* The paper is supported by National Natural Science Foundation Project. (Grant No. 61175122).

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 2–10, 2012.

Section 4. Finally, at the end we conclude our paper in Section 5, and provide suggestions for future work。

## 2 Related Work

The oil and gas SCADA network security defence is essentially a network attack and defence knowledge system[1]. Network attack is the hot research on network attack and defence field.With the wide application of Internet, a number of official and non-governmental organizations have established gradually,such as NIPC,CIAC,CERT and COAST,which are responsible for the research on network attack and track the latest attack technologies.The FIRST conference is held every year to discuss the new technologies about network attack methods.On the contrary,the research on network defence mainly includes the architecture of network attack and defence,model,information hiding[2] and detecting,information analysis and monitoring[3], emergency response and the application of network attack and defence technologies to other industries. The development of the direction is at an early stage, various theories and technologies on the direction arise.But in 1970s and 1980s,network defence system have such many similar functions as to result in great waste. While all countries want to learn the lesson, but it makes waste seems inevitable to develop technology according to their own rhythm and blance the different interests.To build network security, it is not surprising that the identical situation emerges.

Real-world network attack and defense systems consists of a variety of different sub-System, attributes and their relationships.The understanding extent of these factors determines the performance of offensive and defensive knowledge system. In this paper, the formal description model of SCADA network attacking and defence knowledge adoptes Professor Liu Zengliang factor neural ideological framework of the network in knowledge organization , a new factor space of SCADA based knowledge representation and defense attack model is proposed, and the validity of the model is verified.

## 3 The Formal Description Model of SCADA Network Attacking and Defence Knowledge

### 3.1 Factors State Space Based on Object

The object $u$ (comment: italic every variable in this paper) is related to factor $f$, which can view the object $u$ from the point of $f$, and also there is a correspondingly state $f(u)$ associated with it. If U and F are the sets comprising some objects and some factors, and for any $u \in U$, all the factors related to $U$ are in the $F$, ($f \in F$). For a practical problem, we can always assume that there is an approximate matching. For a given matching $(u, f)$, a correlation, R, between the $u$ and $f$ is defined, written as $R(u, f)$. Only When $R(u, f) = 1$ $f$ and $u$ are relevant. So the $u$ space related to $f$ and the $f$ space related to $u$, respectively, can be defined as:

$$D(f) = \{u \in U | R(u, f) = 1\}$$
$$F(u) = \{f \in F | R(u, f) = 1\}$$

Factor $f$ ($f \in F$) can be regarded as a mapping, and function in a certain object $u$ ($u \in U$) to access to certain state $f(u)$. $f: D(f) \rightarrow X(f)$, among them, $X(f) = \{f(u) | u \in U\}$, $X(f)$ is the state space of the $f$.

### 3.2 SCADA Formal Description of Network Attack and Defense Factors Expression

An object is described as a network attack and defense, and the main purpose of this paper is to build a knowledge network. We want to express their knowledge to constitute an integrated analysis model, which builds an organizational structure of the network attack and defense systems, rules of the state and behavior information together so that both the knowledge representation model and a knowledge of the use of model, information status and rules of conduct together. It is both a knowledge representation model and knowledge of the use of model. Constitute a feasible way: the use of factors to the structural fabric of knowledge networks, description of the beginning declarative and procedural knowledge of the organization and packaging, relationship with the structure of the relationship between slot elements to construct knowledge networks in the chain of relationships, so that the whole knowledge network system to form a framework for the node, and to the boundary chain of relational factors[6].

Let $U$ be the considered domain, $SA$ is considered as an offensive and defensive system of SCADA. A real-world SCADA network system consists of a number of different types of subsystems, attributes and their relationship posed. According to different perspectives from perceived and described, it will change its expression as follows:

$$(SA = \{ds, as, fs, ys | ds \in D, as \in A, fs \in F, ys \in Y, s \in S\} <s>)$$

to express the $SA$ offensive and defensive systems of SCADA, $S = \{s\}$ is for a variety of cognitive and describes a collection of ideas, where s describes as a cognitive point of view or perspective.

$A = \{as\}$ is a collection of objects for the behavior of the system of things, $as$ is the system of things can be the object of cognitive behavior, it includes all kinds of things can be perceived, various features actors and behavior of various types of hardware and software, and so on.

$D = \{ds\}$ sets the structure of the system, it is expressed in various types of system behavior pattern of relationships between things. Including pattern of the relationship between the state space, patterns of behavior, state transition pattern of relationships and constraints, and so on.

It is expressed in a variety of explicit knowledge of the attack factors, conditions - feature - the result of inference relations. A typical partial order of causal relationship, and the same relationship between the source and the same results.

$F = \{fs\}$ is a collection of system understanding and description of factors.

The $F$ is expressed attacks in the planning process, and perceived and described the various factors under $s$ point of view a collection of state space, $fs=(f+,aer,f-)$,Were used to represent attack to promote state, attack inhibition state, State of the attacker. $f=(fl, uspw, pc, sev, io, nc, sc)$, $fl$ represents a file variable, $uspw$ state for the user, $pc$ for the process parameters, $sev$ on behalf of system services, $io$ represented input/output parameters, $nc$ for the network connection parameters, $sc$ for the system environment variables.

$Y = \{ys\}$ for the various functions of the system state.

The $Y$ is expressed attack planning factors involved in features state-space, for example .the state space of $fl$ contains examination, upload, download, modify, delete, and so on. $fl= \cup fli(i=1…n)$ is file type. This may correspond to existence of a fl factor space. In order to cognitive and express offensive and defensive systems of SCADA network ($SA$). Need to summarize and abstract of the $SA$. For example, the series has the same characteristics, obey and abide by the rules of the specific acts the same things in the abstract as a cognitive "object", similar entities that have specific traits abstract description into common factors. Offensive and defensive behavior in the specific description of the network performance factors described as a state factor.

### 3.3 Behavior of a Collection of Objects of Things Definition of Equivalence Class Partition

In offensive and defensive of the SCADA network systems, need to view certain things on the behavior of the system to classify ,performance within the domain $U$, $S$ is selected as a cognitive point of view, and $S$ will as the basis and in accordance with an "equivalence relation" to be classified.

[Definition 1]. Let $A$ be a collection, $R$ is a relation on $A$, $R$ is the $S$ point of view under an equivalence relation on $A$. $S$ point of view if the next for $x, y, z \in A$, $R(s)$ to meet

(1) Reflexive $xR(s)x$

(2) Symmetry If $xR(s)\ y$ have $yR(s)\ x$

(3) Transitivity If $xR(s)\ y$, $yR(s)\ z$ then have $xR(s)\ z$

$S$ point of view in the next, $A$ an equivalence relation on $R(s)$ is usually denoted by

$$(x)R(s)\ (y)\ or\ (x)\ R(s)\ (y)\ (x, y \in A)$$

[Definition 2]. The equivalence relation R (S) under the A and O is called an equivalence class, if $a \in A$, $o = \{ y \mid (y)\ R(s)\ (a) \}$

[Definition 3]. If the o is equivalence relation in *R (s)* of a known equivalence class *A*, says the *o* is an abstract object for system *SA* in view of *s*, set *O*= { o | *o* is an abstract object for system *SA* in view of s} , where *O* is an object set for system *SA* in view of *s*.

[Definition 4]. Set *Ai* is subset of the *A*, if *A* = ∪ *Ai, {Ai}* is called a division level of *A*; If still have $A\alpha \cap A\beta = \varphi$ $(\alpha \neq \beta,\ A\alpha、\ A\beta \in A)$,the *{Ai}* called a deterministic division of *A*.

In the network attack and defense system, the division of the object reflects a kind of cognitive and describes view of the things. This view is relate to the levels and relevant of considering problems angle, the object divide into sometimes thick and sometimes fine, this difference division degree is sometimes called the particle size of the system of the classification .

[Definition 5]. Set *R (s1), R (s2)* is the equivalence relation of two different objects division of the classification *A*. If there is ( x ) R (s1) (y) → ( x ) R (s2) (y), said the R (s1) is fine than R (s2), written for R (s1) < < R (s2).

A network attack and defense reality system *SA*, when one of the behaviors is divided into different extent, the relationship between each object, and the factors of the system chooses and the system state of expression by factor will also corresponding change. As people have different degree division of the system into purpose, mainly in order to make it more convenient while research and analyze the problems, research and analysis of the different angle and purpose, can have a variety of different partition to the same system, produce all kinds of different particle size of the object, the introduction of different kinds of system structure and factors description, to create all kinds of different cognitive and description model. Because these cognitive and description model of actual system is the same reflect between them, there are always some contact and has some of the same characteristics, especially, we hope:

(1) The model should have a hierarchical relationship.
(2) In the same level, obtained by different sides of each model can be merged into a comprehensive model.
(3) The nature of the model has a preserving between different levels. For instance, if the original system is topology structure, the topological properties in different levels of the model shall remain unchanged, if the original system is partial sequence structure, it also should have partial order in the various models.

[Definition 6]. Says *M* = < < *O, G* >, *F, X* > *SA* is a cognitive or description model in the view of *S*

If *O* = *{O}* < called objects set in the *M* >, where

*O* is the equivalent clustering of object of *SA* in behavior things under the view of *S*,

*G = {G}* < called structure of *M* >,

*g* is the equivalent transformation of *SA* in behavior things relationship d under the view of *S*,

*F = {f}* < called the cognitive or describe factor sets of *M* >,

*f* is cognitive and describe factors of *SA* which is based in the selection of $f\ o$ under the view of *S*,

*X = {X}* < called factor expression state set of *M* >, and

*X* is a reflected factors system state of *SA* while choose F as the express views under view of *S*.

[Definition 7]. Set *SA = (< A, D >, Fo, Y)* is a practical system, *A ', A* has nature *H*. If in the view of *S*, obtained by *SA* model system

*M = < < O, G >, F, X >*

Make $s : A' \rightarrow s\ (A') \in M$,

*S (A ')* also has properties *H*, said the *M* is the model keep nature *H* of *SA*.

[Definition 8]. Set *s1, s2* for two different cognitive and describe view, to the system *SA*,

Respectively Abstract cognition and describe model

*M1 = < < O1, G1 >, F1, X1 > and M2 = < < O2, G2 >, F2, X2 >.*

If make *s3 = s1* $\odot$ *s2* is the comprehensive *s1, s2*, then

The M3 = M1 $\odot$ M2 = < < O3, G3 >, F3, X3 >= < < O1, O2 $\odot$ G1 $\odot$ G2 > F1 $\odot$ F2, $\odot$ X1 X2 > will call the M1 and M2 comprehensive.

Here, $\odot$ said the comprehensive operational, it sure a equivalence relation R,

to $\forall$ *x, y* $\in$ *A(x) R (s1* $\odot$ *s2) (y) < — > (x) R (s3) (y)*

# 4 Attack and Defense Task Model

## 4.1 The knowledge Model of Network Attacks

The knowledge representation method is the foundation of building knowledge model, using the state space method of factors expresses the complex knowledge of network attack, which will easy to describe the incidence relation between a varieties of knowledge object and make it that can describe fully and accurately the required domain knowledge for further reasoning. In this paper, the composition of a knowledge model of network attack is shown in Figure 1.

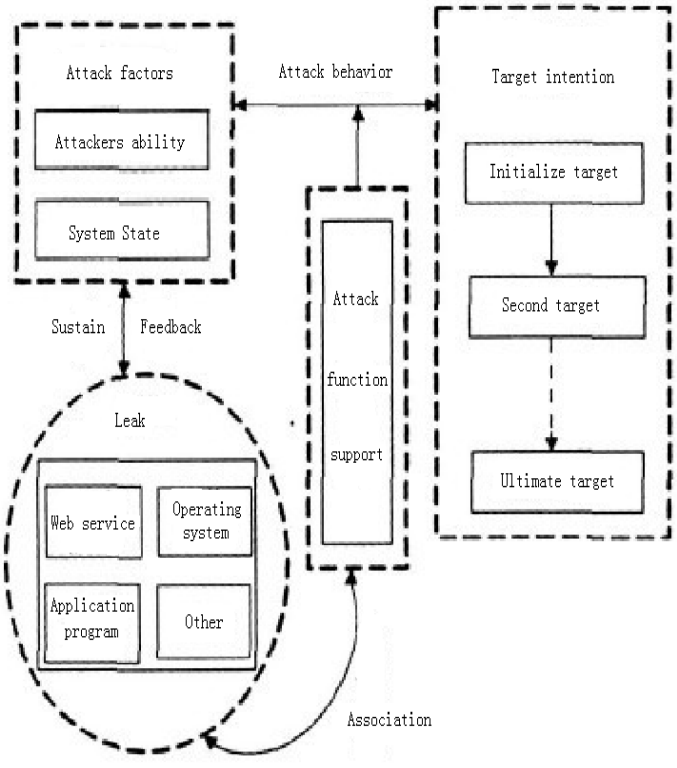



**Fig. 1.** The knowledge model of network attack

The process of network attack is the process of information access and the elevation prompt according to the conditions and objectives. This paper introduces attack target, attack conditions, attack behaviour and support conditions four factors in the attack knowledge model to describe the process of attack. Based on individual state and the knowledge structure of the attacker and according to attack factors, knowledge acquisition costs, it develop appropriate behavioral sequence, obtain operation chain of various of attacks knowledge to achieve the goals, form the operation instruction set of network attack, maximize attack effect and the performance ratio of the loss. Suppose a network structure contains a number of attacks and attack conditions or target-related technologies, knowledge, human factors, according to factor space theory constitutes an attack conditions or attack target of the family $\{as\}$ $s \in S$, in the $\{as\}$ $s \in$ s, each implementation factors $a_s$ for function factors $a_{sf}$, constraint factors $a_{sc}$ and results of factors $a_{se}$ extract. Every function factors through the son function decomposition of orthogonal to ensure that each function independent factors, through the factors determine the state space of factors function evaluation, and achieve the purpose of comprehensive and evaluation.

To the tasks model, time events extraction, analysis and synthesis in model formed to the formed interactive network under different task design model. The effect of the assessment according to the method of DELPHI corresponding index weight assignment,

define information system target protection level for $G = [g1, g2, g3]$, among them, the $g1, g2,$ g3 ∈ [1, 5), 1 is the lowest system defense strength, 5 for the highest defensive intensity. Define against object level for V = [v1, v2 , v3], among them, the v1 and v2 and v3 ∈ [0, *5)*, 0 for attack strength minimum, 5 for the attack strength is the highest.

The object of network attack against the task scheduling mainly according to the following process:

(1) Using the factor to form related state space that can expression attack mission (or child task), this will form factors object working environment.
(2) Collecting information and processing knowledge, composition a group of heuristic rules activities that can perform certain information control and transformation, including reduction rules or operation rule, and to specify conflict solution.
(3) According to certain control strategy implement reasoning activities which is mainly heuristic method of operating rule, search and generate problems or sub problem solutions, including part or the final answer.
(4) According to attack process of events expression to precede problem reduction, and determine their orderly execute and the logical relationship.
(5) The execution environment operation, including updating of the information, additions work environment, etc.

### 4.2 Experimental Verification

The penetration testers validate the model in local area network of SCADA based on CPU? Memory size? Windows2000 professional sp4 operating system to enhance the penetration of privileges. According the definition of network attack knowledge model, the first is to decompose goals intended set of $A_{sef}$= *{fl,uspw,pc, sev, io,nc, sc }*, it can gain target weight $\omega$ *(uspw) = 1*, according to the state of the target *fs* causal selection, status is set to switch factors, $f^+_i = \bigcup_{i=1}^{7} f^+_i$={0, 0, 0, 0, 1, 0, 1}, $f^+_i$= {0,1}, *0* means factors does not exist, *1* means factors exist. Seven kinds of factors are the whole system open to all ports, do not set system password, system open the network service with holes, system has default user account and password, system open distance management functions, shared directory is too large to read and write permissions and existing operating system kernel holes. $f = \bigcup_{i=1}^{4} f_i$={1,1,1,1},$f_i$= {0,1}, four kinds of factors are opening firewall, sharing system set up close, system patch, opening intrusion detection. Through searching current function sets $A_{sfi} = \bigcup_{i=1}^{7} t_i$ contains seven functional attributes factors, achieving the expected functions $\omega(T_f) = 1$, for the attackers ability sets *aer={ access,gues,t user, spuser, root}=(0, 0, 1, 0, 0),* getting feasible against function space $X_{2(t2)}$*=(MS05039-2, IIS-idahack, MS06-035, IPc$, MS-05047, MS07-029).* According to testing, the use of the above attack can be changed the user permissions by increasing the user state and then to access the target attack planning space *X.*

## 5 Conclusion

This paper studies the formal knowledge theory of oil and gas fields SCADA network from the factors state space, factors express, equivalence partitioning, etc, and put forward network attack effect fuzzy evaluation model based on factors knowledge and defense task model using factor neural network theory. The experiments described and verified the attacker's human factors (individual age, professional, Aggressiveness, etc), the network connection factors (internal network, Internet) and the outside environment factors (floods, earthquakes, landslides, etc), the internal factors of the network system (all the port of the system are open, system password was not set, network services with holes was opened in the system, the system included the default user accounts and passwords, remote management function was opened, read-write permissions of the Shared directory was too excessive , kernel loophole of operating system was existed, etc). The experimental results show that the model has good simultaneity and fuzziness. We can use the model to simulate a wider and deeper network attacking and defense of oil and gas SCADA system.

## References

1. Ragsdale, D.J., Surdu, J.R., Carver, C.A.: Information assurance education through active learning.The IWAR Laboratory (2002)
2. De Vivo, M., de Vivo, G.O., Isern, G.: Internet security attacks at the basic levels. Operating Systems Review 32(2) (2002)
3. Teo, L., Zheng, Y., Ahn, G.: Intrusion detection force: an infrastructure for internet-scale intrusion detection. In: Proceedings of the First IEEE International Workshop on Information Assurance (IWIA 2003), Darmstadt, Germany, pp. 73–91 (2003)
4. Zheng, L., Liu, Z., Wu, Y.: Network warfare on the battlefield. Military science press, Beijing (2002)
5. Liu, Z., Liu, Y.: Factor neural network theory and implementation strategy research. Beijing Normal University Press, Beijing (1992)
6. Zhang, S., Tang, C., Zhang, Q., et al.: Based on the efficiency of network attack against method classification and formalism description. Information and Electronic Engineering 2(3), 161–167 (2004)
7. Huang, G., Ren, D.: Based on of both-branch fuzzy decision and fuzzy Petri nets the attackStrike model. Computer Applications 27(11), 2689–2693 (2007)
8. Huang, G., Qiao, K., Zhu, H.: Based on the fuzzy attack FPN graph model and productionAlgorithm. J. Microelectronics and Computer (5), 162–165 (2007)
9. Guo, C., Liu, Z., et al.: The virtual network attack and defense analysis model. Computer Engineering and Applications 44(25), 100–103 (2008)
10. Wang, P.Z., Sugeno, M.: The factors field and back-ground structure for fuzzy subsets. Fuzzy Mathematics 2(2), 45–54 (1982)
11. Guo, C.-X., Liu, Z.-L., Miao, Q.: Network attack planning model and its generating algorithm. Computer Engineering and Applications 46(31), 121–123 (2010)
12. Guo, C.-X., Liu, Z.-L., Zhang, Z.-N., Tao, Y.: Network Attack Knowledge Model Based on Factor Space Theory. Telecommunication Engineering 49(10), 11–14 (2009)

# An Efficient Overload Control Strategy in Cloud

Xiling Sun, Jiajie Xu, Zhiming Ding, Xu Gao, and Kuien Liu

NFS, Institute of Software, Chinese Academy of Sciences, Beijing, China
{xiling,jiajie,zhiming,gaoxu,kuien}@nfs.iscas.ac.cn

**Abstract.** In cloud, service performances are expected to meet various QoS requirements stably, and a great challenge for achieving this comes from the great workload fluctuations in stateful systems. So far, few previous works have endeavored for handling overload caused by such fluctuations. In this paper, we propose an efficient overload control strategy to solve this problem. Crucial server status information is indexed by R-tree to provide global view for data movement. Based on index, a two-step filtering approach is introduced to eliminate irrational server candidates. A server selection algorithm considering workload patterns is presented afterwards to acquire load-balancing effects. Extensive experiments are conducted to evaluate the performance of our strategy.

## 1 Introduction

With the rapid development of Internet technology, cloud computing has emerged as a cutting-edge technique for large scale internet applications [11]. Cloud platforms are popular in big enterprises because of the characteristic called elasticity provision [2]. These platforms allow people to neglect capital outlays of hardware and manage large scale data more effectively and economically [5]. However, heavy workload fluctuations even overload phenomena always cause system problems. So an effective overload control strategy is necessary to guarantee load balance under cloud environment.

In cloud, cost-effective services can be better supported with elasticity provision in response to workload fluctuations [1]. But many overload control challenges related to data movement emerge when system is stateful, intensive and interactive: firstly, it is onerous to manage massive cheap PC in cloud, which prevents us to understand the actual environment for making decisions; secondly, cloud systems may require fast response with high service quality, meaning that overload has to be efficiently processed; thirdly, as data may have workload fluctuation patterns, overload control should consider such characteristic to guarantee the robust performance of server.

In this paper, we propose an overload control strategy to solve above challenges. We use R-tree index, two-step pruning strategy and server selection method to find the best server for data movement. Particularly, we make the following contributions:

- We utilize R-tree index to record crucial information (e.g. location and available workload) of servers. Such index gives us a global view of the whole servers in cloud to make correct data movement decisions.
- We use an index based pruning mechanism to reduce search space, and a set of broadcasting based validation rules to filter out irrelevant server candidate, so that overload handling can be efficiently processed.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 11–18, 2012.

- We use a set of measures related to workload patterns for selecting proper server for data movement, so as to guarantee that unexpected workload spike can be handled in an effective and robust way.

The rest of this paper is organized as follows. Section 2 discusses related work. A model is introduced to set a base for overload control strategy in section 3; Section 4 presents a two-step server pruning strategy; after that, a server selection method is discussed in section 5; Section 6 demonstrates experimental results. At last, a brief conclusion and the future work are mentioned in Section 7.

## 2 Related Work

Previous researches about overload control mainly focus on load shedding because of limited server availability [7]. They believe that sending explicit rejection messages to users is better than causing all users to experience unacceptable resource times [6]. Cloud computing provides an ideal base for handling access deny during unexpected workload spikes [8, 10].

In cloud, simple elasticity provision is welcomed by stateless system in which each server is equally capable of handling any request but is not sufficient to handle overload in stateful system in which only some servers have the data needed by given request [1]. As stateful cloud systems (e.g. Facebook) become popular and have extraordinary developing speed, overload control in such systems also receives increasing attentions. [12] mainly discusses the importance of server numbers that should be added. And [13] pays attention on rebalancing speed. However, these works neglect the problem that where replicated data should be moved in stateful cloud systems.

[1] designs the SCADS Director, a control framework that reconfigures the storage system in response to workload changes. Controller iteration is presented to discuss moved data and servers as receiver. Although they consider the impact on performance come from size of moved data, they do not discuss delay deriving from large scale servers in cloud. In addition, they do not consider the server workload distribution pattern which could be used to support rational spike data movement in cloud.

## 3 Overload Control Model

### 3.1 Overload Control Mechanism

Figure 1 is the mechanism of our overload control on spike data. For each overloaded server, the data that has most workload increase is detected for replication. We use R-tree index to organize sever status information, so as to assist efficient server selection. To avoid unnecessary calculations, a two-step pruning strategy is applied to filter improper servers. Among the remaining server candidates, we optimize data movement by selecting server according to a set of criterions related to workload patterns of data and servers. Data movement is finally executed toward the selected server.

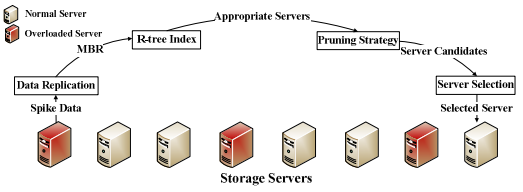



**Fig. 1.** Overload Control Mechanism

### 3.2 Models (Data and Server)

Among features relevant to data movement, we focus on the most important ones including size, location and workload. In our model, the atomic form of data is $d = <ds, dl, DW>$, where $ds$ represents data size; $dl$ is data's longitude and latitude and its form is $dl = (dl_x, dl_y)$; $DW$ is data workload. It is a vector which is composed by $|T|$ elements and its form is $DW = (dw_1, dw_2,\ldots, dw_t,\ldots,dw_{|T|})$, where each $t$ indicates particular unit time interval and $|T|$ is the total number of time intervals.

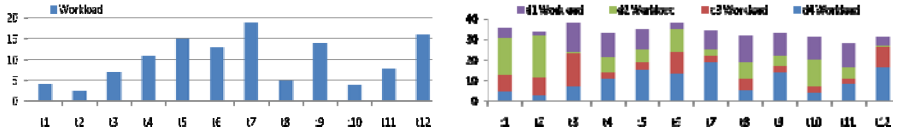



**Fig. 2.** (a) Data Workload (b) Server Workload

Data workload is a straightforward characteristic that determines if there is an overload or not. In this paper, we slice whole time dimension into several time interval units because data workloads always fluctuate along time. Figure 2(a) represents workloads changes among unit time intervals, where $t_i$ is particular time interval and $i$ is an integer in range of [1, 12]. If we neglect workload fluctuations, visiting bottleneck and low resource utilization rate are likely to be derived.

Server's main attributes include the storage space, spatial location and workload. Its workload is the sum of workload of each data stored in this server. Figure 2(b) is an example of server workload fluctuation, where server workload in a time interval $t_i$ is the sum of workloads of four data ($d_1$ to $d_4$) during $t_i$. We aim to ensure server workload to be balanced in different time intervals after data movement.

Generally, one data $d_t$ at a particular time $t$ might have multiple possible workload values $\{dw_{t1}, dw_{t2},\ldots, dw_{tm}\}$. And each possible workload value $dw_{tm}$ has a corresponding probability $p_{tm}$. We use workload expect as finally value of $dw_t$. According to probability theory, expect of random variable $dw_t$ can be calculated from equation (1):

$$E(dw_t) = \sum_{i=1}^{m} dw_{ti}\, p_{ti} \tag{1}$$

# 4 Two-Step Server Pruning Strategy

## 4.1 Index Based Pruning

A massive number of servers in cloud are organized as large scale clusters in different geographical locations far away with each other. This may prevent time-saving and cost-effective of data movement in overload handling [5]. Meanwhile, among all servers, the best-fit one (the server with smallest remaining workload that is big enough to receive moved data) is preferred to use resource rationally.

In this paper, longitude, latitude and remaining workload of servers are set as x-axis, y-axis, and z-axis respectively to build R-tree index. Index based pruning is used to eliminate majority of servers which are improper for data movement due to physical (e.g. spatial and workload) limitations. After the index based pruning, the derived server candidates would be a small portion of servers in whole cloud.

## 4.2 Broadcasting Based Pruning

Some of the servers in R-tree index returned in section 4.1 may not be active due to the network linking to the overloaded server. So we conduct a broadcasting based pruning to validate server candidates. In addition, threshold of storage capacity and data transfer speed are also used to check the qualification of server candidates.

# 5 Server Selection Strategy

In this section, we present a novel server selection strategy to find suitable server for data movement. The ranking of server candidates considers the matching of workload fluctuation between moved data and servers to improve resource utilization.

## 5.1 Measurement on Single Time Interval

For each server candidate, we compute some statistical variables on the new server workload after data movement to guide server selection. Especially, we have to compute the covariance in addition to variance in order to incorporate data correlations [2]. According to Central Limit Theorem (CLT), given a large number of data, we can use the Normal distribution to model the workload distribution pattern of single server after data movement (to this server) to acquire accurate workload demand [9].

Equation 1 in Section 3 could be used to compute data workload expect at a particular time interval. So we need calculate data workload variance in order to use CLT:

$$D(dw_t) = E(dw_t^{\,2}) - [E(dw_t)]^2 \tag{2}$$

Based on the data workload expect and variance, we further use equation 3 to compute server workload expect $\mu_t$ and variance $\sigma_t$ in the time interval $t$.

$$\mu_t = E(sw_t) = \sum_{i=1}^{n} E(dw_{ti}),\ \sigma_t^{\,2} = D(sw_t) = \sum_{i=1}^{n} D(dw_{ti}) + 2\sum_{m \neq n} Cov(dw_{tm}, dw_{tn}) \tag{3}$$

where $n$ denotes the number of data stored in this server candidate, $dw_{ti}$ represents the workload of the $i$-th data in the time interval $t$, and $sw_t$ is the server workload in the time interval $t$. $Cov(dw_{tm}, dw_{tn})$ is the covariance between the workload of the $m$-th data and the workload of $n$-th data. As we do not know the joint probability mass function $p(dw_{tm},dw_{tn})$ which is necessary for calculating $Cov(dw_{tm}, dw_{tn})$, we adopt a linear programming solution, which is proposed in [3] to find the maximum covariance and is more accurate than variance measure in most cases.

Hence, according to CLT, the distribution of server workload in a particular time interval $t$ is approximated by the Normal distribution equation (equation 4).

$$F[sw_t < x] = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \tag{4}$$

where random variable $x$ in Normal distribution is the server workload demand. Therefore, we can calculate workload demand required by all data in one server according to above equations from 1 to 4.

### 5.2 Measurement on Whole Time Dimension

We further seek to measure the server workload characteristic on whole time dimension. Equation 5 is used to compute expect value of workload based on the whole time dimension which can indicate utilization rate of resource [4].

$$E(SW) = \sum_{i=0}^{|T|} Demand(sw_i) / |T| \tag{5}$$

where $|T|$ represents the total number of time intervals, $Demand(sw_i)$ denotes the workload demand which is calculated by equation 4.

$$D(SW) = E(SW^2) - [E(SW)]^2 \tag{6}$$

Then we use equation 6 to find server workload variance in the perspective of whole time dimension. This variable means server workload fluctuation between each time interval. We prefer the value of workload variance on a server in whole time dimension to be small, because it means this server has a balanced workload in different time intervals and rational resource utilization.

To balance all statistical variables above, we definite a distance to synthesize them and to describe the penalty of data movement to a particular serve:

$$Distance(S) = \alpha \times E(SW) + \beta \times D(SW) + \gamma \times HW(SW) \tag{7}$$

where, $\alpha$, $\beta$, and $\gamma$ denote the weights of different factors, and $HW(SW)$ is the highest workload value of servers observed from single time interval. The value of distance (i.e. penalty) between moved data and server candidate is in direct proportion to

expect, variance and the highest workload. The server with the smallest distance is the target server for data movement.

### 5.3 Server Selection Algorithm

Pseudo code for the server selection is shown in Algorithm 1. Server workload fluctuation can be expressed by workload variance (in the whole time dimension), which is calculated by function *compute_variance*() in Line 4. Expect of server workload in all time dimensions is used to avoid overload on candidate servers, and we calculate it by function *compute_exptect*() in Line 5. Also, we calculate the highest workload of servers by function *compute_HW*() in Line 6. Then distance between moved data and server candidate is calculated according to equation 7 in Line 7. At last, we select the server with the minimum distance as the target server to fulfill overload control.

```
Algorithm 1. Server Selection Algorithm
01. distance = ∅// set of distance between moved data and candidates
02. For all server candidate SC do:
03.    distance_i = 0 // distance between moved data and i-th candidate
04.    D = compute_variance() // D: variance of server workload
05.    E = compute_expect() // E: expect of server workload
06.    HW = compute_HW() // HW: the highest workload of server
07.    distance_i =  ×E +  ×D +  ×HW
08.    distance = distance • distance_i
09. Find selected server with the minimum distance
10. Return selected server
```

## 6 Experimental Result

### 6.1 Experimental Settings

All experiments are carried out on a 32-bit server machine running Ubuntu 10.04.2 with Intel(R) Xeon(R) CPU E5520 clocked at 2.27GHz with 1GB RAM. We use simulations to compare our approach with the underload strategy published in [1]. Server expect, server variance, the highest workload, running time, and expect at spike time are selected as comparison indicators. Each experiment is carried out with 10 validations. Simulations are conducted on multiple server numbers, including 1000 (Server-1000 experiment), 10000, 20000, 30000, 40000, and 50000.

### 6.2 Time-Saving Effect of R-tree Index

To illustrate time-saving effect of R-tree index, we first compare the performance of index-based overload control approach with the index-free one. As shown in figure 3 and 4 although a better variance value could be achieved by the index free approach, index-based approach is much more efficient. Given that it is a computational intensive application, index-based strategy is thus superior to the index-free approach because of much less processing time and a balanced workload distribution as well.

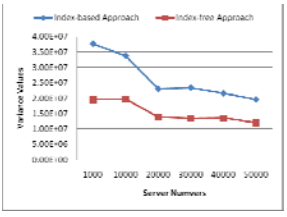



**Fig. 3.** Variance Comparison

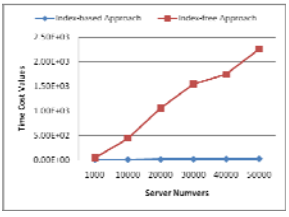



**Fig. 4.** Time Cost comparison

### 6.3 Performance Comparison with Existing Measure

We further compare our approach with the underload strategy proposed in [1]. As shown in figure 5 and 6, our strategy has about 10 percent higher Expect values and 30 percent higher Expect values at Spike Time than those of the underload strategy, which means that the best-fit method of our strategy contributes to increase resource utilization. Although our approach has more sufficient resource utilization, the value of highest workload is almost equals to that of the underload strategy (Figure 7). This is because our strategy tends to achieve balanced workload distribution. According to figure 8, variance values of our strategy are much lower than that of the underload strategy. This indicates data movement under our control strategy results in more balanced workload distribution. Therefore, our overload control strategy is an efficient and balanced solution to resolve unexpected workload spike.

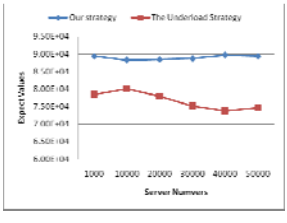



**Fig. 5.** Expect Comparison

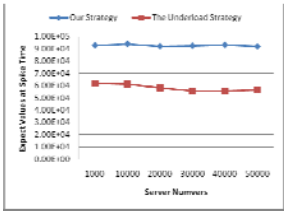



**Fig. 6.** Expect Comparison at Spike Time


**Fig. 7.** Highest Workload Comparison

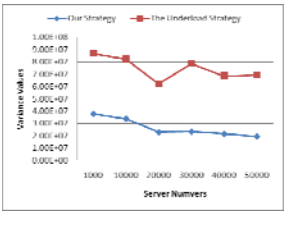



**Fig. 8.** Variance Comparison

## 7 Conclusion and Future Work

In this paper, an efficient overload control strategy is presented to handle unexpected workload spikes in cloud. R-tree is used to organize information of server status and

to eliminate irrelevant servers for reducing search space. To choose the best server for data movement, a set of novel standards are used to achieve balanced workload distribution and improve resource utilization. A serious experimental study is conducted afterwards to evaluate algorithm performances. In the future, we aim to consider more factors related to network to assist server selection in real overload control system.

## References

1. Trushkowsky, B., Bodík, P., Fox, A., Franklin, M., Jordan, M., Patterson, D.: The SCADS director: scaling a distributed storage system under stringent performance requirements. In: FAST, pp. 163–176 (2011)
2. Bodík, P., Fox, A., Franklin, M., Jordan, M., Patterson, D.: Characterizing, modeling, and generating workload spikes for stateful services. In: SoCC, pp. 241–252 (2010)
3. Rolia, J., Zhu, X., Arlitt, M., Andrzejak, A.: Statistical service assurances for applications in utility grid environments. In: MASCOTS, pp. 247–256 (2002)
4. Copeland, G., Alexander, W., Boughter, E., Keller, T.: Data placement in bubba. In: SIGMOD, pp. 99–108 (1988)
5. Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: A view of cloud computing. Communications of the ACM 53(4), 50–58 (2010)
6. Welsh, M., Culler, D.: Adaptive overload control for busy internet servers. In: Proceedings of the 4th Conference on USENIX Symposium on Internet Technologies and Systems, vol. 4, pp. 4–4 (2003)
7. Tatbul, N., Çetintemel, U., Zdonik, B., Cherniack, M., Stonebraker, M.: Load shedding in a data stream manager. In: VLDB, pp. 309–320 (2003)
8. Urgaonkar, B., Shenoy, P., Chandra, A., Goyal, P.: Dynamic provisioning of multi-tier internet applications. In: ICAC, pp. 217–228 (2005)
9. Romano, J., Wolf, M.: A more general central limit theorem for m-dependent random variables with unbounded m. Statistics and Probability Letters 47, 115–124 (2000)
10. Amur, H., Cipar, J., Gupta, V., Ganger, G., Kozuch, M., Schwan, K.: Robust and flexible power-proportional storage. In: SoCC, pp. 217–228 (2010)
11. Foster, I., Zhao, Y., Raicu, I., Lu, S.: Cloud computing and grid computing 360-degree compared. In: Grid Computing Environments Workshop, pp. 1–10 (2008)
12. Chen, J., Soundararajan, G., Amza, C.: Autonomic provisioning of backend databases in dynamic content web servers. In: ICAC, pp. 231–242 (2006)
13. Lim, H., Babu, S., Chase, J.: Automated control for elastic storage. In: ICAC, pp. 1–10 (2010)

# The Geological Disasters Defense Expert System of the Massive Pipeline Network SCADA System Based on FNN*

Xiedong Cao[1], Cundang Wei[1], Jie Li[1], Li Yang[1], Dan Zhang[1], and Gang Tang[2]

[1] Southwest Petroleum University, Chengdu, 610500, P.R. China
[2] Sinopec Southwest Petroleum Branch
cowyco@126.com

**Abstract.** The SCADA system plays an important role in monitoring the long distance operation of mass pipeline network, which may experience huge damage due to landslides geological hazards. It is critical to detect the deformation and displacement of rock to forecast the damage of landslides geological hazards through analyzing detailed information collected by SCADA system. In this paper, we use advanced TDR real-time technology to monitor the factors of rock's inclination, displacement, and humidity, and take advantage of factor neural network (FNN) theory to build a simulation-type factor neural network model. Particularly, based on FNN model, we design an expert system to forecast the potential risks of geological disasters through analyzing the real-time information of the large-scale network in the SCADA system.

**Keywords:** FNN, SCADA System, TDR, Geologic Hazard, Expert System.

## 1 Introduction

In the factor neural network (FNN) theory for information processing systems engineering, the knowledge of the factors expression is its basis, factor neurons and factor neural network are its formal framework. It aims to achieve the storage and application of knowledge and to complete the engineering simulation process of the intelligent behavior.

An expert system is a program system that includes a large number of specialized knowledge and experience, applies artificial intelligence technology and computer technology to a field in order to simulate human experts' decision-making process, such as reasoning and judgment process, and then solves complex problems which need human experts processing. The SCADA system plays an important role in monitoring the long distance operation of mass pipeline network, but landslides geological hazards along the long distance pipeline is a serious threat to the normal operation of the system. In order to timely forecast landslides geological hazards, using advanced TDR real-time technology to monitor the factors of rock, for example, inclination, displacement, humidity and so on. The monitoring information analyzed by

* This paper is supported by National Natural Science Foundation of China. (Grant No. 61175122).

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 19–26, 2012.

the FNN-based expert system can forecast the possible geological disaster and reduce the economic loss.

The paper first apply FNN theory to establish the geological disasters expert system, it can predict the occurrence of landslides in advance, so that people can take timely measures to reduce the destruction of mass pipeline network. Forecasting information present more intuitive, it uses convenient and easy to promote.

The remainder of this paper is organized as follows. The remainder of this paper is organized as follows. In Section 2, introduces the method in this paper and the difference to other method, and expounds the necessity of the establishment of this expert system. In Section 3, this paper tells the SCADA system of the role in the long distance operation of mass pipeline network and landslides for its potential damage, at the same time introduce a monitoring technology-TDR. Then the paper describes the characteristics of landslides and selection of parameters to be monitored in section 4. In the most important section 5, first introduces the theory of FNN and simulation neural network, and then further to establish an expert system prototype, explain how it works and has the function of target.

## 2 Related Work

In previous pipeline transport, the monitoring of landslide most mainly focus on the monitoring technology directly to the mountain, the technology is still relatively traditional, the TDR technique used in this paper has been widely used in foreign countries, while the application is still in the initial stage in China. According to the existing data query, founding very little use of expert system for monitoring the information to do further treatment, in order to arrive at a conclusion more directly, so that ordinary people can also operate using, more conducive to promoting. At the same time, this is the first attempt to use the FNN theory to represent knowledge. It is based on the above, the paper proposes the establishment of this expert system.

## 3 The SCATA System and the TDR Test Technology

The SCATA system is an effective computer software and hardware system in the production process and automation management. It is much mature in the application and technical development. The system is composed by the monitoring center, the communications system and the data acquisition system, and play an important role in the far dynamic system. Through monitoring and controlling the operation equipment, it not only can capture data of some dangerous or unattended special occasions, but also can realize control integration. It is applied to detect geological hazards in the pipeline network for monitoring defense system and to effectively resolve the difficult problem of attending large-scale network.

Landslide is the slope on a certain part of the mountain geotechnical in the function of gravity (including geotechnical itself gravity and groundwater static and dynamic pressure), along some weak structure plane (band) to produce shear displacement and integral to move down slope of the role and phenomenon. The activity time of the

landslide mainly relate to the different kinds of external factors induced landslides. Generally there are the following rules:

- Simultaneity. Some landslide act immediately by effect of inducing factors.
- Hysteretic nature. Some landslides occurred later than the time of the factor-induced effects, such as rainfall, snowmelt, tsunami, the storm surge and human activities. This lag in rainfall trigger type landslide is most obviously.

Since landslide geological hazards have great potential destruction dangerous to the communication optical fiber of the SCADA system , real-time monitoring is necessary for detecting the displacement, humidity of biggest mountain where the optical fiber communication in the SCADA system, and other potential disaster response prediction.

TDR (Time Domain Reflectometry)-the Time Domain reflex testing technology is a kind of electronic measurement technique, it is applied in the various measurement of object form characteristics and spatial orientation. Its principle is that firing pulse signals in the fiber, meanwhile, the reflected signals monitoring, when the optical fiber distortion or meet outside material, the characteristic impedance can change, when test of pulse meet optical fiber characteristic impedance change, it can produce launch wave. Through the comparison of the incident wave and reflected wave, according to the abnormal situation can be judged the current state of the optical fiber. The measurement technology is applied in engineering geology exploration and inspection work, it can detect the rock deformation displacement, humidity etc.

TDR is applied in monitoring the rock displacement and humidity. Firstly, the optical fiber is casted in drilling hole; when the surrounding rock occur displacement, it will shear to optical fiber and the optical fiber will produce distortion. The test pulse of the TDR is launched into optical fiber, when the test pulse meet optical fiber deformation, it will return to produce launch wave, and by measuring the time of arrival of the launch wave and its extent, then the location of the optical fiber deformation and displacement will be known, in order to determine the location of the surrounding rock deformation and displacement. Due to the relative dielectric of the air, water, soil of constant have very big difference, so the moisture content of different layers, the dielectric constant is different, the rate of spread of the TDR signal in them is also different. According to the dielectric constant of the moisture content and rock corresponding relation, it will determine the strata of humidity.

## 4 Geological Disasters Forecast Model

The practice shows that the surface distribution, characteristics and the relationship of rock mass structure decided whether the internal factors of rock mass stability or not, and the rainfall is external factors that influence the stability. The together of internal and external factors determine the degree of the rock mass stability, which are possible deformation of boundary conditions. The main consideration forecast model are rock types ($T$), covering layer of the rock ($C$), slope ($S$), displacement ($D$), friction coefficient between rock ($F$), moisture ($M$), water pressure ($P$), and other factors, then further summarized the fuzzy rules of the expert system.

(1) Geotechnical Type. Geotechnical engineering is a material base of the landslide. Say commonly, all kinds of rock and soil may constitute landslides, including loose structure, and the shear strength and fight decency capacity is low, the nature changed rock, soil, such as loose layer, loess, the red clay, shale, shale, coal formation, tuff, schist, SLATE, the thousand pieces that under the function of the water and the hard and soft rock and the rock strata and composed of landslide slope easily.
(2) Geological tectonic condition. When the composition of soil slope rock is separated into discrete cutting state, it will be possible to slide down. At the same time, the structure surface provides channels for rainfall and water flow into the slope. So all kinds of joints and fissures, level, fault of the development of slope, especially when parallel and vertical slope steep inclination of the slope surface structure and the structure of the soft-hard interbreeding face development, it will most vulnerable to slide.
(3) Topography condition. Only in certain parts of the landscape, have a certain slope, the landslide can occur. General River, lake (reservoir), and the sea, ditch slope, open the front slopes, railway and highway engineering building and the slope is the easily part of the landscape landslide. When the slope is more than 10 degrees and less than 45 degrees or the upper into ring of slope shape is the favorable terrain produce landslide.
(4) Hydrogeology conditions. Groundwater activities play a main role in the landslide formation. It functions mainly displays in: softening of rock, soil, reduce rock, the strength of rock, produce the hydrodynamic pressure and pore water pressure, internal erosion of rock, soil, increasing rock, TuRongChong, float force of permeable rock produce , etc. Especially for sliding surfaces (band) the softening effect and reduce the strength of the most prominent role.

# 5 Expert Systems Based on FNN Theory

## 5.1 The Knowledge of Factors Express

[Definition 1]. In the domain of the $U$, the atomic model of knowledge factors is a triple, $M(\circ)=<\circ,F,X>$, where

$\circ$ is the set of objects of the knowledge description of the $U$,
$F$ is the factors set when the $U$ is used to described the $\circ$,
$X$ is performance status about $F$ when $F$ is used to described $\circ$, and
$X=\{Xo\ (f)\ |f\in F,\ o\in\circ\}$.

[Definition 2]. In the domain of the $U$, the relationship between the mode of knowledge that form of expression for the $R(O)=<RM, M(O), XM>$, where

$RM$ is a knowledge model,
$M(O)$ is said atomic model of knowledge representation in knowledge model, and
$XM$ is expressed structure group states and state transform relations of the atomic model $M(O)$ in $RM$.

The atomic model of the knowledge factor expression gives a set of discrete that is perceived described of the object, this is constituted the basis of knowledge expression

with factors. The relationship mode of knowledge factor expression can associate with various related knowledge or different knowledge expression, this can realize the transformation of the different ways of knowledge and knowledge reasoning. They provide basis of expression and processing of knowledge in using factors neural network.

### 5.2 The Simulation Type of Factors Neural Network

The analogous FNN is based on the analogous factor neuron, and the centered is the object of system domain, the factors as a basis to build functional knowledge storage and use of one's information processing unit, with external matching implicit way to complete the processing of information. The simulation type of factors neurons and its network mainly based on simulating the human brain nerve network system, it simulates the behavior of human intelligence from the outside of things and macro function. The simulation type of factors neural network is the basis of simulation type factors neurons, the key of work is to build a functional analog characteristics of this simulation unit, making it has the need kinds of characteristic, such as associative knowledge storage properties, network stability properties and rapid recall function, etc.

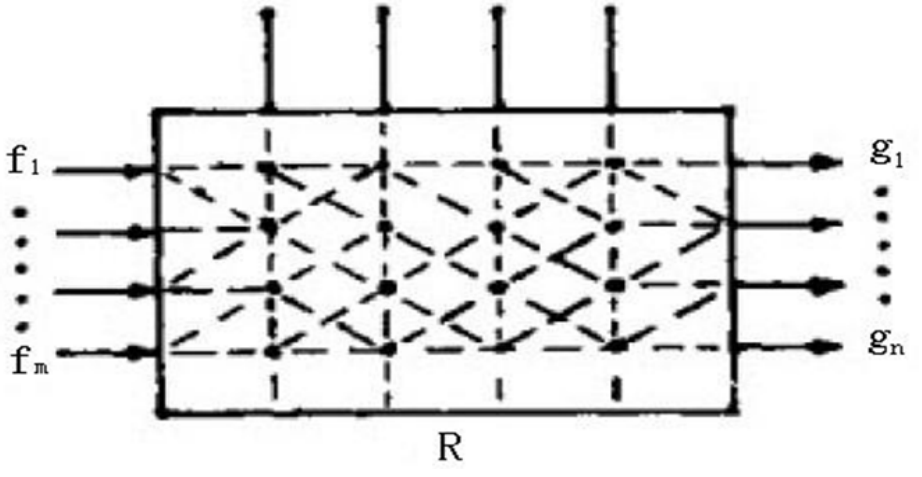



**Fig. 1.** The simulation type of factors neurons model

In Fig. 1, the $f_1 \ldots\ldots f_m$ are input factors that have some connection with ○, each input factors called a perceived channel of the simulation type of factors neurons, the $g_1 \ldots\ldots g_n$ are output factors of the ○, they represent different output response.

$$F_o = \{f_1, f_2, ..., f_m\}$$

$$G_o = \{g_1, g_2, ..., g_n\}$$

$$X_o(F_o) = \{X_o(F_i) \mid i = 1, ..., m\}$$

$$Y_o(G_o) = \{X_o(g_j) \mid j = 1, ..., m\}$$

For one of the simulation type of factors neurons, the external function can be used to express by $Y_o(G_o) = R(X_o(F_o))$, to build the simulation model of neurons within the network module is to try to achieve the purpose of processing this information.

The simulation type of factors neurons achieve the process of intelligent simulation can represent with mathematical formula:

$$Y = F(X, W, T),$$

where

$(X,Y)$ is called the input and output collection mode of the simulation type of factors neurons,

$X$ is called the stimulated or input mode set,

$Y$ is called its corresponding response or output model set, and

$W$, $T$ is controllable parameters of the simulation model of neurons within the network module.

Make the $(X,Y)$ relatively fixed when learning, according to the network characteristics tried to adjust $W$, $T$, to establish the above mapping relationship. Make the $W,T$ relatively fixed when the recalling, the simulation type of factors neurons make $Y$ response according to input $X$.

### 5.3 Expert System

A structure of an expert system is pot in Fig. 2, including 8 parts: Knowledge base, Inference machine, Knowledge Acquisition, Explanatory Mechanism, Blackboard system, Man-machine interface, Interface fro experts and Interface for users. Through the knowledge acquisition interface establish knowledge base, in the using, users will get fact information into the blackboard system, then the inference machine extract

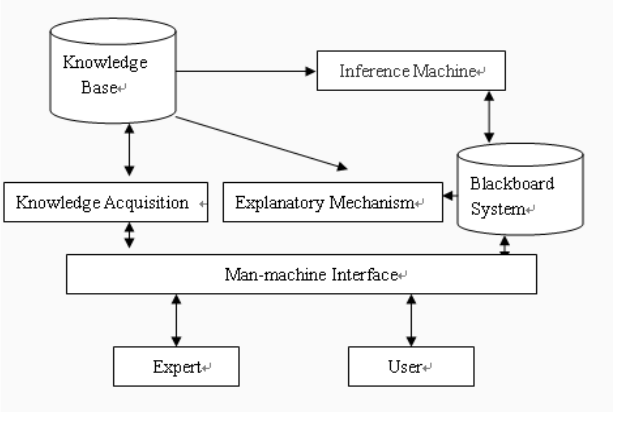



**Fig. 2.** Structure of Expert System

factual information from blackboard system, it matches the knowledge base of rules repeatedly, In this process, middle conclusion will put to the blackboard system, until the system get the final conclusion output, according to user questions, explanatory mechanism explain the final conclusions and the solving process.

Predictions parameters are rock types (*T*), covering layer of the rock (*C*), slope (*S*), displacement (*D*), friction coefficient between rock (*F*), moisture (*M*), and water pressure (*P*).

Thus can be summed up as follows in the form of rules:

if T= Shale and S=80 degrees and M is larger

then Easy appear landslide

When the knowledge base in all of the rules that are lower than output condition,At this point the reasoning used in Multi-stage implies fuzzy reasoning model.

Premise 1 if x is A1 and y is B1, then z is C1,else
Premise 2 if x is A2 and y is B2, then z is C2,else
……
Premise n if x is An and y is Bn, then z is Cn,else
Fact x is A'and y is B'

---

Conclusion z is C'

- When "else" is $\cup$

$$C' = \bigcup_{i=1}^{n} \{ [A' o R(A_i; C_i)] \cap [B' o R(B_i; C_i)] \}$$

- When "else" is $\cap$

$$C' = \bigcap_{i=1}^{n} \{ [A' o R(A_i; C_i)] \cap [B' o R(B_i; C_i)] \}$$

When forecast the landslide in using of landslide parameters and rules of experience, sometimes you can get the status display function or correlation matrix changes, in this case, analytical methods can be used for landslide prediction. However, in actual use the relationship between the parameters of change is very complex, so that to get these shows the relationship is very difficult, in this case, the use of simulation type factor neural network module to simulate and reasoning experience more feasible. For the parameter relation is not easy to show expressed, the principle of treatment is to establish a corresponding simulation type prediction model. Simulation type prediction model forecast landslide mainly to learn the experience, and continuous improvement in using.

The analogous diagnostic mode through the main experience learning to prediction geological disasters, the function entity is chose the forward-type factor neurons as the function entity, it is also set to reverse diagnosis function to show the results, reaction to the people as the basis of reverse validation, continuous improvement in using later. First of all, to establish factors space is based on analysis the monitoring parameters and size analysis of geological disasters, the space included within the output, input canasta system based on FNN is the combination of the modern monitoring technology and modern computer techniques, it is can timely forecasts calculated landslides and other geological disasters, make a prediction to the possibility of impending disaster.

The system has the following advantages:

- First, the system can auto receive the TDR factor information that monitoring through beidou satellite, make a right judgment to the current situation of mountain;
- Second, the system can accept factor information of the monitoring mountain, and also make a right judgment to the current situation of mountain;
- Third, the system can make a detailed forecast to the current situation of mountain; the forecast has time, scale and so on;
- Fourth, the system can record the historical data about the monitoring of the mountain factors, the historical data can store in the knowledge base, this can constantly revised the expert system to improve the accuracy of prediction;
- Fifth, in addition to these above main functions, the expert system also has good man-machine interface.

## 6 Conclusions

The expert system based on FNN theory can make timely forecasts to landslides and other geological disasters through dealing with the deflection, humidity, and other factors of the rock, the factors monitored through by the TDR technology, this will maximum extent to avoid or reduce the SCADA system of destruction from the geological disasters, and provide a guarantee for the SCADA system which plays an essential role in long distance pipeline transport.

## References

1. Liu, Z., Liu, Y.: Factor neural network theory and application. Guizhou Science and Technology Department, Guizhou (1992)
2. Giarratano, J.C.: Expert System: principle. China Machine Press, Beijing (2005)
3. Shi, Y., Zhang, Q.: TDR technology and engineering geology applications
4. Chen, P., Li, J.: Monitoring technology of pipelines using fiber brag grating and application in landslide areas (2010)
5. Tao, G., Fei, L.: Practical Research of Comprehensive Monitoring Means in Landslide Treatment (2010)
6. Kumar, S.: Neural network. Tsinghua University Press, Beijing (2006)
7. Shan, L., Ying, H.: Design of an Early Warning System Based on Wireless Sensor Network for Landslide (2010)
8. Jin, H., Hao, J.: Methods of in-Situ Tests for Final Pile Pressure and Bearing Standard Value. The Chinese Journal of Geological Hazard and Control (2009)
9. Ma, M.: Artificial Intelligence and Expert Systems. Tsinghua University Press, Beijing (2006)
10. Zhang, C., Lou, Z.: Study on Mechanical behavior of nterface between Soil and Rock in complex strata. Yangtze River, 1001–4179 (2010) 17- 0016- 03
11. Mehta, P., Chander, D., Shahim, M., et al.: Distributed Detection for Landslide Prediction Using Wireless Sensor Network. In: First International on Global Information Infrastructure Symposium, pp. 195–198 (2007)
12. Jin, H., Hao, J.: Technique and application of geologic hazard risk semi-quantitative assessment of pipeline. Oil & Gas Storage and Transportation 30(7), 497–500 (2011)

# TDSC: A Two-phase Duplicate String Compression Algorithm

Zhuoluo Yang[1,2], Jinguo You[1,2], and Min Zhou[3]

[1] School of Information Engineering and Automation,
Kunming University of Science and Technology,
Kunming, Yunnan, China
[2] Yunnan Computer Technology Application Key Lab,
Kunming University of Science and Technology,
Kunming, Yunnan, China
[3] Taobao (China) Software Technology Limited Company,
Hangzhou, Zhejiang, China

**Abstract.** Due to current real-time data compression algorithms is not efficient enough, we have proposed a two-phase real-time data compression algorithm which can be very fast in data compression with high compression rate. The algorithm can adapt to both text and binary files. The first phase compresses the file with long common strings into short forms. The second phase compresses the result of the first phase with short bytes in common into the final results. We name the algorithm TDSC Algorithm.

**Keywords:** Compression, Duplicate String, Hadoop.

## 1 Introduction

In many scenarios, especially web based online application scenarios, we need to store and fetch large amounts of data. Sometimes these data are PB scale and we need to put and fetch them in millisecond timescale. In these situations, (1) data can be mostly common in structures (For example, web log usually contains "http://www."). (2) I/O performance is the bottleneck of the system performance while CPU sometimes is not fully made use of. An efficient way to save I/O and enhance performance is real-time data compression.

A real-time compression can compress the data online, so the users can gain the additional time and space efficiencies throughout the data life-cycle. However, most real-time compressions in order to make them running fast are made rather simple even cares less in its compression rate and data structure. It makes it difficult to reduce the I/O throughput effectively.

Due to current real-time data compression algorithms is not efficient enough especially for structured data, we have proposed a two-phase real-time data compression algorithm which can be very fast in data compression with high compression rate and common in use. We have found that a two-phase algorithm can leverage the balance of speed and compression rate. The algorithm can adapt to both text and binary files. We name the algorithm TDSC Algorithm.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 27–34, 2012.

TDSC Algorithm compress the input data block with two phases in order, and decompresses it reversely. The first phase compresses the input file with long common strings into short forms. The second phase compresses the result of the first phase with short bytes in common into the final results. We named the first phase Long Phase and the second phase Short Phase.

## 2 Long Phase

Long Phase represents long common strings that may appear far apart in the input file. The idea comes from Bentley and McIlroy's Data Compression Using Long Common Strings (B&M algorithm) [1]. They have used Karp and Rabin's Efficient pattern-matching algorithms (K&R algorithm) [2] to find long common strings, then feed its output into a standard algorithm that efficiently represents short (and near) repeated strings.

For an input string of length $n$, the original algorithm chooses a block size $m$. It processes every character by updating the fingerprint and checking the hash table for a match. Every $m$ characters, it stores the fingerprint.

We have made three little changes of the B&M algorithm: (1) We have changed the resulting fingerprint to a 55-bit word, saved as a 64-bit unsigned integer, and interpret each factor of the polynomial as an 8-bit unsigned BYTE ranged $[0, 256)$. Because K&R proved that the probability of two unequal strings having the same E-bit fingerprint is near $2^{-E}$, and also $E \times BYTE \leqslant 64bit$, the 64-bit CPU word length. (2) Instead of using "$< start, length >$" where $start$ is the initial position and $length$ is the size of the common sequence. We use the byte "$0xfe$" to identify the encoded block. Because bytes $0xfe$ is unlikely to show in normal text files and it also shows less in binary files. For original $0xfe$ bytes are quoted as $0xfe\,fe$. The next byte after $0xfe$ is *length identification byte (LIB)* which identify the width of $start$ and $length$. We store the $start$ and $length$ sequentially in little endian. (3) We have used two different hash functions to compute table size and hash index.

### 2.1 Data Structure

The data structure of the Long Phase is the hash table which stores the fingerprint $fp$ and block index $pos$.

The $size$ of hash table is computed as *equation* 1 shows.

$$size = \left\lfloor \frac{2n}{m} + 1 \right\rfloor \tag{1}$$

It is implemented as an array, and the index of the array $hash\_index$ is computed as *equation* 2.

$$hash\_index = (fp \;\&\; 0x7fffffff) \bmod size \tag{2}$$

### 2.2 Algorithm

Each block $B$ can be interpreted as a polynomial $b$ and fingerprint of the block can be interpreted as $b$ modulo a large prime. For $d$ is the factor of the polynomial and $m$ is the length of the $B$, We can use Horner's rule to express $b$ as *equation* 3 shows.

$$b = B_{m-1} + d(B_{m-2} + d(B_{m-3} + ... + d(B_1 + dB_0)...)) \tag{3}$$

For $t$ is the fingerprints of the input text, $t_s$ is the current fingerprint, we can calculate $t_{s+1}$ in O(1) complexity. *equation* 4 shows the recurrence of the rolling fingerprint computing.

$$t_{s+1} = d(t_s - d^{m-1}T_{s+1}) + T_{s+m} \tag{4}$$

However, $b$ and $t_s$ may be very large, longer than the CPU word length. So a large prime $q$ is chosen for equivalent fingerprint computing and $f$ are chosen for equivalent fingerprints, show in *equation* 5.

$$f_s \equiv t_s \pmod{q} \tag{5}$$

Because of $q > d$ and $q > T_s$, we make the programmable equivalent *equation* 6.

$$\begin{cases} f_s = \sum_{i=0}^{m-1} T_i d^i \bmod q & s = 1 \\ f_{s+1} = d((f_s \bmod q) - (d^{m-1} \bmod q)T_{s+1}) + T_{s+m} & s > 1 \end{cases} \tag{6}$$

The implementation of the algorithm choose the modulo divisor $q$ a very large prime, so the pseudo hit of the fingerprints in the hash table is extremely small. The Long Phase algorithm expresses in the pseudo code as *algorithm* 1 shows.

We let $h = d^{m-1} \bmod q$, so $h$ will be a constant if block size $m$ is a constant, and *line* 1 of *algorithm* 1 also can be calculated in O(1) complexity by *algorithm* 2.

The *line* 13 of *algorithm* 1 looks up $fp$ in the hash table and encodes a match if one is found.

After checking the block every character in order to ensure that the block with matching fingerprints is not a false match, we greedily extend the match forwards as far as possible and backwards never more than $b1$ characters (or it would have been found in the previous block). If several blocks match the current fingerprint, we encode the largest match among them.

## 3 Short Phase

Short Phase borrows its idea from LZ77 [3] and compresses the result of the Long Phase. Google Snappy [4] is also an LZ77 like compression algorithm. Short Phase encapsulates the Snappy library and make a few changes.

Short Phase hashes the input string every 32 bits (4 bytes). If some hashes match, it writes 4-byte size and current literal bytes once, afterwards it writes copy signature one or more times. The loop continues for 4-byte hash matching, until it reaches the end of the input string or the rest is not enough for the hash matching.

**Input**: $T$= the string to be compressed
**Data**: $f$, $m$, $d$, $q$, *table*
**Output**: $C$ = the compressed result of $T$
$h \leftarrow d^{m-1} \bmod q$;
$f \leftarrow 0$;
**for** $i \leftarrow 0$ **to** $m-1$ **do**
    $f \leftarrow (mf + T_i) \bmod q$;
**end**
**for** $i \leftarrow 0$ **to** $length[T] - m - 1$ **do**
    **if** $i \bmod b = 0$ **then**
        Store $f$ to the hash table.;
    **end**
    $f \leftarrow (d(f - hT_i) + T_{i+m}) \bmod q$;
    **if** $f$ *exists in table* **then**
        **if** $T[i..i+m] = T[table[f] \rightarrow s..table[f] \rightarrow s+m]$ **then**
            Encode $T[i..i+m]$ as $i$;
        **end**
    **end**
**end**

**Algorithm 1.** Long Phase Algorithm

**Input**: $a$, $b$, $n$
**Output**: $a^b \bmod n$
let $< b[k], b[k-1], ..., b[0] >$ be the binary representation of $b$;
$result \leftarrow 1$;
**for** $i \leftarrow k$ **to** 0 **do**
    $result \leftarrow 2 \times result$;
    **if** $b[i] = 1$ **then**
        $result \leftarrow (result \times a) \bmod n$;
    **end**
**end**

**Algorithm 2.** Modular Exponentiation

Google Snappy has a heuristic match skipping. If 32 bytes are scanned with no matches found, start looking only at every other byte. If 32 more bytes are scanned, look at every third byte, etc. When a match is found, immediately go back to looking at every byte. This is a small loss (about 5% performance and 0.1% density) for compressible data due to more bookkeeping, but for non-compressible data (such as JPEG) it's a huge win since the Short Phase quickly "realizes" the data is incompressible and doesn't bother looking for matches everywhere.

### 3.1 Data Structure

The data structure is also a hash table which key is the hash code of the 4-byte string and the value is the offset of the string. The hash table is used for fast encoding of string matching. It is implemented as an array of length of power of 2. Assume $N$ is the length of array, $b$ is the four bytes pattern and $h$ is the hash code of the pattern, equation 7 shows the hash code algorithm.

$$h = b \times 506832829 \div \log_2^{32-\log_2^N} \tag{7}$$

Note that the division of $\log_2^X$ can be implemented as right shift of $X$.

### 3.2 Algorithm

The short phase algorithm can be implemented as *algorithm* 3 shows.

```
1  shift ← 32 − log_2^(table_size+1);
2  next_hash ← getHash(ip, shift) ;
3  repeat
4      hash ← next_hash ;
5      ip ← ip + 4 ;
6      next_hash ← getHash(ip, shift) ;
7      if hash = next_hash then
8          writeLiteral();
9          repeat
10             hash ← getHash(ip, shift) ;
11             writeCopy();
12         until hash ≠ next_hash;
13         if length − ip < 4 then
14             break;
15         end
16     end
17 until false;
18 writeLiteral();
```

**Algorithm 3.** Short Phase Algorithm

The *line* 2 of *algorithm* 3 is to use *equation* 3 to compute the hash code.

Compression searches for matches by comparing a hash of the 4 current bytes with previous occurrences of the same hash code earlier in the 32K block. The hash function interprets the 4 bytes as a 32 bit value, little endian, multiplies by $0x1e35a7bd$ (*equation* 7), and shifts out the low bits.

## 4 Performance Analysis

Both of our two phase implementations of the algorithm stored the entire input file in main memory and scans the input bytes only once. Meanwhile, the algorithm only kept a small block of the data during the compression, so the algorithm could be used to compress the data in disk with low complexity which the only overhead is to keep the small block of the data. And the performance bottleneck of compression would be I/O.

The experiments are done on a laptop with an Intel Celeron U3400 CPU (1.07GHz, 64bit) and Ubuntu 11.10 amd64 operation system. We made our experiment by making three text files into the same one using "$cat \ * .txt \ >$

*text.txt*" in Linux. The three text files ware *Alice's Adventures in Wonderland*, *Paradise Lost* and *As You Like It* from test data of Snappy [4]. The size of combined file was 1185833 bytes. We compressed the text with our TDSC Algorithm and also Snappy and GZip for comparison, the results show in *table* 1 and 2.

**Table 1.** Text File Compression

| Algorithm | Compressed Size | Speed | Time Cost |
|---|---|---|---|
| GZip | 442005 B | 15.91 M/s | 62826947 ns |
| Snappy | 639191 B | 65.18 M/s | 15340582 ns |
| TDSC | 620697 B | 66.85 M/s | 14958797 ns |

**Table 2.** Text File Decompression

| Algorithm | Speed | Time Cost |
|---|---|---|
| GZip | 94.52 M/s | 10574833 ns |
| Snappy | 135.79 M/s | 7363479 ns |
| TDSC | 133.71 M/s | 7448363 ns |

We also tested the binary files. We prepared an Ubuntu 11.10 amd64 ISO file as the input file and compress the it with TDSC, Snappy and GZip. The size of input file is 731.2 mega bytes. The results show in *table* 3 and 4 .

**Table 3.** Binary File Compression

| Algorithm | Compressed Size | Speed | Time Cost |
|---|---|---|---|
| GZip | 719.7 MB | 9.42 M/s | 77.62 s |
| Snappy | 728.3 MB | 39.63 M/s | 18.63 s |
| TDSC | 724.4 MB | 37.62 M/s | 19.43 s |

**Table 4.** Binary File Decompression

| Algorithm | Speed | Time Cost |
|---|---|---|
| GZip | 48.04 M/s | 15.22 s |
| Snappy | 120.87 M/s | 6.05 s |
| TDSC | 112.48 M/s | 6.51 s |

The speed comparison is seen in Fig. 1. We noticed that the performance of TDSC is much better GZip. Even though the short phase encapsulates Snappy, the performance of TDSC is not worse than Snappy. For the long phase has already compressed the input file and the short phase will get the not-so-large input and make better compression rate.

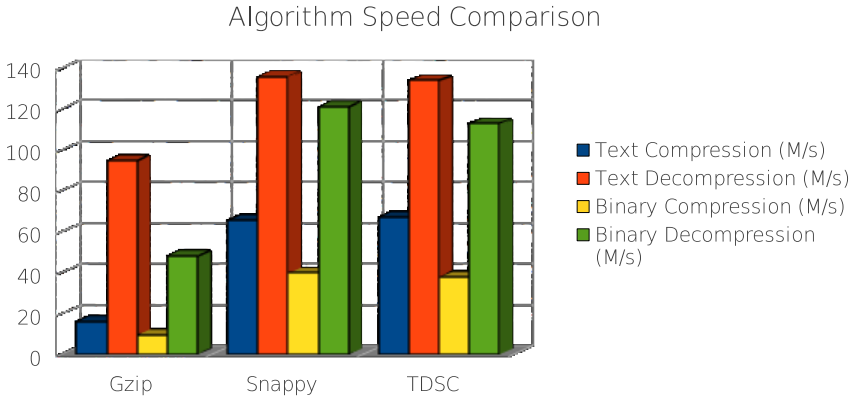



**Fig. 1.** Speed Comparison

## 5 Usage Scenarios

We use the compression algorithm in our distributed columnar storage database based on Hadoop [5]. We also put the algorithm in the column storage which is one of our own features of Pig [6] and Hive [7]. Each column group is similar in structure, so we propose a proper block size of Long Phase to make the put and fetch in real-time. We also open the TDSC source code in `http://code.google.com/p/tdsc` in an Apache License 2.0, so the user can use it anywhere else.

## 6 Summary

We have described a Two-phase Duplicate String Compression (TDSC) algorithm that effectively represents any long common strings that appear in a file first and short common string next. We have extended the fingerprint mechanisms and hash algorithm in real-time compression. We have also developed a new way to store the common string in compression.

Our experiments have showed that the TDSC algorithm had performed very well in the text and binary files than the traditional compression algorithm.

In the future, we will analysis the effect of different parameters in the TDSC algorithm and try to figure out the optimal parameters of different files.

**Acknowledgments.** The research was supported by Yunnan Educational Foundation (09C0109), Natural Science Foundation of Yunnan Province (2010ZC030), Natural Science Foundation of China (71161015). We are grateful to the Basis Development team at Taobao.com for helping with the research.

## References

1. Bentley, J., McIlroy, D.: Data Compression Using Long Common Strings. In: Data Compression Conference, pp. 287–295. IEEE Press, New York (1999)
2. Karp, R.M., Rabin, M.O.: Efficient Randomized Pattern-matching Algorithms. IBM Journal of Research and Development 31(3), 249–260 (1987)
3. Ziv, J., Lempel, A.: A Universal Algorithm for Sequential Data Compression. Information Theory 23(3), 337–343 (1977)
4. Snappy - A fast compressor/decompressor, http://code.google.com/p/snappy
5. Apache Hadoop, http://hadoop.apache.org
6. Apache Pig, http://pig.apache.org
7. Apache Hive, http://hive.apache.org

# Twig Pattern Matching Running on XML Streams

Ziqiang Deng*, Husheng Liao, and Hongyu Gao

College of Computer Science, Beijing University of Technology, Beijing 100124, China
dengzq078@139.com, {liaohs,hygao}@bjut.edu.cn

**Abstract.** Twig pattern matching plays an important role in XML query processing, holistic twig pattern matching algorithms have been proposed and are considered to be effective since they avoid producing large number of intermediate results. Meanwhile, automaton-based approaches are naturally used in filtering XML streams, because Finite State Machines(FSMs) are driven by events which conform to event-based XML parser SAX. In this paper, we proposed a hybrid approach combining FSM and holistic twig matching algorithm to find occurrences of twig pattern in XML streams. That is, we locate the lowest common ancestor(LCA) of return node(s) in twig pattern, decompose the twig pattern according to the LCA, use automaton-based approach for processing the sub twig pattern above LCA, and regular holistic twig pattern matching algorithm for the sub twig pattern below LCA. It only needs to buffer the elements between the start and end tag of LCA. Experiments show the effectiveness of this approach.

**Keywords:** twig pattern matching, XML streams, FSM, XML query processing.

## 1 Introduction

The rich expressiveness and flexible semi-structure of XML makes it the standard format for information exchange in different scenarios. Many Internet applications employ a model of data stream, the characteristics of data stream differ from conventional stored data model can be found in [1]. For XML streams, the features are usually as follows: 1)the data elements arrive as tokens, a token may be a start tag, an end tag or PCDATA for text of an XML element; 2)queries to XML streams are written in XPath or XQuery, 3)the processor is driven by events raised by SAX.

XML data can also be viewed as ordered labeled tree, Figure 1,2(a) depicts a sample XML tree fragment and a query written in XQuery.

It's easy to see that the referred elements in Figure 2(a) form a tree structure as Figure 2(b), in which single line means parent-child(PC) relation, double line means ancestor-descendant(AD) relation. This is the twig pattern. Twig pattern is an abstract representation of query to XML data, given a twig pattern Q and an XML data fragment D, a match of Q in D can be defined as a mapping from nodes in Q to

* Corresponding author.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 35–42, 2012.

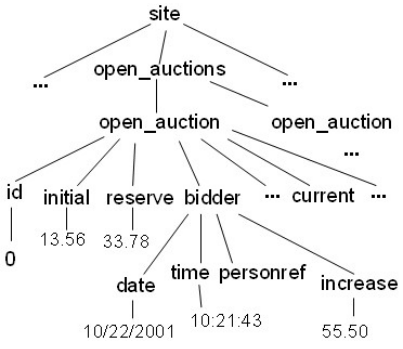



**Fig. 1.** A Sample XML Fragment viewed as labeled tree

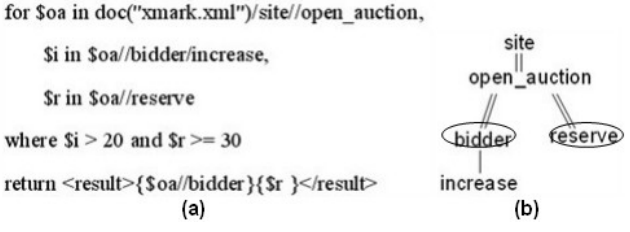



**Fig. 2.** A Sample XQuery Expression and A Twig Pattern Query

elements in D, such that 1)query node predicates are satisfied by the corresponding data elements, 2)the structural(PC/AD) relationships between query nodes are satisfied by the corresponding data elements. The answer to query Q with n nodes can be represented as an n-ary relation where each tuple(d1, ..., dn) consists of the data elements that identify a distinct match of Q in D.

Twig pattern matching is the core operation of XML query processing, and it's been widely studied recently[2-4]. Extensions have been made to twig pattern for processing XQuery expressions, e.g. Generalized Tree Pattern(GTP)[5]; in this paper, we consider the GTP with return node(s) only, which is depicted with circles as in Figure 2(b), thus the answer consists only return nodes(s).

Automaton is naturally suited for pattern matching on tokenized XML streams, it answers the query quickly and doesn't rely on any specific indexes, however, automaton-based approach like YFilter[6] is more suitable for filtering XML stream, when processing twig query with more than one return nodes as in Figure 2(b), it conducts many post processing to compute the combined answers of multiple paths and seems not that natural. On the other hand, regular twig pattern matching algorithms(Twig$^2$Stack[4], TwigList[3] etc.)process twig pattern matching efficiently, yet rely on some indexes(region code, tag streams), while in an stream environment, it's sometimes impossible to build the index before query evaluation because the data has not arrived.

**Motivation.** We observed that: 1)normally, only the return node(s) specified in XQuery need to be stored at processor side, such as *bidder* and *reserve* in Figure 2; 2)XML data fragment usually concerns about some topic, elements with the same tag

appear repeatedly, and queries mainly focus on these elements and their sub elements, such as *open_auction* in the previous example; every time an </open_auction> arrives, it means that all its sub elements arrived already, so the processor can decide whether it satisfies the sub twig query rooted at *open_auction* or not, and make a twig query evaluation to output an answer, or discard the elements arrived yet. We may carry out the query evaluation like this: decompose the twig pattern at *open_auction*, which is an common ancestor of *reserve* and *bidder*, into two sub twig patterns as in Figure 3(a)(b), for sub twig pattern above *open_auction*( Figure 3(a) ), build an NFA like in YFilter, the elements above *open_auction* will drive the NFA to locate the right *open_auction* in the stream, for sub twig pattern below *open_auction*, buffer elements between <open_auction> and </open_auction>, then make a twig pattern evaluation with a regular twig algorithm, thus, every time an </open_auction> arrives, we're able to output an answer if this *open_auction* and its sub elements satisfy the query, this is an incremental process.

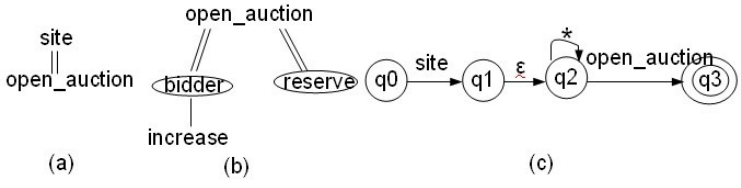



**Fig. 3.** Sub Twig Patterns and NFA Built From Twig Pattern

The rest of this paper is organized as follows, section 2 describes the hybrid approach including the framework, algorithm for constructing NFA from twig pattern, algorithm for NFA execution; section 3 discusses related work, section 4 gives conclusion.

## 2 A Hybrid Approach for Processing Twig Pattern Matching on XML Streams

In this section, we'll give a solution to this problem: Given a GTP G and ordered accessed XML stream S, compute the answer of G on S. The following example in Figure 4 will be used to explain how the algorithm works, for ease of understanding, we depicts the XML data in a tree, actually, the processor will be fed with tokens, i.e. <a1>, <b1>,<a2>,<c1>,</c1>... in order.

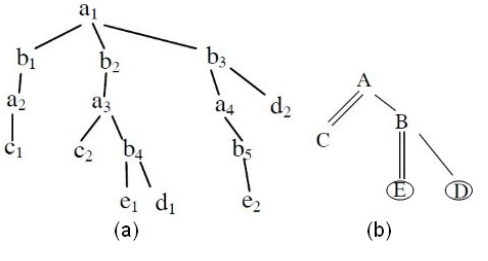



**Fig. 4.** A Sample XML Fragment and A Twig Pattern with Return Nodes

**Definition. Lowest Common Ancestor(LCA):** for node(s) $n_1$, $n_2$,..., $n_i$ of a GTP G, v is the LCA of these nodes if and only if: 1) v is the ancestor of $n_k$ ($1 \le k \le i$), and 2) there exists no node u, such that u is a descendant of v and u is the ancestor of $n_k$ ($1 \le k \le i$).

**Algorithm1: Framework**

Input: a GTP query G, an XML data stream S;

Output: the answer of G against S;

1 LCA <- getLCA(G);

2 decompose G into two parts at LCA;

3 for the sub twig pattern above LCA(*subTPABOVE*), construct an NFA N;

4 for the sub twig pattern rooted at LCA(*subTPBELOW*), do the initial work required later in twig pattern matching algorithm;

5 open S to execute the NFA N;

Algorithm1 is the framework of this solution, it first obtains the LCA of return node(s) using getLCA(). After decomposition, LCA will be a leaf of *subTPABOVE*. Constructing an NFA in line 3 is just like combining several NFAs into a single one in YFilter, every leaf of sub twig pattern corresponds to an accepting state in NFA. The NFA is used to locate the correct LCA element in S, then elements between <LCA> and </LCA> will be buffered, algorithm for NFA execution(Algorithm3) transfers control to a regular twig pattern matching algorithm(we call it evalTwigPattern() ) such as Twig$^2$Stack when the corresponding </LCA> arrives.

We now need an algorithm to construct NFA from a twig pattern. Notice that, there may be several accepting states in the NFA, we say an XML fragment matches twig pattern if it drives the NFA to all of its accepting states.

**Algorithm2 : NFAConstructor**

Input: a twig pattern G(G is not null);

Output: an NFA N that is equivalent to G;

```
1   create initial state q0 of N;
2   create state q1;
3   create transition δ(q0, root[G]) = q1 for the root of G;
4   foreach child Xi of root[G]
5       create new state qx;
6       case the relationship between root[G] and Xi of
7           PC:
8               create transition δ(q1, Xi) = qx;
9           AD:
10              create transition δ(q1, ε) = qx;
11              create transition δ(qx, *) = qx;
```

```
12                  set the type of q_x to "//-child";
13                  create state q_y;
14                  create transition δ(q_x, X_i) = q_y;
15          if X_i is a leaf
16              set the type of new created state(q_x or q_y) to "accepting" and return;
17          process X_i recursively;
```

Algorithm2 first creates initial state $q_0$ for N, then creates transition from $q_0$ for root node of G(line 1-3), next deals with root node's children(line 4-14), notice for AD relation, there's an nil transition to an intermediate state whose type is "//-child" and will be treated specially in NFA execution, see [6].

For the example at the beginning of this section, we obtain two sub twig patterns and an NFA in Figure 5 through Algorithm 1 and 2.

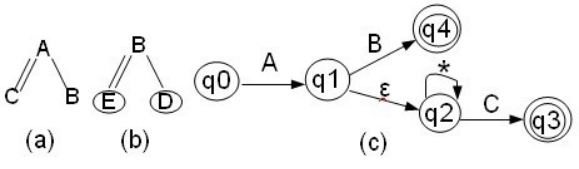



**Fig. 5.** Sub Twig Patterns and NFA Built From Twig Pattern

**Algorithm3 : NFAExecution**

Input: NFA N , sub twig pattern rooted at LCA(*subTPBELOW)*, XML data stream S;
Output: the answer of G against S, G is the original GTP query in Algorithm1;
//*stack* is the run-time stack used to maintain nested structure of XML data
//*active_states* and *target_states* are for state transitions
//*buffered* is the sign whether it needs to buffer the element, i.e. whether in <LCA> and </LCA>, it initiates to FALSE
//*current_tag* is raised by each event(callback method offered by SAX)
// *accepted*[] are symbols used to remember whether an accepting state is covered, initially *accepted*[] are all FALSE
//δis the transition table of N

```
1     add q_0 of N to active_states;
2     push active_states into stack;
3     while not end( S )
4         case current event of
5             START OF ELEMENT:
6                 if buffered = TRUE or ( current_tag = <LCA> and TRUE =
          checkLCA() )
7                     store current element corresponding to current_tag into
          buffer;
8                 if current_tag = <LCA> and TRUE = checkLCA()
9                     buffered <- TRUE;
10                foreach state q_a in active_states
```

```
11                  q_t <-δ(q_a , current_tag);
12                  if q_t != null
13                      add q_t to target_states;
14                  if q_a's type is "//-child"
15                      add q_a to target_states;
16                       q_u <-δ(q_a , ε);
17                       if q_u != null
18                           process q_u recursively for the above steps;
19                       if target_states is empty
20                           push a dumb state into stack;
21                       else
22                           active_states <- target_states;
23                           push active_states into stack;
24                  if there is a q in active_states such that q is an accepting state
25                      accepted[current_tag] <- TRUE;
26              END OF ELEMENT;
27                      pop( stack );
28                      active_states <- top( stack );
29                      if all accepted[] are TRUE and current_tag = </LCA> and
                active_states equals to the active states before <LCA> arrives
30                          evalTwigPattern( subTPBELOW, buffer );
31                          clear buffer;
32                          buffered <- FALSE;
33                      accepted[current_tag] <- FALSE;
```

When XML stream arrives to be parsed, the execution of NFA begins at the initial state $q_0$ which is the only active state for the moment, push $q_0$ into stack. Every time a start tag arrives, there are two conditions(line 6) to buffer the current element: 1)*buffered* is TRUE, this is obvious; 2)*buffered* is FALSE, but *current_tag* = <LCA> and checkLCA() returns TRUE, in this case, *current_tag* is the correct <LCA>, and checkLCA() pre-computes, for each state in current active states, check whether there's a transition leads to an accepting state corresponding to LCA because LCA is a leaf of sub twig pattern above LCA, if there is, checkLCA() returns TRUE, otherwise FALSE. Next operations(line 10-18)are to make transitions, this is like YFilter except for the * checking. After these, if *target_states* is empty, put a dumb state into S, otherwise, set *active_states* to *target_states*, and push them into stack, then check each of the state in *active_states*, if there is an accepting one, set *accepted*[*current_tag*] to TRUE, this indicates that the path whose leaf is current element has been covered. so set *accepted*[*current_tag*] to TRUE.

Every time an end tag arrives, backtrack the NFA to the states it was when the corresponding start tag arrives(line 27), if *current_tag* is the </LCA> corresponding to <LCA> in condition 2 above, e.g. </b2> instead of </b4>, and all *accepted*[] are TRUE, it's time to do the twig pattern matching, then set the corresponding variables to the right value. More results may be produced as the algorithm continues until the end of stream.

we can easily see it evaluates twig pattern matching for b1, b2, b3 and their sub elements, there are no results for b1, and (e1, d1) for b2, (e2, d2) for b3 will be returned.

**Analysis.** The correctness of the solution in this section lies in the fact that Algorithm3 locates the right <LCA> and </LCA> in stream: every time we invoke evalTwigPattern(), the fragment that has been scanned satisfies the sub twig pattern above LCA, line 24, 29 of Algorithm3 ensure this.

The NFA is driven by events when parsing XML stream, its running time is linear with O(|S|), in addition, suppose the running time of evalTwigPattern() over G and buffered elements is fun( |B|, |G| ), so the time complexity of the solution is O(|S| + fun( |B|, |G| )), where |S| is the size of XML stream, |B| is buffer size, |G| is the size of twig pattern.

The maximum size of run time stack S in Algorithm3 is the maximum depth of XML stream; elements between <LCA> and </LCA> are buffered, usually we cannot know in previous the size of buffer in the processor side, However experiments for the given use cases show that in the normal cases, the buffer won't be too large.

**Experiments.** We implemented the idea above on a PC with 2.93 GHz CPU and 2GB RAM running Windows XP and Java 1.6, three typical data sets and 9 queries appeared in our experiments, TwigList[3] is used to implement evalTwigPattern().

Experiments show that it won't buffer too many elements to compute the correct answers for practical queries, and the scalability of memory consumption is good, because buffer size is bound to LCA elements, it is typically just related to the structure/type of document and is independent of the size of documents. Detailed experiments are ignored because of space limit.

## 3 Related Work

[1] is an early overview paper addressing data stream model based on relational data. It talked about some problems and approaches for implementing a general purpose Data Stream Management System(DSMS). Some motivating example applications can also be found in [1]. [10] is a survey focused on XML streams, it shows that people have done many works in this area through some projects.

XFilter[9] pioneered the use of FSMs for XML filtering, YFilter is the subsequent work of XFilter, it exploits commonality among many path queries by merging the common prefixes of the paths so they are processed only once.

Twig$^2$Stack[4] is a representative Algorithm for twig pattern matching, in the first phase, it stores the intermediate results in a complex hierarchical stack structure, then enumerate the query results from the hierarchical stacks. Twig$^2$Stack has good performance and is capable of efficiently processing GTP queries. TwigList[3] comes after and is similar with Twig$^2$Stack, yet it uses a much more simpler data structure and has better performance.

## 4 Conclusion

In this paper, we proposed a solution combining NFA and regular twig pattern matching algorithm for computing answers of twig pattern on XML streams, this solution is based on the observation that LCA of return node(s) in twig pattern captures the semantics of twig pattern and structure of XML data. Experiments show that it won't buffer too many elements to compute the correct answers for practical queries. This solution is helpful for implementing XQuery on XML streams.

## References

1. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and Issues in Data Stream Systems. In: Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS 2002 ), pp. 1–16 (2002)
2. Bruno, N., Koudas, N., Srivastava, D.: Holistic twig joins: Optimal XML pattern matching. In: Proceedings of the ACM SIGMOD International Conference on Managment of Data, pp. 310–321 (2002)
3. Qin, L., Yu, J.X., Ding, B.: *TwigList*: Make Twig Pattern Matching Fast. In: Kotagiri, R., Radha Krishna, P., Mohania, M., Nantajeewarawat, E. (eds.) DASFAA 2007. LNCS, vol. 4443, pp. 850–862. Springer, Heidelberg (2007)
4. Chen, S., Li, H., Tatemura, J., et al.: Twig$^2$Stack: Bottom-up Processing of Generalized-Tree-Pattern Queries Over XML Documents. In: 30th International Conference on Very Large Data Bases, VLDB 2006, Seoul, Korea, pp. 283–294 (2006)
5. Chen, Z., et al.: From Tree Patterns to Generalized Tree Patterns: On Efficient Evaluation of XQuery. In: VLDB 2003, Berlin, Germany (2003)
6. Diao, Y., Altinel, M., Franklin, M., Zhang, H., Fischer, P.: Path sharing and predicate evaluation for high-performance XML filtering. ACM Transactions on Database Systems 28, 467–516 (2003)
7. Diao, Y., Franklin, M.: Query processing for high-volume XML message brokering. In: Proceedings of the 29th International Conference on Very Large Data Bases, vol. 29, pp. 261–272. VLDB Endowment (2003)
8. Josifovski, V., Fontoura, M., Barta, A.: Querying XML streams. VLDB Journal (2004)
9. Altinel, M., Franklin, M.J.: Efficient filtering of XML documents for selective dissemination of information. In: Proceedings of VLDB Conference (2000)
10. Weidong, Y., Baile, S.: A Survey of XML Stream Management. Journal of Computer Research and Development (2009)

# A Novel Method for Extension Transformation Knowledge Discovering

Xingsen Li[1], Zhongbiao Xiang[2], Haolan Zhang[1], and Zhengxiang Zhu[3]

[1] School of Management, Ningbo Institute of Technology, Zhejiang University, Ningbo, China
[2] School of Management, Zhejiang University, Hangzhou, China
[3] Graduate School, National Defence University PLA China, Beijing, 100091, China
{lixs,haolan.zhang}@nit.zju.edu.cn,
xiangzhongbiao@zju.edu.cn, zzx_dl@hotmail.com

**Abstract.** On the foundation of analyzing the existing classification, an acquisition method of extension transformation knowledge based on Decision Tree classification has been proposed The new-bored method re-mines and transforms the decision tree rules to "can't to can, not to yes" strategy which aims to provide targeted decision-making on the transformation of the customer churn by flexible use of the extension set and extension transformation theory. Its practice in a web company has proved that this method is highly feasible, and also has the reference value for other methods research based on Extenics.

**Keywords:** Extension transformation, transformation knowledge, decision tree, Rules Mining customer retention.

## 1 Introduction

Data mining as an important instrument for knowledge discovery has been widespread concerned by scholars to support business strategy [1-2]. Especially, the decision tree classification which has stronger interpretability of classification principle is one of the most commonly used data mining measures[1,3].However, the pattern knowledge obtained from data mining is not practical enough.

In recent years, Extension theory [4] has been initially applied in the field of data mining, and achieved good result. Bibliography [5] has had outlook-style description on extension application in Data mining. Methods were exploited of the corresponding potential knowledge by utilizing properties of matter element including divergence, relevance, and implication, arousing our concerns on potential information category [6]. And then analyzed the existing problem caused within the data mining process based on extension theory, and also proposed a new data mining application based on extension transformation through establishing matter-element set[7]. The conception and assumption of extension classification has been presented [8] and laid the foundation for further research.

The rest of the paper is organized as follows: In section 2 we reviewed extension set and decision tree method for preparing our new method. In section 3 we presented

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 43–50, 2012.

the algorithm design and implementation of transformation knowledge in detail. Section 4 gives a case study and show how this new methodology works. In section 5 we conclude the paper and give future research directions.

## 2 Analysis of Transformation Rules Mining

### 2.1 Three Kinds of Set Theories on Classification

Classification can provide a reference for decision. The existing category mining is based on classical set theory and fuzzy set theory. The classical theory requires each object must belong to one set and only one or the other. The classical collection uses the characteristic function of range {0, 1} to qualitatively describe whether one thing has certain property [5]. The fuzzy set uses the membership function of range {0, 1} to describe the degree of certain things which are undergoing differences during the intermediate transition. Neither classical set nor fuzzy set study the changes among the categories of things, therefore both of them cannot directly describe the conversion between "non" and "is" in certain condition [5]. Frankly speaking, many things can divided into two parts according to P property. The part which does not have the nature of P can be further divided into two categories, one is "can be convert into holding P property" and the other is "cannot be transformed into possessing the nature of P" under certain condition. In the actual production, For instance, unqualified products can turn to be qualified after some processing.

### 2.2 Theory of Extension Set

Extension set reflects the degree of transformation of the nature of things. Its definition can be stated as follows:

**Definition 1.** Set $U$ as universe, $u$ is any element in $U$, $k$ is a mapping of $U$ into real domain $I$, the given transformation of $T=(T_U,T_k,T_u)$ refers to the following equation.

$$\tilde{E}(T)=\left\{(u,y,y')\middle| u\in T_U U, y=k(u)\in I, y'=T_k k(T_u u)\in I\right\}$$

The above set is a extension set of $U$ universe, $y=k(u)$ is correlation function of $\tilde{E}(T)$. $y'=T_k k(T_u u)$ is extension function of $\tilde{E}(T)$. $T_U,T_k,T_u$ is separately the transformation of universe $U$, correlation quasi-function $K$, and element $u$.

On the conditions of $T\neq e$, $\underset{\cdot}{E}_{+}(T)$ is the positive extension domain of $\tilde{E}(T)$, $\underset{\cdot}{E}_{-}(T)$ is the negative domain of, $E_{+}(T)$ is positive stable domain of $\tilde{E}(T)$, $E_{-}(T)$ is negative stable domain of $\tilde{E}(T)$, $J_0(T)$ is extension bounding of $\tilde{E}(T)$.

Positive extension domain indicates that element which initially does not belong to $E$ turn to a part of $E$ after the implementation of $A$ transformation; while the Negative extension domain reflects element which initially does not belong to $E$ remain the same after $A$ transformation; Positive stability domain refers to element which originally belong to $E$ remain belongs to $E$ after the implementation of $A$ transformation, Similarly, Negative stability domain refers to element which originally does not belong to $E$ is still not part of $E$ after $A$ transformation. Extension boundary refers to element which existed at the border of extension transformation and its extension function is zero. Extension boundary describe the qualitative change point which indicates that elements surpass the point will definitely produce qualitative change. T

### 2.3 Decision Tree Method

Decision tree is a tree structure similar to flow chart, where each internal node represents a test on attribute, each branch represents output of test, and each leaf node represents a category. Decision tree as data set classifier resulted in a static subset of classification of data which only describe the characteristics of different branches of leaves, and therefore effective measures aiming to prevent the loss of customers only relied on classification rules are invalid. Using matter element extension set to represent the results of Decision tree mining so that it can transform the static rule set into changing, dynamic extension rule set to dynamic strategy generation.

## 3 The Algorithm Design and Implementation of Transformation Knowledge

### 3.1 The Method of Obtaining Transformation Knowledge

In order to obtain strategies of classification transformation rules turn to change through extension transformation mining On the foundation of decision tree classification rule connecting with extension set theory. Take A, B two types of conversion as example.

Set up A as:

$$\left\{D_{\cdot+}(T)\right\}=\left\{D_i \middle| D_i=\begin{bmatrix} I_i, & d_1, & u_{i1} \\ & d_2, & u_{i2} \\ & d_r & u_{ir} \end{bmatrix}, K_i<0, K_i \bullet K_i(T)<0, i\in J_{D_{\cdot+}}\right\}$$

$J_{D_{\cdot+}}$ is index set of information unit $D_i$ which meet the condition of $K_{ip}<0, K_{ip}\bullet K_{ip}(T)<0$, the later sign is similar, so no long explain

$$\left\{I_{\cdot+}(T)\right\}=\left\{I_i \middle| I_i=\left(O_i, c_j, v_{ij}\right), D_i\in\left\{D_{\cdot+}(T)\right\}, i\in J_{D_{\cdot+}}\right\}$$

Set all of relevant characteristics of $c_j\left(j=1,2,\cdots,m\right)$ and $d_p\left(p=1,2,\cdots,r\right)$ as $\left\{j_0\right\}$ ,change the property value of rules hold the same property in the two types of rule based on $A \leftarrow B$ replacement transformation; there must be transformation set called $T_{AB}$ which make $A \Rightarrow B$ by transforming the rule existed in B but not in A through adding transformation;

The reliability is $\frac{\left|D_{\cdot +}\left(T\right)\right|}{\left|D_{-}\left(T\right)\right|}$ ,support degree is the transformation knowledge of $\frac{\left|D_{\cdot +}\left(T\right)\right|}{\left|D\right|}$ ,which indicates that about $j \in \left\{j_0\right\}$, if $I=\left(O,c_j,v_j\right)$ have $v_j \in V_{\cdot +j}\left(T\right)$, then $D_i=\begin{bmatrix} I_i, & d_1, & u_{i1} \\ & d_2, & u_{i2} \\ & d_r & u_{ir} \end{bmatrix}$ which originally belong to $E$ will turn to not belong to $E$ after the implement of $T$ transformation, among which $V_{\cdot +j}\left(T\right)=\begin{bmatrix} \underset{i\in J_D}{Min}\, v_{ij}, & \underset{i\in J_D}{Min}\, v_{ij} \\ j\in\{\overset{\cdot +}{j_0}\} & j\in\{\overset{\cdot +}{j_0}\} \end{bmatrix}$.

For example:
Through original rules
"Rule 2: (198/14, lift 2.7)
whether using mobile-mail services=0
POINTS <= 6
the length of occupied time> 92
Type = 6
class B [0.925]
Rule 3: (6, lift 2.7)
whether using mobile-mail services = 1
POINTS <= 6
the length of occupied time <= 795
class A [0.875]"
Obtain transformation rule knowledge
"Rule6: (6/9) support：4.25% ID：240-235
Under:
POINTS<=6(same)
92 < occupied time <= 795(same)
Trans:
whether using mobile-mail services =1 to =0
Add: none
class A to B [61.70%]"
ID：240-235 is the source rule number generating transformation rule.

Below the word "under" will list rules existed in category A but not in category B.(add "same" signal in the rear )

Below the word "Trans:" will list rules had the same attribute but different value, and convert the antecedent value among condition category based on target category value

"Add："refers to copy rules existed in B but not in A as an additional transformation condition.

"POINTS<=6"，"92 < the length of occupied time<= 795"and"whether use mobile-mail services=1"are the antecedent of rule knowledge, "whether use mobile-mail services=0" and "none" are the consequent of rule knowledge.

### 3.2 Evaluation Index of Transformation Knowledge

Set the record number of $\{A\} \cup \{B\}$ in database table as $|D|$,set the record number which correspond to antecedent in $\{A\}$ as $F_a$ ,set the record number which correspond to consequent in $\{A\}$ as $R_a$ ,set the record number which correspond to antecedent of transformation rule knowledge $T_{iAB}$ in $\{B\}$ as $F_b$ ,set the record number which correspond to consequent of this in $\{B\}$ as $R_b$ ,set all the record number which meet antecedent of rule set as $F$ and all the record number which meet consequent of rule set as $R$ , set the record number which accord with $A$ rule set in $\{D\}$ as $|D(A)|$ , set the record number which accord with $B$ rule set in $\{D\}$ as $|D(B)|$ .

The accuracy rate of rules:

$$P_{iAB} = (R_b + 1) / (R_a + R_b + 2) \tag{1}$$

anticipative conversion rate

$$T_r = F_a / F \tag{2}$$

the support degree of the rule

$$S = \frac{|F|}{|D(B)|} \tag{3}$$

the reliability

$$R = \frac{|R_b|}{|R|} \tag{4}$$

For instance, In the "Rule6: (6/9) support：4.25% ID：240-235", 6 refers to the record number which meet transformation condition. 9 refer to the record number which meet antecedent of A "under" condition. support：4.25% refers to reliability.

### 3.3 Implementation Steps

1) Read in the original rule set: Take See5 decision tree software for example, the initial rule set saved in .out text file as the form of text file, Rule format as shown in the above example, in which 198 of rule2 represents the record number which meet the rule in training set, 14 represents the record number which does not meet the rule in training set, Predicting accuracy rate=(198-14+1)/(198+2) = 0.925, Enhance degree of lift 2.7=prediction accuracy rate/ the relative frequency of occurrence of such class in training set. Classification rules will be read into database in turn, stored into the rule table.
2) Pretreatment of rule set: Expurgate the same rules generated by rereading in the process, establish keyword of full-text index and so forth.
3) Set mining parameters: Set the following parameters by user:

1）"mining rules transform from class__ into class__", such as class0, and class1,etc, shown in the mentioned example.
2）"the number of rules have the same content >= ", such as “POINTS <= 6”and“92 < the length of occupied time<= 795” in rules 2and rule 3.
3）"the number of rules have the same content<= ”, such as the different value of antecedent in" whether use mobile-mail services" in rule2 and rule3 in the example.
4）"the predicting transformation rate of extension rule>= %", the conversion rate of applying predicting extension rule=the record number consisted with transformation rule in rule set/ all the record number consisted with antecedent of rule set.
4) Tule Mining: Search for rules have many similarity and less discrepancies, by comparing the output generated by transformation rules
5) Rule evaluation index calculation: In order to evaluate the practicality and novelty of extension rule, you should calculate the indicators, such as accuracy rate, predicting transformation rate, support degree and credibility.

### 3.4 Mining Algorithm

The following shows the brief algorithm of transformation knowledge mining:

Input: The result set based on decision tree data mining (two class are respectively represented by A and B), and the minimum record number n from elements of the two set.

Output: matter element of A might transform into strategy of B under the condition of $T_k K(T_R R)$.

Algorithm steps:

(1) The elements in A, B should be respectively represented as multi-dimensional matter element $w_1$ and $w_2$, $R_{1m}$ and $R_{2m}$ analysis indicate the first m matter element in $W_1$ and $W_2$.

(2) The number of matter element for i=0 to A

（3）The number of matter element for j=0 to B

（4）Set integer total equal to 0

（5）Set the dimension of $R_{1i}$ as $i*N$, set the dimension of $R_{2j}$ as $j*N$;

（6）for k=0 to $i*N$

（7）for kk=0 to $j*N$

（8）If $R_{1ir}$ is the K-dimension feature, then value is the same as the value of KK-dimension of $R_{2j}$

（9）total=total+1

（10）End k, kk circulation

（11）If total is greater or equal to the system input value N, then output one transformation knowledge of $R_{1i}$ and $R_{2j}$

（12）End i, j circulation.

## 4 Case Study

A website company own a large number of charge-mail registered users. However, some customers are lost due to the intense competition and other objective reasons. The acquirement of the 245 rule is through applying decision tree data mining algorithm to divide user into "the existing user, the freezing user and the lost user" and predict the user type. However, it cannot acquire knowledge which promotes user transformation from those rules, actually, the freeze user and the normal user can transform into each other in certain condition. Finding transformation knowledge among different users will provide wiser ways.

First of all, import all the decision tree rules into rule base. Then set the parameters (such as transform users from freeze user to normal user) to engage in mining strategy to come out dozens of strategies, the rule 6 of the former section 3.2 is the case in point, from which indicates that users among the scope of POINTS<=6 and the length of occupied time between 92 and 795 can reduce their loss, as long as advising them not to use mobile-mail services. This intuitive transformation knowledge plays a pivotal role on taking effective operational measures.

## 5 Conclusions

The paper briefly analyzes the measures of the acquisition strategy, combined extension theory with research result to come up with strategy knowledge measures of acquiring customer transformation through data mining and extension transformation, and implement through designing algorithm programming. The practicality of this method is confirmed by preliminary test. We found that there are two paths for transformation knowledge mining by combining decision tree method with extension set theory.

1) The indirect rule mining method refers to further digging on the basis of traditional decision tree rule. After generating a static rule set through data mining of decision tree, coming with the second excavation of the rule set.
2) Extension strategy direct mining method refers to directly dig out transformation knowledge on original data base by improving traditional decision tree algorithm.

The paper mainly based on the first path to achieve the acquisition of transformation knowledge, The second path- extension strategy direct mining method get rid of the dependence of decision tree classification rule, which have more practicality and need further research. there would be great application prospection by taking advantage of the result of extension theory research which connect traditional data mining with extension set, and with extension transformation as well as extension logical theory to dig out "can't to can, not to yes" strategy by using methods of extension data mining.

**Acknowledgment.** This research has been supported by grants from National Natural Science Foundation of China (#70871111, # 70921061), Major scientific and technological project in industrial areas (#2010B10024), Bureau of Science and Technology of Ningbo. We would like to thank the researcher CAI Wen, YANG Chun-yan and CHENG Wen-wei for their constructive comments on our early draft. Thanks Ye yukui for her translation.

## References

1. Han, J., Micheline, K.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann (2006)
2. Olson, D., Shi, Y.: Introduction to business data mining, International edition. McGraw-Hill (2007)
3. Nie, G.-L., Zhang, L.-L., Li, X.-S., Shi, Y.: The Analysis on the Customers Churn of Charge Email based on Data mining. In: Sixth IEEE International Conference on Data Mining - Workshops (ICDMW 2006), Hong Kong, China, pp. 843–847 (December 2006)
4. Wen, C.: The Matter-element Model and Its application, 1st edn. Science and Technology Document Publishing House, Beijing (1994) (in Chinese)
5. Lixi, L., Huawen, L., Chunyan, Y.: Study on the Application of Extenics in Data Mining. Engineering Science 6(7), 53–79 (2004)
6. Zhang, Y.-L., He, B.: Potential Information Mining Based on Matter-element Extensibility. Mathematics in Practice and Theory 31(5), 569–575 (2001)
7. Li, X.-S., Shi, Y., Li, A.-H.: Study on enterprise data mining solution based on extension set. Journal of Harbin Institute of Technology 38(7), 1124–1128 (2006)
8. Li, L., Yang, C., Li, H.: Extension Strategy Generating System. Science Press, Beijing (2006) (in Chinese)
9. Li, L., Yang, C., Li, H.: Extension Strategy Generating System. Science Press, Beijing (2006) (in Chinese)

# A Flexible Parallel Runtime for Large Scale Block-Based Matrix Multiplication

Keyan Liu[1], Shaohua Song[2], Ningnan Zhou[2], and Yanyu Ma[1]

[1] Nokia-Simense-Network CTO Research, JiuXianQiao Road 14, Chaoyang District, Beijing
[2] RenMin University of China, No. 59 Zhong Guan Cun Street, Beijing, P.R.C
keyan.liu@nsn.com, ssh192@ruc.edu.cn,
zhouningnan@gmail.com, yanyu.ma@nsn.com

**Abstract.** Block-based matrix multiplication plays an important role in statics computing. It is hard to make large scale matrix multiplication in data statistics and analysis. A flexible parallel runtime for large scale block-based matrix is proposed in this paper. With MapReduce framework, four parallel matrix multiplication methods have been discussed. Three methods use the HDFS to be the storage and one method utilizes the Cloud storage to be the storage. The parallel runtime will determine to use the appropriate block-based matrix multiplication. Experiments have been made to test the proposed flexible parallel runtime with large scale randomly generated data and public matrix collection. The results have shown that the proposed runtime has a good effect to select the best matrix multiplication strategy.

**Keywords:** Matrix Multiplication, MapReduce, Cloud storage, HBase.

## 1 Introduction

Since MapReduce programming model was proposed [1], various MapReduce implementations have been brought out. Apache Hadoop adopted the same architecture with Google's [2]. Due to the scalability, stability, efficiency, simplicity of MapReduce framework, many efforts have been done to solve the scientific problems on distribution computers with MapReduce framework. Phoenix is a shared-memory implementation of MapReduce [3]. GraphLab [4] and Mahout [5] are two processing approaches to parallelize the Machine Learning algorithm with MapReduce. Apache Hama [6] is a distributed computing framework based on BSP (Bulk Synchronous Parallel) computing techniques for massive scientific computations. Hbase, short for Hadoop Database, is an open-source, distributed, column-oriented and KeyValue based data management system [7]. HBase runs on top of HDFS, providing BigTable-like capabilities for Hadoop.

Matrix multiplication is a fundamental kernel for a variety of scientific problems. And thus many parallel algorithms have been proposed to solve large scale matrix multiplication. Block partitioned matrix product is a common technical to parallel the matrix multiplication on distribution computers. In this paper, four approaches are proposed to parallel matrix multiplication on Hadoop. Among the proposed four methods, three methods attempt to solve the Matrix multiplication using different

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 51–59, 2012.

MapReduce strategy and one method uses the emerging cloud storage to implement new MapReduce strategy. The performances of these algorithms are investigated and evaluated, and then a performance model for block based matrix multiplication with MapReduce framework is found.

The rest of the paper is organized as follows. Section 2 presents the four approaches for matrix multiplication with MapReduce framework. The proposed performance model is presented in Section 3. Section 4 describes our evaluation methodology and experimental results and related analyses are presented, followed by conclusions in section 5.

## 2 Parallel Block-Based Matrix Multiplication

### 2.1 Matrix Storage and Notations

For saving storage space, there are different storage formats for sparse and dense matrix. For sparse matrix, we store each non-zero point as a triple (row, column, value); for dense matrix, we store each row of the matrix as a record like (row, [values]).

For two matrices $A$ and $B$, $A$ has dimension $I * K$ with elements a(i, k) for $0 \le \mathrm{i} < I$ and $0 \le \mathrm{k} < K$, accordingly $B$ has dimension $K * J$ with elements b(k, j) for $0 \le \mathrm{k} < K$ and $0 \le \mathrm{j} < J$. Then Matrix $C = A * B$ has dimension $I * J$ with elements c(i, j), and $\mathrm{c(i,j)} = \sum_{\mathrm{k=0}}^{\mathrm{k=K-1}} \mathrm{a(i,k)} * \mathrm{b(k,j)}$. For simplicity, we just assign an uniform block size for each blocks, i.e., we use *ib*, *kb*, *jb* to index the block in matrix $A$, $B$, and $C$. $A[ib, kb]$, $B[kb, jb]$, $C[ib, jb]$ are used to represent the block for each matrices, then $A[ib, kb]$ has dimension $IB * KB$, $B[kb, jb]$ has dimension $KB * JB$, and $C[ib, jb]$ has dimension $IB * JB$. Since $IB$ maybe not divisible by $I$, similar for $KB$ and $JB$, dimension of margin block would less than other block. Let $NIB$, $NKB$, $NJB$ to denote number of blocks for $I$, $K$ and $J$, then

$$NIB = \frac{I-1}{IB} + 1, NKB = \frac{K-1}{KB} + 1, NJB = \frac{J-1}{JB} + 1$$

$$\text{where } 0 \le ib < NIB, 0 \le ib < NIB, 0 \le jb < NJB \tag{1}$$

$a(i, k) \in A[ib, kb]$, $ib * IB \le i < min((ib+1) * IB, I)$ and $kb * KB \le k < min((kb+1) * KB, K)$

$b(k, j) \in B[kb, jb]$, $kb * KB \le k < min((kb+1) * KB, K)$ and $jb * JB \le \mathrm{j} < min((jb+1) * JB, J)$

$c(i, j) \in C[ib, jb]$, $ib * IB \le i < min((ib+1) * IB, I)$ and $jb * JB \le j < min((jb+1) * JB, J)$

Let $C[ib, kb, jb] = A[ib, kb] * B[kb, jb]$, then

$$C[ib, jb] = \sum_{kb=0}^{NKB-1} C[ib, kb, jb] \tag{2}$$

Assuming matrix *A, B and C* have dimension 10*10, then *I*=10, *K*=10 and *J*=10. If we take *IB*=*KB*=*JB*=4, then *NIB*=*NKB*=*NJB*=3, $0 \le ib < 3, 0 \le kb < 3, 0 \le jb < 3$. The blocks partition for matrix *A* is shown in Fig.1.

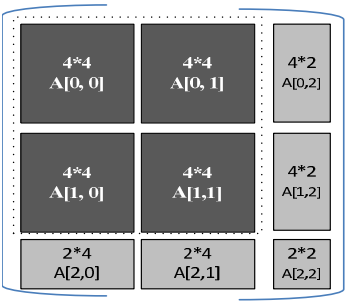



**Fig. 1.** Block-based partition for matrix with dimension 10*10

There are many parameters for MapReduce framework. $M_c$ and $R_c$ are taken to represent the Map/Reduce capacity of the cluster, *M* and *R* to represent Map/Reduce task number for a MapReduce job and *T*, $T_{mapper}$, $T_{reducer}$ to represent execution time for MapReduce job, Map tasks and Reduce tasks respectively.

$$T = T_{mapper} + T_{reducer} \tag{3}$$

$$T_{reducer} = T_{shuffle} + T_{sort} + T_{reduce} \tag{4}$$

$T_{shuffle}$, $T_{sort}$, $T_{reduce}$ denote the execution time for shuffle, sort and Reduce phase respectively.

## 2.2 MapReduce Matrix Multiplication Approaches

Block-based matrix multiplication algorithm can be represented in Fig. 2. In Fig.2, *A* and *B* are split into four blocks respectively, for example$C[0,0] = A[0,0] * B[0,0] + A[0,1] * B[1,0]$. Matrix multiplication is carried out block by block and summed over the output to generate the final result. Based on MapReduce framework, we split the four algorithms into two categories based on which MapReduce phase is response for multiplication.

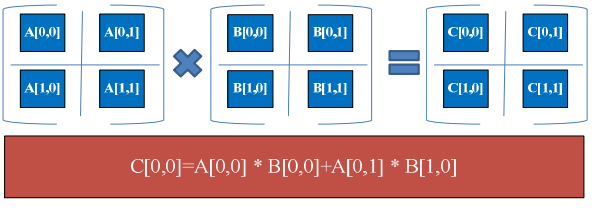



**Fig. 2.** Blocked matrix multiplication algorithm

Since the performance of these strategies affected by many factors, such as network status, data distribution, work balance, for simplification, some assumptions are made through this paper. (1) All data can be filled into memory; (2) Data distribution in each block is uniform; (3) Input data is distributed evenly on each Map task node.

### 2.2.1 Strategy 1

This is one approach in the reduce-side multiplication category. The algorithm works in four steps: (1) each Map fetches input; (2) each Map emits data from *A NJB* times and data from *B NIB* times; (3) each Reducer fetches data from the output of Map; (4) the Reducer does multiplication and summation. In step 3, the Partitioner makes sure that Reducer gets a whole row blocks from matrix *A* and a whole column blocks from matrix *B*. The blocks are sequenced by: *A*[*ib*,0]*B*[0, *jb*]*A*[*ib*,1]*B*[1,*jb*]…*A*[*ib*,*NKB*-1]*B*[*NKB*-1,*jb*], then

$$R_{max} = NIB * NJB \tag{5}$$

In this strategy, *A*[*ib*, *kb*] is required by Reducer which is responsible for computing *C*[*ib*, *jb*] for $0 \leq$ jb $<$ NJB, the Map function has to emit *NJB* times for *A*[*ib*, *kb*]. Similarly, *B*[*kb*, *jb*] needs to be emitted *NIB* times, then we can get

$$D_{size} \propto NJB * I * K + NIB * K * J \tag{6}$$

*IB*, *JB* and *M* are the core parameters for this strategy. *M* controls the concurrency of the Map tasks, *IB* and *KB* determinate the intermediate data size and concurrency of the Reducer tasks. Since Map task has no influence on size of intermediate data, full Map capacity of the cluster should be utilized. If we take *n* to represent the data size fetched in shuffle phase, we can get

$$T_{mapper} \propto D_{size}, T_{shuffle} \propto n$$
$$T_{sort} \propto n * log\,(n), T_{reduce} \propto n^2 \tag{7}$$

If *NIB* * *NJB* increases to α times, $D_{size}$ will increase to β times, $\alpha \geq \beta$ (with equality if and only if A or B is empty). When $R_{max} < R_c$, we can use α times reducer work concurrently. For each Reducers, shuffle phase need less data (β/α times of original), then $T_{shuffle}$, $T_{sort}$ and $T_{reduce}$ will be decreased, Reduce task's execution time will be decreased. With $D_{size}$ increases, execution time of Map task is increased; there are trade-off between increase Reducer concurrency and decrease the intermediate data size. If $R_{max} > R_c$, the Reduce function would run more than once (set $R = R_c$), or the reducer process world run more than once (set $R = R_{max}$) to complete the job. Some computing node (the Reduce task runs on it) will get more immediate data, since the node number for reducer is unchanged, and intermediate data size increased, the total execution time will not be decreased. Based above analysis, the following setting should be taken to achieve best performance for this strategy.

$$M = M_c, \qquad NIB * NJB = \begin{cases} R_c & \text{if when } R = R_c, T_s \geq T_{ic} \\ R' \text{ which let } T_s = T_{ic} & \text{otherwise} \end{cases}$$
$$NIB, NJB \text{ subjct to minimal}(D_{size}) \tag{8}$$

### 2.2.2 Strategy 2

This is another reduce-side multiplication strategy. There are two MapReduce jobs for this strategy. The first job works as following four steps: (1) Maps read input data; (2) Maps emit data from A NJB times and data from *B NIB* times; (3) Reducers fetch data from Map's output; (4) Reducer do multiply to get C[ib, kb, jb]. In step 3, the Partitioner makes sure that Reducer gets a block from *A* and a corresponding block from *B*, then

$$R_{max} = NIB * NKB * NJB \tag{9}$$

The first MapReduce job just gets *C*[*ib*, *kb*, *jb*], so this strategy needs another MapReduce job to sum over *C*[*ib*, *kb*, *jb*]. The second MapReduce job works in two following steps: (1) Map reads and emits *C*[*ib*, *kb*, *jb*]; (2) Reducers sum up these partial results to get *C*[*ib*, *jb*].

For this strategy, the parameters will be taken as equation (10).

$$\begin{cases} IB = I, IK = \dfrac{K}{R_c}, JB = J & \text{if } \dfrac{K}{R_c} \geq 1 \\ IK = 1, \text{select } IB, JB \text{ like strategy 1} & \text{otherwise} \end{cases}$$

$$M = M_c \text{ and } R = R_c \tag{10}$$

### 2.2.3 Strategy 3

This is the map-side multiplication strategy; this strategy does multiply during the map phase. The Map function gets one column from *A* and one row from *B*, then does multiplication and sends the intermediate data to Reducer. Reducer just sums over these partial results. The first MapReduce job works as following two steps: (1) Map reads input data and emits them; (2) Reducer fetches and formats the intermediate data. The second MapReduce job works as following three steps: (1) Map reads input data; (2) Map does multiply and emits partial sum; (3) Reducer sums up the partial results. The following setting should be taken to achieve the best performance for this strategy.

$$R = R_c,$$

$$M = \begin{cases} M_c & \text{when } M = M_c, T_s \geq T_{ic} \\ M' \text{ which let } T_s = T_{ic} & \text{otherwise} \end{cases} \tag{11}$$

### 2.2.4 Strategy 4 in HBase

The two matrices *A* and *B* are stored in the same table of HBase. The table has two column families: matrix-A and matrix-B, there is one column in each column family. The col-A stores the data of matrix *A* and col-B stores the data of matrix *B* in two column families. In a row, there are one element of matrix *A* and one element of matrix *B*. For *A*[*i*][*k*] in matrix *A*, the row key is defined in equation (12). For *B*[*k*][*j*] in matrix *B*, the row key is defined in equation (13).

$$\text{Rowkey}(A, i, k) = [i/IB]\_[k/KB]\_[i\%IB]\_[k\%KB] \quad (12)$$

$$\text{Rowkey}(B, j, k) = [j/JB]\_[k/KB]\_[j\%JB]\_[k\%KB] \quad (13)$$

The matrix $A$ is stored by row and matrix $B$ is stored by column, and then the data can be fetched continuously. The matrix multiply algorithm in HBase works in two steps: (1) Maps fetches input data, and every Map fetches a whole row blocks of matrix A; (2) Map fetches a whole column blocks of matrix B and does multiplication.

## 3 Performance Model

A performance model is proposed to predict the best strategy and parameters for specific input. For the performance model, the computing capacity of the cluster, data sparsity and data distribution of input matrices are needed to collect. As shown in the Fig.3, Hadoop's configuration file is analyzed to get the computing capacity of the distribution cluster. To get the nature of the matrices, some blocks are sampled from input file to predict the sparsity and data distribution of the matrices.

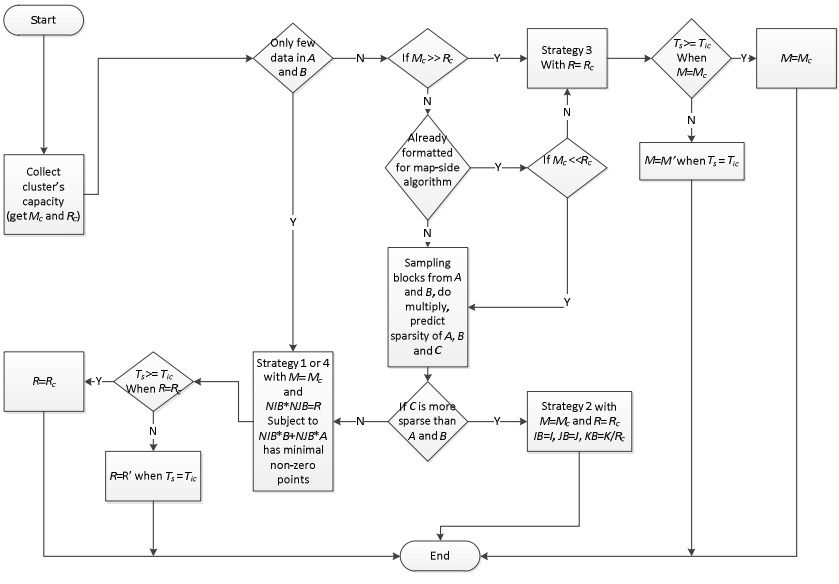



**Fig. 3.** Workflow of the performance model

## 4 Experiments and Evaluation

### 4.1 Experiments Setup

A parallel cluster hardware environment has been setup with 9 HP DL380 servers. Every computing node has two Intel Xeon 3.6GHz CPU processors with 8G main

memory, 125G hard disk, and RHEL4 Linux OS. Hadoop and HBase are configured on the cluster nodes.

### 4.2 Experiments and Results for Randomly Generated Matrices

Randomly generated matrices are used to do experiments. For sparse matrices, each point has same probability to be a non-zero value. To get a stable experiment result, each experiment case has been run twice and the average value is taken as the final result. For strategy 2 and 3, there are two Map-Reduce jobs. *T1* and *T2* are taken to represent execution time for first and second MapReduce job, so do $T_{mapper1}$, $T_{mapper2}$ and $T_{reducer1}$, $T_{reducer2}$.

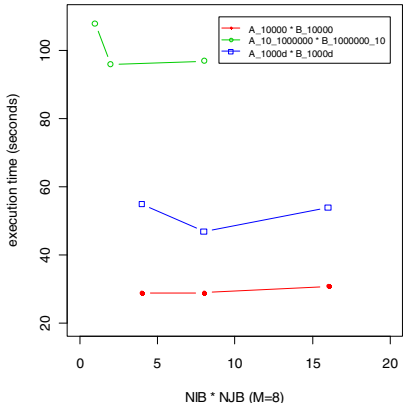



**Fig. 4.** Matrix multiplication with strategy 1

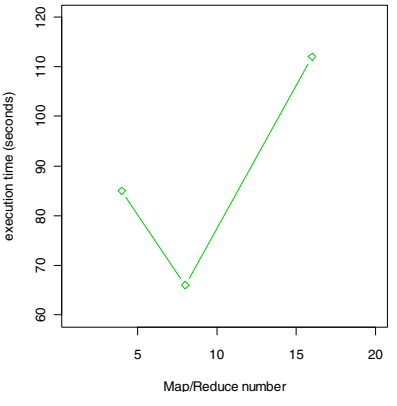



**Fig. 5.** A_10000 * B_10000 with strategy 2

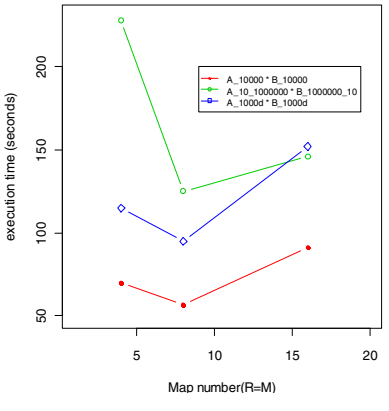



**Fig. 6.** Matrix multiplication with strategy 3

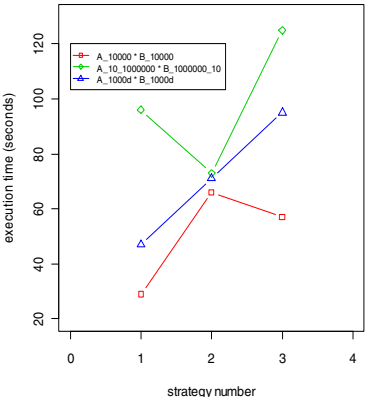



**Fig. 7.** Optimal result for each strategy

As shown in Fig.4, when $R$=8, execution time is minimal for the two experiment data A_1000 * B_1000 and A_10000 * B_10000. When $R = 16 > R_c$, there is no further improvement. For A_10_1000000 * B_1000000_10, this case achieves the optimal result when R=2. For strategy 2, Fig.5 shows the experiment result of strategy 2when $M$=$M_c$=8, $R$=$R_c$=8. Fig.6 presents experiment results for strategy 3, the optimal result can be got when $M$=8. The comparison for these three strategies is shown in Fig.7.

### 4.3 Experiments and Result for Strategy 4 in HBase

For strategy 4, the performance comparison is made with strategy 1. Three randomly generated matrices have been used, 1000*1000, 3000*3000 and 5000*5000. As shown in Table.1, the strategy 4 has a better performance than strategy 1 on 1000*1000 and 3000*3000 data sets. And it has almost equal performance with strategy 1 on 5000*5000 data sets.

**Table 1.** The result for strategy 4 in HBase

| Dimensions | Store | IB | KB | JB | map | reduce | $T_{mapper}$ | $T_{reducer}$ | T |
|---|---|---|---|---|---|---|---|---|---|
| 1000*1000 | HDFS | 250 | 250 | 250 | 16 | 16 | 15s | 30s | 59s |
| 1000*1000 | HBase | 250 | 250 | 250 | 16 | 0 | 52s | 0s | 52s |
| 3000*3000 | HDFS | 500 | 500 | 500 | 36 | 36 | 54s | 3min30s | 4min3s |
| 3000*3000 | HBase | 500 | 500 | 500 | 36 | 0 | 3min3s | 0s | 3min3s |
| 5000*5000 | HDFS | 1000 | 1000 | 1000 | 25 | 25 | 2min | 4min10s | 4min57s |
| 5000*5000 | HBase | 1000 | 1000 | 1000 | 25 | 0 | 5min2s | 0s | 5min2s |

## 5 Conclusion

In this paper, four algorithms have been implemented for block-based matrix multiplication with MapReduce framework. The algorithms are classified to the “Map-side” and “Reduce-side” categories. These algorithms are compared with different characterizes of matrices. Finally a performance model has been proposed to predict the performance for each algorithm based on cluster’s configuration and nature of the matrices. Our contributions can be concluded as follows: 1) the core parameters for each algorithm that highly affects the performance are analyzed; 2) we compared these three algorithms and proposed a performance model to automatically select best strategy and parameters for specify input and distribution computing environment.

## References

1. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI 2004: Sixth Symposium on Operating System Design and Implementation, pp. 137–150 (2004)
2. `http://hadoop.apache.org/`

3. Low, Y., Gonzalez, J., Kyrola, A.: GraphLab: A New Parallel Framework for Machine Learning. In: Conference on Uncertainty in Artificial Intelligence (UAI), pp. 340–349 (2010)
4. http://mahout.apache.org/
5. Seo, S., Yoon, E.J., Kim, J.: HAMA: An efficient matrix computation with the MapReduce framework. In: IEEE CloudCom 2010 Workshop (2010)
6. Norstad, J.: A MapReduce Algorithm for Matrix Multiplication, http://homepage.mac.com/j.norstad/matrix-multiply/index.html
7. http://hbase.apache.org/

# Group Detection and Relation Analysis Research for Web Social Network

Yang Li[1,2,3,⋆], Kefu Xu[1], Jianlong Tan[1], and Li Guo[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, National Engineering Laboratory for Information Security Technologies, Beijing, China, 100190
[2] Graduate School of Chinese Academy of Sciences, Beijing, 100049
[3] College of Computer Science and Information Engineering, Central South University of Forestry and Technology, Changsha, 410004
ly@software.ict.ac.cn

**Abstract.** With the rapid development of web social networks and its integration into our daily lives, interactivity and participatory between people and web have made the web social networks play an important role in the information security, trade relations, community structures, communication behavior and so on. This paper introduces the important significance, the current application, the progress of the study and research on the web social networks from the views of group detection and relation analysis, meanwhile points out the research trend of the web social networks from the evolution, the propagation, multi-scale, link associated with content, and the social computing.

**Keywords:** Web Social Networks, Group Detection, Relation Analysis, Information Propagation, Link Prediction, Social Computing.

## 1 Introduction

With the rapid development of newer media and its application, people have taken part in the abundant social affairs increasingly through internet. Web social networks are entity networks combined by many linking relations, for example, people pay attention to the latest information of their friends and social celebrities from the micro-blog websites such as Twitter and Sina Weibo, and share their new matters with their friends and fans, or people make good friends and play interactive games on the Facebook or RenRen. Web social networks play an important role in people's daily lives, the interactivity and participatory between people and web have promoted the transformation from the social behavior to the web behavior, from the realistic social relations to the web social relations, and from the social information to the web information[1].

⋆ This research was supported by the National Science Foundation of China (NSFC) under Grant No. 61003295, and National High Technology Research and Development Program 863 No. 2011AA010705.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 60–67, 2012.

In recent years, with the development and perfectness of graph theory, probability theory and all kinds of geometry, web social networks also have developed energetically, in which the research on groups detection and relations analysis is particularly important, and the web social networks, widely used in the fields of information security, the trade relations, the social networks services, the groups communication behavior, the information recommendation and so on, is being taken seriously.

## 2 The Research of Group Detection

The important character of web social networks is the community structure of the network, this structure can be denoted a social network with weighted graph. Therefore we introduce some chief algorithms as follow.

### 2.1 Graph Partition Method

The first heuristic method is Kernighan-Lin algorithm (KL)[2]. The algorithm solve graph bipartition problem. But the partition by KL depends heavily on the initial partition of the two groups, therefore, KL is used frequently for subsequent optimization based on other partition algorithms[3]. The other one is spectrum bipartition[4]. The method is based on Laplace matrix so as to solve graph bipartition problem. The flaw of this algorithm is which only divide the graph into two subgraghs, or even subgraphs. Some people introduce an efficient max-flow method[5], and solve the min-cut between two nodes. However graph partition need to confirm the amounts of group and even value in advance, Therefore the method cannot be applied for community detection directly.

### 2.2 Sociology Method

Agglomeration algorithm: Based on the similarity between kinds of peer nodes, we append edges to original empty-network containing $n$ nodes and 0 edge from the node with highest similarity peer nodes. This process can be terminated at every node, the finally formative network can be regarded as the groups of community. However it is apt to find the core of community, and ignore the periphery of community. Partition algorithm: Generally we try to find the peer nodes with lowest similarity, and remove the edge between peer nodes from the concerned network. We can partition gradually entire network into kinds of smaller and smaller sub-community by repeating the process. This process can be terminated at every node equally, and we can get kinds of communities based on current status.

### 2.3 Divided Method

Girvan-Newman (GN) Algorithm: The significant method of splitting is proposed by Girvan and Newman[6]. Specifically speaking, it removes the largest side of the interface (Betweenness) from the network through the iteration and divides

the whole network into many communities. But in the case of unknown number of community, the algorithm can't determine the right number of iterations. Newman Fast Algorithm: Because of the complexity of time in GN algorithm, [7] proposes the fast algorithm on the basis of GN algorithm, it is condensing algorithm based on the thought of greedy algorithm. In these community structures, by choosing the biggest $Q$ value corresponding to the local, people can get the best network community structure.

### 2.4 Overlapping Community Detection

In the actual network, some nodes are often shared by many communities, it means that there is overlap between communities. The Clique Percolation Method (CPM)[8] is the most widely used method in the overlapping community, the main concept of the Method: interior community has higher edge density, and internal edges of communities may form big cliques; but it is nearly impossible that the edges between communities can form bigger cliques. Palla and others propose to define the community by the percolation of $k$-clique, allowing overlap exists between communities[9].

## 3 Analysis and Mining of Relation

### 3.1 Strong and Weak Relation

In "The Strength of Weak Ties"[10] published in 1973, Granovetter argues that scholars often focus on the effect of strong ties, and to some extent, neglect the role of weak ties. In Granovetter's survey, American society is a society with weak relation. Bian YanJie, a Chinese scholar, puts forward an assumption of strong ties[11]. Namely Chinese society is not like the society of the United States with weak ties, but a society with strong ties. That is to say: Bian YanJie's theory on strong ties is a hypothesis in the specific circumstance in China.

SNS distinguished according to the strength of ties. SNS with strong ties: Facebook, Renren, Kaixin001, LinkedIn, Pengyou. Because of a high degree of trust and the acquaintance with each other, the user will be very active if the privacy control is good enough, the spread of information is confined within the circle of friends due to its nature of low spread. SNS with weak ties: Twitter, Sina Weibo. They have a low degree of trust because of the privacy control problems and the broadcast mode. A piece of micro-blog news of celebrities can be propagated to a certain depth.

### 3.2 Core Node Detection

"Center" is one of the focus of social network analysis. At the beginning, the analysts of social networks discuss what kind of power individuals or organizations have in their social network, or how their center position is. This thought is one of the earliest contents they discuss. The early authority node discovery and links are the PageRank Algorithm[12] proposed by Google and the HITS Algorithm[13] proposed by Jon Kleinberg.

At present, the core node of networks is mainly found in the following ways: measuring the importance of nodes based on the static parameter of social network analysis[14], measuring by the segmentation standards in the theory of graph partition[15]. [16] introduces core nodes detection based on frequent itemsets of graph, [17] mines core of consortium based on the Six Degrees of Separation, and describe Shortest Path based on link weight algorithm SPLINE. In [18], it adopts the method of Betweenness centrality in the Random walk to measure the core members. [19] puts forward the LeaderRank algorithm to mine the leaders of social network opinion, the results show that performance is better than the PageRank Algorithm.

### 3.3 Link Prediction

The Link Prediction in the network refers to how to predict the possibility of generating links between two nodes which has not yet generate edge through the known network nodes, the network structures, and others information[20]. In the early studies, Markov chain was used for the predictions of network links and the path analysis. Professor Zhou Tao with his team makes a lot of research results in the field of the link prediction. [21] puts forward two new indexes: resource allocation index and local path index. The new study finds the two indexes have better predictive ability. [22] analyzes the performance of the local path index in detail in the noise intensity and the network modes with controllable network density. [23] proposes two similarity indexes based on local random walk of network.

Another common method it that [24] presents a maximum likelihood estimation algorithm for the prediction of links, this method has better accuracy in dealing with obvious hierarchical organization networks, such as terrorist attack networks and prairie food chains. In [25], the link prediction based on the random block model can get better results than that of the former. Moreover, P.Zhang et al[26] proposes an aggregate ensemble (AE) learning framework for building a robust ensemble model that can tolerate data errors and demonstrates the superior performance, and [27] propose a new Hybrid-Frequent tree, they introduce an innovative idea building more accurate classification models for us to improve accuracy of prediction.

## 4 Development Trend of Research

### 4.1 Community Evolution

Community evolution chiefly studies community state based on the change times, and analyzes the mechanism and cause which result in these changes. Community evolution include the following changes: formation, growth, contraction, merging, splitting, death, and so on. [28] studies firstly the community evolution, they analyze the snapshot based on different times from NEC CiteSeer Database. [29] analyzes systematically the community revolution firstly, the adopted data come from the network of mobile phone in one year and the network of scientist

collaboration based on gelatinous matter field, the experimental result show the small community is stable, but the large one change drastically[30].

Recently, by the theory and evaluation of link prediction, [31,32] compare and evaluate the evolution models separately, meanwhile offer bran-new views and suggestion for evolution models of networks. P.Zhang et al[33,34] proposes a novel Ensemble-tree indexing structure to organize all base classifiers in an ensemble for fast prediction and a novel Lazy-tree indexing struture, which can automatically update themselves by continuously integrating new classifiers and discarding outdated ones. Meanwhile, [35] propose a new LCN-Index based on Map-Reduce framework to handle continuous data stream queries in the Cloud, they open up a new orientation for us to explore the evolution problem of time sequence.

### 4.2 Propagation Dynamics

In the history of human civilization, every propagation of contagion (malaria, smallpox, measles, typhoid) is brought by human civilization; in the other way, every large-scale contagion exert wide and far-reaching impact for human civilization. The social network process of human promotes the flow of human or material, meanwhile accelerate the propagation speed of contagion[36]. The gist of network dynamics is that it can reveal the influence of dynamic process on the network topology structure, and whether or not it can reveal structural character. [37] studies the propagation of contagion on the free-scale network. In 2006, [38] indicates that synchronous dynamics process can reveal the topological scale of network, human begin to discuss the relations between community and network dynamics. [39] analyzes the relation between propagation dynamics and community structure of network, meanwhile indicates the local equilibrium state of propagation process and corresponding relation community structure.

### 4.3 Multi-Scale Problem

Multi-scale Detection of community arouse people's wide concern recently, some famous journals (such as Nature, Science, Pnas) publish these papers about the problems. Multi-scale and hierarchy have close connection but difference. Multi-scale problem pay much greater attention to the topological characteristic of network on different scales, but hierarchy problem concern the hierarchical phenomena.

### 4.4 Link Associated with Content

The most methods of community detection are based on the links between nodes, but ignore the actual content behind theses nodes. [40] introduces a newer algorithm of community detection by combining the similarity of content with the importance of PageRank. The algorithm not only concern the links between webpages but also the focus on the similarity of content. However this algorithm lacks topic detection and tracking, and the interactive evolution between content and links.

### 4.5 Social Computing

The online social networks are booming increasingly, which have generated much social networks containing millions users. Wang Feiyue, from domestic academicians, firstly puts forward the sponsor of "social computing", which develop rapidly. [41] introduces the a thorough review of the state of the art of social computing research and application, its scientific significance and progress, and put forward ACP methodology and practice work of "Artificial Societies" + "Computational Experiments" + " Parallel Execution" based on social model. Finally the thesis outlines the major research tasks ahead of social computing research and the corresponding future prospect.

## 5 Summary and Prospect

Web social networks have been studied widely in all kinds of fields, from the views of group detection and relation analysis, this paper reviews the progress of the web social networks, and introduces the corresponding algorithms. Meanwhile, it prospects some unresolved problems and the research trend of web social networks, including the community evolution, the propagation dynamics, the problems of multiple scales, the relations between content and links, social computing, and so on. When paying attention to these problems, researchers should know how to combine the abundant knowledge and apply to the special web social network (for example: Twitter, Facebook) will be an important research task on the next stage.

**Acknowledgements.** The author Appreciates Shen Huawei(from the Institute of Computing Technology Chinese Academy of Sciences) and Zhou Tao(from the University of Electronic Science and Technology of China)deeply, as well as their valuable suggestions and innovative idea.

## References

1. Yang, S., Sun, L., Cui, P.: Analysis of Web social network. Communications of China Computer Federation 2(7) (February 2011) (in Chinese)
2. Kernighan, B.W., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell System Tech. J. 49(2), 291–307 (1970)
3. Shen, H.: Community structure of complex network, pp. 13–14 (2011) (in Chinese)
4. Barnes, E.R.: An algorithm for partitioning the nodes of a graph. SIAM J. Alg.Disc. Meth. 3(4), 541–550 (1982)
5. Goldberg, A.V., Tarjan, R.E.: A new approach to the maximum-flow problem. Journal of the ACM 35(4), 921–940 (1988)
6. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci 9(12), 7821–7826 (2002)
7. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys.Rev.E 69(6) (2004)

8. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818 (2005)
9. Palla, G., Barabasi, A.L., Vicsek, T.: Quantifying social group evolution. Nature 446, 664–667 (2007)
10. Granovetter, M.: The Strength of Weak Ties. American Journal of Sociology 78, 1360–1380 (1973)
11. Bian, Y.: Find strong relation: indirect relation, network bridge and go for a job in China. Foreign Sociology (2), 50–65 (1998) (in Chinese)
12. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web-search engine. In: Proc 7th International World Wide Web Conference, pp. 146–164. SIGIR, Brisbane (1998)
13. Kleinberg, J.: Authoritative sources in a hyperlinked environment. In: Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677. ACM Press, New Orleans (1997)
14. Freeman, L.C.: Centrality in social networks: Conceptual clarification. Social Networks 1(3), 215–239 (1979)
15. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99(12), 7821–7826 (2002)
16. Song, W., Liu, H., Wang, C., Xie, J.: Core nodes detection based on frequent itemsets of graph. Journal of Frontiers of Computer Science and Technology 04(01), 84–86 (2010) (in Chinese)
17. Tang C., Liu W., Wen F., Qiao S.: Three probes into the social network and consortium information mining. Journal of Computer Applications (9) (2006) (in Chinese)
18. Luo, L.: Community discovery and tracking methods based on core members, pp. 17–18 (2010) (in Chinese)
19. Lu, L., Zhang, Y.-C., Yeung, C.H., Zhou, T.: PLoS ONE, 6(6), e21202 (June 2011)
20. Getoor, L., Diehl, C.P.: Link Mining: A Survey. ACM SIGKDD Explorations Newsletter 7, 3 (2005)
21. Zhou, T., Lu, L., Zhang, Y.-C.: Predicting missing links via local information. Eur. Phys. J. B 71, 623 (2009)
22. Lü, L., Jin, C.-H., Zhou, T.: Similarity index based on local paths for link prediction of complex networks. Phys. Rev.E 80, 046122 (2009)
23. Liu, W.-P., Lu, L.: Link Prediction Based on Local Random Walk. Europhys. Lett. 89, 58007 (2010)
24. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature 453, 98 (2008)
25. Guimera, R., Sales-Pardo, M.: Missing and spurious interactions and the reconstruction of complex networks. Proc. Natl. Sci. Acad. U.S.A. 106, 22073 (2009)
26. Zhang, P., Zhu, X., Shi, Y., Guo, L., Wu, X.: Robust Ensemble Learning for Mining Noisy Data Streams. Decision Support Systems 50(2), 469–479 (2011)
27. Guo, J., Zhang, P., Tan, J., Guo, L.: Mining Frequent Patterns across Multiple Data Streams. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), Glasgow, Scotland, October 24-28 (2011)
28. Hopcroft, J., Khan, O., Kulis, B., et al.: Tracking evolving communities in large linked networks. Proc Natl. Acad. Sci. USA 101, 5249–5253 (2004)
29. Palla, G., Barabsi, A.L., Vicsek, T.: Quantifying the social group evolution. Nature 446(7136), 664–667 (2007)
30. Cheng, X., Shen, H.: Community structure of complex network. Complex Systems and Complexity Science 8(1) (March 2011) (in Chinese)

31. Liu, H., Lu, L., Zhou, T.: Uncovering the network evolution mechanism by link prediction. Sci. Sin. Phys. Mech. Astron. 41, 816–826 (2011) (in Chinese)
32. Wang, W., Zhang, W.: New method of assessing network evolving models based on link prediction. Journal of University of Electronic Science and Technology of China 40(2) (March 2011) (in Chinese)
33. Zhang, P., Li, J., Wang, P., Gao, B., Zhu, X., Guo, L.: Enabling Fast Prediction for Ensemble Models on Data Streams. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2011), San Diego, CA, USA, August 21-24 (2011)
34. Zhang, P., Gao, B., Zhu, X., Guo, L.: Enabling Fast Lazy Learning for Data Streams. In: Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2011), Vancouver, Canada, December 11-14 (2011)
35. Li, J., Zhang, P., Tan, J., Liu, P., Guo, L.: Continuous Data Stream Query in the Cloud. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), Glasgow, Scotland, October 24-28 (2011)
36. Wang, X., Li, X., Chen, G.: Complex network theory and application, vol. 72. Tsinghua University Press (2006) (in Chinese)
37. PastorSatorras, R., Vespignani, A.: Epidemic spreading in scale free networks. Phys. Rev. Lett. 86(14), 3200 (2001)
38. Arenas, A., Diaz-Guilera, A., Perez-Vicente, C.J.: Synchronizat ion reveals to pological scales in complex networks. PhysRev. Lett. 96(11), 114102 (2006)
39. Cheng, X.Q., Shen, H.W.: Uncovering the community structure associated with the diffusion dynamics on networks. J. Stat. Mech, P04024 (2010)
40. Yun, Y., Yuan, F., Liu, Y., Wang, C.: An algorithm for community identification and dynamical addition based on web pages contents similarity and link relation. J. Zhengzhou Univ.(Nat.Sci.Ed.) 43(1) (March 2011) (in Chinese)
41. Wang F., Zeng D., Mao W.: Social Computing: Its Significance. Development and Research Status (July 2010) (in Chinese)

# Worm Propagation Control Based on Spatial Correlation in Wireless Sensor Network*

Wei Guo[1,2], Lidong Zhai[1,**], Li Guo[1], and Jinqiao Shi[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] Beijing University of Posts and Telecommunications
{guowei,zhailidong}@software.ict.ac.cn,
{guoli,shijinqiao}@ict.ac.cn

**Abstract.** Recent threats caused by worms on wireless sensor network are recognized as one of the most serious security threats. Due to the limit of the hardware, there wouldn't many protect technologies applied in the sensor nodes. Therefore, it is necessary to control worm propagation through wireless sensor network. In this paper, we proposed two new parameters (Csc/Fsc) based on spatial correlation to control worm propagation. Both of them can be used as a defense approach, but also an attack technology that achieve some certain effect. A math model is built to analyze the two new parameters and simulated in the Matlab from distribution, speed and location.

**Keywords:** Worm propagation**,** wireless sensor network, network security, spatial correlation.

## 1 Introduction

Wireless sensor network (WSN) have gained worldwide attention in recent years. These sensors are small, with limited processing and computing resources, and they are inexpensive compared to traditional sensors. These sensor nodes can sense, measure, and gather information from the environment and, based on some local decision process, they can transmit the sensed data to the user. [1]

Majority of the sensor network are deployed in hostile environments with active intelligent opposition. Hence security is a crucial issue.

As the special use of wireless sensor network, in our paper, a wireless sensor network is a typically an ad hoc network, which requires every sensor node be independent and flexible to be self-organizing. There is no fixed infrastructure available for the purpose of network management in a sensor network. This inherent feature brings a great challenge to wireless sensor network security. [2]

Also an ad hoc network requires spatially dense sensor deployment in order to achieve satisfactory coverage [3, 4]. Due to high density in the network topology,

* This work is partially supported by 863 National Hi-tech Research and Development Program (2011AA01A103).
** Corresponding author.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 68–77, 2012.

spatially proximal sensors transmitting are highly correlated. [5] So, we proposed two parameters based on spatial correlation to control worm propagation.

The rest of this paper is organized as follows. In Section II, we give background and related work on worms spread modeling. Section III, present a new model to describe worm propagation control based on spatial correlation. In Section IV, we discuss the results of worm propagation simulations. We summarize our work in Section V.

## 2 Related Works

The spreading of worms caused more damages in wireless sensor network than internet. To the Internet, the rapid spread of worm in backbone network may lead to service quality diminished and affect the speed of users connecting to the network. However, to the wireless sensor network, the rapid spread of worm may cause the damage to the physical world. Not only cause economic losses, but also lead people to a danger situation.

Although there have been many works about spatial correlation, it is still have a lot of work to do. James O Berger et al in [6] proposed four spatial correlation model, spherical, power exponential and rational quadratic. Na Li et al in [7] used a new spatial correlation model to analyzed data distortion in WSN.

Despite the worm propagation through wireless sensor network will pose a serious threat, study in this area is still at the initial stage. Pradip De et al in [8] they develop a common mathematical model (SIR) for the propagation in wireless sensor network. Based on this model, they analyze the propagation rate and the extent of spread of a malware over typical broadcast protocols proposed in the literature. Khayam, S.A et al in [9] propose a topology aware temporal and spatial worm propagation model in sensor networks. Although they present a closed form solution for computing the infected fraction of the network, their model assumes a structured grid topology and also does not consider the simultaneous effects of any recovery process on the infection spread.

Recently, the first attack technology ikee.B for smart Bo Gu et al in [10] mention a concept that model can be classified in two categories, namely, the proximity based model and encounter based model. Their model belongs to proximity based model while our work belongs to encounter based model.

## 3 Worm Propagation Model

### 3.1 Network Model

We proposed a worm propagation control model in Wireless sensor network. While there is widespread agreement in the WSN network frame that the nodes include sensor nodes and sink nodes. Information sensed by sensor nodes will aggregate and be processed at the sink nodes, then transmitted to the high level users.

In this paper, sensor nodes are random distributed into a large area with high density and do not move any more if they are deployed in the network. Moreover, all the sensor nodes in the network are isomorphic, which means that all the sensor nodes have the equal ability of information sensing, transmitting and processing. All the sensor nodes are modeled as omnidirectional Boolean sensing, which means that all the nodes in the network have a fixed sensing radius, the sensing area is valid within a circle centered by the node's spatial position [11]. Also all the sensor nodes have a transmitting radius which is double as sensing radius.

We define the above network by four parameters: Area *S*, Number of nodes *N*, sensing radius *Rs*, and transmitting radius *Rt*. The network model is shown as Fig. 1.

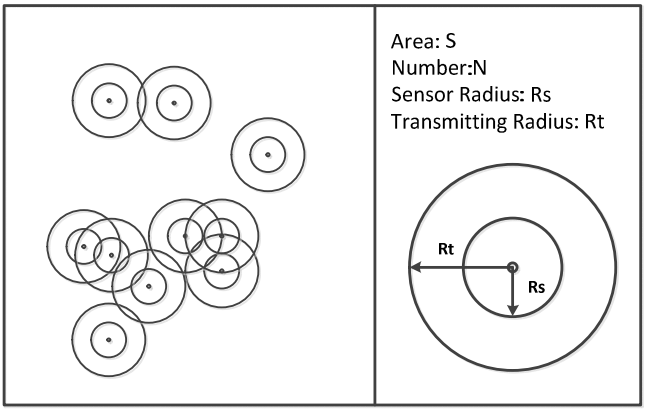



**Fig. 1.** Network Model

### 3.2 Spatial Correlation Parameter

In this part, we proposed two new parameters called Close Spatial Correlation (Csc) and Far Spatial Correlation (Fsc). Before we introduce these two parameters, we have to cite a Spatial Correlation model.

The Spatial Correlation Model sets up by geometry. As shown in Fig 2, the meanings of symbols explained as follows:

- $A_i$ Denotes the round region with center $S_i$ and radius Rs;
- $A_i^j$ Denotes the region which is demarcated by the perpendicular bisector of $S_iS_j$, and next to $S_i$ but belongs to $A_i$;
- $d$ is the distance between $S_i$ and $S_j$;

If $d$ <2Rs, define the correlation as

$$K_\theta(d) = \frac{A_i^j}{A_i - A_i^j} \tag{1}$$

Let $\theta$=2Rs

$$K_{\vartheta}(d)=\begin{cases}\dfrac{\theta^2\cdot\arcsin\sqrt{1-\dfrac{d^2}{\theta^2}}-d\cdot\sqrt{\theta^2-d^2}}{\theta^2(\pi-\arcsin\sqrt{1-\dfrac{d^2}{\theta^2}})+d\cdot\sqrt{\theta^2-d^2}} & ;0\le d\le\theta \\ 0\ ; & d>\theta\end{cases} \tag{2}$$

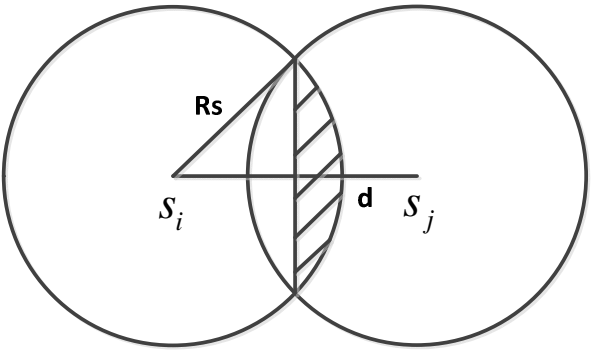



**Fig. 2.** Spatial Correlation

From analyzing the above model, we can inform that $K_{\theta}(d)$ only represents the correlation between two nodes. In this paper, the two new correlation parameters that we proposed try to describe the correlation of every node independently. Csc is summation of one node with all the rest nodes in WSN and Fsc multiply Csc equals 1/1000.

$$\mathrm{Csc}_i=\sum_n K_{\theta}(d)-1 \tag{3}$$

$$\mathrm{Fsc}_i=\frac{1}{Csc_i*1000} \tag{4}$$

If one node's Csc is high, it means that there are many nodes close to it. In Fig1, the node in the high density area will get a high Csc. On the other hand, if one node's Fsc is high, it means that there are few nodes in its sensing area and having a long distance to the node location. In Fig1, the node in the low density area will get a high Csc.

### 3.3 Worm Propagation Model

Scenario: we proposed worm propagation model in wireless sensor network. Assuming that every time a node only transmits with one node and once a node receive the information from an infected node, it will be infected. If a node is infected, it sent information to the highest Csc/Fsc node which hasn't been infected and the distant between the two nodes not more than Rt.

Assuming that at time t the probability of node i being infected is $p_{i,t}$, 1 is a unit of time. Node i is infected at time t can be divided into two cases. First, the node i at time t-1 is infected. Second, node i is infected by an infected node J at time t while node i is not infected at time t-1. Clearly, $p_{i,t}$ is the sum of the above two cases. Note that the probability that an infected node J connecting nodes i and the probability that J is with virus are two independent events. Here we will use mathematical models to describe the above analysis.

First, define several variables. $p_{i,t}$ is the probability that node i is infected at time t, so obviously $p_{i.t-1}$ is the probability that node i is infected at time t-1; $\beta$ is the probability that node i and j connected to each other, noting that $\beta_{ji}$ is the probability that node j connect to node i, and generally $\beta_{ji}$ is not equal to the probability that node i connect to node j.

With the above analysis and variables definitions, the mathematical model we describe is as follows:

$$1-p_{i,t}=(1-p_{i,t-1})\prod_{j\neq i}(1-\beta_{ji}p_{j,t-1}) \tag{5}$$

Due to the way that infected node sent information, we assume that $S_i$ is a set of Csc/Fsc whose nodes can receive the message from Node I and haven't been infected.

$$S_i=\{Csc_j/Fsc_j \mid d_{ij}<2Rt \text{ and } Node_j \text{ is uninfected}\} \tag{6}$$

So，If Node i can be infected by Node j, only the max of $S_j$ equals $Csc_i / Fsc_i$.

$$\beta_{ji}=\begin{cases}0 & Max(S_j)\neq Csc_i/Fsc_i\\ 1 & Max(S_j)=Csc_i/Fsc_i\end{cases} \tag{7}$$

Therefore, the probability expects value that the node i at time t being infected is:

$$E=\sum_{i=1}^{N}p_{i,t}=\| P_t \| \tag{8}$$

Above is a mathematical model of worm propagation control and we solve it by Matlab. The following we will be simulating Csc and Fsc situations based on this model.

## 4 Simulations and Analysis

### 4.1 Simulation Description

The target of our simulations is to find out Csc and Fsc distribution, the speed of worm propagation and the location of the infected node. There are many algorithm of

worm propagation to choose as the control group, but in this paper we choose random worm propagation causes it is easy to achieve and compare the effect with other algorithms. We use the numerical analysis tool, Matlab, to derive theory results from the mathematical model. The following is divided in two parts, Csc and Fsc and analyzes them from distribution, propagation speed and location of infected nodes.

### 4.2 Analysis on Csc Simulation Results

Figure 3-5 respectively show the distribution, speed and location of the Csc worm propagation. In the simulations, we set N=4000, S=1600*1600, Rt=100, Rs=50, to keep a high density of wireless sensor network.

Fig. 3 show the distribution of Csc. It obeys the normal distribution for the reason that nodes in the wireless network are random distributed.

Fig. 4 show the infected nodes expectation of Csc. We can draw the conclusion that the speed of Csc is slower than random. Because every time the node choosing a highest Csc node, it means the node choose a nearby node. In our paper, if surrounding nodes are all infected, this node cannot infect other nodes. This is the reason why Csc is slower than random. So, Csc can be used as a way to slow down the speed of worm propagation.

Fig. 5 show the location of the infected nodes when T=8 in Fig. 4, because the number of infected nodes at T=8 is almost the same and there are enough infected nodes to show the result. In the picture, the area of circle nodes is nearly half of star nodes. It demonstrates that Csc is expert in worm propagation in certain area of the whole wireless sensor network.

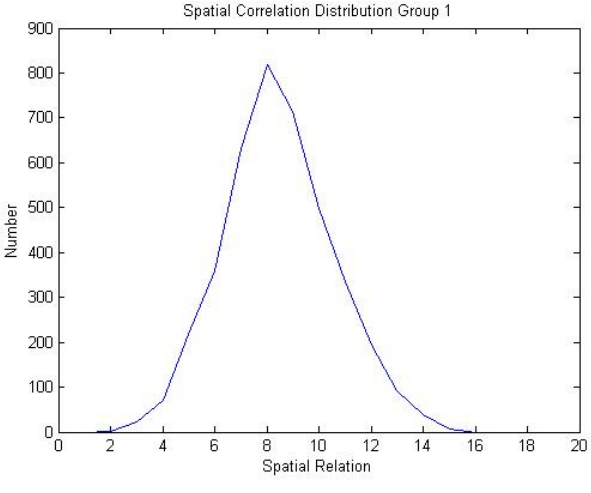



**Fig. 3.** Spatial Correlation Distribution

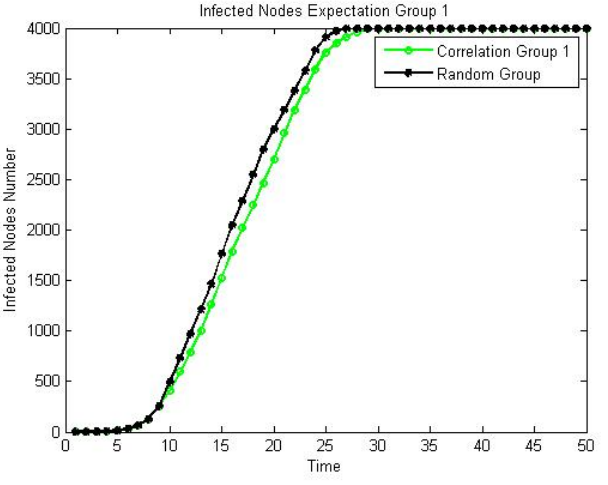



**Fig. 4.** Infected Nodes Expectation

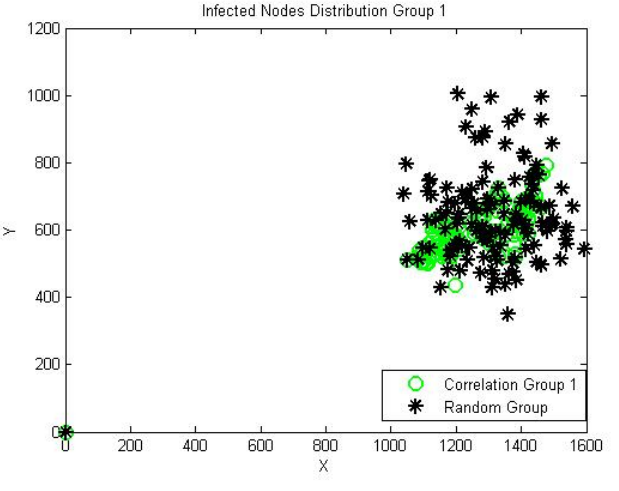



**Fig. 5.** Location of infected nodes

### 4.3 Analysis on Fsc Simulation Results

Figure 6-8 respectively examine distribution, speed and location of the Fsc worm propagation. In the simulations, we set N=2000, S=1000*1000, Rt=100, Rs=50, to get a higher density of wireless sensor network.

Fig. 6 show the distribution of Fsc. Because 1000*Fsc is the reciprocal of Csc, the line is related with normal distribution.

Fig. 7 show the infected nodes expectation of Fsc. We can draw the conclusion that the speed of Fsc is faster than random. Because every time the node choosing a highest Fsc node, it means the node choose a far node. In our paper, if surrounding nodes are not all infected, this node can keep infected other nodes. This is the reason why Fsc is faster than random. So, Fsc can be used as an attack way to accelerate speed of worm propagation.

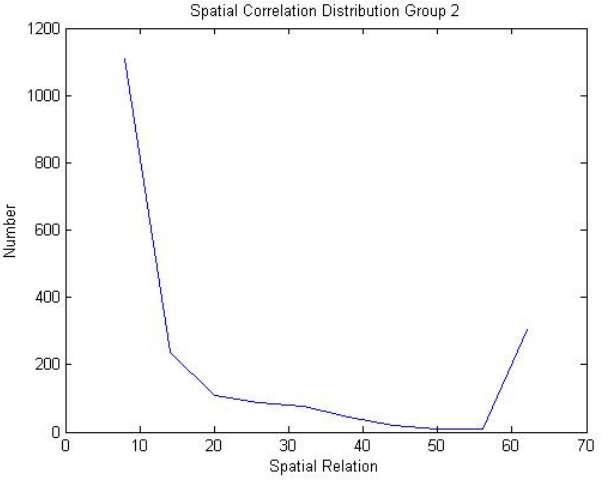



**Fig. 6.** Spatial Correlation Distribution

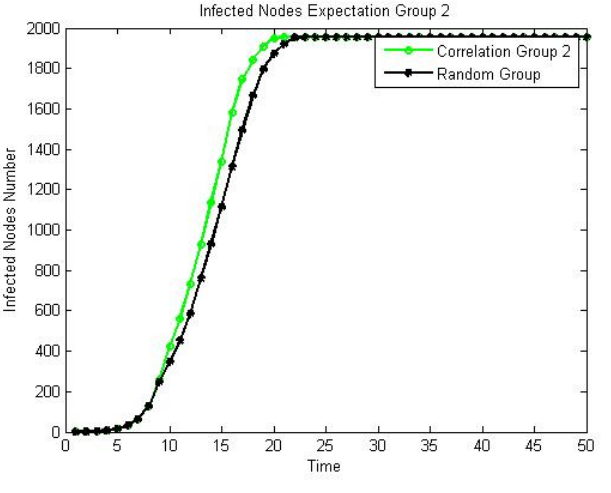



**Fig. 7.** Infected Nodes Expectation

Fig. 8 show the location of the infected nodes when T=8 in Fig. 7, because the number of infected nodes at T=8 is almost the same and there are enough infected nodes to show the result. In the picture, the area of circle nodes is nearly double of star nodes. It demonstrates that Fsc is expert in worm propagation to cover the whole wireless sensor network.

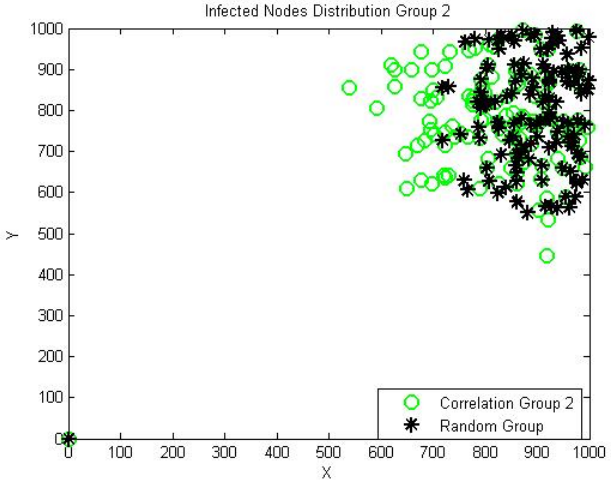



**Fig. 8.** Location of infected nodes

## 5 Conclusion

In this paper, based on the spatial correlation, we proposed two new parameters Csc and Fsc. Worm propagation through Csc is slower than random way, but faster in certain area of the whole wireless sensor network. On the other hand, Fsc is faster than random and is expert in covering the whole wireless sensor network. Both of them can be used as defense policy or an attack technology.

**Acknowledgment.** This work is partially supported by 863 National Hi-tech Research and Development Program (2011AA01A103).

## References

1. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Computer Networks 52(12), 2292–2330 (2008) ISSN 1389-1286, doi:10.1016/j.comnet.2008.04.002
2. Padmavathi, G., Shanmugapriya, D.: A Survey of Attacks, Security Mechanisms and Challenges in Wireless Sensor Networks. International Journal of Computer Science and Information Security (2009) ISSN 1947 5500

3. Clouqueur, T., Phipatanasuphorn, V., Ramanathan, P., Saluja, K.: Sensor deployment strategy for target detection. In: Proceedings of the ACM WSNA 2002, Atlanta, USA (September 2002)
4. Meguerdichian, S., Koushanfar, F., Potkonjak, M., Srivastava, M.B.: Coverage problems in wireless ad-hoc sensor networks. In: Proceedings of the IEEE INFOCOM 2001, Anchorage, AK (April 2001)
5. Berger, J.O., de Oliviera, V., Sanso, B.: Objective Bayesian analysis of spatially correlated data. Journal of the American Statistical Association 96, 1361–1374 (2001)
6. Li, N., Liu, Y., Wu, F., Tang, B.: WSN Data Distortion Analysis and Correlation Model Based on Spatial Locations. Journal of Networks 5 (December 2010)
7. Song, J.-G., Jung, S., Kim, J.H., Seo, D.I., Kim, S.: Research on a Denial of Service (DoS) Detection System Based on Global Interdependent Behaviors in a Sensor Network Environment. Sensors 10, 10376–10386 (2010)
8. De, P., Liu, Y., Das, S.K.: An Epidemic Theoretic Framework for Evaluating Broadcast Protocols in Wireless Sensor Networks. In: IEEE Internatonal Conference on Mobile Adhoc and Sensor Systems, MASS 2007, October 8-11, pp. 1–9 (2007)
9. Khayam, S.A., Radha, H.: A topologically-aware worm propagation model for wireless sensor networks. In: 25th IEEE International Conference on Distributed Computing Systems Workshops, June 6-10, pp. 210–216 (2005)
10. Vuran, M.C., Akan, Ö.B., Akyildiz, I.F.: Spatio-temporal correlation: theory and applications for wireless sensor networks. Computer Networks 45(3), 245–259 (2004) ISSN 1389-1286, doi:10.1016/j.comnet.2004.03.007
11. Hossain, A., Biswas, P.K., Chakrabarti, S.: Sensing Models and Its Impact on Network Coverage in Wireless Sensor Network. In: IEEE Region 10 and the Third International Conference on Industrial and Information Systems (ICIIS 2008), Kharagpur, December 8-10, pp. 1–5 (2008)

# PointBurst: Towards a Trust-Relationship Framework for Improved Social Recommendations

Hongchen Wu[1], Xinjun Wang[1,*], Zhaohui Peng[1], Qingzhong Li[1], and Lin Lin[2]

[1] School of Computer Science and Technology, Shandong University, Jinan 250101, China
[2] China Electronic System Engineering Company, Beijing 100089, China
hc_wu@mail.sdu.edu.cn, {wxj,pzh,lqz}@sdu.edu.cn,
lilyvictory2003@hotmail.com

**Abstract.** With the rapid growth of information on the World Wide Web, social recommendations have appeared as one of the most important roles attracting growing attentions from researchers. Social recommendations enable a form of efficient knowledge for users and help them share contents with others. There are many studies in this area focusing on using trust-relationships in recommendation algorithm, which has become a major trend in recommendation algorithms that used for searching information precisely, feasibly and efficiently, but they neglect how to build the trust-relationships framework at start. In this work, an algorithm, called *PointBurst*, is proposed for building a trust-relationship framework to improve the social recommendations when there is no or too few of available trust-relationships. Here, we first construct a graphical model based on a binary-type vertex relationship, where discusses the explicit and potential connections among users and recommended items. On this basis, we implement a common-used collaborative filtering recommendation algorithm to deal with the situation of enough available trust-relationships existing, and then present *PointBurst*, which builds trust-relationship framework as a supplement. Finally, we crawl through data from three famous recommender websites, i.e., del.icio.us, Myspace and MovieLens and use them in experiments to show that *PointBurst* can suggest relevant items to users' tastes and perform better than collaborative filtering algorithm in precision and stability.

## 1 Introduction

Recently, social recommendations have ushered in numerous studies of networking websites which include personal applications (e.g., online communication, chatting software, and online games) and websites that promote communal interactions (via SNS, Web Log and Micro-Blogging, etc) [1~4]. The methods based on trust-relationships have been under increasing research attention in recommendation algorithms. They also provide interested users with information that is pertinent to their previous interactions with other information, based on their user profiles and

* Corresponding author.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 78–88, 2012.

item descriptions [5]. Two main problems are still existing in the study of recommendation: First, data from web is heterogeneous and it suffers from loss, noise and other issues. Users usually maintain privacy while browsing the internet, making the available data scarce in some respects. Second, the main character of network data is dynamic and this applies to user interests [6] as well as the recommended items themselves [7]. To satisfy users' individual requirements at any moment, the stability of data shall be considered. Recent recommendation algorithms are mainly focusing on using the trust-relationship among users to provide recommendations, ignoring the case of too few of available trust-relationships. Here, we present a friendship recommendation algorithm, *PointBurst*, which is added into collaborative filtering recommendation and directed at building trust-relationship framework.

The rest of the paper is organized as follows. Section 2 reviews related work and puts it into context. Section 3 presents a graphical model based on a binary-type vertex relationship discussing the connections among users and recommended items, which will be used for the construction of the trust-relationship framework. Section 4 presents the implemented recommendation algorithm, and gives our main idea, *PointBurst*, by code. Finally, in Section 5, data crawled from three famous recommender websites, e.g., del.icio.us, Myspace and MovieLens have been used to conduct an experiment showing that our algorithm *PointBurst* is better than previous collaborative filtering recommendation algorithm.

## 2 Related Works

The development of new types of recommendation systems means that the connections among users from networks are closer and they have formed several social communities, known as social networks. Social networks are statistical physical communities with the properties of networked systems, including the Internet, the World Wide Web, and social and biological networks, and have become the major platform for applying recommendation algorithms based on the relationships between users. The usage of trust-relationships for making recommendations, in particular, has been very successful. Schenkel suggested that the basic concept of user trust-relationships is that users tend to believe their close friends more than their acquaintances, and he formally presented the advantages of user trust-relationships in recommendation system evaluation [8]. Golbeck considered trust in social networks as the basis of recommendation algorithms and analyzed the architecture and practical applications of FilmTrust, which is a social network recommender website based on trust-relationships [9]. Machanavajjhala discussed the correlation between privacy and precision in the trust-relationship recommendations found in social networks [10]. Assent used the trust-relationship among users in a social network to "cultivate" a small recommendation system group [11]. Li proposed a novel recommendation method, which leverages the viral marketing in the social network and the wisdom of crowds from endorsement network [12]. Aiming at modeling recommender systems more accurately and realistically, Ma proposed a novel probabilistic factor analysis framework, which naturally fuses the users' tastes and their trusted friends' favors together [13]. Many studies have investigated the

utilization of trust-relationships among users to provide recommendations, but they rarely discuss the construction of trust-relationships before using them or their number is too few. Thus, we mainly focus on trust-relationship framework construction in social networks.

## 3 Graph Model

The entities found in social networks can be cast into a graphical model, which represents the different elements of a social network and their mutual relationships. The model is represented as a network $G = (V, E)$ where $V$ represents a set of vertices and each vertex can either represent a user or an item (a message, a product or others waiting to be recommended to users, such as a film from FilmTrust, a picture from Flickr, or a movie from YouTube). $E$ represents a set of edges. Each edge connects two vertices and it represents a type of relationship between them. Edges can be weighted in different ways, depending on the applications built on top of the graph (via classmates in Facebook, friendship in Myspace, or owning the same movie between users in MovieLens, etc). There are many different types of vertices and edges, but they can be divided into two main types, i.e., relations between vertices of the same type and relations between vertices of the different types.

### 3.1 Relations between Vertices of the Same Type

Relations between vertices of the same type can be divided into two parts: The first group is connections between users, i.e., friendships. The second group is clusters, which represent the relations between items.

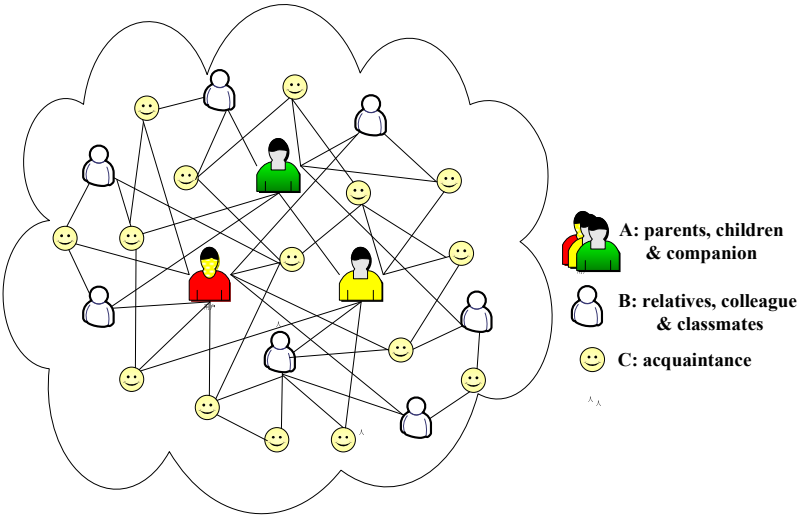



**Fig. 1.** The relationships among users of three levels

**Friendship**
A friend of a user can take many forms. A user might have talked with someone finding that they share a common interest with the user, so they keep them as a friend, formally. A friend can also come from other social network community after an introduction from another friend that the user doesn't know well. If a user *u* has an edge connection with a user *u'*, no matter how *u'* comes, he is defined as a friend of *u*. Friends cab help users find interesting items as a reference within a reasonable amount of time. Friends of a user *u* can be divided into three levels of groups depending on to how close friends are to the user *u* illustrated by Fig. 1.

Generally, a friend from the group *A* is closely connected to the user *u* and they share the same beliefs and hobbies (no risks). Friends from group *B* are rated lower than *A*, and *u* should consider the suggestion carefully. Recommendation from group *C* might be used for reference, but they may be taking a risk when trusting them. The gaps among these groups are not strict. Here we give the formal definition of friendship.

**Friendship of User *u* (User *u'*, String *group*, float *strength score*)**
Each user *u* has a friend list that maintains their data. It holds the set of users $U_u$ that contains all users such as *u'* that have a connection with *u*. *group* represents the level of group that *u'* belongs to, which is represented by *A*, *B,* or *C*. *strength score* indicates the level of proximity to user *u'* in the same group, where a higher score represents a closer relation with user *u*. All this information can help the user *u* to maintain records of friends that are determined by the user himself.

**Cluster**
When two items belong to the same series, type or other form of the same species, they can be placed into a same cluster to show they are alike. Using the items that are contained in the same cluster can easily determine the similar items. Our implemented definition of a cluster is given below:

**Cluster of Item *i* (Item *i'*, float *similar score*)**
For each item *i*, similar items can be determined such as *i'* depending on their clusters. Furthermore, the *similar score* can show the similarity between items *i* and *i'*, where a higher score shows that they are more alike within a same cluster. Furthermore, this may indicate that a user is interested in both or neither of them.

### 3.2 Relations between Vertices of the Different Type

For short, the relation between vertices of the different types is the relation between a user and an item, i.e., tagging.

**Tagging of User *u* (Item *i*, Tag *t*, float *item score*, float *tagging score*)**
Tags are employed to remind users of relevant items. Each tag must be tagged by at least one user, and it must be related to at least one item. The *item score* indicates the attitude of user *u* towards item *i*, where a higher score represents greater interest.

The most important concept is the *tagging score*, which shows how many times *u* has used this tag. A tag can help identify user topics, so the highest score attributed to a tag will be a user's favorite topic.

Our study was mainly focusing on building trust-relationships among users on the base of collaborative filtering recommendation, so, we will first proposed a common-used collaborative filtering recommendation, which can handle the algorithm when there are enough trust-relationships. Then, on this basis, our main idea, *PointBurst*, is provided to deal with the situation of limited amount of trust-relationships using the graphical model listed above.

## 4 The Proposed Methods

Current global popular recommender websites, such as MovieLens and FilmTrust, mainly use collaborative filtering recommendation algorithms. These algorithms shall be generalized to be the foundation of our main idea.

### 4.1 Collaborative Filtering Recommendation

In general, collaborative filtering recommendation systems aim to provide personalized recommendations of items to users based on their previous behavior and other information gathered from item descriptions and user profiles [14]. In this section, the theoretical model of collaborative filtering recommendation algorithm will be given.

Each item in our theoretical model has an *Item Score*. This is computed by *User Score*, which is given by a registered user. It is assumed that $U_x$ represents the set of users who give a score to item *x*. $S_{x\text{-}u}$ represents the score that given to item *x* by user *u* after viewing the item, where $\boldsymbol{u \in U_x}$ and the score ranges from 1 to 5. In this system, 1 indicates that *u* considers that item *x* is of least value. $S_{x\text{-}total}$ is the total score for item *x*, which is computed from all of the scores given by all the users in $U_x$, and can be used to rank items with our algorithm. The top 10 items consist of an item list known as *Item Top 10.* Experienced users give scores to the items and $C_i$ will be set, known as the *Evaluation Weight*, for each of the registered users. When a new registered user gives a score to an item for the first time their $C_i$ will be set at 1, and this increases by +1 whenever they give a score. Higher scores mean that a user has more scoring experience. In general, $S_{x\text{-}total}$ is calculated as follows:

$$S_{x-total} = \left( \sum_{i \in U_x} S_{x-i} \times C_i \right) / \left( \sum_{i \in U_x} C_i \right) \tag{1}$$

$S_{x\text{-}total}$ has to be updated whenever there is a new user *u* scores item *x*, so the new total score of item *x* will be:

$$S_{x-total} = \alpha \times S_{x-u} + (1-\alpha) \times S_{x-toal} \tag{2}$$

where $\alpha$ stands for the weight coefficient, ranging from 0 to 1. It can be computed as follows:

$$\alpha = (2 \times C_u) / \left( \sum_{i \in U_x} C_i + C_u \right) \tag{3}$$

After we get the total score for item *x*, the *Item Top 10* ranking for users can be constructed to view, and also provide recommendation of interest. When $S_{x\text{-}total} > L$ (a user's limitation or preference), the system will automatically recommend item *x* to user *u* (if this user has set a threshold to avoid low level items).

### 4.2 PointBurst: A Recommendation Algorithm towards Building Trust-Relationship Framework

Collaborative filtering recommendation algorithm need to capture more information from users. However, it is common that users trust recommendations of close friends better than the recommendations from system [15]. Also, some users will have little interaction with the system and don't care much for *Items Top 10*, or these users might not give a score after they have viewed items and makes the recommendation system redundant [16]. However, *PointBurst* can help.

This recommendation allows users to maintain a friend list that keeps a record of his common friends. If a user fails to score any items, our algorithm will make a recommendation based on the friend list. The flowchart of our algorithm is shown in Fig.2. When a user makes his query using this platform, collaborative filtering recommendation will begin immediately. If the user doesn't accept this recommendation and has so few of trust friends, *PointBurst* will be activated. The items *I* that selected by user will determine the cluster *C* to which they belong. A record of the items *I'* will be maintained in *C*, and then locate the users *U* who own the items in *I'*. The set of users, *U*, are known as $U_{original}$. After the comparison of similarity, we finally get the set of trusted users, $U_{selected}$. The equation for the similarity is given below:

$$Similarity(u,u') = \alpha \times \frac{\left| \inf set(u \wedge u') \right|}{\left| \inf set(u) \right|} \tag{4}$$

Where $\inf set(u)$ and $\inf set(u \wedge u')$ represents the items selected by user *u* and the items selected by both *u* and *u'*, $\alpha$ represents coefficient and *S* represents a user threshold. When $Similarity(u,u') > S$ ( $u' \in U_{original}$ ), *u'* will be placed into $U_{selected}$. Using the algorithm provided above, the set of trusted users, $U_{selected}$, will be determined. The values of $\alpha$ and *S* will be determined in experimental section. The whole process of our algorithm, *PointBurst*, is given here, and it is supposed that the query user *u* has selected a set of items *I* consisting of $i_1$, $i_2 \ldots i_n$ and *S* represents the threshold of *u* to avoid low level items.

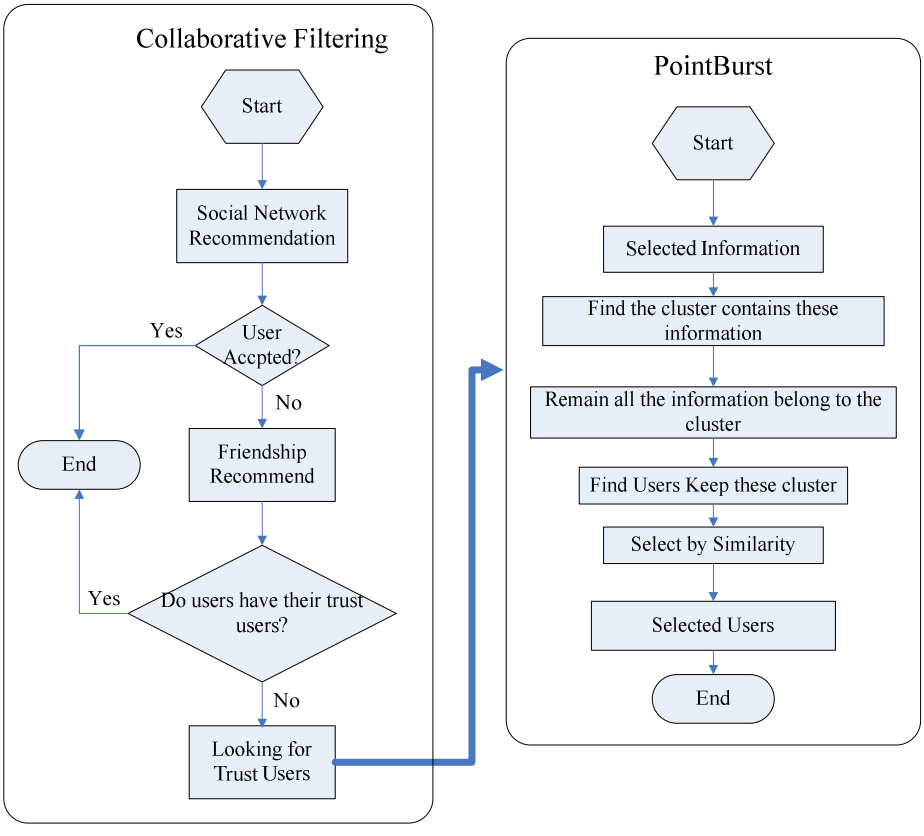



**Fig. 2.** The flowchart consists of collaborative filtering recommendation and *PointBurst*

In this paper, we will conduct a reasonable experiment to show that our algorithm, *PointBurst*, is better than collaborative filtering recommendation in providing recommendations to users' requirements, both in precision and stability.

## 5 Experiments

### 5.1 Data Collections and System Configuration

Before this section, different types of data have been crawled through from three famous information recommender websites: del.icio.us, Myspace and MovieLens, and they are used as the basic data sets in our experiment to show the effectiveness and precision of *PointBurst*. The introductions of these data from them are given below:

1. del.icio.us: The world's largest bookmark website (http://del.icio.us/ ) at presents, and we have crawled 389 users, 754 bookmarks, and 1096 tags from there.

2. Myspace: One of the most influential Social Networking Services, and 347 users and 777 friend relationships have been crawled from this website (http://www.myspace.cn ). Besides, each user has an explicit friend list.
3. MovieLens: One of the most powerful recommender website concentrates on movie-recommended research, and we have gotten 486 users, 605 movies with 893 related marks, and 717 friend relationships. Though some users don't have explicit friends but their connections still exist.

Before giving the comparisons of these data, we must confirm the values of two essential coefficients, which are set in Equ. (4) of section 4, where has presented our equation for similarity.

Here, the values of $\alpha$ and $S$ are determined according to our data from these websites in order to make the precision of recommendation algorithm higher. From the statistical analysis of our current data, almost 62.19% users have their friend list, and if *item score* is set ranging from 1 to 5(higher score means a better evaluation to this item from the user), most of the users will set 3 or 4(almost 46% for each) as their Limitation $S$. Thus $\alpha$ ranges from $8.034\left(3 \div 0.3734\right)$ to $10.713\left(4 \div 0.3734\right)$.

We now set up our main parts of the experiment. Firstly, the three data sets from websites are preprocessed: The friend list of each user in data from del.icio.us has been removed, so they can't get recommendations from friends they have recorded, but to resort to our algorithm; Most of the users, almost 70% in data sets of Myspace, have their friend list without preprocessing, that means the majority of users will get recommendations by collaborative filtering method, except for a small number of users who don't keep a record of friends originally; parts of users' friend lists in data collected from MovieLens are removed, making almost 25% users still keep a record of close friends. All the users from these three data will attend in our experiment, and we have gotten permission from them. In order to compare the precision and stability between *PointBurst* and collaborative filtering recommendation algorithm in these three different situations, three pairs of comparisons are given as follows:

1. *PointBurst* recommendation algorithm based on del.icio.us data set (PBRAD) and collaborative filtering recommendation algorithm based on del.icio.us data set (CFRAD).
2. *PointBurst* recommendation algorithm based on Myspace data set (PBRAM) and collaborative filtering recommendation algorithm based on Myspace data set (CFRAM).
3. *PointBurst* recommendation algorithm based on MovieLens data set (PBRAL) and collaborative filtering recommendation algorithm based on MovieLens data set (CFRAL).

In order to make the differences of experiment results more obvious, we set the value of $\alpha$ =9.5 and $S$=3. All the methods are implemented in Java Language with Myeclipse 6.5 platform and used MYSQL Server 5.0 to generate answers. Experiments are run on PCs with Intel R 2.93GHz CPU and 4G memory under Window 7 operating system.

Precision is the main performance evaluation in our experiment, and the equation of computing precision is given below:

$$\Pr ecision = \frac{S_{TargetHits}}{S_{Total}} \times 100\% \tag{5}$$

where $S_{TargetHits}$ represents the total number of the user $u$ who have chosen the same item with one of the trusted users given by the algorithm, when he logins system for the second time. $S_{Total}$ represents the number of users who board the system for second time.

### 5.2 Experiment Results and Analysis

We first compare the performance of PBRAD and CFRAD in the first part. The time of our algorithm put into usage is 2011-5-31. Because there are small number of users take part in our experiment at start, the results in first 1 month are neglected to make the difference obvious. It is started from 2011-6-30 to 2011-8-31 totally for 60 days to keep the record of the users' number and compute the precision. Over time users' number has increased and they have become more active. The trend graph of PBRAD and CFRAD is drawn as Fig. 3. (a).

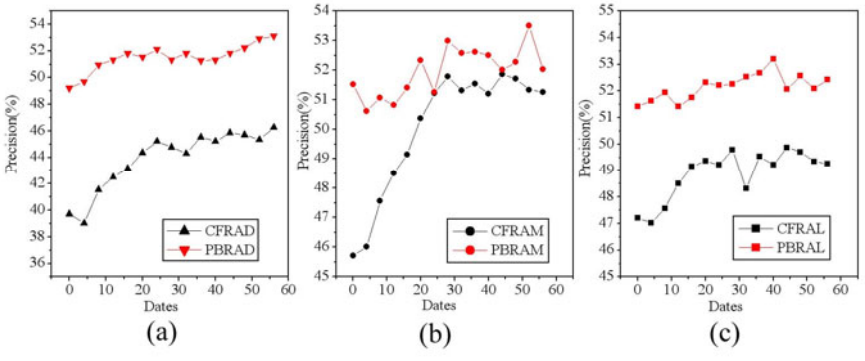



**Fig. 3.** Three pairs of comparisons between collaborative filtering Recommendation and *PointBurst* based on data sets of del.icio.us, Myspace and MovieLens

We have removed the friend list which user has kept a record of, so it is obvious that users cannot receive their interest items from their friends, but *PointBurst* can still build up trust-relationships to provide recommendations, thus the precision of CFRAD is obviously lower than PBRAD. The second part of the experiment compares the performance of PBRAM and CFRAM. The dates of the experiment still range from 2011-6-30 to 2011-8-31 and remain precision as the performance evaluation. The trend graphs are shown in Fig. 3. (b). Most of the users have maintained their friend lists, so the precision of collaborative filtering recommendation is almost as good as *PointBurst* does, but not stable. Furthermore, collaborative filtering has the problem of slow start concluded from result. In the third part of the experiment, the comparison of performance between PBRAL and CFRAL

is operated. Dates also range from 2011-6-30 to 2011-8-31, and precision is used in evaluating the precision in Fig.3. (c). Some users still remain their friend list, but the number of them is too small. The precision of collaborative filtering is a bit higher than experiment part 1 but obviously lower in part 2. Nevertheless, our algorithm, *PointBurst*, achieves higher precision and remaining stable in these three parts, no matter whether the friend list remains or not.

From the comparisons in experiment above, we can apparently conclude that *PointBurst* can act much better performance both in precision and stability. Besides, collaborative filtering has become the main trend of the recommendation algorithm, but there still exist many disadvantages. In this paper, *PointBurst*, aims to supply collaborative filtering recommendation algorithm, has completed the task of building trust-relationship framework and reached the expected goals.

## 6 Conclusion

Social network research and trust-relationship based recommendation are two promising directions in increasing precision and stability of providing users' favorite items. This paper contributed a useful graph model for analyzing the relationships among users and items, and introduced the *PointBurst* algorithm to build the trust-relationship framework as a supplement of previous collaborative filtering recommendation algorithm. From the experiment based on the data crawled from three famous recommender websites, *PointBurst* acted better both in precision and stability than collaborative filtering recommendations and reached the expected achievements.

**Acknowledgement.** This work was Supported by the National Key Technologies R&D Program (Grant No. 2009BAH44B02); the National Natural Science Foundation of China (Grant No.61003051); the Key Technology R&D Program of Shandong Province (Grant No. 2010GGX10114 & 2010GGX10108); and the Independent Innovation Foundation of Shandong University (Grant No.2010TS057).

## References

1. Xu, Y.F., Zhang, L., Liu, W.: Cubic Analysis of Social Bookmarking for Personalized Recommendation. In: Zhou, X., Li, J., Shen, H.T., Kitsuregawa, M., Zhang, Y. (eds.) APWeb 2006. LNCS, vol. 3841, pp. 733–738. Springer, Heidelberg (2006)
2. Royz, S.B., Dasz, G., Yahiay, S., Yuyy, C.: Interactive Itinerary Planning. In: 2011 IEEE 27th International Conference on Data Engineering, ICDE, pp. 15–26. IEEE (2011)
3. Rendle, S., Gantner, Z., Freudenthaler, C., Thieme, L.: Fast Context-aware Recommendations with Factorization Machines. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 331–338. ACM, New York (2011)
4. Smeaton, A., Callan, J.: Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries. In: ACM SIGIR Forum, vol. 35. ACM, Nw York (2001)

5. Setten, M.V.: Experiments with a Recommendation Technique that Learns Category Interests. In: IADIS International Conference – WWW (2002)
6. Castellano, G., Fanelli, A.M., Torsello, M.A.: Dynamic Link Suggestion by a Neuro-Fuzzy Web Recommendation System. In: IADIS International Conference, WWW 2006, pp. 219–226. WWW (2006)
7. Wang, J., Li, Z.W., Yao, J.Y., Sun, Z.Q., Li, M.J., Ma, W.Y.: Adaptive User Profile Model and Collaborative Filtering for Personalized News. In: Zhou, X., Li, J., Shen, H.T., Kitsuregawa, M., Zhang, Y. (eds.) APWeb 2006. LNCS, vol. 3841, pp. 474–485. Springer, Heidelberg (2006)
8. Schenkel, R., Crecelius, T., Kacimi, M., Michel, S., Neumann, T., Parreira, J.X., Weikum, G.: Efficient Top-k Querying over Social-Tagging Networks. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 523–530. ACM, New York (2008)
9. Golbeck, J.: Generating Predictive Movie Recommendations from Trust in Social Networks. In: Stølen, K., Winsborough, W.H., Martinelli, F., Massacci, F. (eds.) iTrust 2006. LNCS, vol. 3986, pp. 93–104. Springer, Heidelberg (2006)
10. Machanavajjhala, A., Korolova, A., Sarma, A.D.: Personalized Social Recommendations Accurate or Private? In: VLDB Endowment, vol. 4. ACM, New York (2011)
11. Assent, I.: Actively building private recommender networks for evolving reliable relationships. In: 25th International Conference on Data Engineering, pp. 1611–1614. IEEE (2009)
12. Li, C.T., Lin, S.D., Shan, M.S.: Exploiting Endorsement Information and Social Influence for Item Recommendation. In: 34th International ACM SIGIR Conference on Research and Development in Information, SIGIR 2011, pp. 1131–1132. ACM, New York (2011)
13. Ma, H., King, I., Lyu, M.: Learning to Recommend with Social Trust Ensemble. In: 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, pp. 203–210. ACM, New York (2009)
14. Konstas, I.: On Social Networks and Collaborative Recommendation. In: 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, pp. 195–202. ACM, New York (2009)
15. Cho, J.H., Kwon, K.: Source Credibility Model for Neighbor Selection in Collaborative Web Content Recommendation. In: Zhang, Y., Yu, G., Bertino, E., Xu, G. (eds.) APWeb 2008. LNCS, vol. 4976, pp. 68–80. Springer, Heidelberg (2008)
16. Dupret, G., Piwowarski, B.: A User Browsing Model to Predict Search Engine Click Data from Past Observations. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 331–338. ACM, New York (2011)

# Analysis Framework for Electric Vehicle Sharing Systems Using Vehicle Movement Data Stream

Junghoon Lee[1], Hye-Jin Kim[1], Gyung-Leen Park[1],
Ho-Young Kwak[2], and Moo Yong Lee[3]

[1] Dept. of Computer Science and Statistics
[2] Dept. of Electric Engineering 690-756, Jeju National University,
Jeju-Do, Republic of Korea
[3] Jinwoo Soft Innovation, Jeju-Do, Republic of Korea

**Abstract.** This paper designs and builds a serviceability analysis framework for electric vehicle sharing systems based on the vehicle movement data stream collected from the taxi telematics system. For the given sharing station distribution and the relocation strategy, our framework can accurately trace the current number of available vehicles in each station using actual travel data consist of the pick-up and drop-off records. Combined with the discrete event simulation, it is possible to measure the service ratio and moving distance. Experiments are conducted to assess the effect of the number of electric vehicles and the access distance to the service ratio for Jeju city area, discovering that up to 91 % service ratio can be achieved with 5 stations and 50 vehicles. In addition, the per-station trace reveals that the relocation strategy must consider the area-specific unbalance between pick-ups and returns, as it significantly affects the service ratio.

## 1 Introduction

Through intelligence and manageability given by information technologies, the modern electricity system called the *smart grid*, can save energy, reduce cost, and improve the power system reliability [1]. Along with the efficiency in power chain from generation to consumption, smart transportation is an important building block of the smart grid, while electric vehicles, or EVs, begin to penetrate into our daily life according to the availability of efficient load management and charging infrastructures. EVs can enhance energy efficiency, limit greenhouse gases, and reduce global warming by replacing the conventional petroleum-combustion vehicles [2]. However, they are still relatively expensive. Hence, EV sharing, necessarily combined with public transportation, is a reasonable business model at this deployment stage [3].

EV sharing can be considered to be an organized short-term car rental, for example, a few hours [4]. A household does not own vehicles but accesses share-use vehicles when it wants. Considering that many houses own vehicles not so frequently used, EV sharing can reduce the number of vehicles, not just improving air quality but also remedying the traffic and parking problem.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 89–94, 2012.

For better convenience, EV sharing systems facilitate one-way rentals, that is, a user can return an EV to a different place he or she has picked up. This system essentially consists of more than two stations, while a user must go to one of them to begin a trip. When a booking request is directed to a station having no EV, it can't be served. The service ratio is the most critical factor for the satisfaction in renters' side.

For the sake of improving the service ratio, it is essential to place the stations in adequate places. The place must be easily reachable from many users and equipped with sufficient charging capacity, having large parking space [5]. Actually, the number of places capable of satisfying such requirements is quite limited especially in an urban area. Hence, it is practical to decide whether to build a station for the set of candidate places to achieve the given performance goal. Next, for the multiple-station system, share-use EVs are desirably relocated to overcome the uneven distribution between stations to improve the service ratio. However, how to place stations and how to relocate EVs are indispensably affected by the actual usage pattern. Without any travel data, above decisions must be made just based on a forecast of customer demand [6].

In the mean time, our taxi telematics system has been collecting the location history of each taxi, tracing its current location and status which indicates whether a passenger is on [7]. Each taxi reports its location, time, speed, status, and the like every one or three seconds according to its status. These spatio-temporal data stream creates useful information on the location and time of pick-ups and drop-offs for a sufficiently long time period. It makes possible to assess and further improve the station distribution and EV relocation strategies. In this regard, this paper is to build a stream data processing framework capable of integrating an EV sharing model, aiming at prompting the deployment of EVs into our daily life.

## 2 Analysis Framework Design

Our target area is Jeju city, which has a well maintained road network which essentially follows the entire coast (200 km) and crisscrosses between the island's major points. In terms of a road network, there are about 18,000 intersections and 27,000 road segments. In Jeju city, combined with global positioning system and radio communication technology, the *Taxi Telematics* system traces the position of taxis, to dispatch the nearest taxi to the customer call point exploiting the latest traffic information. $30 \sim 50$ taxis report their location records every minute, and each record includes taxi ID and status fields in addition to the basic GPS data such as timestamp, latitude, longitude, direction, and speed. When this system is extended to EVs, the report may contain EV-specific information collected from DTG (Digital Tacho Graph) and ECU (Electric Control Unit).

Figure 1 shows the location selection for sharing stations. First of all, the road network of Jeju city is represented by a graph consisting of nodes and links. The area shown in the figure is about 10 km $\times$ 4.6 km. For each road segment, only the two end points are shown. After the analysis of the pick-up point distribution

and easiness-to-install, we have selected 5 candidate points for sharing stations. They are the Jeju international airport, a shopping mall in residential area, Jeju city hall, Jeju national university, and a shopping mall in the tourist hotel area. They are numbered from Station 1 to Station 5 sequentially. Their locations are marked via the map interface and converted to WGS coordinates. Any selection strategy can change the sharing station distribution using this map interface.

Airplanes are the major transportation methods to connect Jeju city to other provinces, as Jeju city is located in an island. Many trips begin and end at the international airport for both residents and tourists. Second, the shopping mall in the residential area is appropriate for a sharing station. Shoppers from various city area visit here due to large parking space and easy reachability. Third, the city hall area includes many public institutions such as the city hall, the bus terminal, governmental offices, and the like. Fourth, in Jeju national university, many students, faculty members, and visitors want to use EVs in addition to the public transportation such as buses. Fifth, another shopping mall in a seaside area gathers a lot of traffic not just for shoppers but tourists, as this area has many restaurants and hotels.

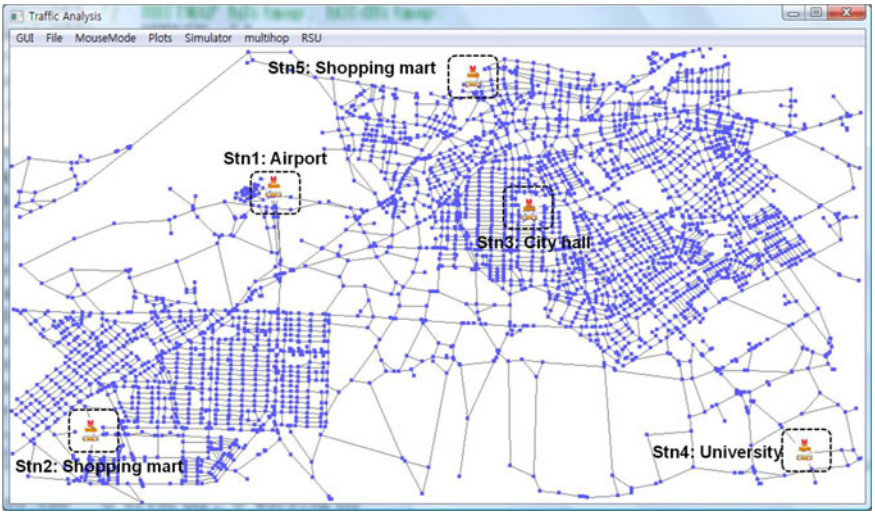



**Fig. 1.** System model

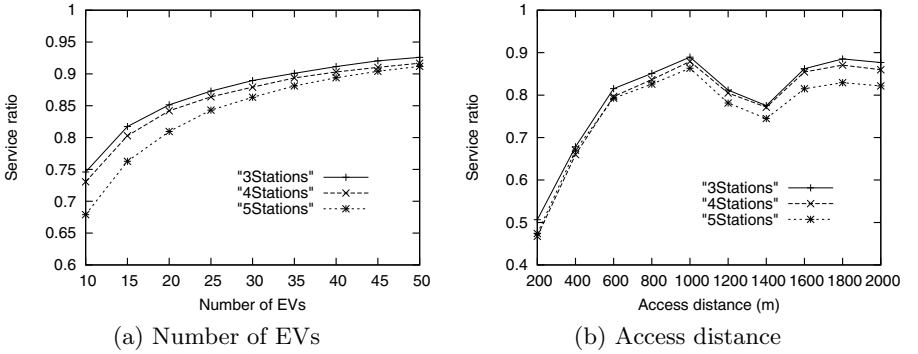
## 3 Analysis Scenarios and Results

The analysis of location history database has found 81,813 trip records during 1.5 month interval. The analysis procedure first assumes that these trip records can give an appropriate demand pattern for EV sharing. For a trip, if its start point is within a given distance bound, this request can be covered by the sharing system. In addition, if there is at least one EV at that station, the request can be served.

The service ratio is the main performance criteria for the assessment of EV sharing infrastructures. Above-mentioned 5 locations are ordered by the traffic load, availability of charging facility, manageability, and many other factors. Hence, if we are to install just 3 stations, it means Station 1, Station 2, and Station 3 are selected. In the following experiment, we set the scenarios of 3, 4, and 5 stations, respectively. At last, we assume that EVs are relocated at every midnight. Even if the analysis framework can implement any relocation strategy, the design of a new EV relocation algorithm is out of scope of this paper. For a pick-up record, if the coordinate marked in its location stamp is serviceable, the EV count of the closest station is decreased. On the contrary, this count increases by one for a drop-off record just like an EV return.

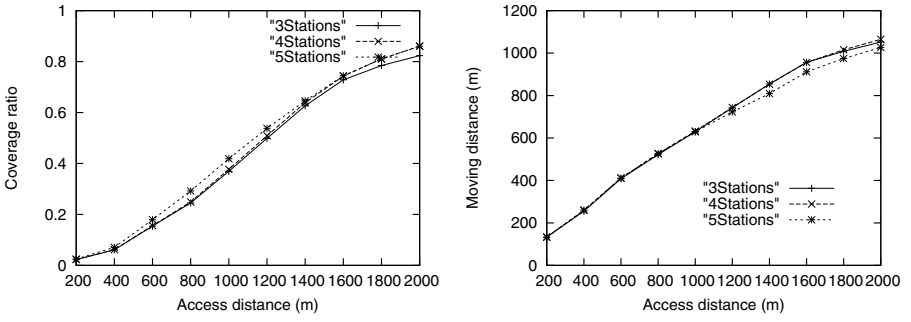
The first experiment measures the service ratio according to the number of EVs, while the results are plotted in Figure 2(a). In this experiment, the access distance, namely, the distance bound, is set to 1,000 $m$. As the total number of EVs is fixed, 3-station case has more EVs per station, showing the highest service ratio. The gap gets smaller according to the increase of the number of EVs. This result indicates that EV pick-up requests are concentrated in the first 3 stations, and the number of EVs for each station is more important for serviceability. Anyway, with 50 EVs, 91 % of pick-ups can be served. The next experiment measures the service ratio according to the access distance, and Figure 2(b) shows the result. In this experiment, the number of EVs is set to 30. Here again, the 3-station case shows the best performance. For the distance interval from 1,000 $m$ to 1,600 $m$, the service ratio is smaller than the adjacent interval. As we just select the closest station, EV booking requests in some area are likely to select the station which has no available EVs.

Figure 3 measures the coverage ratio and the moving distance according to the access distance, respectively. Irrespective of an EV request can be served or not, this experiment checks if a station is reachable from a pick-up point. The more stations, the more requests are covered by the corresponding station distribution. Figure 3(a) discovers that the maximum difference between 5-station and 3-station cases is just 5 %. It is because many trips from Station 4 and Station 5 go to the outside of city area. In addition, Figure 3(b) plots the moving




(a) Number of EVs (b) Access distance

**Fig. 2.** Service ratio analysis

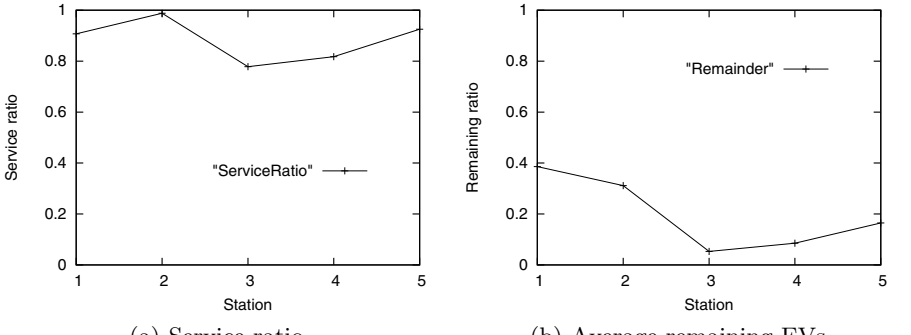



(a) Coverage probability (b) Average access distance

**Fig. 3.** Access distance analysis

distance to an EV station from each pick-up point. When the access distance is less than 1,000 $m$, the moving distance is almost same for the 3 cases. This result indicates EV pick-ups are extremely centered at the sharing stations. Actually, the location history data is obtained from the taxi telematics system, so the public transportation effect is omitted. So, for the 4-station and 5-station scenarios which include Station 4 having much public transportation traffic, the performance enhancement is not significant.

Finally, Figure 4 measures the per-station statistics for the 5-station case. The number of EVs is set to 50, while the access distance is set to 1,000 $m$. For a better service ratio, the balance between the pick-ups and returns is important. Figure 4(a) shows that Station 2 can serve 98 % of requests, while Station 3 serves 77 %. Station 2, the shopping mall in a residential area has constant traffic, so the service ratio is highest. As contrast, the city hall area has unbalanced traffic, that is, many pick-ups but few returns for the most of operation time, making no EV available. Next, Figure 4(b) shows the average number of remaining EVs at midnight just before the relocation. Station 1, the international airport, has 38 %




(a) Service ratio (b) Average remaining EVs

**Fig. 4.** Per-station statistics

of EVs. At evening, there are more departures than arrivals, as tourists take taxi more often. Station 3 has the smallest number of remaining EVs, which is consistent with the low serviceability found in Figure 4(a).

## 4 Conclusions and Summary

EV sharing is a promising business model which can prompt the penetration of EVs into our daily life, not just reducing air pollution and global warming. Not using the forecasted flow but the real-life traffic pattern data obtained from the taxi telematics system which creates spatio-temporal movement data stream, this paper has designed and built a performance analysis framework for EV sharing systems. It can measure the service ratio according to the number of EVs and access distance for the given station distribution and relocation strategy. The experiment discovers that up to 91 % service ratio can be achieved with 5 stations and 50 vehicles. As future work, we are planning to design a relocation scheme which can not only relieve the unbalance between pick-ups and returns but also reduce the relocation distance and delay, aiming at further improving the service ratio.

## References

1. Gellings, C.W.: The Smart Grid: Enabling Energy Efficiency and Demand Response. CRC Press (2009)
2. Ma, Z., Callaway, D., Hiskens, I.: Decentralized Charging Control for Large Population of Plug-in Electric Vehicles: Application of the Nash Certainty Equivalence Principle. In: IEEE International Conference on Control Applications, pp. 191–195 (2010)
3. Hossain, A., Khan, A.: A Framework for Modeling the Design and Operation of Shared Vehicles Systems. In: Annual Transportation Research Forum (2010)
4. Xu, J., Lim, J.: A New Evolutionary Neural Network for Forecasting Net Flow of a Car Sharing System. In: IEEE Congress on Evolutionary Computation, pp. 1670–1676 (2007)
5. Ion, L., Cucu, T., Boussier, J., Teng, F., Breuil, D.: Site Selection for Electric Cars of a Car-Sharing Service. World Electric Vehicle Journal 3 (2009)
6. Wang, H., Cheu, R., Lee, D.: Logical Inventory Approach in Forecasting and Relocating Share-use Vehicles. In: International Conference on Advanced Computer Control, pp. 314–318 (2010)
7. Lee, J.-H., Park, G.-L., Kim, H., Yang, Y.-K., Kim, P.-K., Kim, S.-W.: A Telematics Service System Based on the Linux Cluster. In: Shi, Y., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2007. LNCS, vol. 4490, pp. 660–667. Springer, Heidelberg (2007)

# Research on Customer Segmentation Based on Extension Classification

Chunyan Yang[1], Xiaomei Li[1,2], and Weihua Li[1,2]

[1] Research Institute of Extension Engineering, Guangdong University of Technology, Guangdong, Guangzhou, 510090
fly_swallow@126.com
[2] Faculty of Computer, Guangdong University of Technology, Guangdong, Guangzhou, 510006
lxmdwj@163.com, lw@gdut.edu.cn

**Abstract.** Customer value (CV) can scale the relative essentiality on the customer for an enterprise according to an enterprise as a main body of value and the customer as an object of value at the same time. CV is varied continually in a changing environment. Based on extensible classification method and CV theory, the changing rules of CV can be explored through performing extension transformations, and then the extension classification knowledge on the transfromation of CV can be acquired. Thereby we can carry out the changing segmentation for customers' group. This study can provide a foundation for the enterprise to fomulate the strategy of client relationship management. It can provide a new idea for studying CV and customer segmentation, and also exploit a new applying field for extension data mining.

**Keywords:** customer value (CV), customer segmentation, extension transformation, extension classification, extension data mining.

## 1 Introduction

Customer value (CV) theory in marketing has an important position, and it is the key of customer satisfaction and customer loyalty. The carrier of CV is products and services, so CV can achieve the customers interests and enterprises interests [1]. CV is the essential basis for customer segmentation. Segmentation process based on CV can find the characteristics of customers with different values. Enterprises evaluating CV can find the interests area of customers. Actually there are three research directions in CV study: (1) CV based on customers as the subject of value and the enterprise as the object of value; (2) CV based on an enterprise as the subject of value and the customers as objects of value; (3) CV based on the enterprise and the customers as subjects of value and objects of value at the same time. It is also known as research of CV exchange [1]. CV involved in this paper is the second, that is, research the CV from the perspective of an enterprise.

The CV in an enterprise is continually changing, so the customers' type on CV is also continually changing. The research for CV in the enterprises based on extension

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 95–105, 2012.

classification method [2-4] can obtain varible knowledge of customer type when the enterprises implement transformations on certain evaluating characteristics in the evaluating system. These varible knowledges can provide the basis for customer relationship management of the enterprise.

The current researches of customer segmentation based on CV are mostly static, rarely involved effects based on a variety of transformation. All segmentation methods were described in terms of segmentation dimensions, rationales and the relationship to the purposes of segmentation[5], segmentation algorithm[6-7]. This paper will build the dependent function of evaluating characteristics in the evaluating system [8] and integrated dependent functions based on CV theory [9] and extension set theory [10] to measure CV [11]. Through the implementation of extension transformation to discuss variation law of the CV, obtain extension classification knowledge of the CV based on transformations. It can provide foundation of customer relationship management strategy and new ideas for CV and customer segmentation research.

## 2 Selection of CV Evaluating Index System and the Quantitative Measure of Each Evaluating Index

The central part of CV analysis is customer segmentation to determine the customer group with same demand and values tend. The right customer segmentation can effectively reduce costs, and access stronger and more profitable market penetration. At present the application of customer segmentation methods focus on RFM analysis and CV matrix analysis [1].

There are many kind of CV evaluating index system currently. The reference [8] gives the CV evaluating system according to the current customer value (CCV) and latent customer value (LCV) to evaluate the total value of the customer. This evaluating index system is also consistent with conjugate analysis of the latent and apparent conjugate analysis in Extenics [10], shown in Figure 1.

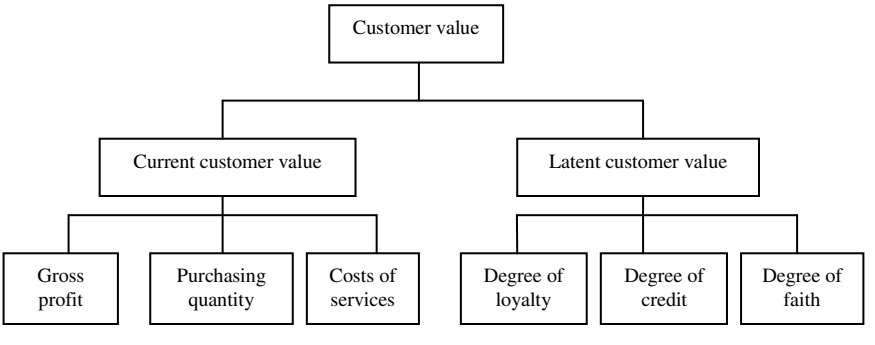



**Fig. 1.** The evaluating index system of customer value

According to the evaluating index system of customer value, we can build customer evaluating characteristic set:

$$D = \{\text{Gross profit}, \text{Purchasing quantity}, \text{Costs of services}, \text{Degree of loyalty}, \text{Degree of credit}, \text{Degree of faith}\} \triangleq \{d_{C1}, d_{C2}, d_{C3}, d_{L1}, d_{L2}, d_{L3}\}$$

**Directions.** For different practical problems, the selected evaluating characteristic set may vary. For example, if you need to add or delete or replace other evaluating characteristics, the following model and integrated dependent function must do the appropriate adjustments.

In order to quantitative study of each CV evaluating characteristic meeting the enterprise requirements degree, we uses dependent functions in Extenics including elementary dependent function, simple dependent function and the discrete dependent function [2]. According to the actual data in the enterprise database, enterprise requirements and industry requirements, suppose the value of customer evaluating characteristics $d_{C1}, d_{C2}, d_{C3}, d_{L1}, d_{L2}, d_{L3}$ are $x_{C1}, x_{C2}, x_{C3}, x_{L1}, x_{L2}, x_{L3}$, build dependent function of customer evaluating characteristics $d_{C1}, d_{C2}, d_{C3}, d_{L1}, d_{L2}, d_{L3}$ are $k_{Cj}(x_{Cj})$ $(j=1,2,3)$ and $k_{Lj}(x_{Lj})$ $(j=1,2,3)$, and calculate their function values.

The above established dependent functions are not normalized, in order to eliminate the impact of different dimension we need to further calculate their standard dependent degrees. Suppose there are $n$ customers' data, the standard dependent degrees can be expressed as:

$$k_{Cji} = \frac{k_{Cj}(x_{Cji})}{\max\limits_{i\in\{1,2,\cdots,n\}} \left| k_{Cj}(x_{Cji}) \right|}, (i = 1,2,\cdots,n;\ j = 1,2,3) \tag{1}$$

$$k_{Lji} = \frac{k_{Lj}(x_{Lji})}{\max\limits_{i\in\{1,2,\cdots,n\}} \left| k_{Lj}(x_{Lji}) \right|}, (i = 1,2,\cdots,n;\ j = 1,2,3) \tag{2}$$

## 3 Selection of Integrated Dependent Functions on CV

Suppose in the customer database customer $u$ corresponding to information element $D$, to make the discussion more general, we assume that the current customer value (CCV) has $s$ characteristic $d_{C1}, d_{C2}, \cdots, d_{Cs}$, and the latent customer value (LCV) has $t$ characteristic $d_{L1}, d_{L2}, \cdots, d_{Lt}$. The integrated dependent function $K_{\mathrm{CV}}(D)$ on the CV consists of the integrated dependent functions $K_{\mathrm{CCV}}(D)$ of the CCV and the integrated dependent functions $K_{\mathrm{LCV}}(D)$ of the LCV. According to various practical problems, $K_{\mathrm{CV}}(D)$、$K_{\mathrm{CCV}}(D)$ and $K_{\mathrm{LCV}}(D)$ use different method to establish. These methods will be described in another paper.

In this paper, we select the following integrated dependent functions:

$$K_{\mathbf{CV}}(D) = K_{\mathrm{CCV}}(D) \wedge K_{\mathrm{LCV}}(D) \tag{3}$$

$$K_{\mathrm{CCV}}(D) = \sum_{j=1}^{s} \lambda_{Cj} k_{Cj}, \quad K_{\mathrm{LCV}}(D) = \bigvee_{j=1}^{t} k_{Lj} \tag{4}$$

In formula (3), the operator $\wedge$ means "or" :

$$k_1 \wedge k_2 = \min\{k_1, k_2\}$$

In formula (4), the operator $\vee$ means "and":

$$k_1 \vee k_2 = \max\{k_1, k_2\}$$

In formula (4), $\lambda_{Cj}$ means the corresponding weights of each evaluation characteristics $k_{Cj}$ and $\sum_{j=1}^{s} \lambda_{Cj} = 1$.

If the implementation of the transformation $\varphi$ causes some characteristic values of CV in customer information-element $D$ occurring conductive transformation $T_\varphi$, i.e., $T_\varphi D = D'$, after the transformation $T_\varphi$, calculate dependent degree (Assuming there is no conductive action between them) and the standard dependent degree, the integrated dependent functions $K_{\mathrm{CV}}(D')$, $K_{\mathrm{CCV}}(D')$ and $K_{\mathrm{LCV}}(D')$ of information-element $T_\varphi D$ are calculated from the corresponding integrated dependent functions. After the transformation, the integrated dependent functions are expressed as:

$$K_{\mathbf{CV}}(D') = K_{\mathrm{CCV}}(D') \wedge K_{\mathrm{LCV}}(D') \tag{5}$$

$$K_{\mathrm{CCV}}(D') = \sum_{j=1}^{s} \lambda_{Cj} k'_{Cj}, \quad K_{\mathrm{LCV}}(D') = \bigvee_{j=1}^{t} k'_{Lj} \tag{6}$$

# 4 Customers Extension Classification on CV

## 4.1 Extension Classification Method

Extension classification is a kind of classification based on the extension transformation, including extension classification based on the direct transformation and extension classification based on the conductive transformation. Since the transformations of CV are generally conductive transformations [11], here we only consider extension classification based on CV conductive transformation. Extension classification methods based on the direct transformation refer to reference [12].

Assumed the domain and dependent functions are unchanged, extension set of CV information-element can be expressed as:

$$\tilde{E}(T) = \left\{(D_i, Y, Y') \middle| D_i \in \{D\}, Y = K(D_i), Y' = K(D'_i)\right\} \tag{7}$$

According to the definition of extension set [2], after the implementation of transformation, the CV information-element set $\{D\}$ in the database according to the integrated dependent function is divided into five domains. Back to the original customer information-element set, the customers are divided into five classifications respectively:

Positive qualitative customers : meet $K_{\mathrm{CV}}(D_i) \le 0$ and $K_{\mathrm{CV}}(D'_i) > 0$;

Negative qualitative customers : meet $K_{\mathrm{CV}}(D_i) \ge 0$ and $K_{\mathrm{CV}}(D'_i) < 0$;

Positive quantitative customers：meet $K_{CV}(D_i)>0$ and $K_{CV}(D_i')>0$;

Negative quantitative customers：meet $K_{CV}(D_i)<0$ and $K_{CV}(D_i')<0$;

Extension boundary customers：meet $K_{CV}(D_i')=0$.

This is a general extension classification. Analysis of specific problems requires specific consideration.

### 4.2 Customer Static Classification Criteria and the Extension Classification Criteria

Suppose the integrated dependent function on CV is expressed as:

$$K_{\mathbf{CV}}(D_i)=K_{\mathbf{CCV}}(D_i)\wedge K_{\mathbf{LCV}}(D_i)=(\sum_{j=1}^{s}\lambda_{Cj}k_{Cji})\wedge(\bigvee_{j=1}^{t}k_{Lji}), \tag{8}$$

According to CV theory, the static classification criteria include:

（1） If $K_{CV}(D_i)>0$， and $K_{CCV}(D_i)>0$ and $K_{LCV}(D_i)>0$, it belongs to the company's high value customers;

（2） If $K_{CV}(D_i)<0$， and $K_{CCV}(D_i)\geq 0$ and $K_{LCV}(D_i)<0$, it belongs to the company's second value customers；

（3） If $K_{CV}(D_i)<0$， and $K_{CCV}(D_i)<0$ and $K_{LCV}(D_i)\geq 0$, it belongs to the company's latent value customers；

（4） If $K_{CV}(D_i)<0$， and $K_{CCV}(D_i)<0$ and $K_{LCV}(D_i)<0$, it belongs to the company's low value customers.

（5） If $K_{CV}(D_i)=0$, it belongs to the company's zero boundary customers.

According to the integrated dependent function and the above classification criteria, we can get two kinds of the integrated dependent degree of each customer and static classification. So we can obtain the following knowledge before the transformation: which customers are high value customers, which customers are second value customers, which customers are latent value customers, which customers are low value customers. And we can calculate the percentage of sample customers based on sum of customers.

Suppose $\varphi$ is active transformation, and $T_{\varphi}$ is the conductive transformation based on CV. Under this transformation, both the CCV and LCV will change. Their values of integrated dependent functions will change accordingly to result in the change of customer classifications.

Based on extension classification method, after the implementation of transformation, information-element set in the database according to the integrated dependent functions ' values of the CCV and LCV are divided into four categories:

（1） Positive qualitative customers：meet $K_{CV}(D_i)\leq 0$ and $K_{CV}(D_i')>0$ ;

（2） Negative qualitative customers：meet $K_{CV}(D_i)\geq 0$ and $K_{CV}(D_i')<0$ ;

（3） Positive quantitative customers：meet $K_{CV}(D_i)>0$ and $K_{CV}(D_i')>0$;

(4) Negative quantitative customers : meet $K_{\mathrm{CV}}(D_i) < 0$ and $K_{\mathrm{CV}}(D_i') < 0$;

(5) Extension boundary customers : meet $K_{\mathrm{CV}}(D_i') = 0$.

According to their different combinations, we can determine the customer's comprehensive classification: positive quantitative high value customers, negative quantitative high value customers, positive qualitative high value customers, the negative qualitative high value customers; positive quantitative latent value customers, negative quantitative latent value customers, positive qualitative latent value customers, the negative qualitative latent value customers; positive quantitative second value customers, negative quantitative second value customers, positive qualitative second value customers, the negative qualitative second value customers; positive quantitative low value customers, negative quantitative low value customers, positive qualitative low value customers, the negative qualitative low value customers; and the extension boundary customers. After calculating their support and confidence respectively, we can find the effect of this transformation to determine which customer is valid on the transformation, which customer is invalid and determine the effect size, etc..Therefore, we can make the customers segmentation.

The positive qualitative change customers as example, its support and confidence are expressed as:

$$\ell = (\text{support, confidence}),$$

In which,

$$\text{support} = \frac{\text{Customers' number in negative domain}}{\text{Customers' sum}}$$

$$\text{confidence} = \frac{\text{Customers' number taking place positive qualitative change}}{\text{Customers' number in negative domain}}$$

The calculation method of other types of support and confidence refer to reference [9].

## 5 Case Analysis

The milk company in a supermarket wants to survey the CV of its customers and impact situation of marketing activities. Randomly 100 customers are selected. Since the company's services cost to each customer have no significant difference, for simplicity, only select four evaluating characteristics: {average monthly gross profit, the average monthly purchases, loyalty, credit} to examine the CV of each customer.

### 5.1 Selection of Evaluating Characteristics Based on Business Conditions and Professional Theory and Establishing Dependent Functions of the Evaluating Characteristics

According to historical data, the company's average gross profit per customer per month is from 10 Yuan to 100 Yuan; Average purchase quantity per month is from 20 boxes to 200 boxes; Maximum of customer loyalty is 5, and minimum is 1; Maximum of customer credit is 5, minimum is 1.

According to the company's situation, in terms of the CCV, the business requirement range of the average gross profit per month is <40,100> (RMB), and optimum is 80 Yuan; Customers that average purchase per month is greater than or equal to 30 boxes meet their business requirements; In terms of the LCV, customers that loyalty or credit is greater than or equal to 3 meet their business requirements. Based on the above information, we can use method in reference [2] to establish the dependent functions of $d_{C1}, d_{C2}, d_{L1}$ and $d_{L2}$ as following:

$$k_{C1}(x_{C1}) = \frac{\rho(x_{C1}, 80, <40,100>)}{D(x_{C1}, <40,100>, <10,100>)} = \frac{\rho(x_{C1}, 80, <40,100>)}{\rho(x_{C1}, <10,100>) - \rho(x_{C1}, <40,100>)}, \tag{9}$$

$$k_{C2}(x_{C2}) = \frac{x_{C2} - 30}{200 - 20} = \frac{1}{180}(x_{C2} - 30) \tag{10}$$

$$k_{L1}(x_{L1}) = \begin{cases} 1, & x_{L1} = 5 \\ 0.5, & x_{L1} = 4 \\ 0, & x_{L1} = 3 \\ -0.5, & x_{L1} = 2 \\ -1, & x_{L1} = 1 \end{cases}, \quad k_{L2}(x_{L2}) = \begin{cases} 1, & x_{L2} = 5 \\ 0.5, & x_{L2} = 4 \\ 0, & x_{L2} = 3 \\ -0.5, & x_{L2} = 2 \\ -1, & x_{L2} = 1 \end{cases} \tag{11}$$

The values of evaluating characteristics can evaluate based on the dependent functions above mentioned.

### 5.2 Establishing Integrated Dependent Functions

According to CV theory and business actual situation, the company's integrated dependent functions of CV, integrated dependent function of the CCV and integrated dependent function of the LCV are expressed as:

$$K_{\mathbf{CV}}(D) = K_{\mathrm{CCV}}(D) \wedge K_{\mathrm{LCV}}(D) \tag{12}$$

$$K_{\mathrm{CCV}}(D) = 0.6k_{C1}(x_{C1}) + 0.4k_{C2}(x_{C2}), \quad K_{\mathrm{LCV}}(D) = k_{L1}(x_{L1}) \vee k_{L2}(x_{L2}) \tag{13}$$

Use these integrated dependent functions, we can get classification of sample customers in the company. All of these show which is company's value customer, which is the company's second value customers, latent value customers, low value customers, and so on. Due to space limitations, we omit to mention.

### 5.3 Selecting the Active Transformation and Obtain the Integrated Dependent Functions after the Transformation

After implementation of the company's marketing activities, we could also calculate the integrated dependent functions' values of CV using the above the dependent functions in order to examine the effects of the integrated dependent degrees based on the transformation, and to determine what kind of customers occur quantitative change on the CV, what kind of customers occur qualitative change, so as to provide the basis for company's marketing activities in the future.

In this case, if the company adopts "ten percent discount to buy a box of milk" activity, after one month we can get the related average data of CV. Using the above dependent functions we can get the integrated dependent degrees of CV after the activity to determine the occurrence of quantitative and qualitative customers.

$$K_{\mathrm{CV}}(D') = K_{\mathrm{CCV}}(D') \wedge K_{\mathrm{LCV}}(D') \tag{14}$$

$$K_{\mathrm{CCV}}(D) = 0.6k_{C1}(x_{C1}) + 0.4k_{C2}(x_{C2}), \quad K_{\mathrm{LCV}}(D) = k_{L1}(x_{L1}) \vee k_{L2}(x_{L2}). \tag{15}$$

### 5.4 Acquiring the Customer Static Classification and Extension Classification According to the Integrated Dependent Functions Values of CV before and after the Transformation.

Comprehensive application of the above integrated dependent functions of CV, according to integrated dependent functions' values before and after the transformation, we can get customer segmentation, gain the customer's static classification and extension classification which can provide the reference of marketing activities and responding to market changes.

(1) Before the company implements the marketing activities , the static classification on CV is:

High value customers' set:

$$E_{+} = \{D_i \,|\, D_i \in U, (K_{\mathrm{CV}}(D_i) > 0) \wedge (K_{\mathrm{CCV}}(D_i) > 0) \wedge (K_{\mathrm{LCV}}(D_i) > 0)\}$$

Second value customers' set:

$$E_{1-} = \{D_i \,|\, D_i \in U, (K_{\mathrm{CV}}(D_i) < 0) \wedge (K_{\mathrm{CCV}}(D_i) \geq 0) \wedge (K_{\mathrm{LCV}}(D_i) < 0)\}$$

Latent value customers' set:

$$E_{2-} = \{D_i \,|\, D_i \in U, (K_{\mathrm{CV}}(D_i) < 0) \wedge (K_{\mathrm{CCV}}(D_i) < 0) \wedge (K_{\mathrm{LCV}}(D_i) \geq 0)\}$$

Low value customers' set:

$$E_{3-} = \{D_i \,|\, D_i \in U, (K_{\mathrm{CV}}(D_i) < 0) \wedge (K_{\mathrm{CCV}}(D_i) < 0) \wedge (K_{\mathrm{LCV}}(D_i) < 0)\}$$

Zero boundary customers' set:

$$E_0 = \{D_i \,|\, D_i \in U, K_{\mathrm{CV}}(D_i) = 0\}$$

According to the specific data in the case we can obtain the following knowledge: which customers are the high value customers, which customers are second value customers, which customers are the latent value customers, which customers are low value customers.

According to the above classification criteria in this case, the high value customers account for 26%, the latent value customers account for 21%, second value customers account for 16%, and the low value customers account for 37%.

In terms of general data mining, this step may be the end of mining. But for the extension data mining, the company must focus to consider the effect after the marketing activities, that is, to consider which customers are quantitative change, which customers are qualitative change, which customers have no effect after the implementation of the transformation. Acquiring this knowledge can provide decision support for the company's marketing activities in the future.

(2) After the company implements the marketing activities $T$, the extension classification on CV is：

Positive qualitative customers' set from second value to high value：

$$\dot{E}_{1+}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)\le 0)\wedge(K_{\mathrm{CCV}}(D_i)\ge 0)\wedge(K_{\mathrm{LCV}}(D_i)<0) \wedge(K_{\mathrm{CV}}(D_i')>0)\wedge(K_{\mathrm{CCV}}(D_i')\ge 0)\wedge(K_{\mathrm{LCV}}(D_i')>0)\}$$

Positive qualitative customers' set from latent value to high value：

$$\dot{E}_{2+}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)\le 0)\wedge(K_{\mathrm{CCV}}(D_i)<0)\wedge(K_{\mathrm{LCV}}(D_i)\ge 0) \wedge(K_{\mathrm{CV}}(D_i')>0)\wedge(K_{\mathrm{CCV}}(D_i')>0)\wedge(K_{\mathrm{LCV}}(D_i')\ge 0)\}$$

Positive qualitative customers' set from low value to high value：

$$\dot{E}_{3+}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)\le 0)\wedge(K_{\mathrm{CCV}}(D_i)\le 0)\wedge(K_{\mathrm{LCV}}(D_i)\le 0) \wedge(K_{\mathrm{CV}}(D_i')>0)\wedge(K_{\mathrm{CCV}}(D_i')>0)\wedge(K_{\mathrm{LCV}}(D_i')>0)\}$$

Negative qualitative customers' set from high value to second value：

$$\dot{E}_{1-}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)>0)\wedge(K_{\mathrm{CCV}}(D_i)>0)\wedge(K_{\mathrm{LCV}}(D_i)>0) \wedge(K_{\mathrm{CV}}(D_i')<0)\wedge(K_{\mathrm{CCV}}(D_i')\ge 0)\wedge(K_{\mathrm{LCV}}(D_i')<0)\}$$

Negative qualitative customers' set from high value to latent value：

$$\dot{E}_{2-}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)>0)\wedge(K_{\mathrm{CCV}}(D_i)>0)\wedge(K_{\mathrm{LCV}}(D_i)>0) \wedge(K_{\mathrm{CV}}(D_i')<0)\wedge(K_{\mathrm{CCV}}(D_i')<0)\wedge(K_{\mathrm{LCV}}(D_i')\ge 0)\}$$

Negative qualitative customers' set from high value to low value：

$$\dot{E}_{3-}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)>0)\wedge(K_{\mathrm{CCV}}(D_i)>0)\wedge(K_{\mathrm{LCV}}(D_i)>0) \wedge(K_{\mathrm{CV}}(D_i')<0)\wedge(K_{\mathrm{CCV}}(D_i')<0)\wedge(K_{\mathrm{LCV}}(D_i')<0)\}$$

Positive quantitative high value customers' set：

$$E_{+}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CCV}}(D_i)>0)\wedge(K_{\mathrm{LCV}}(D_i)>0) \wedge(K_{\mathrm{CCV}}(D_i')>0)\wedge(K_{\mathrm{LCV}}(D_i')>0)\}$$

Negative quantitative second value customers' set：

$$E_{1-}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)<0)\wedge(K_{\mathrm{CCV}}(D_i)\ge 0)\wedge(K_{\mathrm{LCV}}(D_i)<0) \wedge(K_{\mathrm{CV}}(D_i')<0)\wedge(K_{\mathrm{CCV}}(D_i')\ge 0)\wedge(K_{\mathrm{LCV}}(D_i')<0)\}$$

Negative quantitative latent value customers' set：

$$E_{2-}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)<0)\wedge(K_{\mathrm{CCV}}(D_i)<0)\wedge(K_{\mathrm{LCV}}(D_i)\ge 0) \wedge(K_{\mathrm{CV}}(D_i')<0)\wedge(K_{\mathrm{CCV}}(D_i')<0)\wedge(K_{\mathrm{LCV}}(D_i')\ge 0)\}$$

Negative quantitative low value customers' set：

$$E_{3-}(T)=\{D_i \mid D_i \in U,(K_{\mathrm{CV}}(D_i)<0)\wedge(K_{\mathrm{CCV}}(D_i)<0)\wedge(K_{\mathrm{LCV}}(D_i)<0) \wedge(K_{\mathrm{CV}}(D_i')<0)\wedge(K_{\mathrm{CCV}}(D_i')<0)\wedge(K_{\mathrm{LCV}}(D_i')<0)\}$$

Extension boundary customers' set： $\dot{E}_0(T)=\{D_i \mid D_i \in U, K_{\mathrm{CV}}(D_i')=0\}$

The results from this example can find:

1) The extension classification results based on CCV show that only 5% of customers belong to positive qualitative change, no negative qualitative customers, other customers belong to positive quantitative or negative quantitative. These show that the marketing activities have limited effect on CCV, but have some positive effects and no negative effects;

2) The extension classification results based on LCV show there are both positive quantitative and negative quantitative customers, but also there are positive qualitative customers. But the support of qualitative change customers is only 6%. All of results indicate that this marketing activity has a positive role on LCV.

3) The comprehensive classification results based on CV show that the proportion of the high value customers is from 26% to 31%, that there is the qualitative change customers in which 2% come from the second value customers and 4% come from the latent value customers, no from low value customers. There are latent value customers and the second value customers coming from low value customers. In addition, by the data back to the original customer database, we can also find that customers of which age range, occupation and gender belong to qualitative or quantitative change, etc. Due to space limitations, we omit to mention.

Such knowledge obtained from the database has important reference value to develop and take the next step for the company's marketing strategy.

## 6 Conclusion

After the extension classification method is applied to the customers' classification management study, we can master the changing effect of CV and obtain the variable customers' classification knowledge based on all kind of evaluating systems of CV. These can provide the knowledge that different customers have different reaction on marketing activities in order to implement targeted management and gain higher profits. The study has a very important role for enterprises' customer relationship management.

**Acknowledgments.** This paper is supported by the National Natural Science Foundation of China (Grant no. 70671031), the Guangdong Natural Science Foundation (Grant no.10151009001000044).

## References

1. Wang, H.: Customer value analysis based on rough set data mining technology. Economic Science Press, Beijing (2006)
2. Yang, C., Cai, W.: Extension Engineering. Science Press, Beijing (2007)
3. Yang, C., Cai, W.: Recent progress in extension data mining. Mathematics in Practice and Theory 39(4), 134–141 (2009)
4. Yang, C.: Extension Classification Method and Its Application Based on Extensible Set. In: Proceedings of 2007 International Conference on Wavelet Analysis and Pattern Recognition, Beijing, pp. 819–824 (2007)
5. Liu, Y.-Z., Wu, H.: A Summarization of Customer Segmentation Methods. Journal of Industrial Engineerin Engineering Managemen 20(1), 53–57 (2006)
6. Zhang, H.G., Lv, S., Li, W.: Application of Telecom Customer Segmentation Base on Improved C-means Algorithm. Computer Simulation 28(6), 185–188 (2011)
7. Lian, Q., Su, Y.: Investigating the Consumers Segmentation Based on SOM and PSO Algorithm. East China Economic Management 25(1), 118–121 (2011)

8. Qi, J., Su, H.: The evaluation, modeling and decision-making of customer value. Publishing House, BUPT, Beijing (2005)
9. Wang, L., Sheng, Z.: A Summarized Account of the Study on Customer Value Creation. Journal of Guangxi Normal University 41(4), 26–30 (2005)
10. Wen, C., Yong, S.: Extenics: Its significance in science and prospects in application. Journal of Harbin Institute of Technology 38(7), 1079–1086 (2006)
11. Yang, C.Y., Li, W.H.: Research on Customer Value Based on Extension Data Mining. In: Shi, Y., Wang, S., Peng, Y., Li, J., Zeng, Y. (eds.) MCDM 2009. CCIS, vol. 35, pp. 125–132. Springer, Heidelberg (2009)
12. Yang, C.-Y., Li, X.-M., Chen, W.-W., Cai, W.: Extension Data Mining Methods and Computer Implement. Guangdong Higher Education Publishing House, Guangdong (2010)

# An Incremental Mining Algorithm for Association Rules Based on Minimal Perfect Hashing and Pruning

Chuang-Kai Chiou[1] and Judy C.R. Tseng[2]

[1] College of Engineering, Chung Hua University, Hsinchu, 300, Taiwan, ROC
d09524003@chu.edu.tw
[2] Dept. of Computer Science and Information Engineering, Chung Hua University, Hsinchu, 300, Taiwan, ROC
judycrt@chu.edu.tw

**Abstract.** In the literatures, hash-based association rule mining algorithms are more efficient than Apriori-based algorithms, since they employ hash functions to generate candidate itemsets efficiently. However, when the dataset is updated, the whole hash table needs to be reconstructed. In this paper, we propose an incremental mining algorithm based on minimal perfect hashing. In our algorithm, each candidate itemset is hashed into a hash table, and their minimum support value can be verified directly by a hash function for latter mining process. Even though new items are added, the structure of the proposed hash does not need to be reconstructed. Therefore, experimental results show that the proposed algorithm is more efficient than other hash-based association rule mining algorithms, and is also more efficient than other Apriori-based incremental mining algorithms for association rules, when the database is dynamically updated.

**Keywords:** data mining, association rule, incremental mining.

## 1 Introduction

Association rules mining is an important data mining issue. It represents the relationships among items in a given database. The most well-known method for mining association rules is the Apriori algorithm [1]. Many proposed association rule mining algorithms are also Apriori-based [2-4]. Some researchers have tried to find efficient methods to improve Apriori-based algorithms. For instance, FP-Tree [5] and CAT tree algorithms [6] employ special tree structures for mining the frequent itemsets. On the other hand, DHP [2] and MPIP [3] algorithms employ hash structures to reduce the database access times. They are suitable for dealing with the candidates of 2-itemsets ($C_2$), which is the most time-consuming step in association rules mining. Consequently, hash-based association rule mining algorithms are more efficient than Apriori-based algorithms.

Besides, the traditional mining methods focus on mining in a static database (that means the items are seldom changed or updated). In most practical cases, the items in the database are added or updated frequently. Therefore, incremental mining

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 106–113, 2012.

techniques become essential when apply association rule mining in practice. Several incremental mining techniques have been proposed [7-10]. For example, FUP [7], an Apriori-based algorithm, stores the previous counts of large itemsets and examines the newly added transitions with these counts. And then a small number of new candidates were generated. The overall counts of candidates were obtained by scanning the original database. Although FUP dealt with the incremental dataset, the mining efficiency is still poor.

In this paper, we not only employ minimal perfect hashing structure [11] to improve the hash-based mining algorithm but also employ incremental mining technique for realistic practice. Hence, IMPHP (Incremental Minimum Perfect Hashing and Pruning) algorithm is proposed. Two advantages are obtained: 1). each candidate itemset will be hashed into a hash table without collisions and their minimum support vale can be verified directly by a hash function for latter process. 2.) When new items are added, the arrangement of proposed hash structure need not to be re-constructed. We only need to scan the updated parts and add new items into the end of the original hash table. Hence, the efficiency can be improved significantly.

## 2 Relative Work

For evaluating the improvement of hash-based mining algorithm, two hash-based mining algorithms, Direct Hashing & Pruning (DHP) algorithm and Multi-Phase Indexing and Pruning (MPIP) algorithm, will be compared. The details of the algorithms are described in the flowing subsections.

### 2.1 Direct Hashing and Pruning (DHP)

For dealing with the low performance of Apriori-based algorithm, Park et al. proposed DHP (Direct Hashing and Pruning) algorithm [2]. DHP employ hash functions to generate candidate itemsets efficiently, and DHP also employs effective pruning techniques to reduce the size of database. The potential-less itemsets will be filtered out in early stage of candidate generation, and the scanning of database will be avoided. Since DHP does not scan over the database all the time, the performance is also enhanced.

DHP is particularly powerful for finding the frequent itemsets in early stage. It finds 1-itemsets and makes a hash table for $C_2$, and then determines $L_2$ based on the hash table generated in previous stage. However, there is no guarantee that collisions can be avoided. If we use small number of buckets to hash the frequent itemsets, a heavy collision will be occurred. At this time, only few candidate itemsets will be filtered out and the performance might be worse than Apriori algorithm. Enlarging the number of buckets can solve the problem, but the requirement of large memory space will also reduce the performance.

### 2.2 Multi-Phase Indexing and Pruning (MPIP)

Tseng et al. proposed MPIP algorithm to improve DHP algorithm [3]. MPIP employed a minimum perfect hashing function to instead of the hashing function in

DHP. If a hashing function can determine a unique address for each itemset and hash all items without space wasted (That means the length of space is equal to the length of all items), it is said a minimum perfect hashing function [11]. The hashing process is shown as Fig. 1.

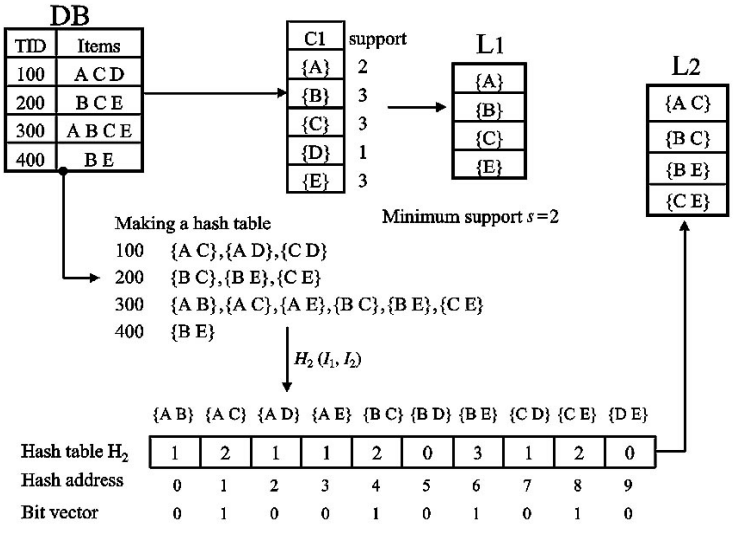



**Fig. 1.** Hashing process in MPIP algorithm

One of the contributions of MPIP is to improve the hashing collision problem in DHP algorithm. In MPIP algorithm, a unique address will be assigned to each itemset. It also promotes the accuracy of the hash table. Each entry in the hash table is used for determining the support value of corresponding itemset. Under such structure, the repeated scanning of database can be avoided. Besides, the Bit Vector in the hash table can filter out the candidate itemset and directly indicate the large itemsets. Hence, once we construct the hashing table, the large itemsets are also obtained.

However, DHP and MPIP cannot deal with the updating transactions where new items are included. In such situation, more collisions may be occurred in DHP. In MPIP, re-constructing the hash table is needed to include the new items. In order to simplify the hashing process and the hash table reconstructing problems, a new algorithm is proposed. In our proposed algorithm, hashing process does not need to adjust. For the updating transaction, we just need to scan the updating parts instead of scanning whole database.

## 3 Incremental Minimum Perfect Hashing and Pruning (IMPHP)

In this paper, an incremental association rule mining algorithm, IMPHP (Incremental Minimum Perfect Hashing and Pruning) is proposed. In this algorithm, hashing

address can be determined by a minimum perfect hashing function and mining with incremental transaction and item is also supported. In this section, we will analyze the regularity in 2-itemsets first and then extend it to the case of 3-item and *k*-itemsets. Finally, the minimal perfect hashing function is obtained.

### 3.1 Minimal Perfect Hashing Function

The notation used in our minimum perfect hashing function is defined as following. Let a k-itemset represent as $\{i_{j1}, i_{j1},\ldots, i_{jk}\}$ where k is an integer. $j_1, j_2,\ldots, j_k$ is the serial number of items and $j_1 < j_2 < j_3 <\ldots< j_k$. Let hash address of k-itemset represent as $F_n(j_1, j_2,\ldots, j_k)$ where n is total number of all items. For example, if $j_1 = 1$, $j_2 = 3$ and $j_3 = 5$ then the hash address of itemset $\{ i_1, i_3, i_5\}$ is represented as $F_n(j_1, j_2 j_3) = F_n(1, 3, 5)$. The minimum perfect hashing function of 2-itemset is list as following:

$$F_n(j_1, j_2) = \begin{cases} 1 & for\ j_1 = 1, j_2 = 2 \\ C_2^{j_2-1} + j_1 & otherwise \end{cases} \quad (1)$$

Secondly, we consider the case of 3-itemset. We can find that the same regularity still be persevered. The hashing result is shown as Fig. 2. All the itemset generated in $n = k$-1 will be included in $n = k$ and their hashing address will be the same.

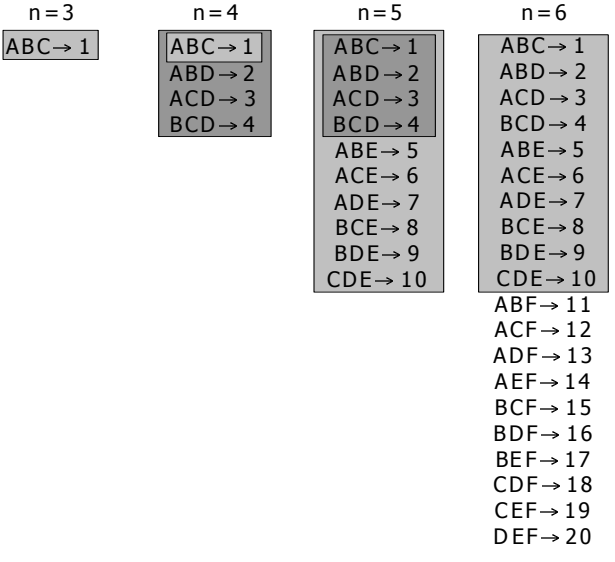



**Fig. 2.** The column-major ordering of 3-itemsets

Therefore we focus on the rest part within $n = k$. The third element in each 3-itemset will be the new added item and the first two elements will be the combinations in previous stage. When we extend $n$ to $k$, the regularity is also available. As $n=k$, the number of 3-itemsets is $C_3^n$. The minimum perfect hashing function of 3-itemset is list as following:

$$F_n(j_1, j_2, j_3) = \begin{cases} 1 & for\ j_1 = 1, j_2 = 2, j_3 = 3 \\ C_3^{j_3-1} + C_2^{j_2-1} + j_1 & otherwise \end{cases} \quad (2)$$

Finally, by repeating these processes, the minimum perfect hashing function of *k*-itemset can be derived and the formula is list as following:

$$F_n(j_1, j_2, ..., j_k) = \begin{cases} 1 & for\ j_1 = 1, j_2 = 2, ..., j_k = k \\ \sum_{m=2}^{k} C_m^{j_m-1} + j_1 & otherwise \end{cases} \quad (3)$$

According to formula (3), an unique address without collision can be determined. For example, let $k$ = 3, $n$ = 6 and $j_1$=A=1, $j_2$=C=3, $j_3$=D=4. The hashing address of $P(A, C, D)$ can be determined by formula (3):

$$F_6(1,3,4) = \sum_{m=2}^{3} C_m^{j_m-1} + j_1 = C_3^{j_3-1} + C_2^{j_2-1} + j_1 = C_3^{4-1} + C_2^{3-1} + 1 = 3$$

### 3.2 IMPHP Algorithm

The flowchart of IMPHP algorithm we proposed is shown as Fig. 3.

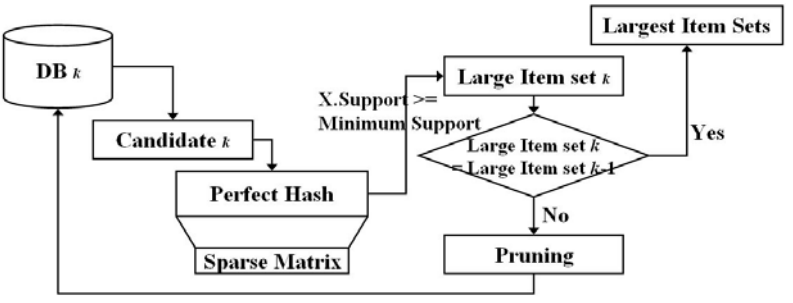



**Fig. 3.** The flowchart of IMPHP algorithm

The detail description is list below:

**Step 1:** Scan a transaction database and find out the 2-itemset combination form of each transaction. For example, $T$= {A, B, C} and the 2-itemset combination form will be {{A, B} {A, C}, {B, C}}. Hashing address of each itemset will be calculated by formula (1) and the support counts will be also increased.

**Step 2:** Repeat Step 1 until all the transactions are hashed into hash table. After selecting the items whose count is large than the minimum support, large 2-itemset ($L_2$) is obtained.

**Step 3:** Prune all transactions whose score is less than the minimum support. For example, assume $L_2$={{A, C}, {A, D}}, $T$ = {{A, B}, {B, C}, {A, C}} and minimum support is 2. Item $A$ in set {A, B} gets 1 point, item $C$ in set {B, C} gets 1 point, and item A and C in set {A, C} get 1 point respectively.

The scores of T will be {A(2), B(0), C(2)}. Then item B will be pruned and item A and C will be preserved. After the pruning process, a pruned transaction database $D_{k+1}$ is obtained for the further mining process. Such operation can reduce the database scanning space significantly.

**Step 4:** Repeat Step 1 to Step 3 and increase the itemset level (3-item, 4-item and etc.) until no new frequent itemset is found.

## 4 Experimental Results

To evaluate the performance of IMPHP, we perform several experiments. The equipment we used is a PC with Intel Core 2 Duo 2.0 GHz processor. Memory space is 1GB. In our experiment, the testing data is generated by the data generator from IBM Almaden Research Center [12] which is widely used in many researches. The meaning of parameters is shown as follows:

*N* : Number of data items
*D*: Number of transactions in the database
*T*: Average length of transactions
*I*: Average size of the potentially frequent itemsets

In this experiment, the original dataset we used is N50.D1K.T5.I2 and we fix the average length of transactions to 5, and average size of the potentially frequent itemsets to 2 (T5.I2). We extend the size of dataset with N50 and D1K for each time and record the performance of the three algorithms-- FUP, MPIP and IMPHP. The result is shown as Fig. 4. In average case, IMPHP improve about 46% than MPIP and about 76% than FUP. When the dataset grows to +N300.D1K, IMPHP improve about 92% than FUP.

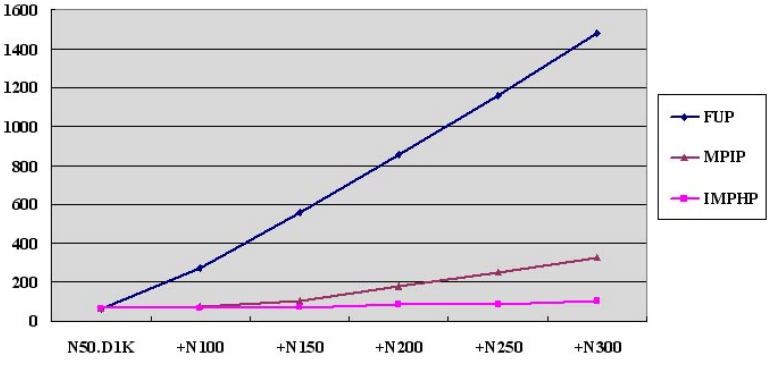



**Fig. 4.** Performance comparisons between FUP, MPIP and IMPHP with T5.I2

When we extend the scale of the previous experiment with 10 times, we can obtain the result as Fig. 5. When the dataset grows to +N500.D10K, IMPHP improve about 84% than FUP.

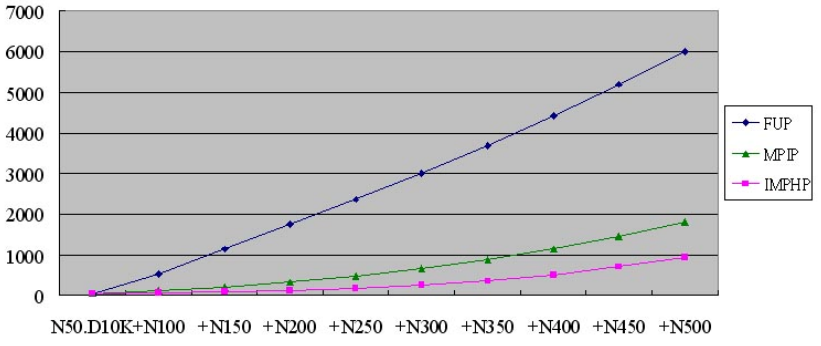



**Fig. 5.** Performance comparisons between FUP, MPIP and IMPHP with T5.I2

## 5 Conclusions

In this paper, IMPHP (Incremental Minimum Perfect Hashing and Pruning) algorithm is proposed for incremental association rule mining. Minimum perfect hashing function is employed in IMPHP to avoid collisions and improved the mining efficiency. Incremental dataset is also supported in this algorithm. That means that IMPHP is also worked when the number and category of dataset are changed.

In additional, the arrangement of proposed hash structure does not need to be reconstructed again when new items are added. All new added items will be placed at the end of the original table. Hence, the efficiency can be improved significantly when it applies for incremental association rule mining.

In order to examine the performance of IMPHP, incremental datasets were used. The number of itemsets is increasing and new items are also added in transaction database. Experimental results show that the proposed algorithm outperforms others. Especially in large variation cases, IMPHP improve about 92%~84% than FUP.

**Acknowledgments.** This study is supported in part by the National Science Council of the Republic of China under contract numbers NSC 99-2511-S-216 -004 -MY3, NSC 98-2631-S-024 -001, and NSC 98-2218-E-216 -003, and by Chung Hua University under contract numbers CHU- NSC 98-2218-E-216 -003 and CHU-NSC 99-2511-S-216 -004 -MY3. The authors would like to thank Mr. En-Sheng Chou for his helps in our experiments.

## References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between the Sets of Items in Large Database. In: Proc. ACM SIGMOD, pp. 207–216. ACM Press, Washington, DC (1993)
2. Park, J.S., Chen, M.S., Yu, P.S.: Using a Hash-Based Method with Transaction Trimming and Database Scan Reduction for Mining Association Rules. IEEE Transactions on Knowledge and Data Engineering 9(5), 813–825 (1997)

3. Tseng, J.C.R., Hwang, G.J., Tsai, W.F.: A Minimal Perfect Hashing Scheme to Mining Association Rules from Frequently Updated Data. Journal of the Chinese Institute of Engineers 29(3), 391–401 (2006)
4. Chiou, C.K., Tseng, J.C.R.: A Scalable Association Rules Mining Algorithm Based on Sorting, Indexing and Triming. In: 2007 International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2257–2262. IEEE Computer Society (2007)
5. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. In: ACM SIGMOD, pp. 1–12. ACM Press, Washington, DC (2000)
6. Cheung, W., Zaiane, R.: Incremental Mining of Frequent Patterns without Candidate Generation or Support Constraint. In: Proc. of the Seventh International Database Engineering and Applications Symposium. IEEE Computer Society (2003)
7. Cheung, D.W., Han, J., Ng, V.T., Wong, C.Y.: Maintenance of Discovered Association Rules in Large Database: An Incremental Updating Technique. In: Proceedings of International Conference on Data Engineering, pp. 106–114. IEEE Computer Society (1996)
8. Pradeepini, G., Jyothi, S.: Tree-based incremental association rule mining without candidate itemset generation. In: Trendz in Information Sciences & Computing (TISC), pp. 78–81. IEEE Computer Society (2010)
9. Dai, B.R., Lin, P.Y.: iTM: An Efficient Algorithm for Frequent Pattern Mining in the Incremental Database without Rescanning. In: Chien, B.-C., Hong, T.-P., Chen, S.-M., Ali, M. (eds.) IEA/AIE 2009. LNCS, vol. 5579, pp. 757–766. Springer, Heidelberg (2009)
10. Shan, S., Wang, X., Sui, M.: Mining Association Rules: A Continuous Incremental Updating Technique. In: International Conference on Web Information Systems and Mining, pp. 62–66. IEEE Computer Society (2010)
11. Chang, C.C.: The Study of an Ordered Minimal Perfect Hashing Scheme. Communications of the ACM 27(4), 384–387 (1984)
12. Agrawal, R.: IBM Research, Almaden Research Center, Computer Science, `http://www.almaden.ibm.com/software/quest/index.shtml`

# LDA-Based Topic Modeling in Labeling Blog Posts with Wikipedia Entries

Daisuke Yokomoto[1], Kensaku Makita[1], Hiroko Suzuki[1], Daichi Koike[1], Takehito Utsuro[1], Yasuhide Kawada[2], and Tomohiro Fukuhara[3]

[1] University of Tsukuba, Tsukuba, 305-8573, Japan
[2] Navix Co., Ltd., Tokyo, 141-0031, Japan
[3] National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

**Abstract.** Given a search query, most existing search engines simply return a ranked list of search results. However, it is often the case that those search result documents consist of a mixture of documents that are closely related to various contents. In order to address the issue of quickly overviewing the distribution of contents, this paper proposes a framework of labeling blog posts with Wikipedia entries through LDA (latent Dirichlet allocation) based topic modeling. More specifically, this paper applies an LDA-based document model to the task of labelling blog posts with Wikipedia entries. One of the most important advantages of this LDA-based document model is that the collected Wikipedia entries and their LDA parameters heavily depend on the distribution of keywords across all the search result of blog posts. This tendency actually contributes to quickly overviewing the search result of blog posts through the LDA-based topic distribution. In the evaluation of the paper, we also show that the LDA-based document retrieval scheme outperforms our previous approach.

**Keywords:** Blog, Wikipedia, Topic Model, LDA, Topic Analysis.

## 1 Introduction

As blog services and blog tools are becoming more and more popular, people have been able to express one's own interests as well as opinions on the Web. Search engines are then used for accessing various information that can be found in the blogosphere, where, given a search query, a ranked list of blog posts is provided as a search result. However, such a search result in the form of a ranked list is usually not helpful for a user to quickly identify blog posts that satisfy his/her information need. This is especially true when, given a search query, the search result is a mixture of blog posts that focus on various contents.

In such a situation, the framework of *faceted search* [1], which has been well studied in the information retrieval community, can be a solution. Based on this observation, [2] proposed a framework of categorizing Japanese blog posts according to their contents, where, given a search query, those blog posts are

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 114–124, 2012.

collected from the Japanese blogosphere. In this framework, the content of each blog post is regarded as a facet of a query keyword, and a facet is automatically assigned to each blog post. This procedure of assigning a facet to a blog post is realized by utilizing Wikipedia entries as a knowledge source and each Wikipedia entry title is considered as a facet label. In its Japanese version, about 770,000 entries are included (checked at November, 2011). Given a query keyword, Wikipedia entries are collected from Wikipedia as candidates of its facets. Then, for each Wikipedia entry, its body text is analyzed as fundamental knowledge source, and terms strongly related to the entry are extracted. Those terms are then used for labelling a blog post with this entry.

One drawback of the framework of [2] is that, when labelling a blog post with Wikipedia entries, it does not exploit the distribution of keywords across all the search result of blog posts, but labels a blog post with Wikipedia entries independently of other blog posts in the whole search result. This is especially disadvantageous considering the purpose of the research, i.e., to quickly overview the search result of blog posts in terms of their contents. In order to overcome this disadvantage of [2], this paper proposes a framework of labeling blog posts with Wikipedia entries through LDA (latent Dirichlet allocation [3]) based topic modeling. More specifically, this paper applies an LDA-based document model [4] to the task of labelling blog posts with Wikipedia entries. Figure 1 illustrates the overall framework of labelling blog posts with Wikipedia entries based on an LDA-based document model (with a query keyword *"global warming"*). In this framework, first, given a query keyword, blog posts that are related to the query keyword are collected. Then, from the collected blog posts, Wikipedia entry titles are extracted. Next, the LDA parameters are estimated with the extracted Wikipedia entries, where the topics that are closely related to the collected blog posts are generated. Those LDA parameters for generated topics are also incorporated into the LDA-based document retrieval scheme [4]. When applying an LDA-based document model to the task of labelling blog posts with Wikipedia entries, we regard each blog post as a query, and the collected Wikipedia entries as the document collection from which one or more documents are retrieved.

One of the most important advantages of this LDA-based document model is that the collected Wikipedia entries and their LDA parameters heavily depend on the distribution of keywords across all the search result of blog posts. This tendency actually contributes to quickly overviewing the search result of blog posts through the LDA-based topic distribution. In the evaluation of the paper, we also show that the LDA-based document retrieval scheme outperforms our previous approach of [2].

## 2 Related Works

In TREC 2009 blog track [5], faceted blog distillation task was studied, where three facets, namely, *opinionated/personal/in-depth* are introduced and participants are required to assign facets to blog feeds. [6] invented a multi-faceted

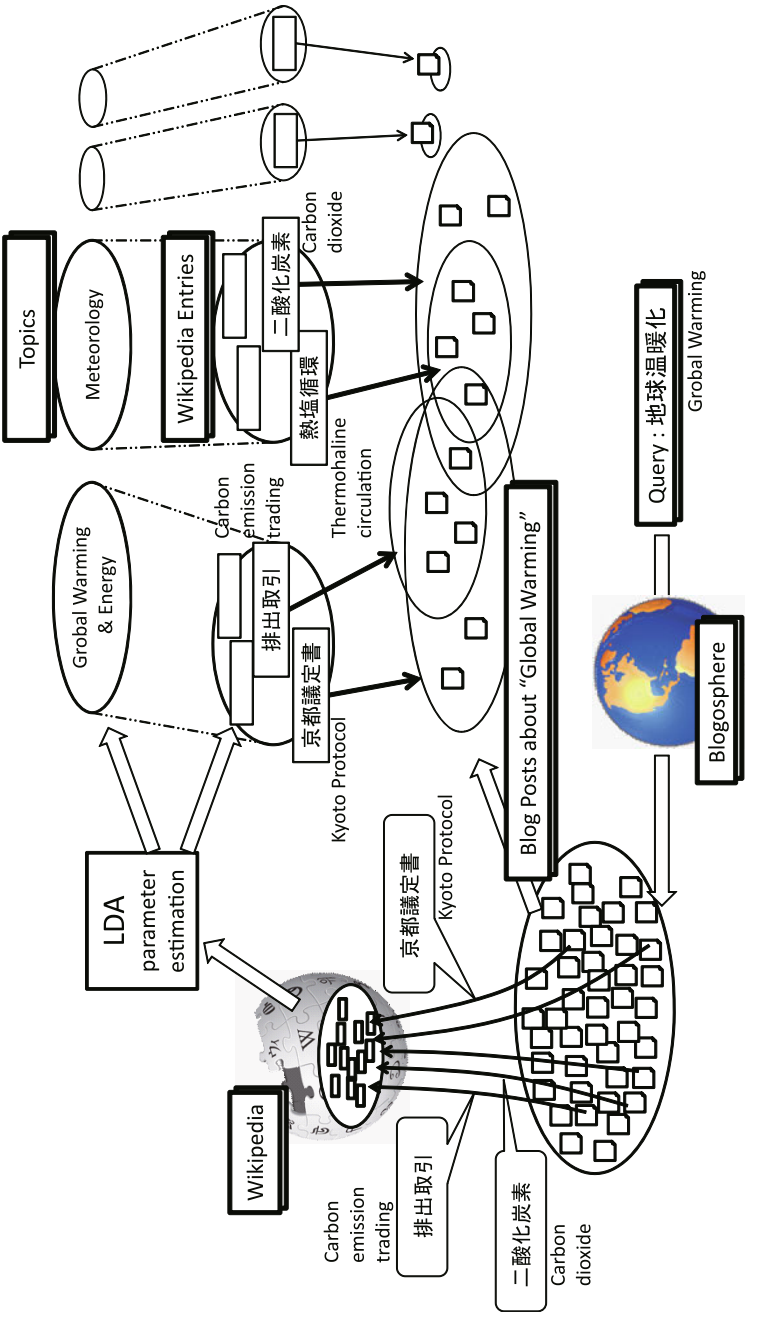



**Fig. 1.** Framework of Labeling Blog Posts with Wikipedia Entries through LDA-based Topic Modeling

blog search framework, where various facets are introduced in terms of topics, bloggers, links, and sentiments. [7] also proposed a framework of generating a faceted search interface for Wikipedia. Compared to those previous works [7,6,5], the proposed method is innovative in that it realizes a novel technique of automatically generating sub-topic oriented facets for blog posts collected from the blogosphere. Our work is related to [7] in that both techniques collect facet candidates from Wikipedia. In [7], it is also presented how to rank facet hierarchies, where the cost of navigation through Wikipedia facet hierarchies is modeled and is utilized in facet hierarchy ranking. However, compared to our technique, that of [7] modeled the cost of navigation only within the Wikipedia facet hierarchy, where the target of navigation is also Wikipedia articles. Our technique is different from that of [7] in that, in our technique, given the set of blog posts collected with an initial query as the target of navigation, facet candidates that are not frequently observed in the collected blog posts are removed. As a future work, it is also possible to introduce the formalization of the navigation cost of [7] into our task.

Another related works include techniques of clustering and summarizing search results [8], or those of clustering search results and assigning cluster labels [9,10,11]. Compared with those techniques on search results clustering, the proposed technique is advantageous in that it is capable of assigning facets to even quite a small number of blog posts, simply because it utilizes Wikipedia as a knowledge source for assigning facets to blog posts.

The technique presented in this paper is also related to previous works on assigning Wikipedia concepts to document clusters (e.g., [12]) and those combining Wikipedia concepts as well as important terms extracted from the cluster content in cluster labeling (e.g., [13]). However, those previous works mostly concentrate on clustering standard document sets such as those of newspaper articles with broad range of topics. In this paper, on the other hand, we focus on extracting facets from Wikipedia, given the set of blog posts collected with an initial query, where the collected blog posts cover much narrower range of contents. Compared with the tasks studied in those previous works, the task of facet categorization of related blog posts studied in this paper is relatively difficult to tackle.

With respect to related works of the LDA-based document retrieval scheme [4], [4] argued that the LDA-based document retrieval scheme overcomes the disadvantages of the pLSI model [14] as well as the cluster-based language model [15].

## 3 Retrieving Blog Posts with a Query Keyword

Given a query keyword $t_0$, this section describes how to retrieve blog posts with $t_0$ as a search query. With this procedure, we intend to collect candidates of blog posts that are closely related to $t_0$.

First, we use an existing Web search engine API, which returns a ranked list of blog posts, given a topic keyword. For the evaluation in section 6, during the period from July to September, 2010, we used the Japanese search engine

"Yahoo! Japan" API[1] for Japanese. Blog hosts are limited to major 8 hosts[2] for Japanese. For each query, this search engine API returns a ranked list of at most 1,000 blog posts. A list of blog feeds is then generated from the returned ranked list of blog posts by simply removing duplicated feeds. From the retrieved blog feeds, we next collect blog posts that include the initial topic keyword $t_0$ in the body text into the set $BP(t_0)$ of blog posts as candidates for those closely related to $t_0$.

## 4 Collecting Wikipedia Entry Titles from Blog Posts

From the set $BP(t_0)$ of collected blog posts, we collect Wikipedia entry titles and construct the set $\mathbb{E}(BP(t_0))$ of Wikipedia entries. Here, we require that the entry $e$ to be collected satisfy that the document frequency $\mathrm{df}(BP(t_0), t(E))$ of the title $t(E)$ of $e$ over the set of collected blog posts $BP(t_0)$ is more than or equal to 10[3].

$$\mathbb{E}(BP(t_0)) = \left\{ E \;\middle|\; \mathrm{df}(\ BP(t_0),\ t(E)\ )\ \geq 10 \right\}$$

## 5 LDA-Based Document Model

### 5.1 Latent Dirichlet Allocation

LDA [3] can be used to model and discover underlying topic structures of discrete data such as text. The formalization of LDA below follows the notation of quantities below:

- $M$: the total number of documents
- $K$: the number of topics
- $V$: vocabulary size
- $\boldsymbol{\alpha}, \boldsymbol{\beta}$: Dirichlet parameters
- $\boldsymbol{\vartheta}_m$: topic distribution for document $m$
- $\boldsymbol{\Theta} = \{\boldsymbol{\vartheta}_m\}_{m=1}^{M}$: a $M \times K$ matrix
- $\boldsymbol{\varphi}_k$: word distribution for topic $k$
- $\boldsymbol{\Phi} = \{\boldsymbol{\varphi_k}\}_{k=1}^{K}$: a $K \times V$ matrix
- $N_m$: the length of document $m$
- $z_{m,n}$: topic index of $n$-th word in document $m$
- $w_{m,n}$: a particular word for word placeholder $[m, n]$

Rough description of the generation process for LDA is as follows:

1. For each topic $k \in [1, K]$, pick a multinomial distribution $\boldsymbol{\varphi}_k$ from a Dirichlet distribution with parameter $\boldsymbol{\beta}$;

[1] http://www.yahoo.co.jp/ (in Japanese).
[2] fc2.com, yahoo.co.jp, yaplog.jp, ameblo.jp, goo.ne.jp, livedoor.jp, Seesaa. net, hatena.ne.jp
[3] We empirically chose this lower bound through preliminary evaluation.

2. For each document $\boldsymbol{w}_m = \{w_{m,n}\}_{n=1}^{N_m}$, pick a multinomial distribution $\boldsymbol{\vartheta}_m$ from a Dirichlet distribution with parameter $\boldsymbol{\alpha}$,
3. For each word placeholder $[m, n]$, pick a topic $z_{m,n}$ from the multinomial distribution $\boldsymbol{\vartheta}_m$,
4. Pick word $w_{m,n}$ for the word placeholder $[m, n]$ from the multinomial distribution $\boldsymbol{\varphi}_{z_{m,n}}$.

Based on the above description, we can write the joint distribution of all known and hidden variables given the Dirichlet parameters as follows:

$$P(\boldsymbol{w_m}, \boldsymbol{z_m}, \boldsymbol{\vartheta_m}, \Phi \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = P(\boldsymbol{\Phi} \mid \boldsymbol{\beta}) \prod_{n=1}^{N_m} P(w_{m,n} \mid \boldsymbol{\varphi_{z_{m,n}}}) P(z_{m,n} \mid \boldsymbol{\vartheta_m}) P(\boldsymbol{\vartheta_m} \mid \boldsymbol{\alpha})$$

And the likelihood of a document $\boldsymbol{w_m}$ is obtained by integrating over $\boldsymbol{\vartheta_m}$, $\boldsymbol{\Phi}$ and summing over $\boldsymbol{z_m}$ as follows:

$$P(\boldsymbol{w_m} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \int P(\boldsymbol{\vartheta_m} \mid \boldsymbol{\alpha}) P(\boldsymbol{\Phi} \mid \boldsymbol{\beta}) \cdot \prod_{n=1}^{N_m} P(w_{m,n} \mid \boldsymbol{\vartheta_m}, \boldsymbol{\Phi}) \, d\boldsymbol{\Phi} d\boldsymbol{\vartheta}_m$$

Finally, the likelihood of the whole data collection $\mathcal{W} = \{\boldsymbol{w}_m\}_{m=1}^{M}$ is product of the likelihoods of all documents:

$$P(\mathcal{W} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{m=1}^{M} P(\boldsymbol{w}_m \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$$

In the evaluation of section 6, we used GibbsLDA++ [16] for LDA parameter estimation, where for the number of topics $K = 50$, $\alpha = 50/K$, and $\beta = 0.1$.

### 5.2 LDA-Based Retrieval of Wikipedia Entries

In our LDA-based framework of retrieving Wikipedia entries given a blog post as a query, let us denote a blog post used as a query as $B \in BP(t_0)$. Also, we denote a Wikipedia entry to be retrieved and considered as a document model as $E$, where each Wikipedia entry $E$ is taken from the set $\mathbb{E}(BP(t_0))$ of Wikipedia entries collected from the set $BP(t_0)$ of blog posts as described in section 4.

The basic approach for using language models for IR is the query likelihood model where each document is scored by the likelihood of its model generating a query $B$,

$$P(B \mid E) = \prod_{w \in B} P(w \mid E)$$

where $E$ is a document model, $B$ is the query and $w$ is a query term in $B$[4]. $P(B \mid E)$ is the likelihood of the document model generating the query terms under

[4] More specifically, we require a query term $w$ in $B$ to be the title of a Wikipedia entry, and also require that the term frequency $freq(B, t(E))$ of the title $t(E)$ of $E$ within $B$ is more than or equal to 3.

the bog-of-words assumption that terms are independent given the documents. Following [4], we specify $P(w \mid E)$ by combining the LDA model $P_{lda}(w \mid E)$ with the traditional linear interpolation of the maximum likelihood estimate $P_{ML}(w \mid E)$ of word $w$ in the document $E$ and the maximum likelihood estimate $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ of word $w$ in the entire collection $\mathbb{E}(BP(t_0))$ as below:[5]

$$P(w \mid E) = \lambda\Big\{\mu P_{ML}(w \mid E) + (1-\mu)P_{ML}(w \mid \mathbb{E}(BP(t_0)))\Big\} + (1-\lambda)P_{lda}(w \mid E)$$

where the LDA model $P_{lda}(w \mid E)$ is given as:

$$P_{lda}(w \mid E) = \sum_{k=1}^{K} P(w \mid \boldsymbol{\varphi_k})P(z_k \mid E)$$

## 6 Evaluation

### 6.1 Evaluation of Wikipedia Entry Ranking

For evaluation, we pick up the 9 query keywords listed in Table 1. For each query keyword $t_0$, Table 1 also shows the number of collected blog posts $|BP(t_0)|$. For each query keyword, we select two or three blog posts that have a high similarity value with the query keyword as the title of a Wikipedia entry, where the similarity is measured as introduced in [2]. In total, we select 20 blog posts for evaluation.

**Table 1.** 9 Query Keywords and the Size of Blog Posts for Evaluation

| Query Keyword $t_0$ | # of Blog Posts $\lvert BP(t_0)\rvert$ |
|---|---|
| smoking | 9,926 |
| organ transplantation | 1,502 |
| global warming | 8,687 |
| medical error | 1,914 |
| population aging | 2,205 |
| Toyota Prius | 5,099 |
| smartphone | 10,039 |
| alcoholism | 2,060 |
| restructuring | 5,007 |

For each of the 20 blog posts for evaluation, we compare the following three methods for ranking Wikipedia entries in terms of labeling the blog post:

[5] When combining $P_{ML}(w \mid E)$ and $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$, we compared the Dirichlet smoothing as employed in [4] with the linear linear interpolation shown here, where the best performance with $\lambda = \mu = 0.7$ outperformed that of the Dirichlet smoothing employed in [4].

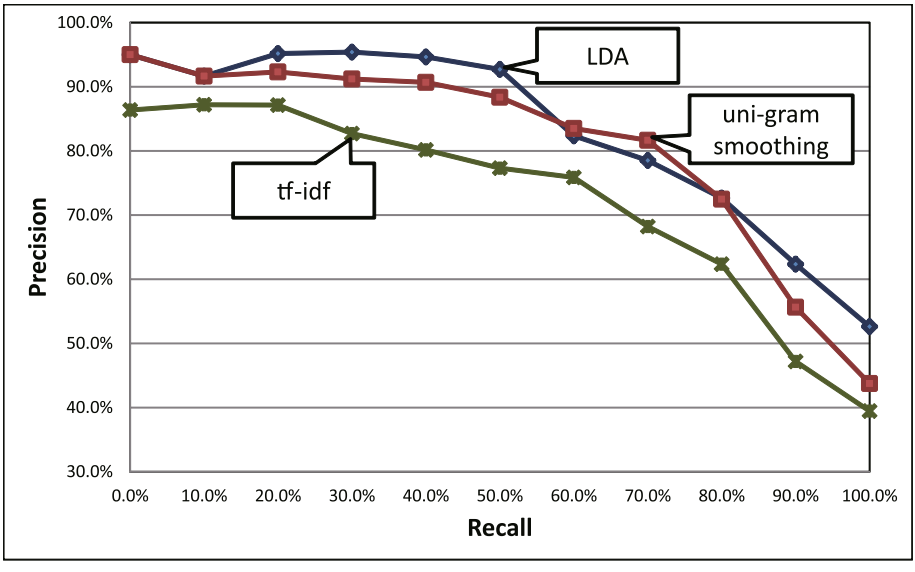



**Fig. 2.** Evaluation Results of Wikipedia Entry Ranking

1. the LDA-based retrieval model presented in section 5.2 (denoted as "LDA").
2. the same as above, except that we specify $P(w \mid E)$ as the linear interpolation of the maximum likelihood estimate $P_{ML}(w \mid E)$ of word $w$ in the Wikipedia entry $E$ and the maximum likelihood estimate $P_{ML}(w \mid \mathbb{E}(BP(t_0)))$ of word $w$ in the entire collection $\mathbb{E}(BP(t_0))$:

$$P(w \mid E) \;=\; \mu P_{ML}(w \mid E) \;+\; (1-\mu) P_{ML}(w \mid \mathbb{E}(BP(t_0)))$$

   where $\mu = 0.7$ (denoted as "uni-gram smoothing").
3. Wikipedia entries are ranked according to the similarity value with the blog post as a query, where, as introduced in [2], the similarity is measured as the inner product of the term frequency vector of the blog post as a query and the inverse document frequency (within the whole Japanese version of Wikipedia) vector of a Wikipedia entry (denoted as "tf-idf").

For those 20 blog posts for evaluation, 60.1 Wikipedia entries are ranked on the average, out of which 12.1 entries are manually judged as correct labels for the query blog post. For the three methods "LDA", "uni-gram smoothing", and "tf-idf", Figure 2 plots the recall-precision curves that are averaged over the 20 blog posts for evaluation[6]. As can be clearly seen from this result, "LDA" constantly outperforms "tf-idf", where their differences are statistically significant at more than half of the 11 recall points at a level of 0.03 in terms of micro average. Considering the result that "LDA" and "uni-gram smoothing" are mostly

[6] For each query blog post, precision at every 11 recall point 0%, 10%, ..., 90%, 100% is measured and averaged over the 20 blog posts.

**Table 2.** An Example of LDA-based Topic Modeling of Wikipedia Entries

| Topic ID | | *T1*: global warming and energy | *T2*: meteorology | *T3*: astronomy | *T4*: politics |
|---|---|---|---|---|---|
| blog post *B1* | labeled Wikipedia entries | — | carbon dioxide, thermohaline circulation | sun, sunspot | — |
| | summary | — | Carbon dioxide does not cause global warming. | Not global warming but global cooling due to sunspot. | — |
| blog post *B2* | labeled Wikipedia entries | fossil fuel, alternative energy | — | — | — |
| | summary | Raise the price of fossil fuel to stop global warming. | — | — | — |
| blog post *B3* | labeled Wikipedia entries | fuel and light expenses, Photovoltaic power generation | — | — | Democratic Party of Japan, manifesto |
| | summary | Fuel and light expenses will greatly increase by introducing Photovoltaic power generation. | — | — | It is not beneficial for Japan to keep the manifesto of the Democratic Party of Japan. |

comparative in their performance, the difference between "LDA" and "tf-idf" is mainly due to whether or not considering the distribution of keywords across all the search result of blog posts.

### 6.2 An Example of LDA-Based Topic Modeling of Wikipedia Entries

For the query keyword *"global warming"*, Table 2 shows topics $z_k$ which have the highest probability value $P(z_k \mid E)$ for at least one Wikipedia entry $E$ which is manually judged as correct. As shown in Table 2, in this case, we have four topics, out of which the one with the topic ID = *"T1: global warming and energy"*: is

most closely related to *"global warming"*, while other topics have rather little relation to *"global warming"*. As can be clearly seen in Table 2, those topics greatly contribute to quickly understanding the contents of blog posts.

We examined three blog posts as the query, where each blog post has one or two corresponding topics as in Table 2. Those topics are then allocated with a Wikipedia entry $E$ which has the highest probability value $P(z_k \mid E)$, where the probability value should be sufficiently large. With this result, it becomes becomes much easier to quickly overview the distribution of topics over the query blog posts.

## 7 Conclusion

This paper proposed a framework of labeling blog posts with Wikipedia entries through LDA-based topic modeling. More specifically, this paper applied an LDA-based document model to the task of labelling blog posts with Wikipedia entries. One of the most important advantages of this LDA-based document model is that the collected Wikipedia entries and their LDA parameters heavily depend on the distribution of keywords across all the search result of blog posts. This tendency actually contributed to quickly overviewing the search result of blog posts through the LDA-based topic distribution. In the evaluation of the paper, we also showed that the LDA-based document retrieval scheme outperformed our previous approach.

## References

1. Tunkelang, D.: Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers (2009)
2. Yokomoto, D., Makita, K., Utsuro, T., Kawada, Y., Fukuhara, T.: Utilizing Wikipedia in categorizing topic related blogs into facets. In: Proc. 12th PACLING, #20 (2011)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Wei, X., Croft, W.B.: LDA-Based document models for ad-hoc retrieval. In: Proc. 29th SIGIR, pp. 178–185 (2006)
5. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC-2009 blog track. In: Proc. TREC 2009 (2009)
6. Fujimura, K., Toda, H., Inoue, T., Hiroshima, N., Kataoka, R., Sugizaki, M.: BLOGRANGER - a multi-faceted blog search engine. In: Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2006)
7. Li, C., Yan, N., Roy, S.B., Lisham, L., Das, G.: Facetedpedia: Dynamic generation of query-dependent faceted interfaces for Wikipedia. In: Proc. 19th WWW, pp. 651–660 (2010)
8. Harashima, J., Kurohashi, S.: Summarizing search results using PLSI. In: Proc. 2nd Workshop on NLPIX, pp. 12–20 (2010)
9. Toda, H., Kataoka, R., Oku, M.: Search result clustering using informatively named entities. International Journal of Human-Computer Interaction, 3–23 (2007)

10. de Winter, W., de Rijke, M.: Identifying facets in query-biased sets of blog posts. In: Proc. ICWSM, pp. 251–254 (2007)
11. Shibata, T., Bamba, Y., Shinzato, K., Kurohashi, S.: Web information organization using keyword distillation based clustering. In: Proc. WI-IAT, pp. 325–330 (2009)
12. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging Wikipedia semantics. In: Proc. 31st SIGIR, pp. 179–186 (2008)
13. Carmel, D., Roitman, H., Zwerdling, N.: Enhancing cluster labeling using Wikipedia. In: Proc. 32nd SIGIR, pp. 139–146 (2009)
14. Hoffman, T.: Probabilistic latent semantic indexing. In: Proc. 22nd SIGIR, pp. 50–57 (1999)
15. Liu, X., Croft, W.B.: Cluster-based retrieval using language models. In: Proc. 27th SIGIR, pp. 186–193 (2004)
16. Phan, X.H., Nguyen, C.T.: GibbsLDA++: A C/C++ implementation of latent Dirichlet allocation (LDA) (2007)

# The Correlation between Semantic Visual Similarity and Ontology-Based Concept Similarity in Effective Web Image Search

Clement. H.C. Leung and Yuanxi Li

Computer Science Department, Hong Kong Baptist University,
Kowloon Tong, Hong Kong
{clement,yxli}@comp.hkbu.edu.hk

**Abstract.** This paper compares the correlations between visual similarity of real-world images and different ontology-based concept similarity in order to find a novel measurement of the relationship between semantic concepts (objects, scenes) in visual domain besides low level feature extraction. For selected concept pairs, we compute their visual similarity and co-occurrence, which is represented by our Probability-based Visual Distance Model (PVDM). Rather than high computational cost of object recognition, by employing the ontology-based concept similarity into query expansion and filtering, the semantic image search and retrieval precision will be much higher. Furthermore, the latent topic will be mapped into images so that users are possible to retrieval the images with satisfying visual characteristic of the target concept.

**Keywords:** visual similarity, ontology, query expansion, image search.

## 1 Introduction

With huge number of images is uploaded on the Web, searching and sharing them in a semantic way is of great significance. Capturing the semantic relationship between concepts becomes a hot research topic recently, since it has wide application on natural language processing, object detection, and multimedia retrieval [1][2][3]. It is important to note that the semantic relationship among images is different from text documents. Besides the relationship of synonym (trousers-pants) and concept similarity (monkey-chimpanzee), it also includes relationships, which has something to do with daily life, such as meronymy (house-door) and concurrence (flower-garden). The concurrence denotes the two or more concepts may appear simultaneously in the senses of real world. These semantic relationships exist independently or together with other relationships in images on the Web. Mining and capturing the relationships of concepts will surely improve the understanding and sharing of Web images, even other multimedia.

The presence of particular objects in an image often implies the presence of other objects. If term $U \rightarrow V$, and if only $U$ is indexed, then searching for $V$ will not return the image in the result, even though $V$ is present in the image. The application of such

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 125–130, 2012.

inferences will allow the index elements $T_i$ of an image to be automatically expanded according to some probability which will be related to the underlying ontology of the application. There are two types of expansion:

(a) Aggregation hierarchical expansion

This relates to the aggregation hierarchy of sub-objects that constitute an object. Associated with each branch is a tree traversal probability$t_{ij}$(Fig. 1) which signifies the probability of occurrence of the branch index given the existence of the parent index. In general, the traversal probabilities of different object classes exhibit different characteristics, with $t_{ij}$>$t'_{mn}$ for $t_{ij}$ belonging to the concrete object class, and $t'_{mn}$ belonging to the abstract object class.

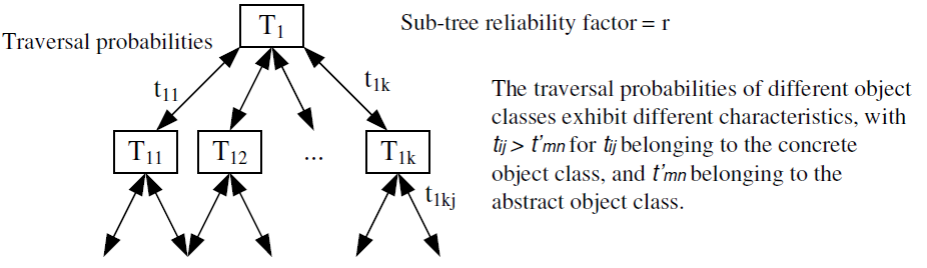



**Fig. 1.** A tree traversal probability $t_{ij}$ which signifies the probability of occurrence of the branch index given the existence of the parent index

(b) Co-occurrence expansion

This relates to the expectation that certain semantic objects tend to occur together. The relevant weighting is expressed as a conditional probability given the presence of other objects. An expansion to associate an image object $O_j$given the presence of object $O_i$ is taken to be indexable when

$$Prob[O_j|O_i] \geq h' \tag{1}$$

where$h'$ is a preset threshold value that depends on the tradeoff between precision and recall performance of the system. More generally, complex probabilistic rules taking the form

$$Prob[O_j|O_1, \ldots, O_n] \geq h' \tag{2}$$

will be applied. The ontology expansion tree is traversed bi-directionally in the course of the expansion. Top-down traversal will lead to an expansion factor > 1, while bottom-up traversal will have an expansion factor < 1 at each level of expansion [4].

The rest of the paper is organized as follows. Section 2 gives more details about the related work. Section 3 elaborates on the different concept distances and visual distance. Section 4 demonstrates comparative experimental results. Section 5 concludes the paper.

## 2 Related Work

Great efforts have been done by experts to create ontology based networks which map concept relationships between knowledge and words. There are mainly three types as follows.

### 2.1 Text-Oriented Concept Relationship

Typical projects are Cyc [5] and the WordNet [6], are trying to create a common knowledge-based network among real word concepts.

Cyc is an artificial intelligence project that attempts to assemble a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning [7].

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications [8].

Because this kind of ontology is not dynamic changing with the information on nowadays Internet, the size of knowledge database is limited and has low scalability.

### 2.2 Webpage-Oriented Concept Relationship

Besides Cyc and WordNet, the relationship between concepts is measured by some distances, such as Google Distance [12] and Wikipedia Distance [13]. This kind of concept ontology is dynamic changing with the human knowledge and information on the Web.

Furthermore, the visual based similarity measurements are employed to mine the semantic concepts or latent topics behind objects images, such as Flickr Distance [14] and LSCOM (Large Scale Concept Ontology for Multimedia) [15]. This kind of measurement provides the semantic correlation between concepts based on real world multimedia database.

## 3 Ontology-Based and Visual-Based Concept Similarity

In order to find relative more suitable concept similarity measurements for continuously changing Web images, as well as to employ an effective query expansion and refinement method in Web image search with concept similarity. We compared the following concept similarity measurement schemes with the real images semantic concept correlations.

### 3.1 Normalized Google Distance

Normalized Google distance (NGD) is proposed [16] to quantify the extent of the relationship between two concepts by their correlation in the search results from Google search engine when querying both concepts, with

$$NGD\ (x,y) = \frac{max\ \{log\ f(x), log(y)\} - log\ f(x,y)}{log\ N - min\ \{log\ f(x), log\ f(y)\}} \tag{3}$$

Where $f(x)$ and $f(y)$ are the numbers of the Web pages returned by Google search engine when typing $x$ and $y$ as the search term respectively, with $f(x,y)$ denotes the number of pages containing both $x$ and $y$.

### 3.2 Flickr Distance

Flickr distance (FD) [14] is another model for measuring the relationship between semantic concepts in visual domain. For each concept, a collection of images are obtained from Flickr, based on which the improved latent topic-based visual language model is built to capture the visual characteristic of this concept. The Flickr distance between concepts $c_1$ and $c_2$ can be measured by the square root of Jensen-Shannon divergence [17,18]between the corresponding visual language models as follows:

$$\mathrm{D}(C_1, C_2) = \sqrt{\frac{\sum_{i=1}^{K} \sum_{j=1}^{k} D_{JS}(P_{z_i c_1} | P_{z_j c_2})}{K^2}} \tag{4}$$

where

$$D_{JS}\left(P_{z_i c_1} \middle| P_{z_j c_2}\right) = \frac{1}{2} D_{KL}\left(P_{z_i c_1} \middle| M\right) + \frac{1}{2} D_{KL}\left(P_{z_j c_2} \middle| M\right) \tag{5}$$

$K$is the total number of latenttopics, which is determined by ment. $P_{z_i c_1}$ and $P_{z_j c_2}$ are the trigram distributions underlatent topic $z_i c_1$and $z_j c_2$respectively, with $M$ representing the mean of $P_{z_i c_1}$and $P_{z_j c_2}$.

### 3.3 Probability-Based Visual Distance Model (PVDM)

We propose a Probability-based Visual Distance Model (PVDM) to measure the visual similarity and correlation of images, and compare the result with the concept similarities generated by Normalized Google Distance and Flickr Distance.

Assume we are going to search images with queryA and query B dependently. A and B are a pair of related concept. The total number of retrieved images are $A$ and $B$, respectively; and the number of relevant images return from search engine are $A^+$and $B^+$. We define the visual similarity ($VS$) of concept A and B as:

$$VS(A,B) = \alpha(A' \middle| B') + (1-\alpha)(B' \middle| A') \tag{6}$$

where

$$A' = \frac{P(A^+ \cap B^+)}{P(B^+)} \tag{7}$$

$$B' = \frac{P(A^+ \cap B^+)}{P(A^+)} \tag{8}$$

The relationship is shown in Fig. 2 as follows:

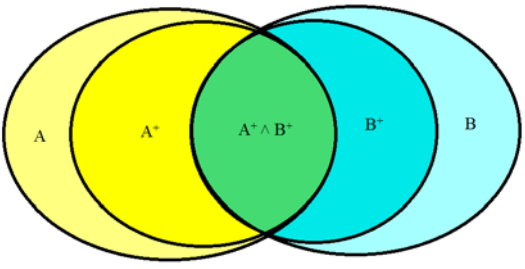



**Fig. 2.** The results distribution of image search for concetp A and B

## 4 Comparative Experiments and Results

We do the comparative experiments on 50+ concept pairs, mining 10,000+ Web images, and calculate the Normalized Google Distance, Flickr Distance as well as the Probability-based Visual Distance. A subset of the experimental results is shown in Fig. 3.

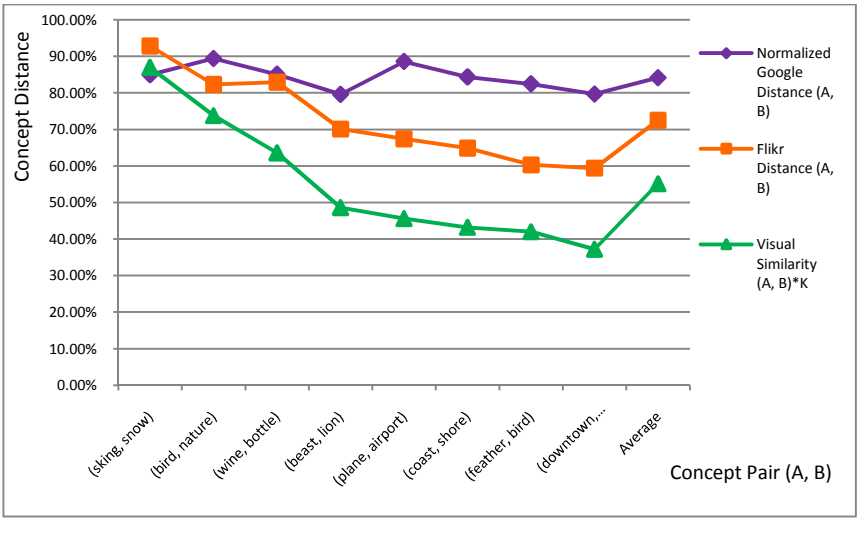



**Fig. 3.** A subset of experimental results

## 5 Conclusion

As seen in the comparative experimental results, the image concept similarity of Flickr Distance is more likely to go with the trends of the image Visual Concept Distance. Which is to say, though Normalized Google Distance is also a dynamic changing concept similarity measurement method, it does a good job for similarity measure

among concepts in textual web page, while it is difficult to capture the real world visual concept in image domain. Thus, to share and search images more effectively, visual similarity measurements which generated directly from multimedia domain, such as Flickr Distance and LSCOM will outperform the concept distance generated from text domain. More comparative experiments are certainly necessary to be done on larger scale database to further prove the conclusion.

## References


1. Datta, R., Dhiraj, J., Jia, L., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (2008)
2. Yu, J., Tian, Q.: Semantic subspace projection and its application in image retrieval. IEEE Transactions on Circuits and Systems for Video Technology (CSVT), 544–548 (2008)
3. Liu, H., Jiang, S., Huang, Q., Xu, C., Gao, W.: Region-based visual attention analysis with its application in image browsing on small displays. In: Proc. of the 15th International Conference on Multimedia (2007)
4. Wong, C.F.: Automatic Semantic Image Annotation and Retrieval. PhD Thesis, Hong Kong Baptist University (August 2010)
5. Lenat, D.B.: CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM 38(11), 33–38 (1995)
6. Miller, G.A., et al.: Wordnet, a lexical database for the english language. Cognition Science Lab, Princeton University (1995)
7. Wikipedia, `http://en.wikipedia.org/wiki/Cyc`
8. CYC Homepage, `http://www.cyc.com/`
9. OpenCyc, `http://www.opencyc.org/`
10. Wikipedia, `http://en.wikipedia.org/wiki/WordNet`
11. Princeton University "About WordNet." WordNet. Princeton University (2010), `http://wordnet.princeton.edu`
12. Cilibrasi, R., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19, 370 (2007)
13. Strube, M., Ponzetto, S.P.: WikiRelate! computing semantic relatedness using wikipedia. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence. AAAI Press (July 2006)
14. Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., Li, S.: Flickr distance. In: MM 2008: Proceedings of the 16th ACM International Conference on Multimedia, New York, NY, USA, pp. 31–40 (2008)
15. Smith, J.R., Chang, S.F.: Large-scale concept ontology for multimedia. IEEE Multimedia 13(3), 86–91 (2006)
16. Enser, P.G.B., Sandom, C.J., Lewis, P.H.: Surveying the Reality of Semantic Image Retrieval. In: Bres, S., Laurini, R. (eds.) VISUAL 2005. LNCS, vol. 3736, pp. 177–188. Springer, Heidelberg (2006)
17. Li, X., Chen, L., Zhang, L., Lin, F., Ma, W.: Image Annotation by Large-Scale Content-Based Image Retrieval. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 607–610 (2006)
18. Wikipedia, `http://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence`

# An Adaptive Design Pattern for Invocation of Synchronous and Asynchronous Web Services in Autonomic Computing Systems

Vishnuvardhan Mannava[1] and T. Ramesh[2]

[1] Department of Computer Science and Engineering,
K L University, Vaddeswaram, 522502, A.P, India
vishnu@kluniversity.in
[2] Department of Computer Science and Engineering,
National Institute of Technology, Warangal, 506004, A.P, India
rmesht@nitw.ac.in

**Abstract.** Asynchronous invocations are needed in the context of distributed object frameworks to prevent clients from blocking during remote invocations. Popular Web Service frameworks offer only synchronous invocations (over HTTP). When a client with asynchronous invocation of Web Service is not supported, client developers have to build asynchronous invocations on top of the synchronous invocation facility. But this is tedious, error-prone, and might result in different remote invocation styles used within the same application. Current autonomic computing application uses synchronous Web Service to manage their resources. In this paper we propose autonomic system by using Design Patterns for Web Service, which is amalgamation of fire and forget, chain of responsibility and Case based reasoning Design Patterns. The system will provide synchronous and asynchronous paradigm based on demand of client. By using our proposed Adaptive Design Pattern for invocation of Web Services previous Web Service systems will update their resources based on client request. Our proposed system satisfies the properties of autonomic system. The pattern is described using a java-like notation for the classes and interfaces. A simple UML class and Sequence diagrams are depicted.

**Keywords:** Web Services, Synchronous invocation, Asynchronous invocations, Web Service composition, Design Pattern, Autonomic system, Service Oriented Architecture (SOA), Web Services and Web Service Description Language (WSDL).

## 1 Introduction

Web Service is defined as an interface which implements the business logic through a set of operations that are accessible through standard Internet protocols. The conceptual Web Services architecture [1] is defined based upon the interactions

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 131–142, 2012.

between three roles: Service Provider, Service Registry and Service Requester. The requester search for suitable Web services in the registry which satisfy his functional and nonfunctional requirements. The requester's Service request sometimes includes multiple related functionalities to be satisfied by the Web Service. In many cases the Web Service has a limited functionality which is not sufficient to meet the requester's complex functional needs. To achieve complex business goals in real world applications, the execution of multiple Web Services should be orchestrated through Service composition.

Design Patterns can be classified using two orthogonal classification schemes. The first option is to classify the patterns according to their purpose: creational, structural, or behavioral. Creational patterns focus on object creation. Structural patterns focus on describing the composition of classes or objects. Behavioral patterns depict the method of interaction and distribute the responsibility of classes or objects. Thus far, we have only identified structural and behavioral adaptation Design Patterns [2].

The second option is to classify the patterns according to their adaptation functions: monitoring, decision-making, and reconfiguration. Monitoring patterns focus on probing components and distributing the information across a network to interested clients. Decision-making patterns focus on identifying when a reconfiguration is needed and selecting a reconfiguration plan that will yield the desired behavior. Reconfiguration patterns focus on safely adding, removing, or modifying components at run time to adapt a program. Thus far, we have identified several Design Patterns for each area.

Autonomic Computing is an initiative started by IBM in 2001 with an ultimate aim to develop computer systems capable of self-management, to overcome the rapidly growing complexity of computing systems management, and to reduce the barrier that complexity poses to further growth. So, the system, to be autonomic, must have the following properties:

- Self-configuring: Automatic configuration of components.
- Self-healing: Automatic discovery, and correction of faults.
- Self-optimizing: Automatic monitoring and control of resources to ensure the optimal functioning with respect to the defined requirements.
- Self-protecting: Proactive identification and protection from arbitrary attacks.

Our proposed system invokes Web Services with different paradigms; system will invoke Web Services either synchronously or asynchronously based on user request. Synchronous invocation means one process executing in a resource don't allocate this resource to other process those are waiting for execution. Asynchronous invocation means all processes are sharing same resource and there is no blocking concept in asynchronous invocation. This system uses Case Based Reasoning for decision making, Fire and Forger pattern used for invoking asynchronous Services, and uses SOA Web Service for invocation, Fig 1 will shows Autonomic System.

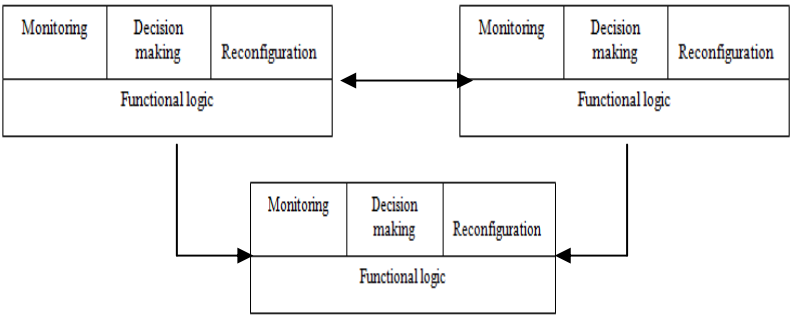



**Fig. 1.** Autonomic computing

## 2 Related work

In this section we summarize some known uses of the asynchrony patterns as related work.

There are various messaging protocols that are used to provide asynchrony for Web Services Uwe Zdun and Markus Voelter [1] paper discuss invocation of Web Services they propose system for asynchronous invoking of Web Services, V.S. Prasad Vasireddy and Vishnuvardhan Mannava [4] paper discuss management of Web Services using synchronous paradigm. Vishnuvardhan Mannava [3] and T. Ramesh paper discuss invocation of Web Services using Design Patterns. In H. Foster, S. Uchitel, J. Magee, J. Kramer [2] paper discuss Web Service composition paradigm based on client request. Andres j. Ramirez, Betty H.C [5] discuss various Design Patterns for designing autonomic system. We use Case Based Reasoning pattern in this paper. In contrast to our approach these messaging protocols do not provide a protocol-independent interface to client-side asynchrony and require developers to use the messaging communication paradigm. Yet these protocols provide a reliable transfer of messages, something that our approach does not deal with. Messaging protocols can be used in the lower layers of our framework.

The Web Services Invocation Framework (WSIF) (Apache Software Foundation, 2002) is a simple Java API for invoking Web Services with different protocols and frameworks, similar to the internal invocation API of Axis. Based on two papers we propose a system, called adaptive Design Pattern for invocation of synchronous and asynchronous Web Services. It provides synchronous or asynchronous Web Services based on requirement of client, this system is perfectly suitable for present systems. By using this system it will upgrade their Services as well as update their modules.

## 3 Objectives of an Asynchronous Invocation Framework in the Context of Web Services

There are a number of issues about Web Services because of the limitations in synchronous invocations. To avoid the work-around of hard coding asynchronous invocations in the client code, we provide an object-oriented framework [7] that can be reused as an extension for existing Web Service frameworks. The framework

design is based on a number of software Design Patterns. Let us summarize the goals of our asynchronous invocation framework:

- ***Better Performance of Client Applications:*** Asynchronous invocations typically lead to better performance of client applications, as idle times in waiting for a blocked invocation to return are avoided.

**- *Simple and Flexible Invocation Model:*** The invocation model must be simple to use by developers. Asynchronous invocations should not be more complicated to use than synchronous invocations. That is, the client developer should not have to deal with issues such as multi-threading, synchronization, or thread pooling. There are different kinds of invocations, including synchronous invocations and various ways to provide asynchronous invocations. All these kinds of invocation should be offered with an integrated invocation model that is easy to comprehend.

**- *Support for multiple Web Services Implementations and Protocols:*** The strength of Web Services is heterogeneity. Thus an asynchronous invocation framework should (potentially) work with different protocols (such as JMS or Secure HTTP) and implementations. An invocation framework that builds on top of an existing Web Service framework automatically integrates the different protocols provided by that Web Service framework.

## 4 Adaptive Design Pattern Template

To facilitate the organization, understanding, and application of the adaptation Design Patterns, this paper uses a template similar in style to that used by Ramirez et al. [2]. Likewise, the Implementation and Sample Code fields are too application-specific for the Design Patterns presented in this paper.

### 4.1 Pattern Name

Adaptive Design Pattern

### 4.2 Classification

Structural, Monitoring and Decision making

### 4.3 Intent

Designing a system for invoking Web Services synchronously or asynchronously based on user request using SOA Web service. We propose new Design Pattern which is an amalgamation of different Design Patterns, Design Pattern that are used for the system: Master-Slave and chain of responsibility.

### 4.4 Motivation

Main objective of Adaptive Design Pattern is to invoke Web Services either synchronously or asynchronously according to user requests. This Pattern invokes Web Service dynamically by using SOA Web Service technique.

### 4.5 Proposed Design Pattern Structure

UML class diagram for the Constrain based composition Design Pattern can be found in Figure 3.

Our proposed system will invoke Web Service either synchronously or asynchronously based on user request. We use three Design Patterns for designing our system they are Case based reasoning, Chain of Responsibility and Fire and Forget. In proposed system, client request all Service providers for WSDL file, all Service providers send WSDL files to client. Client searches for matching constrain in WSDL file, then client chooses a matched Service provider for invoking Web Service. Based on decision generated by case based reasoning client generate XML file. Proposed system uses Cased based reasoning Design Pattern for decision making. Based on the XML file constrains Service provider will invoke the Service. Proposed system invokes Web Service synchronously or asynchronously based on client request. Service repository stores all Web Services in it. The Composition of Services at the Service providers can be realized with the help of this proposed structure of composing the Web Services with SOA example, see Figure 2.

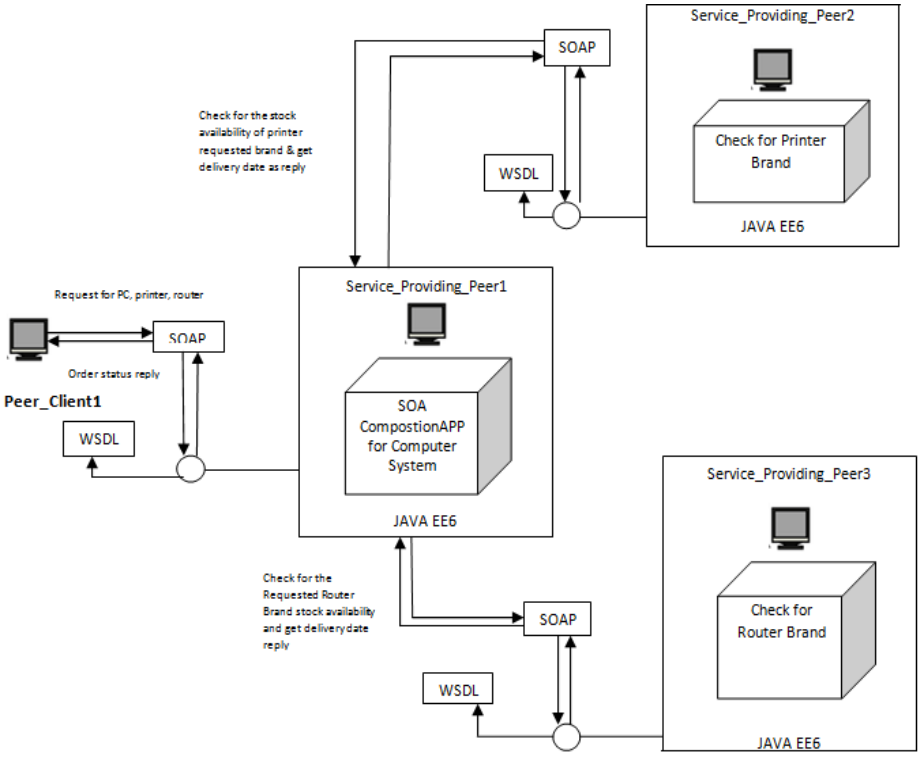



**Fig. 2.** Composition of the Services with Service Oriented Architecture using Web Services

### 4.6 Participants

**(a) Client**

Client class send WSDL file request to service providers, based on WSDL response client uses case based reasoning Design Pattern for finding appropriate plan for composing new Service. Based on decision client generate XML file for service provider.

**(b) Service Provider**

Service provider give response to client based on constrain of user. Service provider composes new Web Service based on XML file generated by client.

**(c) Decision**

This class represents a reconfiguration plan that will yield the desired behavior in the system.

**(d) Fixed Rules**

This class contains a collection of Rules that guide the Inference Engineering producing a Decision. The individual Rules stored within the Fixed Rule scan be changed at run time.

**(e) Learner**

This is an optional feature of the Case based Reasoning Design Pattern.

**(f) Log**

This class is responsible for recording which reconfiguration plans have been selected during execution. Each entry is of the form Trigger-Rule-Decision.

**(g) Rule**

A Rule evaluates to true if an incoming Trigger matches the Trigger contained in the Rule.

**(h) Service Repository**

Service repository will store Web Service in it, based on client request service provider will invoke form service repository, Composed Web Services also stored in service repository.

**(i) Asynchronous Fire**

Asynchronous fire class invoke Web Service asynchronously, class provide Service to all client without any blocking concept of service.

**(h) Forget**

Service is used by several client Asynchronous fire doesn't able to provide Service then forget class remove load of asynchronous pattern, Forger class remove waiting requests of client for Service invocation.

### 4.7 Consequences:

The Adaptive Design Pattern for Web Service invocation offers the following benefits:

**Centralized Administration:** The pattern consolidates one or more Services into a single administrative unit. This helps to simplify development by automatically performing common Service initialization and termination activities. In addition, it centralizes the administration of communication Services by imposing a uniform set of configuration management operations.

**Increased Modularity and Reuse:** The pattern improves the modularity and reusability of communication Services by decoupling the implementation of these Services from the configuration of the Services. In addition, all Services have a uniform interface by which they are configured, thereby encouraging reuse and simplifying development of subsequent Services.

### 4.8 Related Design Patterns:

**REMOTE OBJECT** is a distributed object (here: the Web Service), offered by a server application, that should be reached by the client remotely. Note that a REMOTE OBJECT describes the remote interface, not the actual implementation – thus the Service implementation can well comprise a set of procedures or be a wrapper to a legacy system. The pattern REQUESTER describes how to build up a remote invocation on client side and hand it over to the transport layer of the distributed object framework. Note that clients often do not access the REQUESTER implementation directly, but use a (generated) CLIENT PROXY instead. The CLIENT PROXY offers the interface of the REMOTE OBJECT in the client process and uses the REQUESTER internally to build up the remote invocation.

**POLL OBJECT:** There are situations, when an application needs to invoke an operation asynchronously, but still requires knowing the results of the invocation. The client does not necessarily need the results immediately to continue its execution, and it can decide for itself when to use the returned results. As a solution POLL OBJECTS receive the result of remote invocations on behalf of the client. The client subsequently uses the POLL OBJECT to query the result. It can either just query (poll), whether the result is available, or it can block on the POLL OBJECT until the result becomes available. As long as the result is not available on the POLL OBJECT, the client can continue asynchronously with other tasks.

### 4.9 Applicability

Use the Adaptive Design Pattern for Web Service invocation when:

- Web administrator will use this autonomic system for dynamic composition.
- An application or system can be simplified by being composed of multiple independently developed and dynamically configurable Services; or
- The management of multiple Services can be simplified or optimized by configuring them using a single administrative unit.

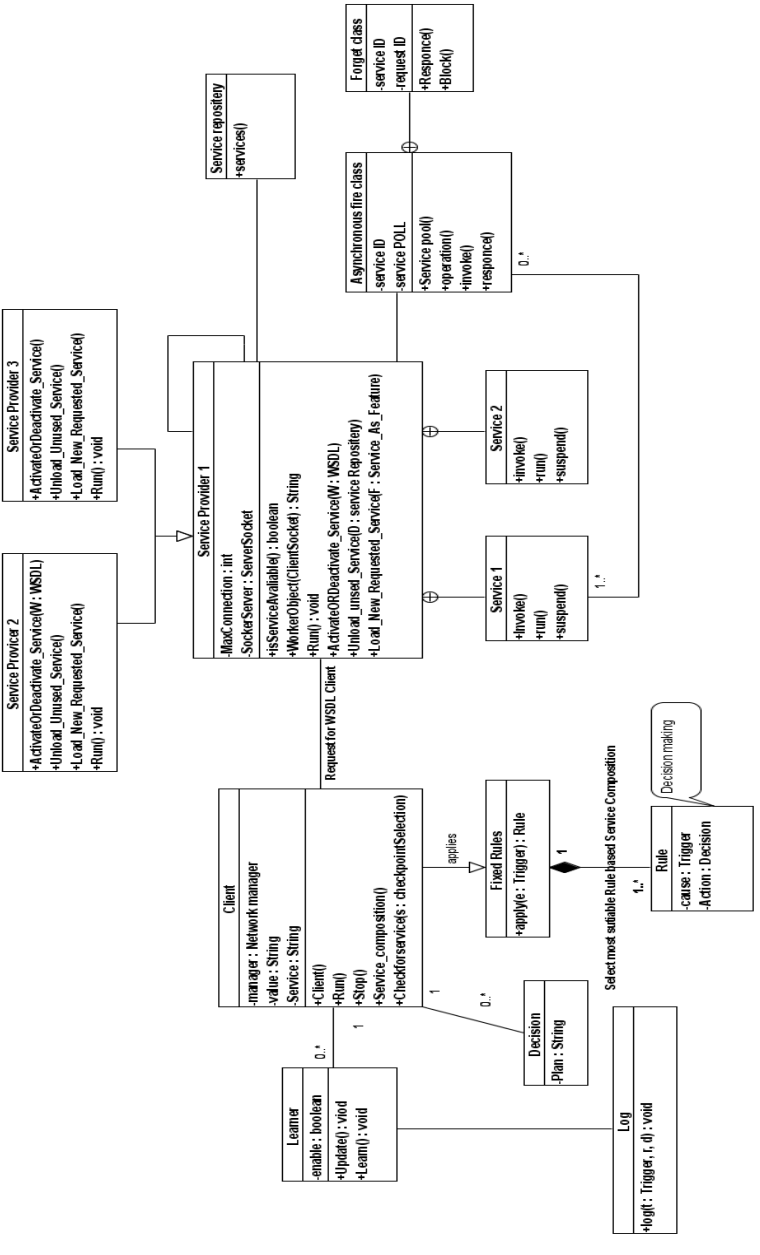



**Fig. 3.** Class Diagram for Adaptive pattern

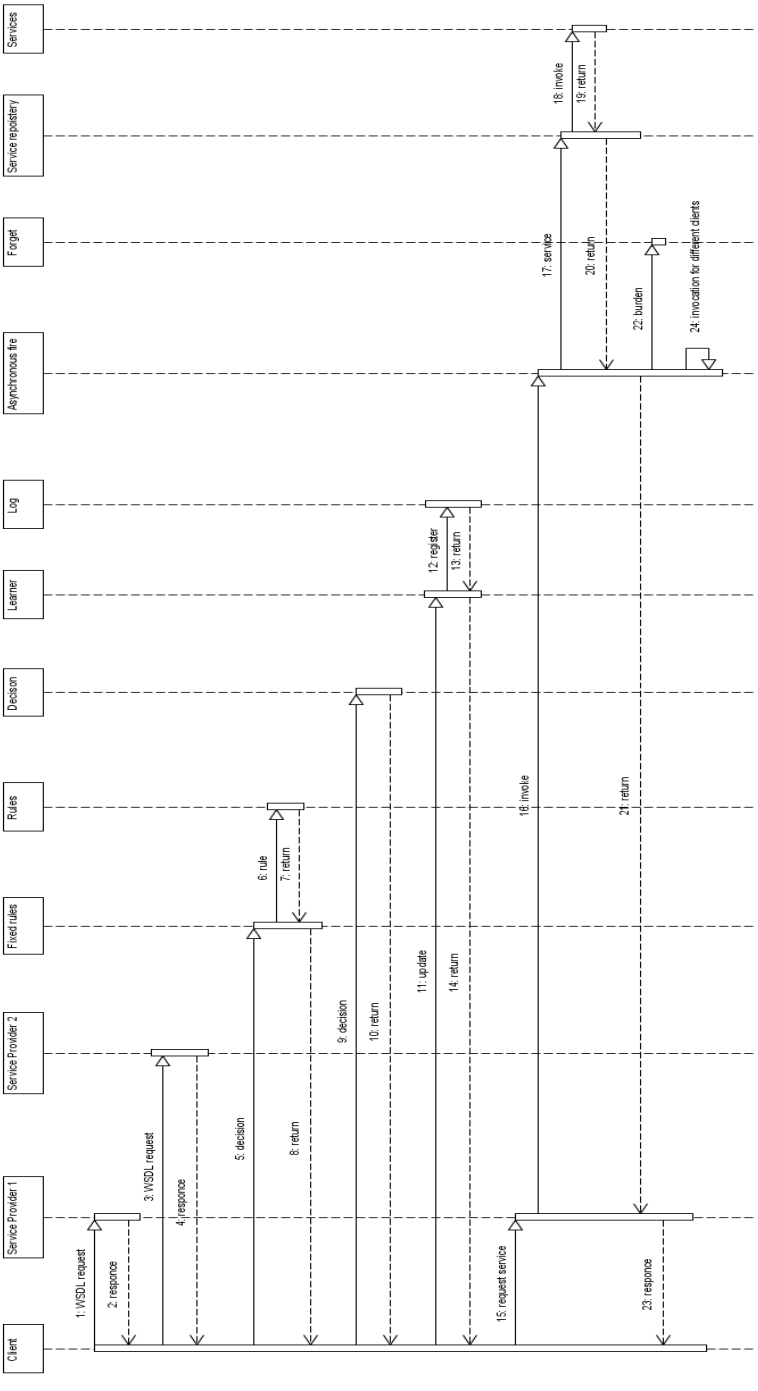



**Fig. 4.** Sequence Diagram for Service Adaptive Design Pattern.

## 5 Interface Definition for Design Pattern Entity

(a) **Client:**

**Client.java**

```
Public class Client
{
  Public invoke( int Serviced){
  }
  Public reload( int Serviced) {
  }
  Public run( int Serviced) {
  }
  Public Service_composition(){
  }
  Public checkforService(){
  }
 Public    Unload_Unused_Service(){
  }
Public
Load_New_Requested_Service      (){
  }
}
```

(c) **Fixed Rules:**

**Fixed Rules.java**

```
Public class Fixed Rules
{
 Public apply (Trigger e) {
 }
}
```

d) **Rule:**

**Rule.java**

```
Public class Rule
{
Trigger t;
Action decision;
}
```

(d) **Learner:**

**Learner.java**

```
Public class SERVICEPROVIDER
{
 Public update(){
  }
```

(b) **Service provider:**

**Service provider.java**

```
 Public class Service provider
 {
  Public isServiceavaliable(){
 }
  Public run(){
 }
 Public
 ActivateORDeactivateService (){
   }
 }
 Public learn (){
 }
 }
```

(e) **Log**

**Log.java**

```
Public class Log
{
 Public log(Trigger e){
 }
 }
```

(f)**Asynchronousfire**

**Asynchronousfire.java**

```
Public class SERVICEPROVIDER
{
 Public Servicepoll(){
 }
 Public operation (){
 }
 Public invoke (){
 }
 Public response (){
 }
}
```

## 6 Case Study

To demonstrate the efficiency of the pattern we took the profiling values using the Netbeans IDE and plotted a graph that shows the profiling statistics when the pattern is applied and when pattern is not applied. This is shown in figure 5.Here X-axis represents the runs and Y-axis represents the time intervals in milliseconds. Below simulation shows the graphs based on the performance of the system if the pattern is applied then the system performance is high as compared to the pattern is not applied.

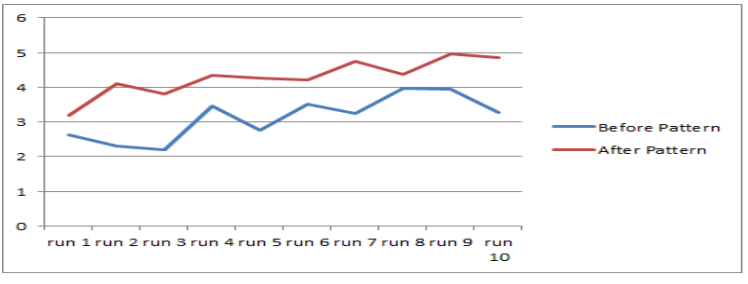



**Fig. 5.** Profiling statistics before applying pattern and after applying pattern

## 7 Conclusion

In this paper we propose a pure client and server approach to provide asynchronous and synchronous invocations for Web Services without necessarily using asynchronous messaging protocols. The framework was designed from a set of patterns of a larger pattern language for distributed object frameworks. The functionalities as well as the performance measurements indicate that the goals of the framework, as introduced at the beginning of this paper, were reached. Proposed system is amalgamation of different Design Patterns; proposed system satisfies all properties of autonomic system. As a drawback, an asynchrony framework on top of a synchronous invocation framework always incurs some overhead in terms of the overall performance of the client application.

**Future Work:** Future aims to develop an autonomous system by applying aspect oriented Design Pattern that will provide synchronous or asynchronous.

## Reference

1. Zdun, U., Voelter, M., Kircher, M.: Pattern-Based Design of an Asynchronous Invocation Framework for Web Services. International Journal of Web Service Research 1(3) (2004)
2. Foster, H., Uchitel, S., Magee, J., Kramer, J.: Model Based Verification of Web Service Compositions. In: 18th IEEE Int'l Conf. Automated Software Eng. (ASE 2003), pp. 152–161 (2003)

3. Mannava, V., Ramesh, T.: A Novel Event Based Autonomic Design Pattern for Management of Web Services. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) ACITY 2011. CCIS, vol. 198, pp. 142–151. Springer, Heidelberg (2011)
4. Prasad Vasireddy, V.S., Mannava, V., Ramesh, T.: A Novel Autonomic Design Pattern for Invocation of Services. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) CNSA 2011. CCIS, vol. 196, pp. 545–551. Springer, Heidelberg (2011)
5. Ramirez, A.J., Betty, H.C.: Design Patterns for developing dynamically adaptive Systems. In: 5th International Workshop on Software Engineering for Adaptive and Self-Managing Systems, Cape Town, South Africa, pp. 29–67, 50, 68 (2010)
6. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1995)
7. Pree, W.: Design Patterns for Object-Oriented Software Development. Addison-Wesley, MA (1994)
8. Crane, S., Magee, J., Pryce, N.: Design Patterns forBinding in Distributed Systems. In: The OOPSLA 1995 Workshop on Design Patterns for Concurrent, Parallel, and Distributed Object-Oriented Systems, Austin, TX. ACM (1995)
9. Cheng, S.-W., Garlan, D., Schmer, B.: Architecture-based self-adaptation in the presence of multiple objectives. In: ACM, International Workshop on Self-Adaptation and Self-Managing Systems, New York, p. 2 (2006)
10. Hwang, S.Y., Lim, E.P., Lee, C.H., Chen, C.H.: On Composing a Reliable Composite Web Service: A Study of Dynamic Web Service Selection. In: Proc. IEEE Int'l Conf. Web Services, pp. 184–191 (2007)

# Mining Tribe-Leaders Based on the Frequent Pattern of Propagation

Zhaoyun Ding*, Yan Jia, Bin Zhou, and Yi Han

School of Computers, National University of Defense Technology
410073 Changsha, China
{zyding,jiayan,binzhou,yihan}@nudt.edu.cn

**Abstract.** With the rapid development of new social networks, such as blog, forum, microblog, etc, the publication and propagation of information become more convenient, and the interactions of users become more active and frequent. Discovering the influencers in the new social network is very important for the promotion of products and the supervision of public opinion. Most of the previous research was based on the method of mining influential individuals, while the tribe-leaders were neglected. In this paper, a new method of mining tribe-leaders is proposed based on the frequent pattern of propagation. First, a method of changing the diffusion trees is proposed to overcome the problem of multi-pattern in propagation, where the information propagation trees are changed into a connected undirected acyclic graph. Then, a new frequent subgraph mining method called Tribe-FGM is proposed to improve the efficiency of graph mining by reducing the scale of pattern growth. Experiments are conducted on a real dataset, and the results show that Tribe-FGM is more effective than the method of Unot. Finally, we validate the effectiveness of our method by comparing it with the repost algorithms, where the experimental results indicate that the tribe-leaders with our method are consistently better than that of repost algorithms in both the one-step and multi-step coverage.

**Keywords:** Social network, Frequent pattern, Influence, Tribe-leaders, Influencers, Microblog, Graph mining.

## 1 Introduction

Recently, online social networks have obtained considerable popularity and are now among the most popular sites on the Web, such as blog, forum, microblog, etc. Online social networks play an important role for the spread of information since a piece of information can propagate from one node to another through a link on the network in the form of "word-of-mouth" communication. Therefore, Social network sites have become one of the several main sites where people spend most of their time [1].

* The work was supported by the Key Project of National Natural Science Foundation of China under Grant No. 60933005, NSFC under Grant No. 60873204, "863" under Grant No. 12505, and "863" under Grant No. 2011AA010702.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 143–153, 2012.

Their massive popularity has led to the viral marketing of content, products, or political campaigns on the sites. For instance, if we know there are a small number of "leaders" who set the trend for various actions, targeting them for the adoption of new products or technology could be profit-able to the companies. Also, if we know there are some leaders who set the trend for public opinions, we can target them to control the propagation of the public opinion. So, it is important to discover influential individuals for the promotion of products and the supervision of public opinions.

A number of recent empirical papers have addressed the matter of diffusion on networks in general, and particularly the attributes and roles of influencers. In general, influencers are loosely defined as individuals who disproportionately impact the spread of information. Interest among researchers and marketers alike has increasingly focused on whether or not diffusion can be maximized by seeding a piece of information or a new product with certain special individuals, often called "influentials" or simply "influencers" [2, 3]. Also, other research has mainly focused on discovering influential individuals such as leaderships [4] and ranking web documents [5, 6] and users [7].

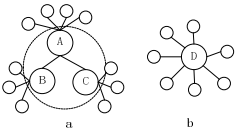



**Fig. 1.** The tribe-leaders and the influential individuals

Most of the previous research was based on the mining of influential individuals, while the tribe-leaders were ignored neglected. A Tribe-leader is a set of influential individuals who exchange information actively and frequently. Clearly, every tribe leader is an influential individual and all tribe leaders form a connected undirected acyclic graph. Tribe-leaders play a very important role in setting the trend for advertisement and public opinions. If an influential individual in the tribe-leader is affected, it will immediately diffuse the information to other influential individuals. This means the tribe-leader will diffuse the information to other nodes in the social network through all the influential individuals in the tribe-leader. Obviously, the tribe-leaders have better ability to diffuse information than a single influential individual. A single influential individual in the tribe-leader may not necessarily have a strong ability to diffuse information as other influential individuals, but the whole ability of a tribe-leader is stronger than that of other influential individuals. For example, the tribe-leader {A, B, C} in figure 1(a) has better ability to diffuse information than the influential individual D in figure 1(b). Therefore, mining the tribe-leaders is more important than discovering the influential individuals.

To discover the tribe-leader, we use the frequent pattern of propagation which was neglected in previous research. The propagation of information in the social network constructs diffusion trees, and the influential individuals in some diffusion trees diffuse the information frequently through their interactions. For instance, figure 2 shows four diffusion trees about a topic in the microblog. In the four diffusion trees, influential individuals A, B, C diffuse information frequently through the interactions of themselves. So, the set of {A, B, C} is a tribe-leader.

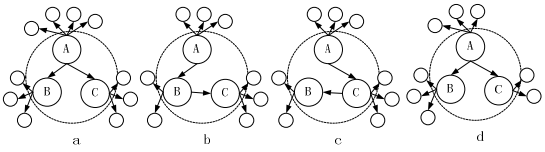



**Fig. 2.** The four diffusion trees about a topic in the microblog

In this paper, we use the frequent pattern of propagation in the online social network. First, a method of changing the diffusion trees is proposed to overcome the problem of multi-pattern in propagation, where the information propagation trees are changed into a connected undirected acyclic graph. Then considering its support and strength, a new frequent subgraph mining method called Tribe-FGM is proposed to improve the efficiency of graph mining by reducing the scale of pattern growth. Finally, we validate the effectiveness of our method by comparing it with the repost algorithms in the real dataset of sina microblog from china. Experimental results indicate that the tribe-leaders with our method are consistently better than that of repost algorithms in both the one-step and multi-step coverage.

## 2 Problem Definition

A social graph is an undirected graph G = (V, E) where the nodes are users. There is an undirected edge between users u and v representing a social tie between the users. The tie may be an explicit repost relationship in the microblog.

The propagation of information in the social network will construct diffusion trees $T = (V, E, \sum, L, r)$ where the nodes are users. There is a directed edge between users u and v representing a diffused path between the users. The alphabet $\sum$ is a set of labels, and the mapping label: $L: V \cup E \to \Sigma$ is called a labeling function. We define the label of edges as the name of information diffused. For instance, the propagation of information is a post in the microblog. The alphabet $r$ represents the root node of diffusion trees. For every node v $\in$ V, there is a unique path UP (v) = (v0 = r, $v_1$. . . $v_d$) (d $\geq$ 0) from the root r to v. Let u and v be nodes. If (u, v) $\in$ E then u is a parent of v, or v is a child of u. If there is a path from u to v, then u is an ancestor of v, or v is a descendant of u. A leaf is a node having no child.

**Property 1:** diffusion trees $T = (V, E, \sum, L, r)$ are rooted unordered trees, where there are no nodes with same labels in a diffusion tree.

**Property 2:** diffusion trees $T = (V, E, \sum, L, r)$ are connected directed acyclic graph.

The propagation of multi-information in the social network will construct multi-diffusion trees called diffusion forest $F = \{T_1, T_2, ..., T_n\}$. For example, the propagation of multi-post about a topic in the microblog will construct a diffusion forest.

In general, influencers are loosely defined as individuals who disproportionately impact the spread of information. Unfortunately, this definition is fraught with ambiguity regarding the nature of the influence in question, and hence the type of individuals who might be considered special. In light of this definitional ambiguity,

we note, however, that our use of the term influencer corresponds to a particular and somewhat narrow definition of influence, specifically the user's ability to write the post which diffuses directly through the social network graph.

**Definition 1** (influence). Give a diffusion tree $T = (V, E, \sum, L, r)$; the influence of a node u in the diffusion tree is defined as $\inf luence_u = |L(u)|$, where $L(u)$ is a set whose elements are those nodes which are linked with node u directly.

In the diffusion forest $F = \{T_1, T_2, ..., T_n\}$, a user may write lots of posts which diffuse directly through the social network graph in multi-diffusion trees such as node A in figure 2. So, influential individuals are defined as those users who have high influence and appear in multi-diffusion trees. For example, influential individuals about a topic in microblog are those users who write lots of posts about a topic and most of the posts have high influence. So, two thresholds $\sigma$ and $\psi$ are assigned, where the threshold $\sigma$ represents the frequency of a user appearing in the multi-diffusion trees and the threshold $\psi$ represents the influence of the user. According to property 1, we can infer that a user appears only once in a diffusion tree. So, the frequency of a user appearing in the multi-diffusion trees is the num of diffusion trees which contain this user.

**Definition 2** (influential individuals). Give a social network graph G = (V, E) and the diffusion forest $F = \{T_1, T_2, ..., T_n\}$ from the social network graph, and two thresholds $\sigma$ and $\psi$, a user v $\in$ V is an influential individual iff:

$$|L(T_i)| \geq \sigma, v \in T_i \ and \ \inf luence_{v,T_i} \geq \psi, v \in T_i \tag{1}$$

The set $L(T_i)$ represents the set of multi-diffusion trees containing the user v. The formula $\inf luence_{v,T_i}$ represents the influence of the user v in the diffusion tree $T_i$.

A Tribe-leader is a set of influential individuals who exchange information actively and frequently. Clearly, every leader is an influential individual and a tribe-leader is a connected undirected acyclic graph. So, three thresholds $\sigma$, $\psi$ and $\zeta$ are assigned, where the threshold $\sigma$ represents the frequency of a tribe-leader appearing in the multi-diffusion trees, the threshold $\psi$ represents the influence of the user in this tribe-leader, and the threshold $\zeta$ represents the whole influence of all users in this tribe-leader.

**Definition 3** (tribe-leader). Give a social network graph G = (V, E) and the diffusion forest $F = \{T_1, T_2, ..., T_n\}$ from the social network graph, and three thresholds $\sigma$, $\psi$ and $\zeta$, a set of user $L(v_i)$ is a tribe-leader iff:

$$\begin{aligned} &\exists G_{tribe} = (V_{tribe}, E_{tribe}) \ (V_{tribe} = L(v_i) \text{ and } \forall v_i, v_j \in V_{tribe}, \exists (v_i, v_j) \in E_{tribe}) \\ &and \ |L(T_i)| \geq \sigma, L(v_i) \subseteq T_i \ and \ \inf luence_{v,T_i} \geq \psi, v \in T_i, v \in L(v_i) \\ &and \ \sum_{v_k \in L(v_i)} \inf luence_{v_k,T_i} \geq \zeta, v_k \in T_i \end{aligned} \tag{2}$$

Give a social network graph G = (V, E) and multi-diffusion trees from the social network graph, the discovery of tribe-leaders is formalized as a problem of frequent pattern mining. Intuitively, we can discover tribe-leaders through the method of

frequent subtree mining. However, in fact, a tribe-leader exist multi-pattern of propagation. For instance, there are three patterns of propagation in the users' set {A, B, C} in figure 3. The three users A, B and C exchange information actively and frequently, and also three of these users are influential individuals. According the definition 3, we can infer the users' set {A, B, C} is a tribe-leader. The methods of frequent subtree mining only consider single paths of propagation, but the multi-pattern of propagation in the social network is neglected. So, some of members in the tribe-leader will be neglected for specific support. For instance, given the support thresholds 50%, in the figure 3, we can infer the set {A, B} and the set {B, C} are two tribe-leaders according the method of frequent subtree mining not considering the multi-pattern of propagation in the social network. More, if we give the support thresholds 100%, the tribe-leader {A, B, C} will be neglected. To overcome the problem of the multi-pattern of propagation, we change the diffusion trees into connected undirected acyclic graphs, which contain two steps: eliminating the direction of the diffusion trees and adding some edges which exist in the social network and also whose nodes are affected in the spread of information into the diffusion graphs.

For instance, we can get connected undirected acyclic graphs called diffusion graphs b, c and d in figure 4 after eliminating the direction of the diffusion trees. The two diffusion paths $A \to B$ and $B \to A$ are eliminated by eliminating the direction of the diffusion trees. Given the support thresholds 100%, we can infer the set {A, B} is a tribe-leader, where the results are better than that of the diffusion trees without eliminating the direction.

Next, we can add some edges which exist in the social network and also whose nodes are affected in the spread of information into the diffusion graphs. For instance, in figure 4(b), node B and node C are affected by node A; moreover, there is an edge between node B and node C in the social network in figure 4(a). So, we can add an edge into the diffusion graph in figure 4(b) and then get the changed diffusion graph in figure 5(b). With the same method, after adding the edge into the diffusion graph in figure 4(c) and figure 4(d), we can get the changed diffusion graphs in figure 5(c) and figure 5(d). After adding the edges into the diffusion graphs in figure 5, we can find that the multi-diffusion paths are eliminated. Thus, given the support thresholds 100%, we can infer the set {A, B, C} is a tribe-leader, where the results are better than that of the diffusion graphs without adding some edges into the diffusion graphs.

In the next section, we will introduce the detailed methods to change the diffusion trees into connected undirected acyclic graphs.

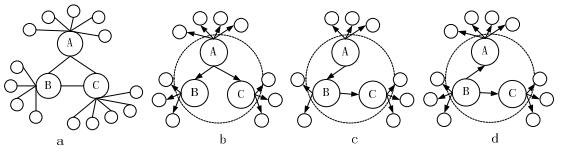



**Fig. 3.** The three patterns of propagation in the user set {A, B, C}

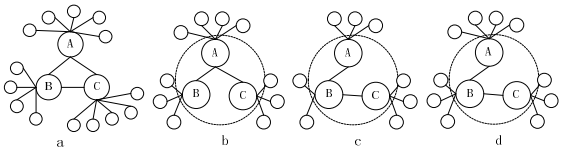



**Fig. 4.** The diffusion graphs after eliminating the direction of diffusion trees

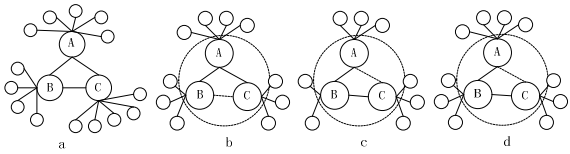



**Fig. 5.** The changed diffusion graphs after adding some edges into diffusion graphs

# 3 The Methods of Changing Diffusion Trees

## 3.1 Eliminating the Direction of Diffusion Trees

To overcome the problem of multi-pattern in propagation, we first eliminate the direction of diffusion trees and change the diffusion trees into connected undirected acyclic graphs. To eliminate the direction of diffusion trees, the diffusion trees are formalized as adjacency matrixes.

$$A[u,v] = \begin{cases} 1 & \exists edge(\mathrm{u},\mathrm{v}) \\ 0 & else \end{cases} \tag{3}$$

According to the calculation of matrixes, we can eliminate the direction of diffusion trees. We define the diffusion graphs after eliminating the direction of diffusion trees as $GT = (V, E, \sum, L)$. Also, the diffusion graphs are formalized as adjacency matrixes Q, that is $Q = A + A^T$.

**Property 3:** the diffusion graphs $GT = (V, E, \sum, L)$ are connected undirected acyclic graphs, and are subgraphs of the social network graph G = (V, E).

## 3.2 Adding Some Edges into Diffusion Graphs

To eliminate multi-diffusion paths in the diffusion graphs, we add into the diffusion graphs some edges which exist in the social network and also whose nodes are affected in the spread of information.

According to section 4.1, the diffusion graphs are defined as $GT = (V, E, \sum, L)$ and the diffusion graphs are formalized as adjacency matrixes Q. Also, we define a sequence table $ST$ to store the labels of nodes. At the same time, we define the social network graph G = (V, E), adjacency matrixes M, and the sequence table S. And we define the changed diffusion graphs $GF = (V, E, \sum, L)$, adjacency matrixes N, and the sequence table $SF$.

# 4 Tribe-FGM

To discover the tribe-leaders, a new algorithm called Tribe-FGM is proposed in this paper. The algorithm of Tribe-FGM is a new method of frequent subgraph mining, which uses the support and the strength to improve the efficiency of graph mining by reducing the scale of pattern growth.

First, we give the definition of support to measure the frequency of the tribe-leaders in the changed diffusion graphs.

**Table 1.** The algorithm of Tribe-FGM to discover the tribe-leaders

**Algorithm 2.** The algorithm of Tribe-FGM to discover the tribe-leaders

**Input.**

1) The set of changed diffusion graphs D, the influence of each user $\inf luence_{ij}$ .

2) The thresholds of support min_sup , strength min_$str$ , $\psi$ and $\zeta$ , and the DFS code s.

**Output.** The set of tribe-leaders S.

**Method**.

1) $S \leftarrow \phi$ .

2) Calling the function of memberPruning ( min_$str$ , $\psi$ , inf $luence_{ij}$ , D).

3) Calling the function of tP_gS ( min_sup , min_$str$ , $\psi$ , $\zeta$ , inf $luence_{ij}$ , D, s).

4) **procedure** memberPruning( min_$str$ , $\psi$ , inf $luence_{ij}$ , D)

a) Computing the frequency f of users whose $\inf luence_{ij} \geq \psi$ .

b) For each changed diffusion graph $GF$ in the set of D, do

c) for each member in the changed diffusion graph, do

d) if $f / |D| <$ min_$str$ for a user

e) Remove this user

f) End for

g) End for

h) Return the new set of changed diffusion graphs D.

5) **procedure** tP_gS( min_sup , min_$str$ , $\psi$ , $\zeta$ , inf $luence_{ij}$ ,D,s)

a) Inserting the s into the S, and $C \leftarrow \phi$ .

b) Finding all the edges e which are most right expanded $s\Diamond_r e$ , after scanning the D.

c) Inserting the $s\Diamond_r e$ into the C.

d) Sorting the C according to the DFS.

e) For each $s\Diamond_r e$ in the C, do

f) Computing the frequency f1 of users in the $s\Diamond_r e$ whose $\sum_{v \in s\Diamond_r e} \inf_v(Q) \geq \zeta$

g) Computing the frequency f2 of users in the $s\Diamond_r e$

h) if $f1 / f2 \geq$ min_$str$ $and$ $f2 / |D| \geq$ min_sup

i) tP_gS( min_sup , min_ $str$ , $\psi$ , $\zeta$ , inf $luence_{ij}$ ,D, $s\Diamond_r e$ )

j) End for

k) Return

6) End

**Definition 4** (support). Given the set of changed diffusion graphs $Graph = \{GF_1, GF_2, ..., GF_n\}$ and a tribe-leader T, the support $Supp_{Graph}(T)$ of the tribe-leader T in the set of changed diffusion graphs is defined as follows.

$$Supp_{Graph}(T) = \frac{|\{Q(c_i) | T \subseteq Q(c_i) \& Q(c_i) \in Graph\}|}{|Graph|} \tag{4}$$

The $|Graph|$ represents the size of the set of changed diffusion graphs. The range of the support is from 0 to 1. The support represents the frequency of the tribe-leaders appearing in the set of changed diffusion graphs. It also represents the activity of users in the tribe-leaders. The higher the support is, the more active the users in the tribe-leaders are.

To measure the influence of tribe-leaders, we give the definition of strength. Given a specified support, we can get a set of frequent subgraphs. The sum of users' influences in different subgraphs is different. For example, the sum of users' influences from A, B, C in figure 3(b) is 10, and the influences of individual users are 4, 3 and 3 respectively. Also, the sum of users' influences from A, B, C in figure 3(c) is 8, and the influences of individual users are 3, 2 and 3 respectively. The strength represents the frequency of the tribe-leaders whose whole influence is higher than the specified thresholds.

**Definition 5** (strength). Given the set of changed diffusion graphs $Graph = \{GF_1, GF_2, ..., GF_n\}$, a tribe-leader T, and the specified thresholds $\psi$ and $\zeta$, the strength $Stre_{Graph,\psi,\zeta}(T)$ of the tribe-leader T in the set of changed diffusion graphs is defined as follows.

$$Stre_{Graph,\psi,\zeta}(T) = \frac{|\{Q | \sum_{v \in Tribe} \inf luence_v(Q) \geq \zeta, \inf luence_v(Q) \geq \psi, T \subseteq Q, Q \in Graph\}|}{|\{Q | T \subseteq Q, Q \in Graph\}|} \tag{5}$$

The formula $\sum_{v \in Tribe} \inf_v(Q) \geq \zeta$ represents the whole influence of a tribe-leader, and the formula $\inf_v(Q) \geq \psi$ represents the influence of an individual user in the tribe-leader. The strength represents the frequency of the tribe-leaders whose whole influence is higher than the specified thresholds. The range of the strength is from 0 to 1. For example, given the threshold $\zeta$ as 10 and the threshold $\psi$ as 2, the strength of the tribe-leader {A, B, C} in figure 5 is 66.7%. Given the threshold $\zeta$ as 8 and the threshold $\psi$ as 2, the strength of the tribe-leader {A, B, C} in figure 5 is 100%.

Based on the algorithm of gSpan [32], we propose a new algorithm of frequent subgraph mining called Tribe-FGM which uses the support and the strength to improve the efficiency of graph mining by reducing the scale of pattern growth. Traditional methods of frequent subgraph mining such as gSpan did not consider the strength and would add some low influence nodes into the candidate thus increase the size of searching space. So, a new algorithm called Tribe-FGM is proposed.

Most of the nodes whose influence is low are pruned by the function of memberPruning (). The experimental results in section 6 indicate that only a small number of users are influential individuals, most of the users only participate in the discussion of topics, and many non-influential individuals form the "the long tail". So, most of the nodes are pruned by the function of memberPruning () and the scale of pattern growth will be reduced. The time complexity of the function of memberPruning () is $O(n \wedge 2)$ . The support and the strength are considered in the function of tP_gS () to improve the efficiency of graph mining by reducing the scale of pattern growth.

## 5 Experiments

First, we select a real dataset of sina microblog from China for experiments. We get the data through the search API of sina microblog. The dataset consists of two parts: one is about the topic of "Earthquake" which contains about 0.9 million posts and 0.6 million users, and the other is about the topic of "Two Sessions of NPC and CPPCC" which contains about 0.31 million posts and 0.21 million users.

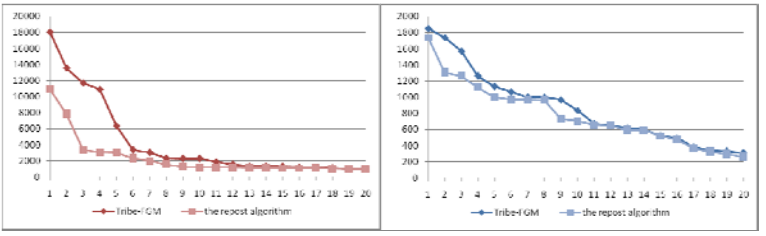
A post reposted about a topic will construct a diffusion tree, and lots of diffusion trees will be constructed for a specific topic. The frequency of different posts reposted is different, and the size of diffusion trees is different. About 71429 different diffusion trees are constructed for the topic of "Earthquake" and 18485 different diffusion trees are constructed for the topic of "Two Sessions of NPC and CPPCC".

The effectiveness of Tribe-FGM is validated by comparing it with the repost algorithms. To evaluate the effectiveness of our Tribe-FGM algorithm, we introduce the metrics of "One-Step Coverage" and "Multi-Step Coverage".

**Metric 1. One-Step Coverage:** Given a set of nodes in a network, one-step coverage is defined as the number of nodes that are directly influenced by this set of nodes. In the microblog, the one-step coverage of an influential individual is measured as how many users directly repost this user.

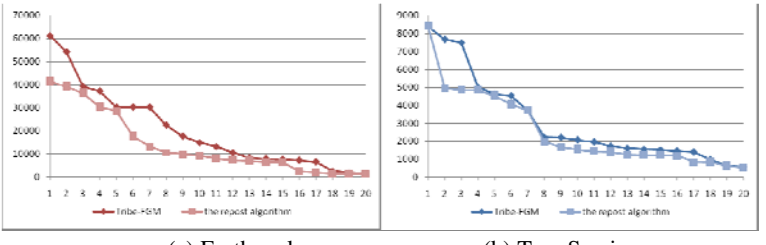
**Metric 2. Multi-Step Coverage:** Given a set of nodes in a network, multi-step coverage is defined as the number of nodes that are either directly or indirectly influenced by this set of nodes. In the microblog, the multi-step coverage of an influential individual is measured as how many users either directly or indirectly repost this user through N hops.

In the following experiments, we give the thresholds as follows: $\zeta = 8$ , $\psi = 5$ , $\min\_str = 80\%$ , $\min\_\sup_{\text{earthquake}} = 0.7\%$ and $\min\_\sup_{two\ sessions} = 1\%$ . We compare the effectiveness of the "one-step coverage" and the "multi-step coverage" with N=1 and N=2 of top-20 influential individuals identified by the algorithm of Tribe-FGM and the repost algorithm. Figure 6 illustrates how the one-step and multi-step coverage change with the rank of identified influential individuals. The experimental results indicate that the influential individuals in tribe-leaders with the algorithm of Tribe-FGM are consistently better than that of the repost algorithm in both the one-step and the multi-step coverage.



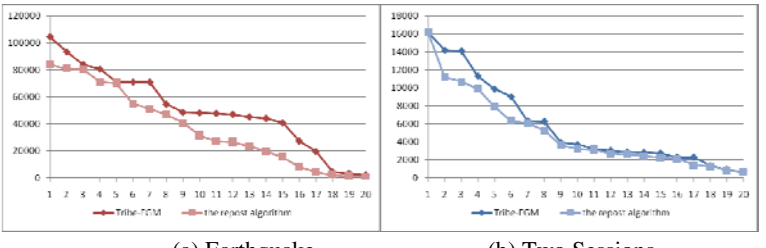

(a) Earthquake (b) Two Sessions

(1) One-Step Coverage




(a) Earthquake (b) Two Sessions

(2) Multi-Step Coverage with N=1




(a) Earthquake (b) Two Sessions

(3) Multi-Step Coverage with N=2

**Fig. 6.** How the one-step and multi-step coverage change with the rank of identified influential individuals. The horizontal ordinate represents the rank of identified influential individuals, and the vertical ordinate represents the one-step and multi-step coverage change with the rank of identified influential individuals.

## 6 Conclusion

Most of the previous research was based on the mining of influential individuals, while the tribe-leaders were neglected. In this paper, a new method of mining tribe-leaders is proposed based on the frequent pattern of propagation. To overcome the problem of multi-pattern in propagation, we change the diffusion trees into connected undirected acyclic graphs, which contain two steps: eliminating the direction of the diffusion trees, and adding into the diffusion graphs some edges which exist in the social network and whose nodes are affected in the spread of information. To discover the tribe-leaders, a new algorithm called Tribe-FGM is proposed in this paper. The

algorithm of Tribe-FGM is a new method of frequent sub-graph mining, which uses the support and the strength to improve the efficiency of graph mining by reducing the scale of pattern growth. Finally, in the experiments, we validate the advantages, performance and effectiveness of the new method which we propose.

In this paper, the influence is narrowly defined as the user's ability to write the post which diffuses directly through the social network graph. However, the influence of a user in the microblog is determined by multi-dimension, such as the novelty of his posts, the sensitivity of his posts, and so on. In the future, we will discover the influential individuals based on different characteristics of the users. Moreover, the diffusion paths of information are determined by the explicit relationships of repost, but some implicit relationships are ignored in this paper. In the future, we will mine the implicit relationships through the similarity of contents and the behaviors between two users.

## References

1. Ahn, Y., Han, S., Kwak, H., et al.: Analysis of topological characteristics of huge online social networking services. In: Proc of the 16th International World Wide Web Conference, pp. 835–844. ACM, New York (2007)
2. Keller, E., Berry, J.: The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy (2003)
3. Weimann, G.: The Influentials: People Who Influence People. State University of New York Press, Albany (1994)
4. Wu, D., Ke, Y., Yu, J.X., et al.: Leadership discovery when data correlatively evolve. World Wide Web Journal 14, 1–25 (2011)
5. Page, L., Brin, S., Motwani, R., et al.: The PageRank Citation Ranking: Bringing Order to the Web (1998)
6. Yu, W., Zhang, W., Lin, X., et al.: A Space and Time Efficient Algorithm for SimRank Computation. In: Proc of the 12th International Asia-Pacific Web Conference (2010)

# Research on Semantic Label Extraction of Domain Entity Relation Based on CRF and Rules

Jianyi Guo[1,2], Jun Zhao[1], Zhengtao Yu[1,2], Lei Su[1,2], and Nianshu Jiang[1]

[1] The School of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650051, China
[2] The Institute of Intelligent Information Processing,
Computer Technology Application Key Laboratory of Yunnan Province,
Kunming 650051, China
{gjade86,s28341}@hotmail.com, {zhaojun5831,jns1986}@163.com,
ztyu@bit.edu.cn

**Abstract.** For the vast amounts of data on the Web, this paper presents an extraction method of semantic label of entity relation in the tourism domain based on the conditional random fields and rules. In this method, firstly making use of the ideas of classification in named entity recognition, semantic items reflecting entity relations are seen as semantic labels in the contextual information to be labeled, and identify the semantic label with CRF, then respectively according to the relative location information of the two entities and semantic label and rules, the semantic labels are assigned to the associated entities. The experimental results on the corpus in the field of tourism show that this method can reach the F-measure of 73.68%, indicating that the method is feasible and effective for semantic label extraction of entity relation.

**Keywords:** Entity relation, Semantic label, Conditional random fields, Feature template.

## 1 Introduction

The automatically extraction of semantic relations is an important part of information extraction, information retrieval and the knowledge base building [1]. ACE (Automatic Content Extraction) defines the entity relation extraction task is to identify and describe the relationship. That is to say, the entity relation extraction process includes: 1) are there any relationship between two entities? what the relationship type is; 2) giving the semantic description (semantic tags) of relationship characteristics, thus a complete description of the relationships between entities can be gotten. From the application point, the entity object is only marked the name, so as to meet the demand of the accuracy and effectiveness of information retrieval. Semantic label of entity relation is to give the two related entities a detailed semantic description, however, if only to know the relationship between two entities

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 154–162, 2012.

whereas do not know what semantic relations, which ultimately can not meet the application requirements. For example, in this sentence "Bill Gates is the CEO of Microsoft", taken by entity relation extraction, we can only get the employment relationship existed between entity "Bill Gates" and entity "Microsoft", but do not know what the more accurate semantic relationship is. If the specific duty "CEO" is seen as semantic tag to indicate their semantic relationships, the relationship between the pair of associated entities is made to be more explicit. Therefore, the entity relationship is extended to identify the semantic labels—"key words" which can distinguish the different relationships. However, most traditional researches only stay in the stage of finding relationship and determining the relationship types, such as Text-TO-Onto[2], Text2Onto[3], DOGGLE[4] just to find an associated pair of concepts, and do not give the semantic label of concept relationship. To solve this problem, OntoLearn[5] predefines a relatively large available set of relationship predicates, through the match subject and object on WordNet and FrameNet to identify the most suitable relationship predicate from a predefined set as the corresponding relationship label. The method proposed by J. Villaverde[6] is based on the analysis of syntactic structures and dependencies among concepts existing in a domain-specific text corpus to find and mark up the relationship, which limited the POS of semantic tags to verb. In Chinese, in addition to the verb to express the relationship, the noun and other words can also express the relationship information, which is more complex compared to English. About the researches on semantic label of Chinese entity relation, Liu Kebin[7] from Shanghai Jiaotong University proposes an algorithm to solve the expansion to binary entity relation based on a combination of rules and semantic information, which obviously takes a certain degree of human; Sun Xiaoling[8] from Dalian University of Technology proposes an extraction method of relationship description words on the basis of TongYiCi CiLin. The above methods to get the semantic label of entity or concept relation are mostly dictionary and rule-based methods. Learning for labeling, Chen Yiheng[9] from Harbin Institute of Technology uses the singular value decomposition (SVD) results in LSI (latent semantic index) method and neurons vector relationship after training with SOM (self-organizing map) to automatically access to the text category label. In the face of the Web's growing mass of data, common dictionary is often unable to meet the needs of semantic items in specific fields, so machine learning is combined to access to semantic labels fast and efficiently. As the semantic information depends on the context, therefore, this paper presents semantic tag extraction method of entity relation with a combination of machine learning and rules to define semantic label extraction process includes label recognition and label description. Firstly using the thought of categories ,see semantic labels as entities and identify the semantic label with conditional random fields. Then from view of POS and location specificity of domain semantic tags, extract the corresponding semantic labels of related entity pairs by adding the appropriate rules. The method can better capture the statistical regularities of natural language, and can better describe the traits of natural language, which has good portability. In the field of tourism, this method not only extracts granular relationship type between entities in specific domains, and further obtains concrete semantic information of entity relation.

## 2 Semantic Tag Extraction of Entity Relation Based on Conditional Random Fields and Rules

In this paper, under the background of tourism field, taking into account the characteristics of the field of tourism, divide the relationship types to four major categories and seven sub-relationship types. Different relationships with different types have description in different categories, so extraction methods are not the same. Common types of relationships can be better to find the semantic labels using of machine learning algorithms, whereas, to the semantic tags of unusual relationships, we need to add appropriate rules to obtain them. This article uses the method of combination of statistics and rules to extract semantic tags.

### 2.1 The Selection of Conditional Random Fields

The sentence in the domain-specific text "[Shangri-La], year-round mild climate, [maximum temperature] bq is [25 degrees], the weather of spring is warm and drought, the weather of autumn is cool and wet, the weather of winter is cold and dry, the weather of summer is hot and rainy, [the average annual temperature] bq is [14.7 ℃ ] ". We find that [the maximum temperature] bq can be seen as the semantic label between [Shangri-La] and [25 degrees] (here "bq" means "label", likewise, [the average annual temperature] bq can be seen as the semantic label between [Shangri-La] and [14.7 ℃]), which can be identified as a named entity. The conditional random fields CRF[10] for named entity recognition is superior to other methods available to effect, of which the idea is to calculate an undirected graph model of conditional probability of the output nodes under the given condition of input nodes, and is often used as the segmentation and label for sequence data. This problem of semantic tag recognition can be defined as the mark of sequence, that is to determine whether observed words belong to predefined characteristics sets, which is just consistent with the advantages of sequence mark with CRF, so conditional random field model is selected to identify the semantic label.

### 2.2 The Process of Semantic Label Extraction of Domain Entity Relation

In this paper, the method identifies semantic labels of entity relation with CRF and combines the related entity pairs and the corresponding semantic labels on the basis of rules. The process as shown in Fig. 1. The first step, firstly segment words and tag POS of words and other pretreatments on raw corpus, secondly perform feature selection and formulate feature template according to the semantic label characteristics of entity relation in tourism domain, then use of CRF on the training corpus for training the classifier to construct a classifier of semantic tag identification, finally, test on the test corpus with classifiers. The second step, process raw corpus used in the first step, then extract entity relation based on binary classifier and reasoning [11]. The third step, the POS of semantic labels of entity relation in tourism domain are not only the verb, and some are nouns or compound nouns, so according to location information of

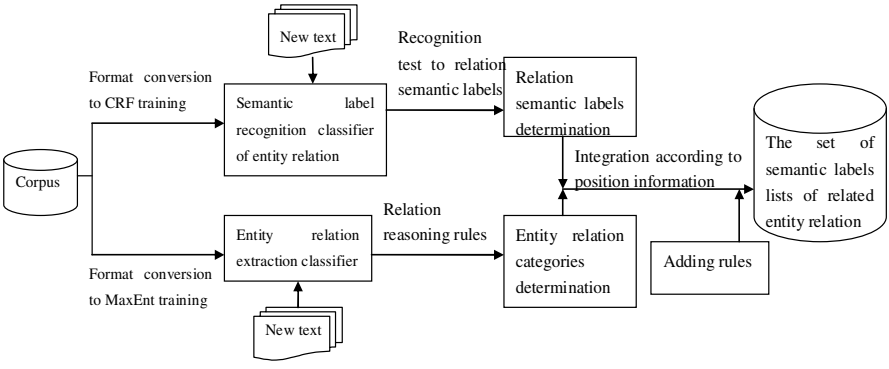



**Fig. 1.** The extraction process of semantic label of entity relation

the two related entities and semantic labels, extract the semantic tags of common types of entity relation, and then add rules to extract semantic tags of some special kinds of entity relation, that is a combination of related entity pairs and corresponding semantic labels.

### 2.3 CRF Corpus Annotation

In the semantic labels recognition study of domain entity relation, we need to do some different treatments according to possible segmentation granularity of various types of named entities. As named entities in the field of tourism mainly relate to place names and other attractions, the method of marking semantic labels adopts commonly used BMEO mark of CRF model, which references to mark method of named entity recognition in the literature [12]. Firstly, the semantic labels in text are marked "bq", however not a semantic label without any mark. Secondly, in specific operations of marking semantic tags, take BMEO annotation methods, the so-called BMEO way refers to an input unit marked with one of B (begin), M (middle), E (end) and O (other).

### 2.4 Feature Selection and Formulate Feature Template

**Feature Selection.** The best advantage of Conditions Random Fields (CRFs) [13] model is the ability to fully integrate the words, terms and part of speech information in the context. For the semantic tags, most of the words circle around the vocabulary are auxiliary words, interjection, verb or quantifier, it is difficult to distinguish whether they are domain terms under the above circumstance. In the tourism field of Yunnan, in order to take advantage of these information and border information that is external characteristics, drawing on the experiments of literature [12], set the sliding window size as 2, which achieved good results through experiments. To increase description of the context information, each above atom feature selection needs the following four positions offsets, such as -2, -1, 1and 2. Learning from experiments in

literature [12], this paperdetermines the atomic feature used to identify the semantic labels, consider only one factor for each feature, hence call them atomic features.

**Formulate Feature Template.** Semantic labels are not only compact and relatively fixed combination of words or phrases, but also have strong domain features. The semantic tag of single word is most a verb, noun, and the combination of complex semantic labels has certain patterns most of which are the following combination patterns, such as adjective-noun, noun-noun composition or nominal composition. For example, "The [planning / n area / n] bq of / u [Purple river mountain / ns Scenic / n] jd is [43 / m s.q.km / q] m.", "planning area" appears as a composite semantic label. Semantic label has its own external features which are the existed boundary features of semantic tags, including the anterior sector, posterior sector, both anterior and posterior sector. For the semantic tags, in the process of accessing to them we must make full use of characteristics within their own words and boundary information to obtain. Therefore, from the view of specific phenomenon and regular pattern of language, not only to consider the characteristics information of semantic terms within the word itself, but also consider the boundary external feature information of semantic tags, then develop a template with the relative location information through the context and part of speech information. When selecting the above features, the atomic feature template of semantic tag recognition has been formulated.

### 2.5 The Integration Process of Entity Relation Extraction and Semantic Tag Recognition

The semantic labels of entity relation, as description information of important significance for semantic relations between entities, can identify entity relation types more detailed types. Therefore, this paper carried out the extraction study of semantic labels while research work of entity relation extraction in the field is performing. As the recognition model of semantic tag and entity relation extraction model were respectively obtained, the identified semantic label was needed to be assigned to the appropriate entity pair, which was an integration of identified semantic tag and extracted related entity pair. The concrete integration process is as follows:

1) Segment words and mark part of speech on free text corpus in the field of tourism, and process them to meet the required format of conditional random fields.
2) Call respectively named entity recognition models, model_file1 and model_file2, and use the commands CRF toolkit provides on processed corpus in the first step to perform named entity recognition and semantic tag recognition.
3) The corpus will be divided by a single sentence, and then respectively extract automatically out the beginning and end position of each entity and semantic label from a single sentence by the program.
4) Arrange and freely combinate two entities for all entities, and extract all the features.
5) Respectively according to location information of the two entities and corresponding semantic tags, place the common types of semantic labels after the appropriate related entities.

```
0 E1:[崇圣寺三塔公园]jd E2:[洱海]jd bq:面向 1:1 2:1 3:6 4:6 5:0 6:13 7:8 8:8 9:8 10:8 11:2
 12:8 13:2 14:7 15:历史 16:的 17:内 18:, 19:, 20:面向 21:, 22:占地
0 E1:[大理三塔苑酒店]ht E2:[23000多平方米]m bq:占地面积 1:2 2:3 3:3 4:7 5:0 6:16 7:8 8:8 9:8
 10:2 11:1 12:8 13:2 14:1 15:0 16:0 17:座落 18:于 19:占地 20:面积 21:, 22:是
0 E1:[崇圣寺三塔公园]jd E2:[苍山]jd bq:背靠 1:1 2:1 3:6 4:6 5:0 6:10 7:8 8:8 9:8 10:2 11:8
 12:8 13:2 14:7 15:历史 16:的 17:内 18:, 19:背 20:靠 21:, 22:面向
Eg:[大理三塔苑酒店]ht座落于距今已有1300多年历史的[崇圣寺三塔公园]jd内，[距]/bq[大理古城]jd
[1.5公里]m，[背靠]/bq[苍山]jd，[面向]/bq[洱海]jd，[占地面积]/bq[23000多平方米]m，是国内为数
不多的庭园式[三星级]m酒店。
```

**Fig. 2.** The integration results of semantic tag recognition and entity relation extraction

6) And then add rules to extract specific types of semantic tags and place them after the corresponding entity pair.

Finally, the integration results of semantic tag recognition and entity relation extraction are shown in Fig. 2. The term 'bq' pointed to in each token is the corresponding semantic label between E1 and E2 in current token, and can clearly describe the relation significance of related entity pair of E1 and E2. In table 2, jd, ht and m respectively represents attraction, hotel and quantifier.

## 3 Experimental Setup and Results Analysis

### 3.1 Preprocess Corpus

Since there is no authoritative corpus in specific fields, to the relation extraction tasks for the field of tourism, the corpus are built by us in the field of tourism. The corpus crawled from the Web with Nutch are 1000 free documents in Yunnan tourism, which are not standardized. And the content of each text is relatively small and fragmented, so some work has been done to pretreat the corpus.

### 3.2 Experimental Methods and Results Analysis

We performed related experiments on the corpus built by us in the tourism field. The experiments were divided into two phases, the first phase of the experiments was to train pretreated corpus using a fixed set of training corpus, and then identified semantic label of entity relation and diagramed the statistical results. The CRF experiment results of phase 1 have been shown in Table 1. The second phase, based on the relation extraction experiments with the approach of binary classification and reasoning [11], the experimental 1 used location information of two related entities and semantic tags to assign each semantic label to the corresponding entity pairs. The results of experiment 1 have been shown in table 2. Experiment 2 added rules on the basis of experiment 1. The results of experiment 2 as shown in table 4. The comparing results of the two experiments in phase 2 have been shown in Table 2. Finally, the semantic label extraction results of entity relation were shown in Table 3.

**Table 1.** The semantic label recognition test results of entity relation in tourism domain

| Open/ Closed | The total number of ntities | Number of Identification | Correct number | Precision (%) | Recall (%) | F-score (%) |
|---|---|---|---|---|---|---|
| Closed test | 5025 | 4994 | 4924 | 98.60 | 98.00 | 98.30 |
| Open test | 3310 | 3265 | 2877 | 88.12 | 86.92 | 87.52 |

**Table 2.** The test results of semantic labels extraction of entity relation in field of tourism

| | Precision(P)/% | Recall(R)/% | F-score(F)/% |
|---|---|---|---|
| Experiment 1:Use of location information (without rules) | 47.83 | 73.33 | 57.90 |
| Experiment 2:Use of location information and rules | 60.87 | 93.33 | 73.68 |

**Table 3.** The semantic labels extraction results of entity relation in field of tourism

| Entity 1 | Entity 2 | Major Types of Entity Relation | Semantic Labels | Subtypes of Entity Relation |
|---|---|---|---|---|
| Shangri-La | 14.7°C | Property-limited relation | Average temperature | Limited attraction property |
| Shangri-La | 25 degree | Property-limited relation | Maximum temperature | Limited attraction property |
| Yunnan Meideng Hotel | Four-star | Property-limited relation | Star | Limited hotel property |
| Yunnan Meideng Hotel | 21 acres | Property-limited relation | Covers an area of | Limited hotel property |
| Dali Three Pagodas Hotel | Chongsheng Three Pagodas Garden | Structure affiliation relation(AFF) | Is located | The hotel is located at the attraction |
| Chongsheng Three Pagodas Garden | Dali Ancient City | Geographic location relation(GEO) | be apart from | Where attractions are located in |

Table 1 shows the semantic label was seen as a named entity to identify with CRF, of which predictive ability achieved the desired results. In the closed test, the F value of semantic labels recognition reached 98.30, and it also achieved good results in the open test. The observed results can be seen from table 2 that the result has been greatly improved since adding rules, because some relation semantic labels can not be gotten only by machine learning algorithm and location information. It shows, with the effective combination of statistics and rules, the experiment can achieve better extraction results.

## 4 Conclusion

To give entity relation the semantic labels could help a computer to compute and understand more effectively for user tasks, and can be used for knowledge discovery, vertical search and other applications, which could allow users to quickly and easily identify what they need. This article proposes a semantic label extraction method of entity relation in specific field with the combining CRF and rules-based. Experimental results show that CRF can effectively combine contextual information to find important single and complex semantic vocabularies, which has a good performance on semantic tags recognition. Further with the vocabulary characteristics of describing entity relation in tourism field, we can find semantic labels of some common types of entity relation through location information of related entities and associated semantic labels. On this basis, we targetedly add rules to special types of semantic vocabulary, which makes the overall extraction performance has been improved significantly.

**Acknowledgements.** This paper is supported by National Nature Science Foundation (No. 60863011), The Key Project of Yunnan Nature Science Foundation (No. 2008CC023), Yunnan Young and Middle-age Science and Technology Leaders Foundation (No. 2007PY01-11).

## References

1. Aone, C., Ramos Santacruz, M.: REES: A Large-scale Relation and Event Extraction System. In: Proc. of the 6th Applied Natural Language Processing Conference, pp. 76–83. ACM Press, New York (2000)
2. Maedche, A., Staab, S.: Mining Ontologies from Text. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 189–202. Springer, Heidelberg (2000)
3. Cimiano, P., Völker, J.: Text2Onto:A Framework for Ontology Learning and Data – Driven Change Discovery. In: Montoyo, A., Muńoz, R., Métais, E. (eds.) NLDB 2005. LNCS, vol. 3513, pp. 227–238. Springer, Heidelberg (2005)
4. Sugiura, N., Shigeta, Y., Fukuta, N., Izumi, N., Yamaguchi, T.: Towards On-the-Fly Ontology Construction - Focusing on Ontology Quality Improvement. In: Bussler, C.J., Davies, J., Fensel, D., Studer, R. (eds.) ESWS 2004. LNCS, vol. 3053, pp. 1–15. Springer, Heidelberg (2004)
5. Navigli, R., Velardi, P., Cucchiarelli, A., Neri, F.: Quantitative and Qualitative Evaluation of the Ontolearn Ontology Learning System. In: ECAI Workshop on Ontology Learning and Population (2004)
6. Villaverde, J., Persson, A., Godoy, D., Amandi, A.: Supporting the discovery and labeling of non-taxonomic relationships in ontology learning. In: Expert Systems with Applications, pp. 10288–10294 (2009)
7. Liu, K., Li, F., Liu, L., Han, Y.: Implementation of a Kernel-Based Chinese Relation Extraction System. Journal of Computer Research and Development 44(8), 1406–1411 (2007)
8. Sun, X.-L., Lin, H.-F.: Relations Extraction and Structure Mining of Personal Network. Microelectronics & Computer 25(9), 233–236 (2008)

9. Chen, Y., Qin, B., Liu, T.: Search Result Clustering Method Based on SOM and LSI. Journal of Computer Research and Development 46(7), 1176–1183 (2009)
10. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields:Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. of the 18th International Conference on Machine Learning, San Francisco, USA (2001)
11. Lei, C., Guo, J., Yu, Z., Zhang, S., Mao, C., Zhang, C.: The Field of Automatic Entity Relation Extraction based on Binary Classifier and Reasoning. In: Third International Symposium on Information Processing, pp. 327–331 (2010)
12. Guo, J., Xue, Z., Yu, Z., Zhang, Z., Zhang, Y., Yao, X.: Named Entity Recognition for the Tourism Domain. Journal of Chinese Information Process 23(5), 47–52 (2009)
13. Della Pietra, S., Della Pietra, V., Lafferty, J.: Inducting features of random fields. IEEE Transaction on Pattern Analysis and Machine Intelligence 19(4), 380–393 (1997)

# Recognition of Chinese Personal Names Based on CRFs and Law of Names

Zheng Lvexing[1], Lv Xueqiang[1,2], Liu Kun[1,2], and Du Yuncheng[1,2]

[1] Chinese Information Processing Research Center, Beijing Information Science & Technology University, Beijing 100101, China
[2] Beijing TRS Information Technology Co., Ltd., Beijing, China
zhengulxin@163.com

**Abstract.** Recognition of chinese personal names becomes a difficult and key point in chinese unknown word recognition. This paper explored the context boundary of names and the law of names. The context boundary of names is concentrated, which can reduce recognition errors brought up by Forward Maximum Matching Segmentation; from real text corpus, we discover that names begin with surname, dislocation characters of names reach 70.83%, and rare characters of names reach 9.42%. This paper improved the Forward Maximum Matching Segmentation and implemented a name recognition test based on CRFs, which was combined with the surname, the context boundary, the dislocation character and the rare character. The open test shows that recall reaches 91.24% from BakeOff-2005.

**Keywords:** Chinese personal names, Segmentation, CRFs, Dislocation character, Rare character.

## 1 Introduction

Chinese unknown word recognition is very important to Chinese word segmentation and syntactic study. The number of Chinese personal names account for 48.6% [1] of Chinese unknown words, so recognition of names is important in natural language processing. The major difficulties [1] of Chinese personal name recognition are that: name segmentation errors and complex structure.

At present, Base on statistical method is an effective way to recognize names, and statistics models which are commonly used include Hidden Markov Model [3] and Max Entropy Model [4], but this method can't recall names of segmentation error and is difficulty to recognize the names which consist of non-common characters. Moreover the HMM requires independence assumptions and the MEM has label bias problem.

This paper provided a solution to these problems-- Recognition of Chinese Personal Names Based on CRFs and Law of Names. The solution depended on the context boundary of names to reduce segmentation error that can't recall the names. It analyzed a large number of the chinese character tagging[8] from real text corpus to find the law of Chinese names (dislocation character and rare character). By the law of names, it could recall names which consist of non-common characters. Finally we solved sequence labeling problem by CRFs [2].

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 163–170, 2012.

## 2 Recognition of Chinese Personal Names Based on CRFs and Law of Names

### 2.1 Segmentation System

Chinese Personal names often consist of the surname and the first name. The surname is in front of the first name. The first name is composed of 1-2 characters. And the context boundary of names is concentrated. This paper established a set of surnames [5], a set of prior-boundary words [5] and a set of posterior-boundary words [5]. Therefore, the segmentation system is composed of three sets of words and forward maximum matching method [9]. According to statistics, there are main three kinds of segmentation errors.

#### 2.1.1 Posterior-Boundary Word and the Last Character of Name Form a Word

Because the first name is composed of 1-2 characters, it considers two ways. When the each longest word is matched by forward maximum matching method, if the prior word is surname(1 characters) or the prior word is one character that the prior-prior word is surname(2 characters), remove the first character form the word. The rest of sentence matches posterior-boundary words by forward maximum matching method. If matching successfully, the word divides into the first character and posterior-boundary word. Follow figure 1:

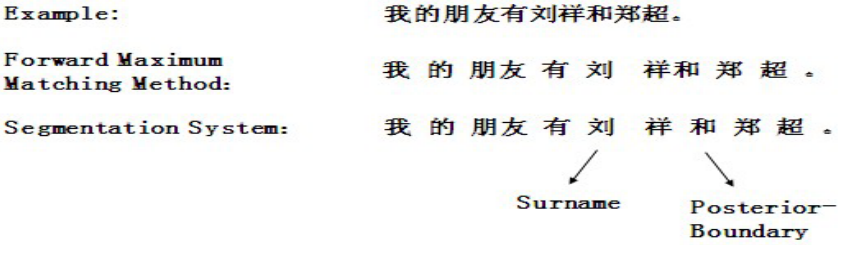



**Fig. 1.** Segmentation System

#### 2.1.2 Surname and the First Name Form a Word

When the each longest word is matched by forward maximum matching method and the word is two character, if the first character of word is surname, The prior word matches prior-boundary words. If matching successfully, the word divides into two characters. Follow figure 2:

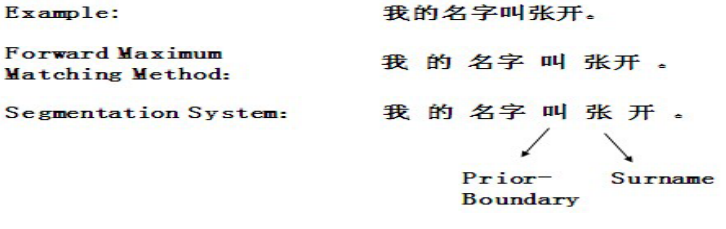



**Fig. 2.** Segmentation System

### 2.1.3 The First Name(2 Characters) Form a Word

When the each longest word is matched by forward maximum matching method and the word is two characters, if the prior word is surname. The prior-prior word matches prior-boundary words. If matching successfully, the word divides into two characters. If match fails, the rest of sentence matches posterior-boundary words by forward maximum matching method. If matching successfully, then the word divides into two characters. Follow figure 3, 4:

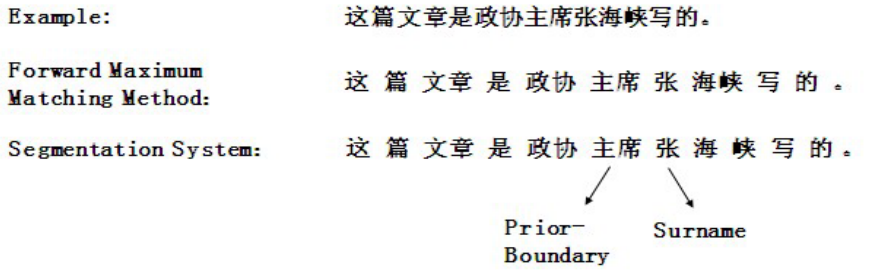



**Fig. 3.** Segmentation System

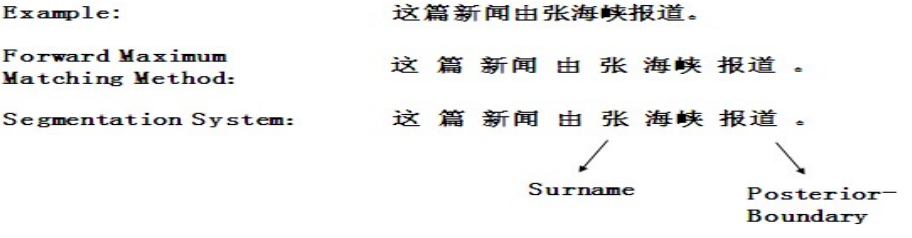



**Fig. 4.** Segmentation System

## 2.2 Role Tagging

Based on January 1998 People's Daily corpus, this paper defined the character commonly used in text, the rare character and the dislocation character.

1. The character commonly used in text is appeared 10 times or more in the People's Daily corpus.

2. The rare character is appeared 9 times or less in the People's Daily corpus.

3. The dislocation character belongs to the character commonly used in text, and the proportion of forming a word by itself[8] is less than 10%. Moreover, because the first name is composed of 1-2 character, the dislocation character only closely follows the surname.

The corpus statistics shows: the corpus is composed of 4555 character, 2028 character can become the dislocation character (44.52%), 1692 character can become the rare character (37.15%). Two kinds of character can occupy up to 81.67%, so can cover the range of characters which maybe a name. Other statistics as figure5:

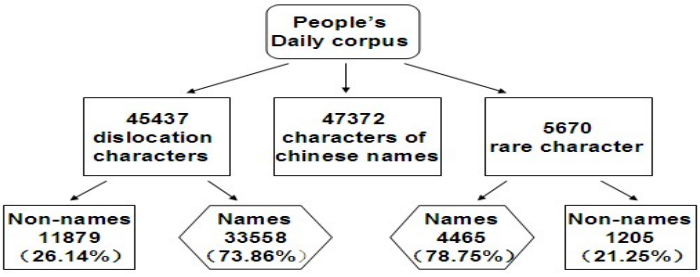



**Fig. 5.** The Corpus Statistics

Analysis of these data: 47372 characters of all names have 33558 dislocation characters and 4465 rare characters. Dislocation characters and rare characters are characters of names as high as 80.25%. And dislocation character or rare character is non-name about a quarter. So they are major component of names, also can be a difference between non-name and name.

Therefore, role tagging [1] of names is consisted of the surname, the context boundary, the dislocation character and the rare character. Role tagging table of names follow Table 1:

**Table 1.** Role tagging table

| Code | Define | Example |
|---|---|---|
| X | Surname | 张 天 平 老 师 |
| L | Prior-boundary | 学 生 主 席 郑 海 |
| F | Posterior-boundary | 记 者 红 兰 光 摄 |
| C | Both prior-boundary and posterior-boundary | 红 兰 光 和 张 天 平 |
| H | All except X, L, F, C | |
| Z | Rare character | 李 岚 清，樊 如 钧 |
| U | Dislocation character | 陈 雁，张 严 |
| N | All except Z,U | 张 严，樊 如 钧 |
| E | Character of name | 张 严，林 如 心 |
| G | All except L,F, E | |

## 2.3 Recognition Process

CRFs training and testing use CFR++ 0.53[2]. Recognition process mainly consists of three processes.

### 2.3.1 Label People's Daily Corpus by Role Tagging Table 1

CRFs toolkit required that corpus must be formed training text.

1. According to forward maximum matching method and Xue Nianwen [8] four-bit word, the chinese characters of sentences label tagging, then it get the first layer of the label.
2. According to a set of the surnames, a set of prior-boundary words, a set of posterior-boundary words and role tagging of X, L, F, C, H, and then it get the two layer of the label.
3. According to role tagging of Z, U, N, and then it get the three layer of the label.
4. According to a set of prior-boundary words, a set of posterior-boundary words and role tagging of E, L, F, C, G, and then it gets the four layer of the label and forms the training text.

### 2.3.2 Feature Template of CRFs

Building feature template is a complex process, and it ultimately achieves satisfactory test results. This paper uses two feature templates in experiments. Follow table2, 3:

**Table 2.** Feature template 1

| Single feature template | Composite feature template |
|---|---|
| U0:%x[-2,0] | U0:%x[-2,2]/%x[-2,0] |
| U1:%x[-1,0] | U1:%x[-1,2]/%x[-1,0] |
| U2:%x[0,0] | U2:%x[0,2]/%x[0,0] |
| U3:%x[1,0] | U3:%x[1,2]/%x[1,0] |
| U4:%x[2,0] | U4:%x[2,2]/%x[2,0] |
| | U5:%x[-2,2]/%x[-2,3] |
| | U6:%x[-1,2]/%x[-1,3] |
| | U7:%x[0,2]/%x[0,3] |
| | U8:%x[1,2]/%x[1,3] |
| | U9:%x[2,2]/%x[2,3] |
| | U10:%x[-2,2]/%x[0,3] |
| | U11:%x[-1,2]/%x[0,3] |

**Table 3.** Feature template 2

| Single feature template | Composite feature template |
|---|---|
| U00:%x[-2,0] | U10:%x[-2,2]/%x[-2,0] |
| U01:%x[-1,0] | U11:%x[-1,2]/%x[-2,0] |
| U02:%x[0,0] | U12:%x[0,2]/%x[0,0] |
| U03:%x[1,0] | U13:%x[1,2]/%x[1,0] |
| U04:%x[2,0] | U14:%x[2,2]/%x[2,0] |

### 2.3.3 Recognize Names

First, the text need to be formed the test text which CRF + + toolkit required, and then it uses the tool to predict, finally recognize names.

1. The text needs segmentation by the segmentation system of this paper.
2. According to forward maximum matching method and Xue Nianwen [8] four-bit word , the chinese characters of sentences label tagging, then it get the first layer of the label.
3. According to a set of surnames, a set of prior-boundary words, a set of posterior-boundary words and role tagging of X, L, F, C, H, and then it get the two layer of the label.
4. According to role tagging of Z, U, N, and then it gets the three layer of the label and forms the test text.
5. It predicts the fourth layer of the label by the CRFs tool; finally Chinese characters that consist of continuous E are names.

## 3 Experiments

Date of the open test come form the Second International Chinese Word Segmentation Bakeoff. The experimental results are shown in table 4:

**Table 4.** Experimental results

| classification | Close test 1 | Open test 1 | Open test 2 | Open test 3 |
|---|---|---|---|---|
| Training data sources are the People's Daily (Bytes) | January 1998 (8,621K) | January 1998 1-20 (5,552K) | January 1998 (8,621K) | January 1998 (8,621K) |
| Testing data sources (Bytes) | January 1998 of People's Daily (8,621K) | January 1998 of People's Daily 21-31 (3,069K) | Bakeoff-2005 Testing data (336K) | Bakeoff-2005 Testing data (336K) |
| Feature template | Template 1 | Template 1 | Template 1 | Template 2 |
| Number of names | 16380 | 7352 | 776 | 766 |
| Number of recognition names | 16188 | 7213 | 966 | 724 |
| Number of correct recog-nition names | 15941 | 6801 | 708 | 638 |
| Precision | 98.47% | 94.29% | 73.29% | 88.12% |
| Recall | 97.32% | 92.51% | 91.24% | 82.22% |
| F-value | 97.89% | 93.39% | 81.76% | 85.07% |

Precision = Number of correct recognition names / Number of recognition names *100%;

Recall = Number of recognition names / Number of names *100%;

F-value = Accuracy rate * Recall rate *2/ (Precision + Recall)*100%;

The experimental data shows:

1. Precision, recall and F-value in the closed test can be up to 98.47%, 97.32% and 97.89%,

2. Open test 2 compares to closed test 1 and open test 1. The reason that affects recognition results is mainly caused by two aspects. The segmentation errors of names are main reason. The segmentation system bases on a set of the surnames, a set of prior-boundary words and a set of posterior-boundary words. It can reduce errors, but is unable to avoid. Example for "张 海峡 的 人生 充满 喜剧 色彩。", "的" bellows "张海峡", but isn't in boundary library and results in that "海峡" can't separate. Then the model is difficult to identify such names; Many common words are cut apart by segmentation system. For example: "容易". Due to "容" is the surname, and when it meets the condition of segmentation system, "容易" is divided into two characters and "容" or"易" maybe dislocation character. Then the CRFs recognizes "容 易"as a name, and decline the precision. However, this can be avoided by subsequent processing.

3. Open test 2 uses the feature template 1 which introduces the dislocation character and the rare character. The recall of open test 2 compared with the open test 3, increases by 9.02%. Precision can increase by subsequent processing. So this paper is aimed at introducing features of name characters which is practicable.

## 4 Conclusions and Future Work

This paper depended on the context boundary of names to reduce the name segmentation errors, and explored law of the names (Dislocation character and rare character). From the experimental results, this method is practicable, and names without surname also can be recognized. The next step will extend law of Chinese person-name and explores law of foreign names.

**Acknowledgement.** The research work is supported by National Natural Science Foundation of China(60872133), Beijing Municipal Natural Science Foundation(4092015), The National Key Technology R&D Program (2011BAH11B03) and Scientific Research Common Program of Beijing Municipal Commission of Education (KM201110772021).

## References

1. Zhang, H., Liu, Q.: Automatic Recognition of Chinese Person Name Based on Role Tagging. Chinese Journal of Computers 27(1), 86–91 (2004)
2. CRFs++0.53[OL].[2009], http://sourceforge.net/projects/crfpp/
3. Zhou, G., Su, J.: Named Entity Recognition using an HMM-based Chunk Tagger. In: Proc. of the 40th Annual Meeting of the ACL, Philadelphia, pp. 473–480 (2002)

4. Oliver, B., Josef, O.F., Hermann, N.: Maximum Entropy Models for Named Entity Recognition. In: Proc. of the Conference on Computational Natural Language Learning, Edmonton, Canada, pp. 148–151 (2003)
5. Zhang, Y., Xu, B., Cao, Y., Zong, C.: Research of Chinese Person Names Identification Based on Surname. Computer Engineering and Applications (4), 62–65 (2003)
6. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields Probabilistic Models for Segmenting and Labeling Sequence Data. In: International Conference on Machine Learning (2001)
7. Xiang, X.: Chinese Named Entity Recognition based on Conditional Random Fields. Xiamen University (2006)
8. Xue, N.: Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing 8(1), 29–48 (2003)
9. Liang, N.: CDWS the Mordern Printed Chinese Distinguishing Word System. Journal of Chinese Information Processing 1(2), 44–52 (1987)

# An Academic Search and Analysis Prototype for Specific Domain

Zhiqiang Gao, Yaocheng Gui, Man Zhu, and Zhisheng Huang

[1] School of Computer Science and Engineering,
Southeast University, Nanjing 211189, P.R. China
{zqgao,yaochgui,mzhu}@seu.edu.cn
http://iws.seu.edu.cn
[2] Division of Mathematics and Computer Science, Faculty of Sciences,
Vrije University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam,
The Netherlands
huang@cs.vu.nl
http://www.cs.vu.nl/~huang/

**Abstract.** There exist several powerful and popular academic search engines, such as Microsoft Academic Search, Google Scholar and CiteSeerX, etc. However, query answering is now being required by users in addition to existed keyword and semantic search. Academic search and analysis is based on techniques of keyword and semantic search, and implements part of the query answering functions. It can provide but not limited to the following services: ranking researchers, mining university and company relationships, finding research groups in an affiliation, evaluating importance of research communities, recommending similar researchers, and so on. This paper introduces an academic search and analysis prototype ALL4ONE for specific domains of information science, computer science and telecommunication science. It is focused on solving the two major difficulties of affiliation name and researcher name disambiguation, as well as domain specific large scale ontology construction.

**Keywords:** Semantic Search, Query Answering, Name Disambiguation, Information Extraction, Ontology Learning.

## 1 Introduction

There are two major difficulties for domain specific academic search and analysis, i.e. affiliation name and researcher name disambiguation, as well as domain specific large scale ontology construction.

**(1) Name Disambiguation**
Because there are no international standards for publishing research papers, there are various *incompleteness*, *errors* or *ambiguities* in published papers. For example: 1) *Information missing.* We test 149,804 research papers from ACM digital library, and find that 72,218 of them are published without researcher affiliations. 2) *Name ambiguity.* Author name may refer to various researchers

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 171–178, 2012.

in published papers. For example, "Zhang, X." may be recognized as "Zhang, Xiang" or "Zhang, Xi". This situation becomes worse when author name is abbreviated, especially for Chinese researchers. Because a lot of them share one common family name, such as "Zhang", "Wang", etc. 3) *Large numbers of co-authors.* Some papers contain hundreds of authors. Given as an example, the paper "Measurement of CP-violating asymmetries in B0 decays to CP eigenstates" from EI digital library contains 647 authors. Existence of these research papers increases the difficulty for researcher name disambiguation by co-author features.

We have tested several academic search engines for person search, such as Microsoft Academic Search (MAS) [1] in April 2011. We used the researcher names of our university (Southeast University). The experimental results are not satisfiable. Part of the results is given in Table 1. MAS performs well when researcher name is rarely used, such as "Qi, Guilin" and "Qu, Yuzhong". The results are useless if the researcher name is commonly used. For example, "Gao, Zhiqiang" is a very common Chinese name, MAS returns that "Gao, Zhiqiang" of Southeast University has 65 papers in total. But, we know that among the 65 papers, only 8 papers are correct. The rest papers are referred to at least three other researchers with the same name. In Table 1, the first two columns are search conditions while the third column is the number of returned papers from MAS. The last column is the correct number of research papers written by the researchers from Southeast University.

**Table 1.** Researcher Name Ambiguity in Microsoft Academic Search

| Researcher | Affiliation | # Total Papers | # Correct Papers |
|---|---|---|---|
| Gao, Zhiqiang | Southeast University China | 65 | 8 |
| Li, Huiying | Southeast University China | cannot find | NA |
| Hu, Wei | Southeast University China | cannot find | NA |
| Cheng, Gong | Southeast University China | 26 | 15 |
| Zhang, Xiang | Southeast University China | cannot find | NA |
| Qi, Guilin | Southeast University China | 64 | 64 |
| Qu, Yuzhong | Southeast University China | 47 | 47 |

**(2) Ontology Construction**

Ontology construction is the bottleneck for academic search and analysis. The difficulties comes from: 1) There are few usable ontologies or classification systems. To the best of our knowledge, in computer science, there is ACM Computer Classification System 98 (ACM CCS 98) [2]. However, There is no such classification systems for information science and telecommunication science. Although there are some thesaurus and vocabularies in the Web. Additionally, there are only about 1000 nodes in ACM CCS 98, and it can not be used directly.

[1] http://academic.research.microsoft.com/

[2] http://www.acm.org/about/class/1998/

2) Ontologies for academic search and analysis must be tuned to fit the developments of specific domain,which means that we need to build large scale ontology. Given as an example, the depth of ACM CCS 98 is five, and we can not find the nodes like "semantic web", "decision tree learning", "probabilistic graphical model",etc.

The remainder of this paper is organized as follows. Section 2 describes the name disambiguation algorithms, and Section 3 illustrates ontology construction framework. Section 4 demonstrates major academic search and analysis services, followed by related works and conclusions in section 5 and section 6.

## 2 Name Disambiguation

Name disambiguation is also know as *name correspondence*, *name identification* or *name co-reference*. In this section, we introduce affiliation name disambiguation and researcher name disambiguation.

### 2.1 Affiliation Name Disambiguation

Affiliation names are published in research papers, patents, projects, as well as personal web pages of researchers. One affiliation name may be "National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China". This string contains affiliation name and other irrelevant details. In which, "Southeast University" is the affiliation name, "National Mobile Communications Research Laboratory" is a laboratory in the affiliation, and "Nanjing 210096, China" contains the city name, the post code, and the country name. Another affiliation name may be "Southeast University". Although the second is a sub-string of the first, we know that they denotes the same affiliation "Southeast University".

Affiliation names may be separated by ",", and we call this kind of structure as the standard representation of an affiliation name. Unfortunately, not all of the affiliation names are represented in the standard way. Various problems may occur, such as: 1) *Separation mistake.* The standard separator may miss or be placed in wrong positions. Given as an example, the string "Southeast University National Mobile Communications Research Laboratory 210096 Nanjing China" misses all the separators. 2) *Name abbreviation.* For example, the string "Dept. of Radio Eng., Southeast Univ., Nanjing, China" denotes "Southeast University", while the word "university" is abbreviated as "Univ.". Another kind of abbreviation is like "UCLA Computer Science Department", which denotes "University of California at Los Angeles". 3) *Branch school:* Some of the affiliation names contain geography names. For example, the string "Department of Materials Science and Engineering, University of California, Los Angeles, CA 90095-1595" may be identified as "University of California". While the "University of California" is a university system which contains 9 independent members. The affiliation name in this example should be "University of California at Los Angeles". 4)*Name alias.* Some affiliations have more than one name, for example, "Peking University" and "Beijing University" denote the same affiliation.

The algorithm for affiliation name disambiguation is given below: 1) Most of the affiliation names are given in standard representation. Through computing repetitive strings, we extract a set of affiliation names. 2) We clean up the set of affiliation names , and check the *branch school* problems and the *name abbreviations* problems by using a manually constructed knowledge base. 3) We match un-recognized affiliation names with a set of feature sequences, which are extracted by excluding stop words from affiliation names, such as "the", "of", "at" and so on. We extracts totally 12,177 distinct affiliation names, and the experimental precision reaches 98.5%.

### 2.2 Researcher Name Disambiguation

Different researchers may use the same name in their publications. This situation frequently occurs among Chinese researchers, whose English names are translated from Chinese characters. Many Chinese characters have the same English translation. As a result, one English name often corresponds to multiple Chinese names. in addition, researcher names may be represented in several abbreviated formats. Identifying different researchers with the same English names, and identifying different names referred to the same researchers, are the two main tasks in researcher name disambiguation.

The standard structure of a researcher name representation is as "Mitchell, Tom M.". In which "Mitchell" is the family/last name of the researcher, "Tom" is the given/first name and "M." is the middle name or its abbreviation. However,there are various kinds of researcher name representations in research papers, including:1) *Name abbreviation.* The middle name and the given name may be represented by the first character of the name string such as "Mitchell, T. M.", and in some cases the middle name may be omitted, such as "Mitchell, T.". We have to identify whether "Mitchell, Tom" and "Mitchell, T." refer to the same researcher. 2) *Inverse name.* When the name string is represented without separators such as "Zhang Xi". It is difficult to recognize which is the family name and which is the given name. Because in China and some other Asian countries the first part is family name, while in European and USA the last part is family name. Further, family name database in China contains both "Zhang" and "Xi", and this makes the name identification much more difficult.

The algorithm for researcher name disambiguation is described as below: 1) For identifying different names referred to the same researcher, we normalize the name representation. Family names are recognized by using American surname database and Chinese family name database. 2) For identifying different researchers with the same English name, we cluster the normalized names by assuming that there is only one researcher in an affiliation in a specific domain. When researcher names do not have corresponding affiliation names, we assume that the same group of co-authors can be distinguished. 3) We find that some famous researchers always publish papers in more than one affiliation. We consider the publication time of their publications. If no obvious contradiction is found, we cluster these researcher names together.

We have extracted totally 12,681,416 distinct researcher names. We select more than one hundred researchers from our university (Southeast University, SEU) to form a test data set. It reaches the precision of 95%. Our prototype performs better than MAS, and part of the experimental results are given in Table 2.

**Table 2.** Researcher Name Disambiguation Between Our Prototype ALL4ONE and Microsoft Academic Search (MAS)

| Search Terms | | MAS Results | | ALL4ONE Results | |
|---|---|---|---|---|---|
| **Researcher** | **Affiliation** | **# Total Papers** | **# Correct Papers** | **# Total Papers** | **# Correct Papers** |
| Gao, Zhiqiang | SEU | 65 | 8 | 13 | 13 |
| Li, Huiying | SEU | cannot find | NA | 6 | 6 |
| Hu, Wei | SEU | cannot find | NA | 13 | 13 |
| Cheng, Gong | SEU | 26 | 15 | 15 | 15 |
| Zhang, Xiang | SEU | cannot find | NA | 8 | 8 |
| Qi, Guilin | SEU | 64 | 64 | 30 | 30 |
| Qu, Yuzhong | SEU | 47 | 47 | 58 | 58 |

## 3 Ontology Construction

It is well known that ontology learning can hardly generate an ontology that can be used directly. Domain experts should participate in the process of ontology construction. Ontology construction is divided into three steps, i.e. term extraction, subsumption relation learning and ontology integration.

### 3.1 Term Extraction

Extracting research interests/areas from personal web pages of researchers have been done in our previous work [1], and we have found that only a small part of researchers published their personal web pages. Therefore, we turn to academic publications such as research papers and patents to extract terms in which research interests/areas are represented indirectly. We build the corpus with titles and abstracts of research papers and patents. We leverage the techniques of automatic term recognition (ATR) to extract terms from the corpus. We implement a $C$-value method to combine linguistic and statistical information to extract nested multi-word terms [6].

The ATR algorithm is divided into the following steps: 1) The corpus is tagged by GATE [3]. We leverage a rule based linguistic filter to extract noun phrases as candidate terms. Most of the previous works agree that terms are noun phrases linguistically [7]. 2) We measure the termhood of candidate terms by $C$-value, and rank these candidate terms according to $C$-value. 3) We cut off the long tail of the candidate terms which occurs only once or twice. 4) Candidate terms are evaluated by domain experts. About 110,000 candidate terms are chosen, and Table 3 shows the top 30 extracted candidate terms.

[3] http://gate.ac.uk/

**Table 3.** Top 30 Extracted Candidate Terms

| | Term | Frequency |
|---|---|---|
| 1 | neural network | 48965 |
| 2 | monte carlo | 33803 |
| 3 | finite element | 35620 |
| 4 | sensor network | 18613 |
| 5 | information system | 24099 |
| 6 | genetic algorithm | 21860 |
| 7 | control system | 26064 |
| 8 | numerical simulation | 30214 |
| 9 | wireless sensor | 15051 |
| 10 | wireless sensor network | 9880 |
| 11 | electron microscopy | 41275 |
| 12 | management system | 16280 |
| 13 | differential equation | 27947 |
| 14 | monte carlo simulation | 13356 |
| 15 | wireless network | 14090 |

| | Term | Frequency |
|---|---|---|
| 16 | web service | 9707 |
| 17 | field theory | 12983 |
| 18 | information retrieval | 14188 |
| 19 | natural language | 13459 |
| 20 | signal processing | 17713 |
| 21 | software engineering | 13004 |
| 22 | support vector | 12030 |
| 23 | image processing | 13910 |
| 24 | computer simulation | 19215 |
| 25 | communication system | 15687 |
| 26 | finite element method | 11915 |
| 27 | support vector machine | 8073 |
| 28 | artificial neural network | 8481 |
| 29 | semantic web | 8358 |
| 30 | mobile robot | 9410 |

### 3.2 Subsumption Relation Learning

We need to construct subsumption relations between terms accepted by domain experts. There are three kinds of sources that are used to learn subsumption relations: 1) Controlled terms provided by digital libraries. Some digital libraries provide controlled terms for researchers to tag their publications, and users can query by terms and get the comments and subsumption relations between them. 2) Book contents given by online bookstores. Book introductions are usually available on online bookstores, such as Amazon.com [4]. We crawl the book introductions guided by the classification of bookstores and extract the subsumption relations between terms. 3) Wikipedia[5] pages. We query Wikipedia by terms and analyze the matched pages. Some of the Wikipedia pages give the sub-domains of its topic, which can be considered as subsumption relations between terms. 4) Cycles are detected automatically and appropriate concepts and subsumption relations between terms are chosen with the help of domain experts.

### 3.3 Ontology Integration

The taxonomy provided by domain experts is a classification system of 1435 nodes in total with 282 internal nodes from Level 0 to Level 2 and 1153 leaf nodes from Level 3. We match between the nodes of the given taxonomy and the terms extracted from the corpus. Finally, we construct the large scale domain specific ontology for ICT domain, which contains 7814 nodes (concepts) in total.

## 4 Experiments

The academic search and analysis prototype provides 5 major queries in addition to keyword and semantic search: *1)Query for distribution of research interests*

[4] http://www.amazon.com/
[5] http://www.wikipedia.org/

*on geography maps; 2)Query for the evolution of research interests of affiliations, 3)Query for evolution of research interests of researchers; 4)Query for co-author relations; 5)Query for distribution of collaboration.*

## 5 Related Works

Etzioni et al. developed KnowItAll [8], a system that aims to automate the tedious process of extracting large collections of facts from web in an autonomous, domain-independent, and scalable manner. Cimiano et al. proposed PANKOW [9] which is a method employing an unsupervised, pattern-based approach to categorize instances with regard to an ontology. C-PANKOW [10] has overcome several shortcomings of PANKOW. Text-to-Onto is a semi-automated system that uses linguistics and statistics-based techniques to perform its ontology learning tasks [12,13]. Navigli and Velardi [4] developed a system called OntoLearn, trying to extract ontology from web sites. Dolby et al. [3] provide a process to automatically extract a domain specific vocabulary from unstructured data in enterprise. Speretta et al. [2] studied the distribution of the small set of tagged publications in the CiteSeer collection over ACM's Computer Classification System. Smeaton et al. [5] provided a content analysis of all papers published in the proceedings of SIGIR conferences in 25 years until 2002 using information retrieval approaches.

## 6 Conclusions

We introduce an academic search and analysis prototype All4One for specific domains of information science, computer science and telecommunication science. There are totally 12,177 distinct affiliations, 12,681,416 distinct researchers, 7,800 concepts and 100 million facts in the knowledge base. We put forward affiliation name and researcher name disambiguation algorithms, which performs better than Microsoft Academic Search. We suggest a framework for ontology construction to integrate taxonomy provided by domain experts and terms as well relations extracted from the corpus of academic publications.

**Acknowledgement.** We gratefully acknowledge funding from the National Science Foundation of China under grants 60873153, 60803061, and 61170165.

## References

1. Gao, Z.Q., Zhu, W.Y., Qu, Y.Z., Huang, Z.S.: Analyzing Distribution and Evolution of Research Interests by Term Extraction and Ontology Learning. In: Proceedings of the 9th International Conference on Web-Age Information Management (2008)
2. Speretta, M., Gauch, S., Lakkaraju, P.: Using CiteSeer to Analyze Trends in the ACM's Computing Classificatin System. In: Conference on Human System Interaction (HSI), Poland, pp. 571–577 (2010)

3. Dolby, J., Fokoue, A., Kalyanpur, A., Schonberg, E., Srinivas, K.: Extracting Enterprise Vocabularies Using Linked Open Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 779–794. Springer, Heidelberg (2009)
4. Navigli, R., Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. Comput. Linguist. 30(2), 151–179 (2004)
5. Smeaton, A.F., Keogh, G., Gurrin, C., McDonald, K., Sodring, T.: Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century? SIGIR Forum 36(2), 39–43 (2002)
6. Frantzi, K.T., Ananiadou, S., Mima, H.: Automatic recognition of multi-word terms: the C-value/NC-value method. Int. J. on Digital Libraries (JODL) 3(2), 115–130 (2000)
7. Kageura, K., Umino, B.: Methods of Automatic Term Recognition: A Review. Terminology 3(2), 259–289 (1996)
8. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-Scale Information Extraction in KnowItAll. In: Proceeding of 13th World Wide Web Conference (2004)
9. Cimiano, P., Handschun, S., Staab, S.: Towards the Self-Annotating Web. In: Proceedings of the 13th World Wide Web Conference (2004)
10. Cimiano, P., Ladwig, G., Staab, S.: Gimme' The Context: Context-driven Automatic Semantic Annotation with C-PANKOW. In: Proceedings of the 14th World Wide Web Conference (2005)
11. Guha, R., Mccool, R., Miller, E.: Semantic Search. In: Proceeding of the 12th World Wide Web Conference, New York, NY, USA (2003)
12. Cimiano, P., Staab, S.: Learning Conception Hierarchies form Text with a Guided Agglomerative Clustering Algorithm. In: Proceedings of Workshop on Learning and Extracting Lexical Ontologies with Machine Learning Methods, Bonn, Germany
13. Maedche, A., Staab, S.: Discovering Conceptual Relations from Text. In: Proceedings of the 8th International Conference on Conceptual Structures, Darmstadt, Germany (2000)

# Learning Chinese Entity Attributes from Online Encyclopedia

Yidong Chen, Liwei Chen, and Kun Xu

Peking University, Beijing, China
{chenyidong,clwclw88,xukun}@pku.edu.cn

**Abstract.** Automatically constructing knowledge bases from free online encyclopedias has been considered to be a crucial step in many internet related areas. However, current research pays more attention to extract knowledge facts from English resources, and there is less work concerning other languages. In this paper, we describe an approach to extract entity attributes from a free Chinese online encyclopedia–HudongBaike[1]. We first identified attribute-value pairs from HudongBaike pages that are featured with InfoBoxes, which in turn can be used to learn which attributes we should pay attention to for different HudongBaike entries. We then adopted a keyword matching approach to identify candidate sentences for each attribute in a plain HudongBaike article. At last, we trained a CRF model to extract corresponding values from these candidate sentences. Our approach is simple but effective, and our experiments show that it is possible to produce large amount of <S,P,O> triples from free online encyclopedias which can be then used to construct Chinese knowledge bases with less human supervision.

**Key words:** online encyclopedia, knowledge base, entity attribute extraction,

## 1 Introduction and Related Work

Recent years have witnessed the unprecedented development of world wide web (WWW), which subsequently creates unforeseen volume of data, most of which are unstructured while few are manually structured but not directly machine-readable. How to effectively manage and make these data usable and accessible has become a crucial task in many research areas, including information retrieval (IR), natural language processing (NLP), semantic web[1], database and so on.

A key challenge emerged here is knowledge base construction. Different from plain text documents, knowledge bases not only store knowledge facts as a list of concepts or records in documents, but also explicitly code the meaning of concepts and relationship between concepts. Knowledge bases are usually arranged in a computer-readable form for ease of performing automatic deductive reasoning, which enables improving the quality of semantic information retrieval.

[1] http://www.hudong.com/

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 179–186, 2012.

Despite of different scale or domain, most existing knowledge bases are manually created, e.g., WordNet[2], Cyc or OpenCyc[3], GeneOntology[4] and so on. However, manually constructing knowledge bases is a timing consuming and labor intensive task thus expensive to update or adapt to new domains. More importantly, compared to the increasing demand in real world, these are of a small scale and continuous human effort is needed to keep them up to date.

With the development of Web 2.0, there are more and more collectively created online encyclopedias emerging in a specific domain or open domain, e.g., Wikipedia[5], which is edited and maintained by a huge but collaborative open community of volunteers, currently accommodates over 3 million articles (until June 2011) and is still growing. Compared to unstructured free text, these online encyclopedias are edited in a more formal style and have been manually structured by volunteers in the aim of better interpreting target concepts. For example, each article is supposed to explain one topic or concept from different aspects and usually is fertilized with links to related articles or external resources. Sometimes volunteers even create a summary box, Infobox(shown as right of figure 1), containing different attributes for an article which provides a clear and concise look for the target.

The deep involvement from the volunteer community makes online encyclopedias easier to extract knowledge facts and further set up relations. Recent research has focused on community-based ontology building. They try to extract knowledge facts from English version of Wikipedia and build a semantic web accordingly which has benefited many related research[6][7].

YAGO[6] and DBpedia[7] are two huge semantic knowledge bases based on Wikipedia. They extract knowledge facts from the structured content-categories and Infoboxes. Categories provides "is_a" relations while Infoboxes provides other $< S, P, O >$ triples.

Fei Wu, Raphael Hoffmann, and Daniel S. Weld developed Kylin[8][9] system. They not only use the structured content, but also try to extract triples from the unstructured text of Wikipedia. In their work, articles with Infobox are used as training data to learn the correspondence between summarized entries in Infobox and the document text. The trained models are then applied to extract knowledge facts from the articles without Infoboxes. All these efforts mainly focused on the English part of Wikipedia.

In China, there are also several organized online encyclopedias, such as HudongBaike, BaiduBaike and SosoBaike , and knowledge base construction also attracts more and more attention like [10][11]. However there hasn't been any work ever concerning building a web scale semantic knowledge base like YAGO in Chinese language, especially constructed in a near automatic fashion with less human involvement. Our work is aiming to solve this problem. We will focus on how to extract knowledge facts from Chinese online encyclopedias which can be used to construct a Chinese semantic knowledge base. Our model first extract concept facts from InfoBoxes in HudongBaike pages, learn the correspondence between InfoBoxes and its corresponding text document, and finally extract concept facts from plain HudongBaike pages. We show that it is pos-


Abbeville County 刘德华

**Fig. 1.** An InfoBox from Wikipedia; An InfoBox from HudongBaike.

sible to extract knowledge facts from Chinese online encyclopedia pages with less supervision. In the rest of the paper, we will first describe how we extract knowledge triples from Infoboxes (Section 2) and then introduce how we learn to process the pages that do not contain Infoboxes (Section 3). In Section 4, we describe our experimental setup and results; we conclude in Section 5.

## 2 Structured Data Extraction

As we discussed before, compared to other resource on the Web, most online encyclopedias are collaboratively edited and filtered by volunteers, thus contain less noises and more regular structures, which are considered to be much easier to extract knowledge facts. Especially, similar to Wikipedia, many HudongBaike pages are also featured with a structured summary box, *Infobox*, which can be conveniently parsed into attribute-value pairs. In this section, we first introduce how to extract category information from HudongBaike pages and parse their Infoboxes into attribute-value pairs.

After examining the style of HudongBaike pages, we find that it is feasible to extract category information and Infobox (if exists) directly from the HTML

source code. Below is part of the source code for the Infobox illustrated in Figure 1:

```
    <div id="docinfotemplettable" class="module m-t10">
  <table class="table">
    <tbody
      <tr>
        <td align="right">中文名: </td>
        <td style="width:190px;">刘德华</td>
      </tr>
      <tr>
        <td align="right">籍贯: </td>
        <td style="width:190px;">中国香港</td>
      </tr>  ···
```

It is easy to find the location of Infobox in the page source code by searching the keyword "docinfotemplettable". And using tags like $< tr >$, $< td >$, we can efficiently get $< S, P, O >$ triples like <刘德华, 籍贯, 中国香港>.

Similar approaches are applied on category information and alias information which provide "is_a" and "has_alias_name" attributes.

## 3 Unstructured Data Extraction

Structured Infoboxes can be efficiently extracted with a high precision. However, there are limited number of pages featuring an Infobox, while most HudongBaike pages do not have a summary box thus can not be processed directly. In this section, we describe how to extract $< S, P, O >$ triples from the plain HudongBaike pages which can not benefit from an Infobox to get more hints.

The Kylin[8, 9] system tried to extract information from plain text of wikipedia first. Its method was a self-supervised fashion which uses the structured data extracted from wikipedia's Infoboxes as training data to train machine learning models, which are then used to extract structured data from web pages that do not have an Infobox. Their training data and test data are of almost the same form. The only difference is whether they have infoboxes. We use similar idea to deal with the plain HudongBaike pages.

### 3.1 Deciding Category Attribute

Before we process those plain pages, the first step is to identify what attributes we need to focus for each category, since different categories of subjects have different attributes. For example, subjects in the "人物" category have attributes like "姓名", "籍贯", "性别", "职业", etc., while subjects in the "国家" category have "国土面积", "人口", "官方语言", etc. as their attributes. We need to first decide what attributes we are interested in and will extract from the given article.

Due to the collective nature of the online encyclopedia data, different concepts from the same category might be annotated with different attributes. In

our work, a statistical approach is used to make the decision. We select all articles with Infobox from a given category. Then we count the times of each attributes occurs in the Infoboxes. For a certain attribute A, if it is not extracted in any parent categories of category C, and it occurs in more than 20% Infoboxes of category C, then it will be extracted in all articles of category C.

### 3.2 Deciding Candidate Sentence

After choosing proper attributes for each category, given a plain HudongBaike article, we need to locate appropriate candidate sentences to extract each attribute. In other words, for each sentence in the article, we need to decide whether it contains information for certain attributes and if yes, we should find out which attribute we are interested in. For example, the first sentence of an article is "xxxx年xx月，xxx出生于浙江杭州.", which is known to be a common description for attributes "出生年月" and "籍贯", thus will be parsed according to these two attributes, respectively.

Here, we explore different classification approaches to judge whether a sentence is a candidate sentence for a target attribute. For each target attribute within a category, a sentence classifier is trained to learn whether a sentence will be parsed for the target attribute in the current category. As a result, we need plenty of labeled sentences for training purposes.

*Training Data* : Recall that we have extracted large amount of attribute-value pairs from Infoboxes, which enables us to obtain labeled sentences automatically while avoiding costly human annotation. If an attribute occurs in an article's Infobox, we try to trace back its candidate sentence in the article. Then this candidate sentence will be used as a positive training example for this attribute's classifier.

For example, the triple <刘德华, 籍贯, 中国香港> has been extracted from the Infobox of article "刘德华". We score all sentences in this article according to the given triple. In our scoring, the object "中国香港" is the most important element. Besides, the subject "刘德华" is more important than the attribute (predication) "籍贯". For each sentence, the score is calculated as:

$$Weight = (match\_S + 1) * (match\_P + 2) * match\_O$$

$match\_S + 1$, $match\_P + 2$ and $match\_O$ are designed to emphasize different contributions from objects, predicates and subjects as:

- if the subject or its alias name appears in the sentence, $match\_S = 1$, otherwise 0,
- if the predication or its synonym appears in the sentence, $match\_P = 1$, otherwise 0,
- if the object appears in the sentence, $match\_O = 1$, otherwise 0,

If a sentence doesn't contain the target object, we think it is not a candidate for this attribute. If more than one sentence contain the object, the one that

contains the target subject is considered to be more likely a candidate sentence than the one which only contains the predicate. Finally, we choose the sentence with the highest score (not 0) to be the candidate sentence.

**Sentence Classifier**: However, it is still not easy to pick up an all-round classifier, since it is difficult to collect proper negative training examples for each classifier. The problem can be seen as an unbalanced text classification problem. We tried using random sentences as negative training examples, and our experiments show that traditional classifiers like Naive Bayes or SVM can not work well in this situation when using bag of words and POS tags as features. One possible reason is about the ill quality of our negative training examples that may contain positive examples since we perform a very strict criterion to automatically collect positive data which actually has a low recall.

In our work, we care more about how precise our extracted attribute-value pairs are. We finally choose a simple but effective approach to build classifiers which are based on positive training examples only. After performing word segmentation and POS tagging, we first calculate word frequency for all words appearing in the positive training examples and choose top $n$ keywords. We only choose nouns and verbs which appear at least half of the highest frequency of all the nouns and verbs. We then search the keywords and predicates (the attributes) in all testing sentences. If a sentence contains one of these words, it will be labeled as a candidate sentence for the corresponding attribute. In our experiment, we find that most keywords extracted from positive examples are the predicates' synonym or have a close relation to the predicates. This approach achieves a high precision but a slightly low recall. In our work, it is worth to make such a tradeoff from recall to precision since our aim is to extract accurate facts from plain articles.

### 3.3 Finding Attribute Value

After locating candidate sentences from the plain articles, our model should parse the candidate sentences into attribute-value pairs. For example, "xxxx年xx月, xxx出生于浙江杭州" is a candidate sentence for attribute "籍贯", our task is to extract the object "浙江杭州" to match with the target attribute "籍贯". After this step, we will get a completed <S,P,O> triple since the subject is the concept name that the current article discusses.

Conditional random fields (CRF model) has been recognized to perform well in sequential labeling tasks. Our setting can also be formulated in a sequential labeling format. We choose CRF++[2] in our experiments.

In CRF practice, there are usually two problems to solve. One is how to choose proper features, the other is how to get training data. Traditional textual features for English language are not suitable for Chinese text since in Chinese there are not letter case or word segments at all. We explored different features, and eventually decide to use four kinds of features: POS tags, named entities, whether in quotes and whether in angle quotes. POS tags contain noun, verb,

[2] http://crfpp.sourceforge.net/

adjective, adverb, etc. Named entities include person, location and organization names. Our experimental results show that the last three features work well for the objects.

Training data is easy to get by reusing the sentence classifiers' training set. Each sentence in the training set corresponds to a predicate and an object extracted from Infobox. It is turned to be a training example for CRF model by highlighting its object.

## 4 Experiments

In our experiment, we use ICTCLAS50[12], developed by Institute of Computing Technology, Chinese Academy of Sciences, as our word segmentation tool, and build the CRF model based on CRF++[13].

Our structured content extraction has a high precision of more than 95%. Most errors are caused by the web pages' own mistakes or editing errors in Infoboxes.

We perform pilot experiments on plain HudongBaike articles, focusing on category "国家", "人物" and its sub-category "演员". We choose articles without Infoboxes as our test data. We manually annotated more than 200 testing examples for the purpose of evaluation. We summarize the results in the Table 1:

**Table 1.** Extraction results for plain HudongBaike articles.

| category | Infobox # | Test # | attribute | precision | Recall |
|---|---|---|---|---|---|
| 人物 | 2054 | 100 | 出生年月 | 86.4% | 73.1% |
| | | | 籍贯 | 75.5% | 58.7% |
| | | | 毕业院校 | 76.4% | 75.0% |
| 国家 | 159 | 33 | 官方语言 | 86.9% | 81.1% |
| | | | 面积 | 70.8% | 60.7% |
| | | | 首都 | 79.0% | 53.6% |
| 演员 | 326 | 91 | 代表作品 | 90.8% | 65.0% |

As we expected, our model achieves high precisions in all three categories, some even goes up to 90%. We find that if the attribute values have an obvious form, it will achieve a higher precision. For example, the values for attribute "出生年月" are always time words and in attribute "代表作品", its values always contain "《》", these two attributes get precisions of 86.4% and 90.8%, respectively. It is not surprising that the precision of attribute "官方语言" is higher than that of "籍贯", "毕业院校" and "首都", since values for the latter attributes could be various named entities, their abbreviations or even out-of-vocabulary words while the former only contain different language names which are quite common and limited in number.

## 5 Conclusions and Future Work

In this paper, we present an approach to extract entity attributes from Chinese online encyclopedias. We first identified attribute-value pairs from HudongBaike pages that are featured with InfoBoxes, which in turn can be used as training data to parse plain HudongBaike articles. We extracted typical keywords for each category from the training data and then adopted a keyword matching approach to identify candidate sentences. We trained a CRF model to extract corresponding values from the selected sentences. This simple but effective approach can produce large amount of <S,P,O> triples which can be then used to construct Chinese knowledge base. However, in our experiments, we found that the quality of our extracted knowledge facts are greatly effected by the performance of Chinese word segmentation and named entity recognition tools. As we known, knowledge bases are considered to improve the performance of Chinese word segmentation and NER. An interesting direction will be to investigate how we can jointly push those tools with our knowledge base and in turn improve the quality of our extraction work.

## References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American (May 2001)
2. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. MIT Press (1998)
3. Matuszek, C., Cabral, J., Witbrock, M., DeOliveira, J.: An introduction to the syntax and content of Cyc. In: AAAI Spring Symposium (2006)
4. http://www.geneontology.org
5. http://www.wikipedia.org
6. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web 6(3), 203–217 (2008)
7. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web (7), 154–165 (2009)
8. Wu, F., Weld, D.S.: Automatically Refining the Wikipedia Infobox Ontology. In: Proceedings of the 17th International Conference on World Wide Web, pp. 635–644. ACM, New York (2008)
9. Wu, F., Weld, D.S.: Autonomously semantifying Wikipedia. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management. ACM, New York (2007)
10. Qu, Y., Cheng, G., Ji, Q., Ge, W., Zhang, X.: Seeking knowledge with Falcons. Semantic Web Challenge (2008)
11. Shi, F., Li, J., Tang, J., Xie, G., Li, H.: Actively Learning Ontology Matching via User Interaction. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 585–600. Springer, Heidelberg (2009)
12. http://ictclas.org/
13. http://crfpp.sourceforge.net/

# A Mathematical Model for Human Flesh Search Engine

Lei Zhang[1], Yuankang Liu[1], and Juanzi Li[2]

[1] Graduate School at Shenzhen, Tsinghua University
Shenzhen, P.R. China
zhanglei@sz.tsinghua.edu.cn, yk-liu09@mails.tsinghua.edu.cn
[2] Department of Computer Science and Technology
Tsinghua University,Beijing, P.R. China
ljz@keg.cs.tsinghua.edu.cn

**Abstract.** Human Flesh Search Engine (HFSE) is a phenomenon of massive researching using Web media such as blogs and forums with a purpose of exposing personal details of perceived misbehaviors. With the increasing efficiency and convenience of the Web, Human Flesh Search Engine is becoming more and more powerful and able to discover information which is "mission impossible" by other conventional means. Existing research work focuses on legal or privacy issues of this emerging tool, while we aim at building a mathematical model to understand the evolution of the search process and hence to evaluate the power of this massive collaboration intelligence. Viewing the initiator and target of a search campaign as source/destination nodes in the social network, a HFSE searching is modeled as a probabilistic flooding routing algorithm in the graph. Typical Human Flesh Search Engine cases and simulation-based experiments are used to evaluate the validity of the model and provide new insights for HFSE.

**Keywords:** Human Flesh Search Engine, social network.

## 1 Introduction

Human Flesh Search Engine (HFSE) is literally translated from the Chinese phrase "RenRouSouSuo", which means a kind of mass searching campaign using human resource collaboration on the Web. Unlike conventional search engines, like Google, which may be used to target almost anything, HFSE shows special interest on the targets: public misbehaviors.

Among the numerous cases of HFSE, the uncovering of fake pictures of "South China Tiger" is probably the most famous and typical one. In October 2007, afarmer published over 30 pictures of a wild "South China Tiger"(SCT), a kind of endangered animal which hasn't been seen in wild since 1980's. Netizens noticed that all tigers in all the pictures looked the same in postures, shades and stripes on the fur. Their collaborative work proved that the proclaimed "South China Tiger" was actually a flat 2D paper tiger and pointed out the original source of the tiger came from a Spring-Festival Picture published 10 years ago. The event shows the great power of crowds of netizens online and even attracted the attention of the Science magazinetwice, first to publish the photos[21], then to confirm their fabrication[22].

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 187–198, 2012.

It took the netizens more than a year to win the victory over the faked "South China Tiger" in 2007. However, only one year later, in October 2008, the netizens only used less than[13] hours to find out the identity and full public/personal information of the target when a video of a middle-aged man harassing a teenager girl was posted on the Web.

Given the significant efficiency of HFSE, hiding personal information on the Web becomes extremely difficult if you unfortunately become the target of a HFSE campaign. Therefore, there have been numerous discussions of the legal and privacy issues of the problem. The Times describes it as Chinese vigilantes hunting victims by interviewing some HFSE targets who received hate mails and threats after their personal information is exposed on the Web[10]. Xinhua News Agency even views it as a virtual lynching and warns people to behave themselves appropriately to avoid possible HFSE punishment[1].CNN goes one step further and considers people participating HFSE as lynching mob [5].

It is necessary to build a mathematical model for the following purposes: First, toacquire betterunderstanding of the structure and evolution procedure of HFSE; Second, to provide a scheme to evaluate the power of HFSE and estimate the power grown trend in the future; Third, to identify parameters affecting the success of a HFSE campaign; Fourth, to evaluate HFSE vulnerability against careless of maliciousmisleading information.

Despite the fruitful discussions on legal issues of HFSE, very little work has been done on approaching a mathematical model for it. Research work on social search[12][27], crowd-sourcing[14], and some attempts on theoretical interpretation of the Six-Degree of Separation [28] offers valuable experience and reference on how to model similar complex problems in a massive collaboration environment.

By collecting data of several typical HFSE cases in China's top forums, Tianya[26]and MOP[17], we gain a better understanding of HFSE campaigns and divide them into the following five phases:

1. Ignition: This is the phase when the very first post goes online with partial information of the target. The post must contain reasons for calling for a HFSE campaign as well.
2. Infection: With appropriate energy, the HFSE campaign can kick off very fast with many netizens commenting, and forwarding the post to other forums.
3. Fading: Netizens are very easy to be distracted by other sensational events on the Web. Therefore, it is extremely important for a HFSE campaign to keep itself refreshed and attracting the eyeballs of netizens.
4. Re-ignition: Most successful HFSE campaigns have the re-ignition phase against the fading effect. When a follow-up participant manages to discover new personal information of the target, name, address, affiliations for example, the fading phase will be stopped and the thread starts to attract more netizens again.
5. Success/failure: A HFSE campaign can end up with two different results. Each time a new re-ignition is performed, HFSE is getting closer to the hidden target. It can successfully hit the target with a series of re-ignition phases. Normally, the final re-ignition is performed by people who arewithin very close proximity to the target, say friends, colleagues, former/current classmates, and etc. If a HFSE campaign fails to start any re-ignition phase, it is highly possible that it will end up with a failure.

By modeling the online social community as a network graph, the initiator of a HFSE campaign is viewed as the source node and the target becomes a hidden destination node in the graph. Lacking information about the target, the only feasible way of finding it out is to flood the message to all neighbors in the graph. With neighbors forwarding to their neighbors, the solution of HFSE becomes a flooding routing algorithm in the graph. Compared to traditional flooding routing[24]and gossiping algorithms [3]in computer networking, a HFSE flooding is different in the following aspects.

- First, a node in traditional flooding will always forward the message to all its neighbors, while a participating node in HFSE flooding is only performing the forwarding with probability p ($p < 1$).
- Second, the energy of a traditional flooding never fades, while a HFSE flooding keeps losing power during message propagation unless a re-ignition is generated.
- Third, nodes in a traditional flooding are computers or other hardware devices which never cheat, while participating nodes in a HFSE are netizens online who are prone to make mistakes carelessly or even maliciously to mislead the whole campaign.
- Fourth, a traditional flooding will always reach the destination as long as the graph is connected, while a HFSE flooding's success or failure depends on the interplay of many affecting parameters.

The rest of the paper is organized as follows. Section 2 gives a brief survey on several related research topics which can be used for HFSE analyze; Section 3 contains a summarized study on several typical cases of HFSE, highlighting main characteristics and typical evolution procedures based on trace data collected from top forums. A mathematical model for HFSE is then proposed in Section 4, formulating the problem as a restricted flooding routing algorithm in the social network. The model is then verified and evaluated by experiments and simulations in Section 5. Finally, Section 6 concludes the paper and points out possible future work directions.

## 2 Related Work

Traditional research on search engines only focuses on keywords and ranking algorithms. While social search considers the individual searcher behavior and even interactions among multiple searchers and study the impact of human factors on search results [9][12]. Using data from two "Small-World" experiments, measurements of 162'368 message chains yield an average chain length at 6-7 in [9] which comply with the theory of six-degree separation. [12]uses a sociology approach, surveying Amazon users and analyzing their pre-search, during-search, and after-search behaviors. Results show that "social search", explicit or implicit, may facilitate the process of information seeking.

Like HFSE, massive collaboration is also common in large open source software development communities where a large system is divided into many smaller pieces

and outsourced to a crowd of people. The phenomenon is also called crowd-sourcing [14]and is studied with a modified scale-free network model in [11].

Topology is one of the key factors in analyzing HFSE. In [1], the structures of three social networks, Cyworld, MySpace, and Orkut, are studied and compared. Research on other social networks, Facebook, ArnetMiner, and Tencent QQ, is done in [28] to approach a mathematical model for justifying the theory of six-degree of separation.

In the experiments conducted to verify the "Small-World" hypothesis, many researchers adopt similar approaches of sending messages to friends and have them forwarded continuously along the chain to reach a target individual. The process shares some similarity with the HFSE campaign. Although most papers only consider the social network overlay, [16] argues that the underlay geographic network should not be neglected.Compared to the HFSE probabilistic flooding routing algorithm, geographic routing is more like LAR [15] or ZRP [13]algorithms which use location information to restrict the flooding area during routing.

Gossip algorithms [3][23]have the same purpose of reaching an unknown destination as HFSE. Messages in gossip algorithms are passed from the current node to a randomly chosen neighbor. It is discovered that the averaging time of a gossip algorithm depends on the second largest eigenvalue of a doubly stochastic matrix characterizing the algorithm with random-walk transition probabilities.

HFSE can be viewed as a distributed information collection process from multiple forums and sources. This is similar to BitTorrent[7] kind of algorithms in P2P networks where a large crowd of users try to download independent pieces of a big file and exchange them to recover the whole file.Fluid Model is introduced in [18] to study the scalability, performance and efficiency of BitTorrent-like P2P file-sharing mechanism.

Flooding routing algorithms [24] are often used in mobile ad-hoc networks and P2P networks for message communication or query forwarding.Performance analysis of flooding-based resource discovery approaches in distributed environments can be found in [8]. Probabilistic flooding algorithm [20]is obtained if intermediate nodes in the network only forward message/query with probability p instead of 1.Performance analysis of probabilistic flooding can be found in [19]. The approaches of their analysis are helpful for the performance and scalability analysis of HFSE.

To the best of our knowledge, no work has been done to approach a mathematical model for the HFSE problem. But the approachof modeling social interaction as particle swarms [4]in the analysis of multidimensional search space deals with a similar problem of identifying uncertain target in an unknown problem space with implicit massive collaboration. Also, the model of BubbleStorm[25]used in P2P exhaustive search analysis share the same characteristic of growing search area as HFSE.

## 3 Case Studies

In this Section, we study typical HFSE cases by analyzing data collected from China's top hot forums. Five major phases of a typical HFSE campaign are summarized based on the traces.

### 3.1 The Forums

The booming power of HFSE can be partially attributed to the increasing number of netizens in China, which doubles itself every two years and reaches 338 million in July 2009 (see Fig.1). Unlike western Web users who prefer browsing news and dealing emails, Chinese netizen's first stop on the Web are Instant Messaging, and Forum/BBS browsing[6]. For most sensational HFSE cases, their very first posts were published on forums instead of news portals or blogs.

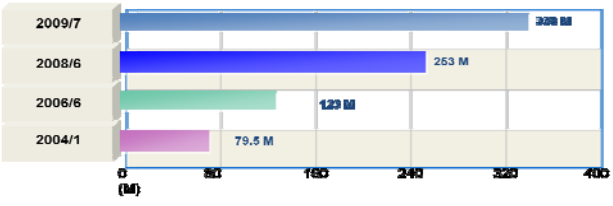



**Fig. 1.** China's netizen population boom in recent years

Therefore, our data collection of HFSE traces focuses on the following two forums where most HFSE campaigns are initiated:

- Tianya.cn: a forum with 62 million registered users;
- Mop.com: a forum with 7.9 million registered users.

### 3.2 Data Collection Method

First we use the script language (PERL) crawling the pages from Tianya.cnand Mop.com forums. Then we use HTML-Parser and Regular Expression technology to extract the information we needed. The information include user's name, user's id, reply content, reply number and reply time.

### 3.3 Typical HFSE Cases

There are lots of HFSE cases in the net world, give some graphs. In this paper, we focus on two HFSE cases, one is the SCT event, the other is Lin harassing event.

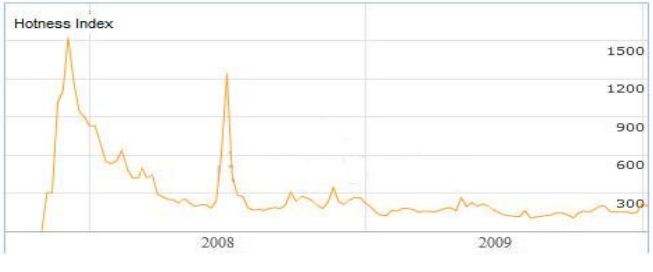



**Fig. 2.** Hotness index curve over time of the South China Tiger HFSE event

As shown in Fig.2, the South China Tiger event developed over one year to succeed with several re-ignition phases. Each re-ignition re-fuels the HFSE campaign and generates a new spike in the hotness index curve.

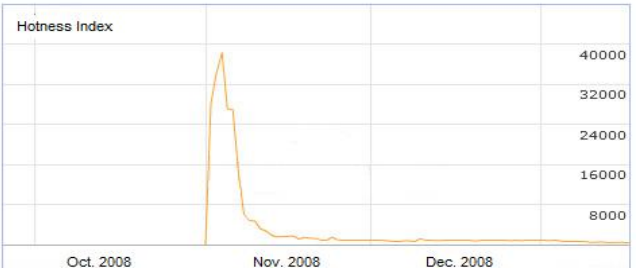



**Fig. 3.** Hotness index curve over time of the Lin Harassing HFSE event

Different from the South China Tiger event as shown in Fig.2, the Lin Harassing event is much simpler. It finishes within several days and succeeded in finding out the target. There is no re-ignition phase in this event simply because the first phase of ignition is so strong that it infected almost every node in the whole network.

## 4 The Model

We develop our mathematical model for the HFSE based on our observation made from the typical case studies. The model includes a network modeling of the social communities involved in the HFSE campaign and a probabilistic flooding algorithm simulating the HFSE search process.

### 4.1 The Network Model

We define the HFSEsocial network as a graph G= (V, E), where V is the set of the people involved in the HFSE campaign and E is the set of edges connecting them.

### 4.2 The Probabilistic Flooding Algorithm

The probabilistic flooding algorithm is shown as follow:

**Algorithm 1.** HFSE probabilistic flooding algorithm

```
Input:
(V, E): the network
S: source node starting the search
T: target node to be reached
Procedure:
Search begins from S;
Repeat
```

```
{Forward to neighbors with probability p;
  Re-ignition with probability q;
  }
Search stops if T reached or no more nodes are involved.
```

# 5 Simulation Results

## 5.1 Success Rate

In this set of simulation experiments, we are interested in examining the success rate of a random HFSE campaign by looking at different combination of simulation parameters. Considering the efficiency of the simulation, we choose a population size of 652995 which is large enough to accommodate all our data traces.

We use visualization tools to illustrate the steps of the search progress. Without loss of generality, we assume that the search always starts from the center point of the network. We don't prefix the target of the search in the simulations and allows the search to develop on its own. At the beginning of each simulation, all nodes are not involved in the search (shown in green). With the infection going on, involved nodes are increasing (shown in black). Finally, when the simulation ends because of running out of energy, we look at the proportion of all infected nodes in the whole population of the network. Considering the fact that the target of the HFSE could be random in the network, the success rate of a given HFSE campaign could be defined as that ratio.

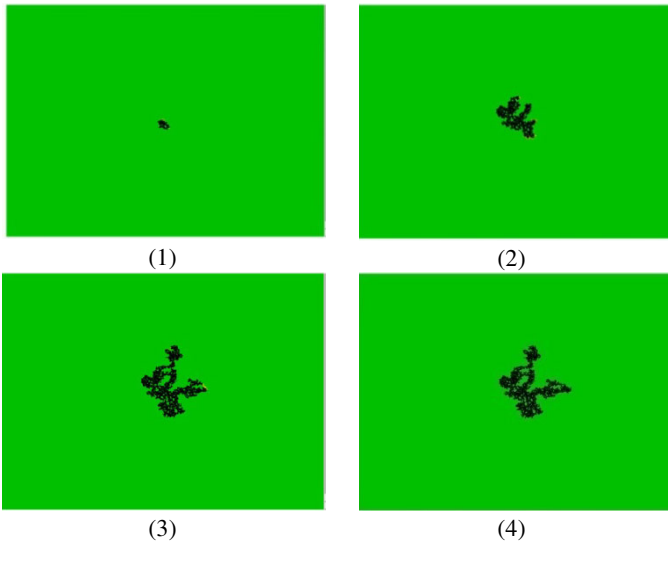



**Fig. 4.** Simulation process ($p$=0.49)

We look at the following 4 typical simulation results:

(1). $p$=0.49 which represents a highly possible FAILURE;
In fact, most simulations performed under this parameter setting failed to take off and the search stops after only infecting several nodes (less than 0.1% of the whole population). Fig.4 shows one of the only noticeable results which managed to cover 1.7% of the nodes.When p is even smaller, no search will take off at all.

(2). $p$=0.50 which represents a 50/50 SUCCESS/FAILURE;
As shown in Fig.5, HFSE campaign will have a impact throughout the network and results in 43.4% people in the network to be involved in the search.

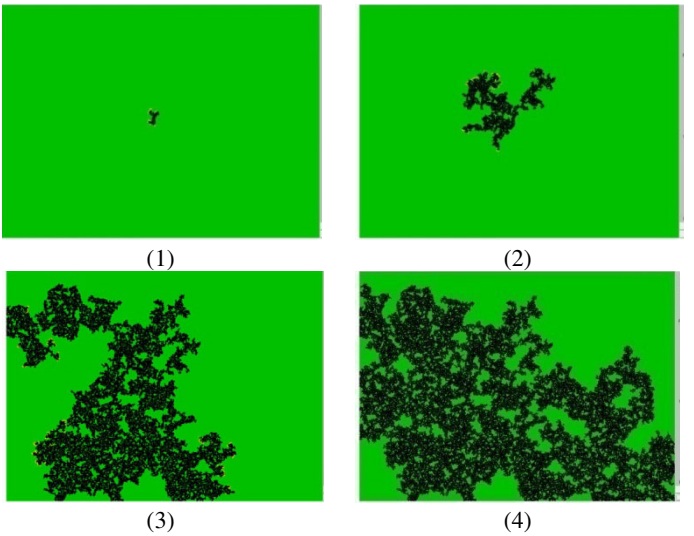



**Fig. 5.** Simulation process ($p$=0.50)

(3). $p$=0.55 which represents an almost definite SUCCESS;
As shown in Fig. , this case is different from the previous ones in the following two aspects. First, it infects neighbors' almost isotropic ally and the search moves on steadily along different dimensions and directions. Second, it results in a fully coverage of the network and involves 88.7% of the whole population.

### 5.2 Vulnerability Analysis

In this subsection, we examine the vulnerability of the HFSE search and try to Fig. 7 out feasible solutions to defend oneself from being reached if he/she unfortunately becomes the target.

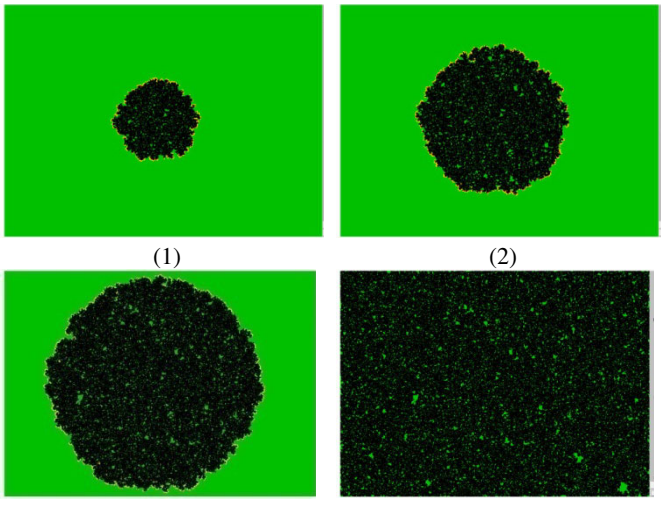



**Fig. 6.** Simulation process ($p$=0.55)

We developed two different strategies:

(1). Setting obstacles
This strategy simulates the different hiding measures of deleting posts from forums, asking friends and relatives not to participate in the HFSE and other possible actions.

(2). Misleading directions
This strategy simulates the solution of providing wrong or misleading information to HFSE participants and hoping the probabilistic flooding would be directed along the wrong way and wrong direction.

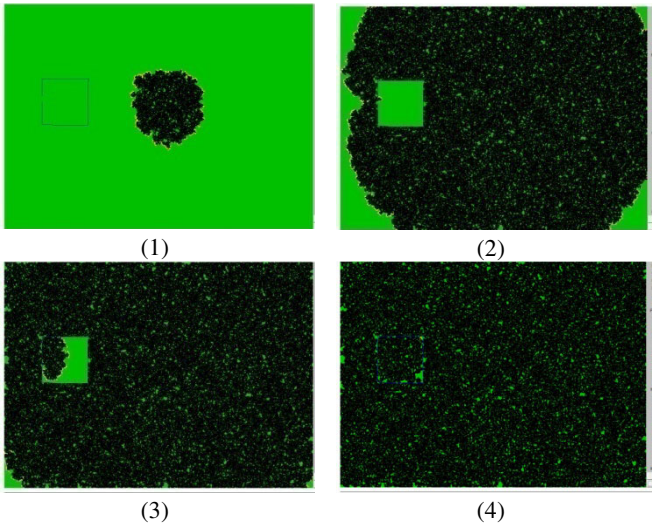



**Fig. 7.** Setting obstacles ($p$=0.55)

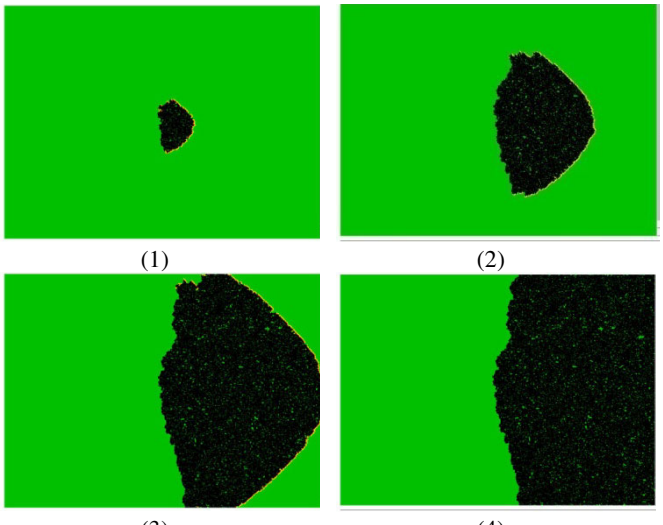



**Fig. 8.** Mislead the HFSE campaign ($p$=0.55)

Fig. 7(1) shows a HFSE target setting obstacles around himself but still leaving one small opening to the left of the square. This works for a while but the search will eventually find him if the probability parameters of the search are not going down.

Fig. 8 shows a promising future of the second strategy. We simulate the effect of the misleading information by adding a directional factor to the probabilistic forwarding of the messages. In this case, the direction is set to the right since the target is located to the left of the ignition start point. Result shows that the target is safe after the energy of the search dies out.

## 6 Conclusions

In this paper, weexamined several typical HFSE cases and developed a mathematical model which can describe the ignition, infection and re-ignition phases of HFSE campaigns. Simulation experiments are conducted to verify the model. Results show that the success/failure of a HFSE campaign depends highly on the forwarding probability of an involved node to its neighbors.

We also examine the vulnerability of HFSE. Simulation results and analysis show that providing misleading information can direct the HFSE along the wrong way and cause the whole campaign to fail. Other strategies like setting obstacles to prevent message spreading are not working.

Future research work can be done to gain better understanding on the target and take that into consideration in the model.

## References


1. Ahn, Y.-Y., Han, S., Kwak, H., Eom, Y.-H., Moon, S., Jeong, H.: Analysis of topological characteristics of Huge Online Social Networking Services. In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007 (2007)
2. Bai, X., Ji, S.: Human Flesh Search Engine: an Internet Lynching, Xinhua English 2008-07-04, http://english.sina.com/china/1/2008/0704/170016.html
3. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Gossip Algorithms: Design, Analysis and Applications. In: Proceedings of IEEE Infocomm, Miami, vol. 3, pp. 1653–1664 (March 2005)
4. Clerc, M., Kennedy, J.: The Particle Swarm—Explosion, Stability, andConvergence in a Multidimensional Complex Space. IEEE Transactions on Evolutionary Computation 6(1) (February 2002)
5. CNN, From Flash Mob to Lynch Mob, June 5 (2007)
6. CNNIC (China Internet Network Information Center). The 24th Statistical Report on Internet Development in China (July 2009), http://www.cnnic.net.cn/uploadfiles/pdf/2009/10/12/114121.pdf
7. Cohen, B.: Incentives Build Robustness in BitTorrent. In: Workshop on Economics of Peer-to-Peer Systems (2003)
8. Dimakopoulos, V.V., Pitoura, E.: On the Performance of Flooding-Based Resource Discovery. IEEE Transactions on Parallel and Distributed Systems 17(11), 1242–1252 (2006)
9. Evans, B.M., Chi, E.H.: Towards a Model of Understanding Social Search. In: Proceedings of the ACM 2008 Conference on Computer Supported Cooperative Work (2008)
10. Fletcher, H.: Human flesh search engines: Chinese vigilantes that hunt victims on the web. The Times (2008-06-25), http://technology.timesonline.co.uk./tol/news/tech_and_web/article4213681.ece
11. Gao, Y., Madey, G.: Towards understanding: a study of thesourceforge.net community using modeling and simulation. In: SpringSim 2007: Proceedings of the 2007 Spring Simulation Multiconference (2007)
12. Goel, S., Muhamad, R., Watts, D.: Social Search in "Small-World" Experiments. In: WWW 2009: Proceedings of the 18th International Conference on World Wide Web, pp. 701–710 (2009)
13. Haas, J.: A new routing protocol for the reconfigurable wireless networks. In: Proc. of IEEE 6th International Conference on Universal Personal Communications, pp. 562–566 (1997)
14. Howe, J.: Wired 14.06: The Rise of Crowdsourcing, http://www.wired.com/wired/archive/14.06/crowds.htm
15. Ko, Y.B., Vaidya, N.H.: Location-Aided Routing (LAR) in Mobile Ad-Hoc Networks. Wireless Networks 6(4), 307–321 (2000)
16. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic Routing in Social Networks. Proceedings of the National Academy of Sciences (2005)
17. MOP forum, http://www.mop.com
18. Qiu, D., Srikant, R.: Modeling and performance analysis of BitTorrent-like peer-to-peer networks. ACM SIGCOMM Computer Communication Review 34(4), 367–378 (2004)
19. Oikonomou, K., Stavrakakis, I.: Performance Analysis of Probabilistic Flooding Using Random Graphs. In: WOWMOM 2007, pp. 1–6 (2007)
20. Sasson, Y., Cavin, D., Schiper, A.: Probabilistic Broadcast for Flooding in Wireless Mobile Ad-Hoc Networks. In: Proc. IEEE WCNC, New Orleans, Louisiana, USA (2003)

21. Science Magazine.: Rare-Tiger Photo Flap Makes Fur Fly in China. Science 318, 893, November 9 (2007)
22. Science Magazine.: End of a Tiger's Tale. Science 321, 32, July 18 (2008)
23. Shah, D.: Network gossip algorithms. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3673–3676 (2009)
24. Stojmenovic, I., Lin, X.: Loop-Free Hybrid Single-Path/Flooding Routing Algorithms withGuaranteed Delivery for Wireless Networks. IEEE Transactions on Parallel and Distributed Systems 12(10), 1023–1032 (2001)
25. Terpstra, W.W., Kangasharju, J., Leng, C., Buchmann, A.P.: BubbleStorm: Resilient, Probabilistic, and Exhaustive Peer-to-Peer Search. In: SIGCOMM 2007, Kyoto, Japan, August 27–31 (2007)
26. Tianya forum, `http://www.tianya.cn`
27. Watts, D.J., Dodds, P.S., Newman, M.E.J.: Identity and Search in Social Networks. Science 296(5571), 1302–1305 (2002)
28. Zhang, L., Tu, W.: Six Degrees of Separation in Online Society. In: Proceedings of the WebSci 2009: Society OnLine, Athens, Greece, March 18-20 (2009)

# A New Dynamic ID-Based Remote User Authentication Scheme with Forward Secrecy

Chun-Guang Ma[1], Ding Wang[1,2,*], Ping Zhao[1], and Yu-Heng Wang[3]

[1] College of Computer Science and Technology, Harbin Engineering University, 145 Nantong Street, Harbin City 150001, China
wangdingg@mail.nankai.edu.cn
[2] Automobile Management Institute of PLA, Bengbu City 233011, China
[3] Golisano College of Computing and Information Sciences, Rochester Institute of Technology, 102 Lomb Memorial Dr., Rochester, NY 14623, USA

**Abstract.** Forward secrecy is one of the important properties of remote user authentication schemes to limit the effects of eventual failure of the entire system when the long-term private keys of one or more parties are compromised. Recently, Tsai et al. showed that Wang et al.'s dynamic ID-based remote user authentication scheme fails to achieve user anonymity and is vulnerable to user impersonation attack, and proposed an enhanced version to overcome all the identified flaws. In this paper, however, we will point out that, Tsai et al.'s scheme still suffers from the denial of service attack and cannot provide forward secrecy. To remedy these security flaws, we propose an enhanced authentication scheme, which covers all the identified weaknesses of Tsai et al.'s scheme and is more suitable for mobile application scenarios where resource constrained and security concerned.

**Keywords:** Password-based, Authentication protocol, Non-tamper resistant, Smart card, Cryptanalysis, Denial of service attack.

## 1 Introduction

With the large-scale proliferation of internet and network technologies over the last couple of decades, more and more electronic transactions for mobile devices are implemented on Internet or wireless networks. In electronic transactions, remote user authentication in insecure channel is an important issue. Smart cards have been widely used in many e-commerce applications and network security protocols due to their low cost, portability, efficiency and cryptographic properties. Smart card authentication is based on different techniques such as passwords, digital certificates and biometric technology. Among these techniques, password is the most commonly used authentication technique to authenticate users on the server due to its simplicity and convenience. Except efficiency and convenience, there are also many other desirable properties of a secure remote authentication scheme, such as freedom of choosing passwords, mutual authentication, user anonymity and forward secrecy.

* Corresponding author.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 199–211, 2012.

Recently, Since Chang and Wu [1] introduced the first remote user authentication scheme using smart cards in 1993, there have been many smart card based authentication schemes proposed [2-6]. In most of the previous authentication schemes, the user's identity is transmitted in plaintext over insecure networks during the authentication process, which may leak the identity of the logging user once the login messages were eavesdropped, hence user privacy is not preserved. The leakage of the user identity may also cause an unauthorized entity to track the user's login history and current location. In many cases, it is of utmost importance to provide anonymity so that the adversary cannot trace user activity. Therefore, user anonymity is an important feature that a practical authentication scheme should achieve.

As noted by Blake-Wilson et al. [7], forward secrecy is an admired security feature for authentication protocols with session keys establishment. Particularly, forward secrecy is a property concerned with limiting the effects of eventual failure of the entire system. It indicates that, even if the long-term private keys of one or more entities are compromised, the secrecy of previous session keys established by honest entities should not be affected and thus the previous sessions shall remain secure. Hence, a sound authentication scheme should achieve this important property.

In 2004, Das et al. [8] first introduced the concept of dynamic ID-based authentication scheme to resist ID-theft and thus to achieve user anonymity. However, in 2005, Chien and Chen [9] pointed out that Das et al.'s scheme fails to protect the user's anonymity, so they proposed a new one. In 2009, to overcome the security pitfalls of Das et al.'s scheme, Wang et al. [10] also proposed a dynamic ID-based authentication scheme, and claimed that their scheme is more efficient and secure while keeping the merits of Das et al.'s scheme. Later on, Tsai et al. [11] pointed out that Wang et al.'s scheme fails to provide user anonymity as claimed and cannot withstand user impersonation attack, and further proposed an enhanced version to eliminate the identified defects.

In this paper, however, we will demonstrate that Tsai et al.'s scheme fails to provide the property of forward secrecy, and suffers from the denial of service attack and insider attack. In addition, their scheme also has some practical pitfalls. To conquer the identified flaws, a robust authentication scheme based on the secure one-way hash function and the well-known discrete logarithm problem is presented.

The remainder of this paper is organized as follows: in Section 2, we review Tsai et al.'s authentication scheme. Section 3 describes the weaknesses of Tsai et al.'s scheme. Our proposed scheme is presented in Section 4, and its security analysis is given in Section 5. The comparison of the performance of our scheme with the other related schemes is shown in Section 6. Section 7 concludes the paper.

## 2 Review of Tsai et al.'s Scheme

For reader's convenience, we first briefly review Tsai et al.'s scheme [11] before demonstrating its weaknesses. Their scheme consists of four phases: the registration phase, the login phase, the verification phase and password update phase. For ease of presentation, we employ some intuitive abbreviations and notations listed in Table 1.

**Table 1.** Notations

| Symbol | Description |
|---|---|
| $U_i$ | $i^{th}$ user |
| $S$ | remote server |
| $ID_i$ | identity of user $U_i$ |
| $P_i$ | password of user $U_i$ |
| $x$ | the secret key of remote server $S$ |
| $n$ | a large prime number |
| $g$ | a primitive element in Galois field $GF(n)$ |
| $h(\cdot)$ | collision free one-way hash function |
| $\oplus$ | the bitwise XOR operation |
| $\|\|$ | the string concatenation operation |
| $A \Rightarrow B : M$ | message $M$ is transferred through a secure channel from $A$ to $B$ |
| $A \rightarrow B : M$ | message $M$ is transferred through a common channel from $A$ to $B$ |

### 2.1 Registration Phase

Let ($x$, $y = g^x \bmod n$) denote the server $S$'s private key and its corresponding public key, where $x$ is kept secret by the server and $y$ is stored inside each user's smart card. The registration phase involves the following operations:

*Step R1.* $U_i$ chooses his/her identity $ID_i$ and password $P_i$.
*Step R2.* $U_i \Rightarrow S$: $\{ID_i, P_i\}$.
*Step R3.* On receiving the registration message from $U_i$, the server $S$ computes $N_i = h(P_i \| ID_i) \oplus h(x \| ID_i)$.
*Step R4.* $S \Rightarrow U_i$: A smart card containing security parameters $\{N_i, y, n, g, h(\cdot)\}$.

### 2.2 Login Phase

When $U_i$ wants to login to $S$, the following operations will be performed:

*Step L1.* $U_i$ inserts his/her smart card into the card reader, and inputs $ID_i$ and $P_i$.
*Step L2.* The smart card computes $h(x \| ID_i) = N_i \oplus h(P_i \| ID_i)$, $C = g^k \bmod n$, $CID_i = ID_i \oplus h(g^k \| T_1) \bmod n$, and $B_i = h(CID_i \| C \| h(x \| ID_i) \| y \| T_1)$, where $T_1$ is the current timestamp and $k$ is a random number.
*Step L3.* $U_i \rightarrow S$: $\{CID_i, B, C, T_1\}$.

### 2.3 Verification Phase

After receiving the login request message from user $U_i$ at time $T_2$, the server $S$ performs the following operations:

*Step V1.* $S$ verifies whether $(T_2 - T_1) \leq \Delta T$. If the verification fails, $S$ rejects the login request.
*Step V2.* $S$ computes $ID_i = CID_i \oplus h(C^x \| T_1) \bmod n$ and $B' = h(CID_i \| C \| h(x \| ID_i) \| y \| T_1)$, and then compares the computed $B'$ with the received $B$. If they are not equal, $S$ rejects the request.
*Step V3.* $S$ computes $SK = h(h(x \| ID_i) \| T_1 \| B \| CID_i \| T_2)$ and $D = h(SK \| h(x \| ID_i) \| T_1 \| T_2)$.

*Step V4.* $S \rightarrow U_i$: $\{D, T_2\}$.

*Step V5.* Upon receiving the reply message $\{D, T_2\}$ at the time $T_3$, $U_i$ verifies whether $(T_2 - T_1) \leq \Delta T$ . If the verification fails, $U_i$ terminates the session.

*Step V6.* $U_i$ computes $SK = h(h(x \| ID_i) \| T_1 \| B \| CID_i \| T_2)$ and $D' = h(SK \| h(x \| ID_i) \| T_1 \| T_2)$, and then compares the computed $D'$ with the received $D$. If they are not equal, $U_i$ terminates the session.

*Step V7.* After authenticating each other, $U_i$ and $S$ use the same session key $SK$ to secure ensuing data communications.

### 2.4 Password Change Phase

When $U_i$ wants to change the old password $P_i$ to the new password $P_i^{new}$, $U_i$ insert his/her own smart card into card reader. The smart card computes $N_i^{new} = N_i \oplus h(P_i) \oplus h(P_i^{new})$ and updates $N_i$ with the new $N_i^{new}$.

## 3 Cryptanalysis of Tsai et al.'s Scheme

In this section, we will first show that Tsai et al. have made a mistake when designing the login phase, and then we will demonstrate that their scheme fails to achieve forward secrecy, and suffers from denial of service attack and insider attack. Moreover, several practical pitfalls in their scheme are also pointed out.

### 3.1 Failure to Achieve Forward Secrecy

Let us consider the following scenarios. Supposing the server $S$'s long time private key $x$ is leaked out by accident or intentionally stolen by an adversary $A$. Once the value of $x$ is obtained, with previously intercepted messages $\{CID_i^j, B^j, C^j, T_1^j, T_2^j\}$ transmitted during $U_i$'s $j$th authentication process, $A$ can derive the session key $SK^j$ of $S$ and $U_i$'s $j$th encrypted communication through the following method:

**Step 1.** Computes $L_1 = (C^j)^x \bmod n$ , as $C^j$ and $x$ are known.

**Step 2.** Computes $ID_i = CID_i^j \oplus h(L_1 \| T_1^j)$ , as $CID_i^j$ and $T_1^j$ are previously intercepted.

**Step 3.** Computes $SK^j = h(h(x \| ID_i) \| T_1^j \| B^j \| CID_i^j \| T_2^j)$ .

Once the session key $SK^j$ is obtained, the whole $j$th session will become completely insecure. Therefore, the property of forward secrecy is not provided.

### 3.2 Denial of Service Attack

The password change phase of Tsai et al.'s scheme is insecure like that of Wang et al.'s scheme. If an attacker manages to obtain the smart card of legitimate user $U_i$ for a very short time, he can change the password of user $U_i$ as follows:

*Step 1.* The attacker inserts $U_i$'s smart card into a card reader and initiates a password change request.

*Step 2.* The attacker submits a random string $R$ as $U_i$'s original password and a new string $P_i^{new}$ as the targeting new password.
*Step 3.* The smart card computes $N_i^{new}=N_i \oplus h(R) \oplus h(P_i^{new})$ and updates $N_i$ with $N_i^{new}$.

Once the value of $N_i$ is updated, legitimate user $U_i$ cannot login successfully even after getting his/her smart card back because the value of $h(x \| ID_i)$ cannot be valid and thus $U_i$'s login request will be denied by the server $S$ during the verification phase. Hence, denial of service attack can be launched on the user $U_i$'s smart card.

It should be noted that, although this vulnerability seems too basic to merit discussion, it cannot be well remedied just with minor revisions. To conquer this vulnerability, a verification of the authenticity of the original password before updating the value of $N_i$ in the memory of smart card is essential. And thus, besides $N_i$, some additional verifier(s) or parameter(s) should be stored in the smart card, which may introduce new vulnerabilities, such as offline password guessing attack and user impersonation attack. Therefore, only radical changes can eliminate this vulnerability.

### 3.3 Insider Attack

In many scenarios, the user uses a common password to access several systems for his convenience. If the user registers to the server with plaintext password, an insider of server can impersonate user's login by abusing the legitimate user's password and can get access to the other systems [2].

In the registration phase, $U_i$'s password $P_i$ is submitted in plaintext to $S$, and thus it can be easily learned by the insider of $S$. If $U_i$ uses this $P_i$ to access several servers for his/her convenience, the insider of $S$ can impersonate $U_i$ to access other servers. Hence, it's an insecure factor to commit plain password to the server, and Tsai et al.'s scheme is susceptible to insider attack.

### 3.4 Some Practical Pitfalls

In practice, the length of $ID_i$ and $h(y^k \| T_1)$ are much smaller than that of $n$, performing a modular operation on the value of $ID_i \oplus h(y^k \| T_1)$ as done in Step L2 of the login phase is meaningless but only to increase the computation overhead. The right way is to first compute $Y=y^k \bmod n$ and then derive $CID_i=ID_i \oplus h(Y \| T_1)$, and the same is the case with the derivation of $ID_i$ in Step V2 of the verification phase.

Since clock synchronization is difficult and expensive in existing network environment, especially in wide area networks, these schemes employing timestamp mechanism to resist replay attacks is not suitable for use in distributed networks or large-scale application environments. What's more, these schemes employing timestamp may still suffer from replay attacks as the transmission delay is unpredictable in real networks [12].

## 4 Our Proposed Scheme

According to our analysis, three principles for designing a sound password-based remote user authentication scheme are presented. First, user anonymity, especially in

some application scenarios, (e.g., e-commence), should be preserved, because from the identity $ID_i$, some personal secret information may be leaked about the user. In other words, without employing any effort an adversary can identify the particular transaction being performed by the user $U_i$. Second, a nonce based mechanism is often a better choice than the timestamp based design to resist replay attacks, since clock synchronization is difficult and expensive in existing network environment, especially in wide area networks. Finally, the password change process should be performed locally without the hassle of interaction with the remote authentication server for the sake of security, user friendliness and efficiency [3]. In this section, we present a new remote user authentication scheme to overcome the security flaws described in previous section.

### 4.1 Registration Phase

The server $S$ generates two large primes $p$ and $q$ and computes $n = pq$, then chooses a prime number $e$ and an integer $d$, such that $ed \equiv 1 \bmod (p\text{-}1)(q\text{-}1)$. Finally, the server $S$ finds an integer $g$, which is a primitive element in both Galois field $GF(p)$ and $GF(q)$, and make the values of $n$, $e$ and $g$ public, while $p$, $q$ and $d$ are only known to server $S$. The registration phase involves the following operations:

**Step R1.** The user $U_i$ first chooses his/her identity $ID_i$, $P_i$ and a random number $b$, and then computes $PW_i = h(b \| P_i)$.

**Step R2.** $U_i \Rightarrow S$: $ID_i$, $PW_i$.

**Step R3.** On receiving the registration message from $U_i$, the server $S$ chooses random value $y_i$ and computes $N_i = h(ID_i \| PW_i) \oplus h(d)$, $A_i = h(PW_i \| ID_i)$, $B_i = y_i \oplus ID_i \oplus PW_i$ and $D_i = h(h(ID_i \| y_i) \oplus d)$. Server $S$ chooses the value of $y_i$ corresponding to $U_i$ to make sure $D_i$ is unique to each user. The server $S$ stores $y_i \oplus h(h(d) \| d)$ and $ID_i \oplus h(d \| y_i)$ corresponding to $D_i$ in its database.

**Step R4.** $S \Rightarrow U_i$: A smart card containing security parameters $\{N_i, A_i, B_i, n, e, g, h(\cdot)\}$.

### 4.2 Login Phase

When $U_i$ wants to login the system, the following operations will perform:

**Step L1.** $U_i$ inserts his/her smart card into the card reader and inputs $ID_i^*$ and $P_i^*$.

**Step L2.** The smart card computes $A_i^* = h(PW_i^* \| ID_i^*)$ and verifies the validity of $A_i^*$ by checking whether $A_i^*$ equals to the stored $A_i$. If the verification holds, it implies $ID_i^* = ID_i$ and $P_i^* = P_i$. Then, the smart card chose a random number $N_u$ and computes $y_i = B_i \oplus ID_i \oplus PW_i$, $h(d) = N_i \oplus h(ID_i \| P_i)$, $CID_i = h(ID_i \| y_i) \oplus h(h(d) \| N_u)$, $C_1 = N_u^e \bmod n$, Otherwise, the session is terminated.

**Step L3.** $U_i \rightarrow S$: $CID_i$, $C_1$.

### 4.3 Verification Phase

After receiving the login request from $U_i$, server $S$ performs the following operations:

**Step V1.** The server $S$ decrypts the random number $N_u$ from $C_1$ using its private key $d$, then computes $D_i^* = h(CID_i \oplus h(h(d) \| N_u) \oplus d)$ and finds $D_i$ corresponding to $D_i^*$ in its database. If there exists no matched $D_i$, the request is rejected. Otherwise, server $S$ extracts $y_i \oplus h(d \| h(d))$ and $ID_i \oplus h(d \| y_i)$ corresponding to $D_i^*$ from its database. Now the server $S$ computes $y_i$ from $y_i \oplus h(h(d) \| d)$ and $ID_i$ from $ID_i \oplus h(d \| y_i)$ because the server $S$ knows the value of $d$. Then, the server $S$ generates a random number $N_S$ and computes the session key $SK = h(ID_i \| y_i \| N_u \| N_s \| CID_i)$, $C_2 = h(y_i \| ID_i \| SK)$.

**Step V2.** $S \rightarrow U_i$: $N_s$, $C_2$.

**Step V3.** On receiving the reply message from $S$, $U_i$ computes $SK = h(ID_i \| y_i \| N_u \| N_s \| CID_i)$, $C_2^* = h(y_i \| ID_i \| SK)$ and compares $C_2^*$ with the received value of $C_2$. This equivalency authenticates the legitimacy of the server $S$, and $U_i$ goes on to compute $C_3 = h(N_u \| N_s \| y_i \| ID_i \| SK)$.

**Step V4.** $U_i \rightarrow S$: $C_3$.

**Step V5.** Upon receiving $C_3$ from $U_i$, the server $S$ first computes $C_3^* = h(N_u \| N_s \| y_i \| ID_i \| SK)$ and then checks if $C_3^*$ is equal to the received value of $C_2$ . If this verification holds, the server $S$ authenticates the user $U_i$ and the login request is accepted else the connection is terminated.

**Step V6.** The user $U_i$ and the server $S$ agree on the common session key $SK$ for securing future data communications.

### 4.4 Password Change Phase

In this phase, we argue that the password change phase should be performed locally without interaction with the authentication server for the sake of security, user friendliness and efficiency. In addition, the user's smart card must have the ability to detect the failure times. Once the number of login failure exceeds a predefined system value, the smart card must be locked immediately to prevent the exhaustive password guessing behavior. This phase involves the following steps.

**Step P1.** $U_i$ inserts his/her smart card into the card reader and inputs $ID_i$, the original password $P_i$ , the new password $P_i^{new}$.

**Step P2.** The smart card computes $A_i^* = h(PW_i^* \| ID_i^*)$ and verifies the validity of $A_i^*$ by checking whether $A_i^*$ equals to the stored $A_i$. If the verification holds, it implies $ID_i^* = ID_i$ and $P_i^* = P_i$. Otherwise, the smart card rejects.

**Step P3.** The smart card asks the cardholder to resubmit a new password $P_i^{new}$ and computes $N_i^{new} = N_i \oplus h(ID_i \| h(b \| P_i)) \oplus h(ID_i \| h(b \| P_i^{new}))$, $A_i^{new} = h(h(b \| P_i^{new}) \| ID_i)$and $B_i^{new} = y_i \oplus ID_i \oplus h(b \| P_i^{new})$. Thereafter, smart card updates the values of $N_i$, $A_i$ and $B_i$ stored in its memory with $N_i^{new}$, $A_i^{new}$ and $B_i^{new}$.

## 5 Security Analysis

Recent research results have shown that the secret data stored in the common smart card could be extracted by some means, such as monitoring the power consumption [13] or analyzing the leaked information [14]. Schemes based on the tamper resistance assumption of the smart card are vulnerable to some types of attacks, such as user impersonation attacks, server masquerading attacks, and offline password guessing attacks, etc., once an adversary has obtained the secret information stored in a user's smart card [5]. Hence, a desirable scheme should put aside any special security features that could be supported by a smart-card, and simply assume that once a smart-card is stolen by an adversary, all the information stored in it are known to the adversary. In the following, we will analyze the security of the proposed scheme under the assumption that the secret information stored in the smart card can be revealed, i.e., the secret information $b$, $N_i$, $A_i$ and $B_i$ can be revealed. Consequently, $h(d)$ can also be obtained by a malicious privileged user $U_k$, as $h(d)=N_k \oplus h(h(b||P_k)||ID_k)$, where $N_k$ and b is revealed, and the malicious user $U_k$ knows his own identity $ID_k$ and password $P_k$ corresponding to his smart card.

The security of our proposed authentication scheme is based on the secure hash function and the difficulty of the large integer factorization problem. As summarized in Refs. [15] and discussed in Section 1, the following criteria are important for evaluating smart card based remote user authentication schemes in terms of security.

**(1) User Anonymity:** Suppose that the attacker has intercepted $U_i$'s authentication messages ($CID_i$, $C_1$, $C_2$, $C_3$). Then, the adversary may try to retrieve any static parameter from these messages, but $CID_i$ and $C_1$, $C_2$, $C_3$ are all session-variant and indeed random strings due to the randomness of $N_u$. Accordingly, Without knowing the random number $N_u$, the adversary will face to solve the large integer factorization problem to retrieve the correct value of $h(ID_i||y_i)$ from $CID_i$, while $h(ID_i||y_i)$ is the only static element in the transmitted messages. Hence, the proposed scheme can overcome the security flaw of user anonymity breach.

**(2) Offline Password Guessing Attack:** Suppose that a malicious privileged user $U_i$ has got $U_k$'s smart card, and the secret information $b$, $N_k$, $A_k$ and $B_k$ can also be revealed under our assumption of the non-tamper resistant smart card. Even after gathering this information and obtaining $h(d)=N_k \oplus h(h(b||P_i)||ID_k)$, the attacker has to at least guess both $ID_k$ and $P_k$ correctly at the same time, because it has been demonstrated that our scheme can provide user anonymity. It is impossible to guess these two parameters correctly at the same time in real polynomial time, and thus the proposed scheme can resist offline password guessing attack with smart card security breach.

**(3) Stolen Verifier Attack:** In the proposed protocol, only the server $S$ knows private secret $d$ and stores $y_i \oplus h(h(d)||d)$ and $ID_i \oplus h(d||y_i)$ corresponding to $D_i$ in its database. Although a malicious privileged user can compute $h(d)$ in the way described above, he/she does not have any technique to find out the value of $d$, nor can he/she calculates $y_i$ corresponding to other legitimate user. Therefore, the proposed protocol is secure against stolen verifier attack.

**(4) User Impersonation Attack:** As $CID_i$ , $C_1$ and $C_3$ are all protected by secure one-way hash function, any modification to these parameters of the legitimate user $U_i$'s authentication messages will be detected by the server $S$ if the attacker cannot fabricate the valid $CID_i^*$ and $C_3^*$. Because the attacker has no way of obtaining the values of $ID_i$, $P_i$ and $y_i$ corresponding to user $U_i$, he/she cannot fabricate the valid $CID_i^*$ and $C_3^*$. Therefore, the proposed protocol is secure against user impersonation attack.

**(5) Server Masquerading Attack:** In the proposed protocol, a malicious server cannot compute the session key $SK = h(ID_i||y_i||N_u||N_s||CID_i)$ and $C_2 = h(y_i||ID_i||SK)$ because the malicious server does not know the values of $ID_i$ and $y_i$ corresponding to user $U_i$ , and has to solve the large integer factorization problem to retrieve $N_u$. Therefore, the proposed protocol is secure against server masquerading attack.

**(6) Replay Attack and Parallel Session Attack:** Our scheme can withstand replay attack because the authenticity of authentication messages ($CID_i$, $C_2$, $C_3$) is verified by checking the fresh random number $N_u$ and/or $N_s$. On the other hand, the presented scheme resists parallel session attack, in which an adversary may masquerade as legitimate user $U_i$ by replaying a previously intercepted authentication message. The attacker cannot compute the agreed session key $SK$ and valid $C_3$ because he does not know the values of $N_u$, $ID_i$ and $y_i$ corresponding to user $U_i$. Therefore, the resistance to replay attack and parallel session attack can be guaranteed in our protocol.

**(7) Mutual Authentication:** In our dynamic ID-based scheme, the server authenticates the user by checking the validity of $C_3$ in the access request. We have shown that our scheme can preserve user anonymity, so user $ID_i$ is only known to the server $S$ and the user $U_i$ itself. we have proved that our scheme can resist user impersonation attack. Therefore, it is impossible for an adversary to forge messages to masquerade as $U_i$ in our scheme. To pass the authentication of server $S$, the smart card first needs $U_i$'s identity $ID_i$ and password $P_i$ to get through the verification in Step L2 of the login phase. In this Section, we have shown that our scheme can resist offline password guessing attack. Therefore, only the legal user $U_i$ who owns correct $ID_i$ and $P_i$ can pass the authentication of server $S$. On the other hand, the user $U_i$ authenticates server $S$ by explicitly checking whether the other party communicating with can obtain the correct session key $SK = h(ID_i||y_i||N_u||N_s||CID_i)$ and compute the valid $C_2$ or not. Since the malicious server does not know the values of $N_u$, $ID_i$ and $y_i$ corresponding to user $U_i$, only the legitimate server can compute the correct session key $SK$ and $C_2$. From the above analysis, we conclude that our scheme can achieve mutual authentication.

**(8) Denial of Service Attack:** Assume that an adversary has got a legitimate user $U_i$'s smart card. The smart card checks the validity of user identity $ID_i$ and password $P_i$ before the password update procedure. Since the smart card computes $A_i^* = h(h(b||P_i^*)||ID_i^*))$ and compares it with the stored value of $A_i$ in its memory to verify the legality of the user before the smart card accepts the password update request, it is not possible for the adversary to guess out identity $ID_i$ and password $P_i$ correctly at the same time in real polynomial time. Accordingly, once the number of login failure exceeds a predefined system value, the smart card will be locked immediately. Therefore, the proposed protocol is secure against denial of service attack.

**(9) Online Password Guessing Attack:** In this type of attack, the attacker pretends to be a legitimate client and attempts to login to the server by guessing different words as password from a dictionary. In the proposed scheme, the attacker first has to get the valid smart card and then has to guess the identity $ID_i$ and password $P_i$ corresponding to user $U_i$. It is not possible to guess out identity $ID_i$ and password $P_i$ correctly at the same time in real polynomial time. Therefore, the proposed protocol can thwart online password guessing attack.

**(10) Forward Secrecy:** In our scheme, the session key $SK$ is generated with the contribution of identity $ID_i$ and security parameter $y_i$, thus the attacker cannot compute the previously generated session keys without knowing the correct value of $ID_i$ and $y_i$ corresponding to user $U_i$, even the attacker knows the server $S$'s long time private key $d$. As a result, our scheme achieves forward secrecy.

## 6 Performance Analysis

We compare the performance and security features among the relevant password-based authentication schemes and our proposed scheme in this section. The comparison results are depicted in Table 2 and 3, respectively.

**Table 2.** Performance comparison among relevant authentication schemes

| | Our scheme | Tsai et al. [14] (2010) | Chung et al. [4] (2009) | Horng et al. [5] (2010) | Kim et al. [6] (2011) |
|---|---|---|---|---|---|
| Total computation cost | $2T_E$+$14T_H$ | $3T_E$+$10T_H$ | $4T_E$+$12T_H$ | $7T_E$+$4T_S$+$8T_H$ | $3T_E$+$6T_H$ |
| Communication overhead | 1536 bits | 2560 bits | 2656 bits | 2432 bits | 1664 bits |
| Storage cost | 3456 bits | 2176 bits * | 3232 bits | 3328 bits | 1280 bits |

* It's likely that a parameter was missed out when Tsai et al. designed the registration phase.

**Table 3.** Security features comparison among relevant authentication schemes

| | Our scheme | Tsai et al. [14] | Chung et al. [4] | Horng et al. [5] | Kim et al. [6] |
|---|---|---|---|---|---|
| Preserving user anonymity | Yes | Yes | Yes | Yes | No |
| Resistance to offline password guessing attack | Yes | Yes | Yes | Yes | Yes |
| Resistance to stolen verifier attack | Yes | Yes | Yes | Yes | Yes |
| Resistance to user impersonation attack | Yes | Yes | Yes | Yes | Yes |
| Resistance to server masquerading attack | Yes | Yes | Yes | Yes | Yes |
| Resistance to replay attack | Yes | Yes | Yes | Yes | Yes |
| Resistance to parallel session attack | Yes | Yes | Yes | Yes | Yes |
| Resistance to denial of service attack | Yes | No | Yes | No | No |
| Resistance to online password guessing attack | Yes | Yes | Yes | Yes | No |
| Resistance to password disclosure to server | Yes | No | Yes | Yes | Yes |
| Mutual authentication | Yes | Yes | Yes | Yes | Yes |
| Forward secrecy | Yes | No | Yes | No | Yes |

Since the login phase and verification phase are executed much more frequently than the other two phases, only the computation cost, communication overhead and storage cost during the login and verification phase are taken into consideration. Note that the identity $ID_i$, password $P_i$, random numbers, timestamp values and output of secure one-way hash function are all 128-bit long, while $n$, $e$, $d$ and $g$ are all 1024-bit long. Let $T_H$, $T_E$, $T_S$ and $T_X$ denote the time complexity for hash function, exponential operation, symmetric cryptographic operation and XOR operation respectively. Since the time complexity of XOR operation is negligible as compared to the other three operations, we do not take $T_X$ into account. Typically, time complexity associated with these operations can be roughly expressed as $T_E >> T_S \geq T_H >> T_X$ [16-17].

In our scheme, the parameters $N_i$, $A_i$, $B_i$, $n$, $e$ and $g$ are stored in the smart card, thus the storage cost is $3456$ $(=3*128+3*1024)$ bits. The communication overhead includes the capacity of transmitting message involved in the authentication scheme, which is 1536 $(=4*128+1024)$ bits. During the login and verification phase, the total computation cost of the user and server is $2T_E+14T_H$. The proposed scheme is more efficient than Chung et al.'s scheme, Kim et al.'s scheme and Horng et al.'s scheme. As compared to Tsai et al.'s scheme, our scheme requires less computation cost and communication overhead; to conquer all the identified security flaws, the increase of some additional storage is reasonable and unavoidable.

In particular, since smartcards are usually characterized as resource-constrained and low-powered devices, computation cost at the user end is always regarded as a key criterion for smartcard-based schemes. In our scheme, the computation cost at user end is $T_E+9T_H$ during the login and verification phases. Clearly, $T_E$ is the most time-consuming operation and contributes the main overhead at user end. What needs further investigation is that, in practice, the encryption exponent $e$ of this exponential operation is often very limited, such as $e$=3 and $e$=7, and a widely accepted encryption exponent is $e=2^{16}+1$ [18]. As the studies in [17,19] suggest, when the encryption exponent $e$ is much smaller than the modular $n$, the time taken for pure computation is significantly shorter than that of common exponential operation with big exponents, and thus it is acceptable to conduct one such exponential operation in resource-limited environments, e.g., smart card based applications.

Table 3 gives a comparison of the security features of the proposed scheme with the other password-based authentication schemes. Our improved scheme and the scheme proposed in [4] can provide all eleven security feature, while the schemes presented in [5] and [6] are susceptible to several threats. It is clear that our scheme achieves the highest security strength with negligible decrease of performance as compared to other relevant schemes with non-tamper resistant smart cards.

It should be noted that, in our scheme, server $S$ maintains an account-database, which contains users' security parameters for authentication. If the adversary performs any unauthorized modifications on the account-database, the data will become inconsistent and the system may be crumbled. Thus, special security measures should be taken to eliminate such risks. Fortunately, the countermeasure is not complicated: $S$ can routinely and frequently make offsite backup of the account-database and check the consistency, and restore the account-database by using the offsite backup when necessary. Thus, there is a trade-off between performance and functionality in our scheme, while this trade-off is inevitable for authentication schemes with provision of sound reparability [4].

## 7 Conclusion

In this paper, we have shown that, besides some practical pitfalls, Tsai et al.'s scheme suffers from denial of service attack and fails to provide forward secrecy. As to our main contribution, a robust dynamic ID-based authentication scheme is proposed to remedy these identified flaws, the security and performance analysis demonstrate that our presented scheme achieves all of the security requirements with high efficiency, and thus our scheme is more secure and efficient for practical application environment. Remarkably, our scheme eliminates several hard security threats that are difficult to be solved in the previous scholarship at the same time. Since our security analysis is still scenario-based, in future work, we will perform a more rigorous security analysis of our scheme by employing some suitable formal methods.

**Acknowledgements.** This research was supported by the National Natural Science Foundation of China (NSFC) under Grants No. 61170241 and No. 61073042.

## References

1. Chang, C.C., Wu, T.C.: Remote password authentication with smart cards. IEE Proceedings-E 138(3), 165–168 (1993)
2. Ku, W.C., Chen, S.M.: Weaknesses and improvements of an efficient password based remote user authentication scheme using smart cards. IEEE Transactions on Consumer Electronics 50(1), 204–207 (2004)
3. Liao, I.E., Lee, C.C., Hwang, M.S.: A password authentication scheme over insecure networks. Journal of Computer and System Sciences 72(4), 727–740 (2006)
4. Chung, H.R., Ku, W.C., Tsaur, M.J.: Weaknesses and improvement of Wang et al.'s remote user password authentication scheme for resource-limited environments. Computer Standards & Interfaces 31(4), 863–868 (2009)
5. Horng, W.B., Lee, C.P., Peng, J.: A secure remote authentication scheme preserving user anonymity with non-tamper resistant smart cards. WSEAS Transactions on Information Science and Applications 7(5), 619–628 (2010)
6. Kim, J.Y., Choi, H.K., Copeland, J.A.: Further Improved Remote User Authentication Scheme. IEICE Transactions on Fundamentals 94(6), 1426–1433 (2011)
7. Wilson, S.B., Johnson, D., Menezes, A.: Key Agreement Protocols and Their Security Analysis. In: Darnell, M.J. (ed.) Cryptography and Coding 1997. LNCS, vol. 1355, pp. 30–45. Springer, Heidelberg (1997)
8. Das, M.L., Saxena, A., Gulati, V.P.: A dynamic ID-based remote user authentication scheme. IEEE Transactions on Consumer Electronics 50(2), 629–631 (2004)
9. Chien, H.Y., Chen, C.H.: A remote authentication scheme preserving user anonymity. In: IEEE AINA 2005, pp. 245–248. IEEE Computer Society, Los Alamitos (2005)
10. Wang, Y.Y., Kiu, J.Y., Xiao, F.X.: A more efficient and secure dynamic ID-based remote user authentication scheme. Computer Communications 32(4), 583–585 (2009)
11. Tsai, J.L., Wu, T.C., Tsai, K.Y.: New dynamic ID authentication scheme using smart cards. International Journal of Communication Systems 23(12), 1449–1462 (2010)
12. Gong, L.: A security risk of depending on synchronized clocks. ACM Operating System Review 26(1), 49–53 (1992)

13. Kocher, P., Jaffe, J., Jun, B.: Differential Power Analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999)
14. Messerges, T.S., Dabbish, E.A., Sloan, R.H.: Examining Smart-Card Security under the Threat of Power Analysis Attacks. IEEE Transactions on Computers 51, 541–552 (2002)
15. Tsai, C.S., Lee, C.C., Hwang, M.S.: Password Authentication Schemes: Current Status and Key Issues. International Journal of Network Security 3(2), 101–115 (2006)
16. Schneier, B.: Applied cryptography,protocols, algorithms, and source code in C, 2nd edn. John Wiley and Sons Inc., New York (1996)
17. Wong, D.S., Fuentes, H.H., Chan, A.H.: The Performance Measurement of Cryptographic Primitives on Palm Devices. In: Proceedings of ACSAC 2001, pp. 92–101. IEEE Computer Society, Washington, DC (2001)
18. Mao, M.B.: Modern Cryptography: Theory and Practice. Prentice Hall PTR, New Jersey (2004)
19. Potlapally, N.R., Ravi, S., Raghunathan, A., et al.: A study of the energy consumption characteristics of cryptographic algorithms and security protocols. IEEE Transactions on Mobile Computing 5(2), 128–143 (2006)

# Reconstructing Unsound Data Provenance View in Scientific Workflow*

Hua Hu[1], Zhanchen Liu[1], and Haiyang Hu[1,2]

[1] School of Computer Science HangZhou DianZi University, HangZhou, China
[2] State Key Laboratory for Novel Software Technology, Nanjing University

**Abstract.** The view of data provenance provides an approach of data abstraction and encapsulation by partitioning tasks in the data provenance graph (DPG) of scientific workflow into a set of composite modules due to the data flow relations among them, so as to efficiently decrease the workload consumed by researchers making analysis on the data provenance and the time needed in doing data querying. However, unless a view is carefully designed, it may not preserve the dataflow between tasks in the workflow. Concentrating on this scenario, we propose a method for reconstructing unsound view. We also design a polynomial-time algorithm, and analyze its maximal time complexity. Finally, we give an example and conduct comprehensive experiments to show the feasibility and effectiveness of our method.

**Keywords:** scientific workflow, data provenance graph, sound composite module.

## 1 Introduction

Technological advances have enabled the capture of massive amount of data in many different domains, taking us a step closer to solving complex problems such as global climate change and uncovering the secrets hidden in genes. The running of scientific workflow will produce large amounts of intermediate data and results, the management and analysis of those data will help scientists analysis complicated and multi-steps scientific experiment. The data provenance in scientific workflow system collects the history of data product, including original data source, intermediate data products and the production steps of the application of data products. Due to the complexity of scientific workflow system and the enormous of the data provenance which produced in the running of scientific workflow system, how to query and manage those data effectively is becoming one of important issues for scientists. In order to hide unrelated data provenance details and reduce the size of the data provenance graph, now many research work is aimed at constructing data provenance view of scientific workflow.

* This work is partially supported by the National Natural Science Foundation of China under Grant No.60873022, 60903053, 61003047, the Natural Science Foundation of Zhejiang Province (Z1100822), the Open Foundation of State Key Laboratory for Novel Software Technology of Nanjing University (KFKT2011B07).

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 212–220, 2012.

The main contribution of the paper is as follows:

1. We propose a method of detecting unsound view in data provenance and develop an efficient algorithm for correcting unsound view into sound view.
2. The time complexity of the algorithm for correcting unsound view into sound view is $o(n^2)$.

The rest of the paper is organized as follows: related work is reviewed in Section 2; Section 3 introduces the data provenance graph model and defines the construction of sound composite modules; Section 4 presents the design of an effective algorithm Experimental results and analysis are discussed in Section 5, followed by summary in Section 6.

## 2 Related Work

In recent years, the querying of data provenance graph in Scientific Workflow have been paid more attention in academic circles and proposed a large number of technologies to reduce the complexity of the querying of data provenance. Researchers put forward many methods about querying and processing large amounts of data of graph [1-3]. The existing research achievement mainly includes processing inquires problem on a large number of small size graph, including subgraph inquires [4], reachability inquires [5], label constrain inquires[6]. Biton et al. proposed the concept of user view and gave details of the usage of user view in the querying of data provenance in [7]. Based on [7], they optimized the construction of the user's view and proposed a better algorithm that can reduce the time complexity in the data provenance graph which only contains series and parallel structure in [8]. Sun et al. put forward a concept of sound view and developed two new algorithms for constructing sound views based on the concept of sound view in [1]. Those two algorithms can correct unsound view effectively.

On the other hand, the researcher put forward many methods about the querying of certain graph, such as [4-6,9-11], the related research achievements are mainly focused on the querying of subgraph [4], the querying of reachability of graph [5], the querying of label constraint reachability of graph [6]. The querying of uncertain graph include the management of uncertain graph data [12] and so on.

## 3 A Base Model

In this section, we first introduce some basic concepts about the algorithm. The definition of data provenance graph in scientific workflow is as follows:

**Definition 1.** *The data provenance graph in scientific workflow can be expressed as a directed acyclic graph* $G_{DP}=(V_G, E_G, L_G)$*, among them* $V_G=\{v_1,\ldots, v_N\}$,$(n=|V_G|)$, $V_G$ *stand for a set of task in workflow;* $E_G=\{e_{ij}\}$ *stand for a set of data flow between tasks;* $L_G$ *defines the label of* $V_G$ *and* $E_G$ .

The definition of the view of data provenance graph is as follows:

**Definition 2.** *The view of data provenance graph G can be represented as* $W=\{V_W, E_W, \pi_W, L_W\}$, *the view divides G into a couple of non-intersect composite tasks set* $V_W$: $\{P_1, P_2,...,P_n\}$, $\forall P_i,P_j \in V_W$, $P_i \cap P_j=\Phi$, $\bigcup_{i=1}^{n} P_i = V_G$, *and the edges set in G,* $E_G$ *is defined by a surjection from* $E_G$ *to* $E_W$: $\Psi$: $E_G \rightarrow E_W$, *in other words,* $\forall v_p,v_q \in V_G$, *if 1)* $v_p \in P_i$, $v_q \in P_j$ $(P_i \neq P_j)$, *and there is edge* $e_{pq} \in E_G$, *then* $(P_i,P_j) = \Psi(e_{pq}) \in E_W$; *2) if* $v_p \in P_i$, $v_q \in P_j$ $(P_i=P_j)$, *then* $\Psi(e_{pq}) = \varepsilon$.

In this paper, for any composite nodes set, $P_i \in V_W$ can be called composite task or composite module and for any node $v_i \in V_G$ can be called atomic task. For any two graph $G_i=\{V_{G_i}, E_{G_i}\}$, $G_j=\{V_{G_j}, E_{G_j}\}$, if $V_{G_j} \subseteq V_{G_i}$ and $E_{G_j} \subseteq E_{G_i}$, then it is called that $G_j$ is included in $G_i$ in this paper.

**Definition 3.** *A Composite task is sound in data provenance graph if and only if : (1)* $\forall P_i, P_j \in V_W$, *if there is a path between* $P_i$ *and* $P_j$ *then there must be a corresponding path* $\Psi^{-1}(L_1)*C_1, \Psi^{-1}(L_2)*C_2 \ldots\ldots \Psi^{-1}(L_p)*C_p$ *in uncertain data provenance graph* $G=\{V_G, E_G, \pi_G, L_G\}$, *which* $C_m (1 \le m \le p-1)$ *is a internal path in composite task. (2)* $\forall v_i,v_j \in V_G$, *if there is a path* $l_1,l_2,\ldots,l_p$ *from* $v_i$ *to* $v_j$, *then there must be a path* $\Psi(l_1),\ldots,\Psi(l_p)$ *in* $V_W$.

## 4 Our Algorithm

In this section, we described how to detect unsound composite task and how to reconstruct unsound composite task. During the reconstruct of unsound composite task, we list the combined standard of different types of nodes. Finally, we integrate the rule of combined standard of different types of nodes into a comprehensive algorithm.

### 4.1 Preliminaries

In this section, we describe the usage of several concepts which is used in this paper.

**Definition 4.** *Given a composite task* $P=\{V_P,E_P\}$, *and* $\exists v_i \in V_p$, $\forall v'_w \in P_i$, *the pre-order set of P can be noted as* $H_P(v_i)=\{ v_j \mid v_j \in v_p\}$, *there is a path in P:* $\rho= v_j \ldots v_i\}$; *especially, if* $(v_i,v_j) \in E_p$, *we call* $v_j$ *is the preorder with one edge of* $v_i$ *in P, and we can noted first-order preceding of* $v_i$ *in P as* $H_P(v_i,1)$.

**Definition 5.** *Given a composite task* $P=\{V_p,E_P\}$, *and* $\exists v_i \in V_P$, $\forall v'_w \in P_i$, *the postorder set of P can be noted as* $Q_p(v_i)=\{v_j \in V_P$, *there is a path in P:* $\rho= v_j \ldots v_i\}$; *especially, if* $(v_i,v_j) \in E_p$, *we call* $v_j$ *is the preorder with one edge of* $v_i$ *in P, and we can noted first-order posterior of* $v_i$ *in P as* $Q_p(v_i,1)$.

Based on Definitions 4 and 5 we can get the definition of input node set and output node set in *P*.

**Definition 6.** *For a composite task $P=\{V_p,E_P\}$ in data provenance graph G, the input node set is $IS(P)=\{v_i|v_i\in V_p$, and $H_p(v_i,1)\neq H_G(v_i,1)\}$.*

**Definition 7.** *For a composite task $P=\{V_p,E_P\}$ in data provenance graph G, the output node set is $OS(P)=\{v_i|v_i\in V_P$ and $Q_p(v_i,1)\neq Q_G(v_i,1)\}$.*

Obviously, if $v_i$ is a output node in $P$ , then $v_j\in Q_G(v_i,1)$ and $v_j\in v_p$.

**Definition 8.** *For a composite task $P=\{V_p,E_P\}$ in data provenance graph G, for any $v_i$ node in P, the input node set of $v_i$ in P is $is_P(v_i)=\{v_j|\ v_j\in H_G(v_i)\cap IS(P)\}$ .*

### 4.2 Unsound Composite Module Detecting

According to Definition 3, if the view $W=\{V_W, E_W, L_W\}$of $G$ is sound if and only if all the composite tasks in $G$ is sound. ie. if $\forall P\in V_W$ , $\forall v_i\in OS(P)$ , then $is_p(v_i)=IS(P)$. So we can redefine definition 3 as follows:

**Definition 9.** *The view $W=\{V_W, E_W, L_W\}$ of G is sound if and only if all the composite tasks in W is sound. In other words, if $\forall P\in V_W$ , $\forall v_i\in OS(P)$ , then $is_p(v_i)=IS(P)$.*

According to the definition, we can know that for any composite tasks $P$, if $\square v_i\in OS(P)$ and $is_p(v_i)\in IS(P)$, then $P_i$ is unsound composite tasks. We adopt the *Unsound Task Detection Algorithm* (UTDA) in [1] to detect the unsound composite tasks. When input data provenance graph, UTDA first check whether each composite task is sound in the view according to the function *IsSound*; if the composite task is unsound, then use the algorithm of *ConstructSoundView* to reconstruct unsound composite task(we will dwell on the algorithm of *ConstructSoundView* in 4.3 and 4.4).

### 4.3 The Standard of Node Combine

The following is the standard of all kinds of node can be merged.

**Rule 1:** *For any node $P_i\in V_p$, if $\exists v_A\in H_G(P_i,1)$ and $v_A\in v_p$, if $Q_G(v_A,1)=\{P_i\}$ ,then $P_i$ can be merged with $v_A$ into a new composite task.*

**Rule 2:** *For any node $P_i\in V_p$, if $\exists v_A\in Q_G(P_i,1)$ and $v_A\in V_p$, if $H_G(v_A,1)=\{P_i\}$, then $P_i$ can be merged with $v_A$ into a new composite task.*

According to Rule1, Rule2, we develop a comprehensive algorithm called NCNC (in Fig. 1).

**Algorithm 1**: NCNC (no circle-node checking)
**Input**: $P_i$ and a preceding vertex of $P_i$ $v_A$ ;
**Output**: return true or false;
1. Compute the $H_G(P_i,1)$of $P_i$ and $Q_G(v_A,1)$ of $v_A$.
2. if( $H_G(P_i,1)$={ $v_A$ } )
3. return true;
4. if($Q_G(v_A,1)$={ $P_i$ })
5. return true;
6. return false;

**Fig. 1.** Algorithm of no circle-node checking

**Rule 3:** *For the node $P_i \in Vp$, if there is a node $v_A \in H_p(P_i,1)$, $v_A \in v_p$, for any node $\forall v_j \in H_p(P_i,1)$-$\{v_A\}$,we have $v_j \in v_p$ and $Q(v_j,1)=\{v_A,P_i\}$ then $P_i$ and $v_A$ can be merged into a new composite task.*

The pseudo-code (in Fig.2) of Rule 3 is shown in algorithm 2 ( **CNC**).

**Algorithm 2**: The CNC algorithm
**Input**: $P_i$ and $H_P(P_i,1)$;
**Output**: return true or false;
1. if (( $\forall v_j \in H_P(P_i,1)$, $Q_G(v_j,1)>1$) and ($\exists v_k \in H_P(P_i,1)$, $Q_G(v_j,1)>2$)) then
2. return false
3. else
4. for each $v_A$ in $H_P(P_i,1)$ do
5. beMerged ← true; length ++;
6. if ( $\forall v_j \in H_P(P_i,1) \setminus \{v_A\}$ and $Q_G(v_j,1)=\{v_A, P_i\}$ ) then
7. return true;
8. else
9. return false;
10. return false;

**Fig. 2.** Algorithm of circle node checking

The combine standard of the nodes which have common preorder set is as follows: Given a composite task $P$, if a part of it's subgraph $P_T=\{v_i\} \in V_p$ , if $H_p(P_T,1)= \bigcup H_G(v_i)$, $\forall v_i \in PT$, then we can merge $P_T$ and $H_G(P_T,1)$ into a new composite task, where $H_G(P_T,1)$ is the preorder of $P_T$.

**Rule 5:** *For the composite task P, if a part of it's subgraph set $P_T=\{v_i\} \in V_p$, if $\forall v_i, v_j \in P_T$ then $H_p(P_T,1)=H_G(v_j,1)=H_G(v_i,1) \subseteq P$, and for $\forall v_j \in H_G(v_i,1)$, $Q_G(v_i,1) =P_T$, $\forall v_B \in H_G(v_i,1)$, $v_B \in OS(P)$, $\forall v_i \in PT$, $v_i \in IS(P)$, we can merge $P_T$ and $H_G(v_i,1)$ into a new sound composite task $P_c$ and the new $P_c$ is a sound composite task.*

**Algorithm 3**: CPNC
**Input:** $PT$ and $H_P(PT,1)$;
**Output**: return true or false;
1. if ($\forall v_i \in H_P(PT,1)$and $Q_G(v_i,1) \neq PT$)
2. return false;
3. else if ($\forall$each $v_B \in H_G(v_i,1)$ and ($v_B \in OS(P)$)
4. return false;
5. else if ($\forall v_i \in PT$ and $v_i \in IS(P)$)
6. return false;
7. else
8. return true;

**Fig. 3.** Algorithm of common preorder node checking

**Algorithm 4**: ConstructSoundView
**Input**: unsound $P_i$ and $G$;
**Output**: a set of sound nodes { $v_1$ , $v_2$ ,…, $v_k$ };
1. Computer the input node set $IS$ ($P_i$), output node set $OS$ ($P_i$)
2. Compute $H_G(v_k,1)$ and $Q_G(v_k,1)$ for each $v_k$ in $P_i$
3. push each $v_k$ in $OS$ ($P_i$) into a queue $q$
4. Sort {$v_k$ }in $q$ due to the number of their {$H_G(v_k,1)$}
5. while ($q$ != NULL)
6. pop all the elements from $q$ *to* $q'$;
7. find those elements which share a common set of parents: $\Omega(P_i) = \{PT_1, PT_2, \ldots, PT_n\}$;
8. for each $PT_j$ in $\Omega(P_i)$
9. if it can be merged with $v_i \in PT_j \cup H_P(v_k,1)$ using the **CPNC** then
10. merge them into a set $P_c$ and push them into $q$;
11. if they can't be merged then
12. for each $v_k$ in $PT_j$
13. put the $H_P(v_k,1)$ into $h$;
14. push all nodes in $PT_j$ and $h$ into $q$;
15. pop all elements from $q$ and put them into $s$;
16. for each element $v_k$ in $s$
17. if can be merged with $H_P(e_k,1)$ using the **CNC**
18. merge $v_A$ and $H_P(e_k,1)$into $v_k$ and push $v_k$ into $q$;
19. if they can't be merged then
20. for each $v_k$ in $H_P(e_k,1)$
21. if $v_k$ and $v_A$ *of* $H_P(e_k,1)$ can be merged with **NCNC** then
22. merge them into $v_k$ and put $v_k$ into $q$;
23. if $q = q'$ then reconstruction is completed.
24. return $q$.

**Fig. 4.** Algorithm of constructing unsound composite task

According to Rule 5, we develop **CPNC** algorithm (in Fig. 3), in this algorithm we judge whether a set of nodes and their common preorder nodes with one edge can be merged together.

### 4.4 The Reconstruction of the View

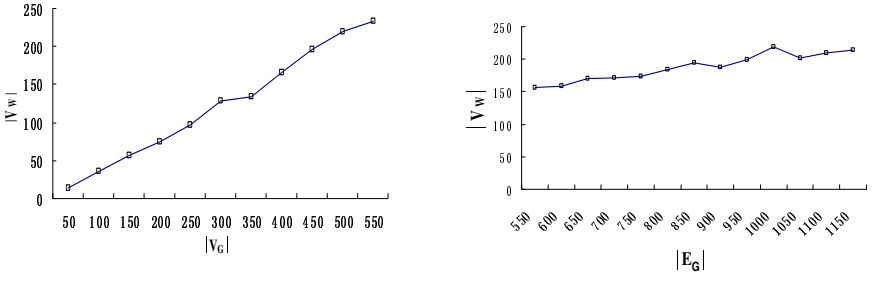
Based on the above analysis, we develop an algorithm of the reconstruction of unsound composite task in data provenance graph (in Fig. 4). In this algorithm we first find the input node set and output node set of unsound task, and put them into $IS(P_i)$ and $OS(P_i)$ separately. Then we push $OS(P_i)$into queue $q$. We can check the nodes in the queue and their preorder node with one edge use NCNC, **CNC** and **CPNC**, then push the new composite task into queue. When there are no nodes in $q$ can be merged with other nodes , the algorithm will be terminated.

## 5 Experimental Evaluation

We randomly generated a lot of graph. We divide those graphs into several composite tasks and the number of composite tasks is 70% of the total number nodes of graph.

### 5.1 Effect of the Nodes and Edges on the Number of Sound Views

In the first experiment, as shown in Fig. 5(a), we can see that the number of the edges $|E_G|$ is twice as the number of nodes($|E_G|=2|V_G|$). From Fig. 5 we can summarize that the higher the number of the nodes the more the composite tasks, but there is no big change of the composite tasks with the increase of the edges. That is because the number of the composite tasks depends much on the number of the nodes rather than the edges.

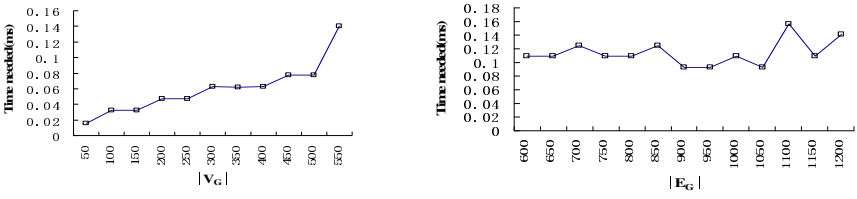



(a) Number of composite tasks generated (b) Number of composite tasks generated

**Fig. 5.** Impact of $|V_G|$and impact of $|E_G|$

### 5.2 Effect of the Nodes and Edges on the Construction Time

In this experiment, we test the effect of nodes and edges on the construction time. In Fig. 6(a), the number of the edges $|E_G|$ is twice as the number of nodes ($|E_G|=2|V_G|$). From Fig. 6(a) we can see that when $|V_G|$ increased, the construction time will increase. But there is no big change in the construction time when the threshold is changed. In Fig. 6(b), the number of the nodes $|V_G|$ is 550, when the number of edges increased, we get the construction time of sound composite tasks. From this figure we can see there is no big change of the construction time when $|E_G|$ increase.

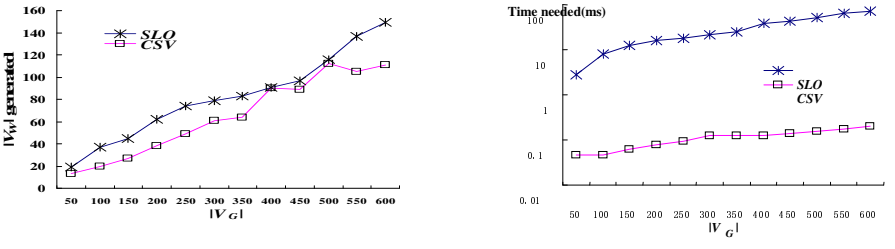



(a) Impact of $|V_G|$ (b) Impact of $|E_G|$

**Fig. 6.** Time needed to generate sound view

### 5.3 The Comparison of SLO and CSV

In the final experiment, as shown in Fig. 7, we compared the *CSV* of this paper with Strong Local Optimum algorithm (*SLO*) in [1]. The number of the edges $|E_G|$ is twice as the number of nodes ($|E_G|=2|V_G|$). In Fig. 7, we can see that in the same workflow graph the number of sound composite tasks $|V_W|$ in *CSV* is less than *SLO*. This indicates that *CSV* is more effective than *SLO* in finding the relationship of the data flow. On the other hand, the construction time of *CSV* is less than *SLO* obviously. That is because the time complexity of *SLO* is $O(n^3)$ [1] and the time complexity of *CSV* is $O(n^2)$.




(a) Comparison of *SLO* and *CSV* on $|V_G|$ (b) Comparison of *SLO* and *CSV* on time needed

**Fig. 7.** Comparison of *SLO* and *CSV*

## 6 Conclusions and Future Work

This paper is committed to how to reconstruct unsound composite tasks, and develop an effective algorithm for the reconstruction of unsound composite tasks. We put forward the concept of data provenance graph and the sound composite tasks in data provenance graph. We developed a nice method for reconstructing the unsound composite tasks. Finally, we give an example and conduct comprehensive experiments to show the feasibility and effectiveness of our method.

In our future work we will focus on how to optimize the construction time and decrease the number of the view. For future research, we note that detecting and resolving unsound composite task in uncertain data provenance graph is still a big challenge.

## References

1. Sun, P., Liu, Z.Y., Susan, D., Chen, Y.: Detecting and resolving unsound workflow views for correct provenance analysis. In: Cetintemel, U., Zdonik, S.B., Kossmann, D., Tatbul, N. (eds.) The ACM SIGMOD International Conference on Management of Data, pp. 549–562. ACM, Rhode Island (2009)
2. Zou, Z.N., Li, J.Z., Gao, H., Zhang, S.: Mining frequent subgraph patterns from uncertain graphs. Journal of Software 20, 2965–2976 (2009)
3. Chui, C.-K., Kao, B., Hung, E.: Mining Frequent Itemsets from Uncertain Data. In: Zhou, Z.-H., Li, H., Yang, Q. (eds.) PAKDD 2007. LNCS (LNAI), vol. 4426, pp. 47–58. Springer, Heidelberg (2007)
4. Hintsanen, P., Toivonen, H.: Finding reliable subgraphs from large probabilistic graphs. In: Daelemans, W., Goethals, B., Morik, K. (eds.) Data Mining and Knowledge Discovery, vol. 17, pp. 3–23. Springer, Tucson (1997)
5. Cheng, J., Yu, J., Lin, X.: Fast computing reachability labelings for large graphs with high compression rate. In: Kemper, A., Valduriez, P., Mouaddib, N., Teubner, J., Bouzeghoub, M., Markl, V., Amsaleg, L., Manolescu, I. (eds.) The 11th International Conference on Extending Database Technology, pp. 193–204. ACM, Nantes (2008)
6. Jin, R., Hong, H., Wang, H.X., Ruan, N., Xiang, Y.: Computing label-constraint reachability in graph databases. In: Elmagarmid, A.K., Agrawal, D. (eds.) The ACM SIGMOD International Conference on Management of Data, pp. 123–134. ACM, Indianapolis (2010)
7. Biton, O., Davidson, S.B., Khanna, S., Roy, S.: Optimizing user views for workflows. In: Ronald, F. (ed.) The 12th International Conference on Database Theory, pp. 310–323. ACM, Saint-Petersburg (2009)
8. Biton, O., Boulakia, S.C., Davidson, S.B., Hara, C.S.: Querying and managing provenance through user views in scientific workflows. In: The 24th Int'l Conf. on Data Engineering, pp. 1072–1081. IEEE, Cancun (2008)
9. Zhou, S.G., Yu, Z.C., Jiang, H.L.: Concepts, issues, and advances of searching in graph structured data. Communication 3, 59–65 (2007)
10. Shasha, D., Wang, T.L., Guigno, R.: Algorithmics and applications of tee and graph searching. In: Franklin, M.J., Moon, B., Ailamaki, A. (eds.) The 21st ACM SIGMOD-SIGART Symposium on Principles of Database Systems, Madison, pp. 39–52 (2002)
11. Yan, X., Yu, P.S., Han, J.: Graph indexing: A frequent structure based approach. In: Weikum, G., König, A.C., Deßloch, S. (eds.) The 2004 ACM SIGMOD International Conference on Management of Data, pp. 335–346. ACM, Paris (2004)
12. Gao, H., Zhang, W.: Research status of the management of uncertain graph data. In: Communications of the China Computer Federation, pp. 31–36 (2009)

# A Process Distance Metric Based on Alignment of Process Structure Trees

Xiaodong Fu[1], Kun Yue[2], Ping Zou[3], Feng Wang[1], and Kaifan Ji[1]

[1] Yunnan Provincial Key Lab. of Computer Technology Application,
Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
xiaodong_fu@hotmail.com, {wangfeng,jkf}@cnlab.net
[2] School of Information Science and Engineering,
Yunnan University, Kunming 650091, China
kyue@ynu.edu.cn
[3] Faculty of Management and Economics,
Kunming University of Science and Technology, Kunming 650093, China
zoup@ynnu.edu.cn

**Abstract.** For various applications in today's service-oriented enterprise computing systems, such as process-oriented service discovering or clustering, it is necessary to measure the distance between two process models. In this paper, we propose a quantitative measure to calculate the distance or similarity between different block-structured processes. We first transform each process into a process structure tree, and then calculate the process distance based on the alignment of two process structure trees. The proposed distance metric satisfies four distance measure properties, i.e., non-negativity, identity of indiscernible, symmetry and triangle inequality. These properties make the distance metric can be used as a quantitative tool in effective process model management activities. We illustrate the methodology with examples, by which its features are shown. Moreover, experiment study shows that the proposed method is feasible.

**Keywords:** Business process, Process distance, Process structure tree, Tree alignment.

## 1 Introduction

In today's dynamic business world, success of an enterprise increasingly depends on its ability to react to changes with its environment in a quick, flexible and cost-effective way. Thus, it is indispensable to automate business collaboration within and across organizations to increase their competiveness and responsiveness to the fast evolving global economic environment. The widespread of mobile technologies has further resulted in an increasing demand for the support of business collaboration across multiple platforms anytime and anywhere. Along this trend a variety of process and service support paradigms as well as corresponding specification languages (e.g., SOA and WS-BPEL) have emerged to build process-aware information systems (PAISs) [1].

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 221–232, 2012.

With the increasing adoption of PAISs, large process model repositories have emerged. Typically, the models in such repositories are realigned to real-world events and demands through adaptation on a day-to-day basis. In large companies, such process repositories can easily contain thousands of process models [2]. Systematic methods of analyzing and improving the process asset is getting necessary, since business processes are continuously accumulated by those repositories [3]. Therefore, discovering and clustering process-oriented services has attracted more and more attention in the recent years [4]. Clearly, there is a need for effective calculation of distances or similarities between process models to facilitate certain model management activities. First, similar models as well as the corresponding business operations can be integrated into one process. This is interesting not only for refactoring the model repositories, but also for facilitating the integration of business operations in a merger scenario. Second, the reference models of an ERP system vendor could be compared to company processes. By this way, organizations could easily decide the packages that match their current operations best. Third, multi-national enterprises can identify specialized processes of some national branch, which no longer comply with the procedures defined in the company-wide reference model using a distance or similarity measurement. It is definitely that all these mentioned process model management activities are beneficial to provide useful and timely service to end users in more effective way.

Some papers have studied the process distance or similarity problem. The approaches presented in [3-5] measure the distance between two process models by determining their difference with respect to dependencies among activities. However, it does not intuitively take the structural aspects of process models into consideration. Another kind of process distance or similarity measure is based on the observed behavior as captured in execution logs [6]. These approaches can only be used in the context of processes have been executed. The measure provided in [7] is based on behavioral inheritance rules and it can only provide a binary answer to whether two models are the same or whether one model is a sub-model of the other. Reference [8] evaluates the distance and similarity between two process models based on high-level change operations (e.g., to add, deletes or move activities). This work can measure similarity in terms of a unique number but it doesn't take the nesting control structures of process models into account.

A block-structured process model [9] is constructed based on a set of predefined building blocks, i.e., activities, sequences, branching, and loops constitutes with unique start and end nodes. Block-structured process can be used in most of process modeling situation. Major process modeling languages (such as WS-BPEL) provide building blocks for structured modeling. Comparing with non-block-structured process models, block-structured models are easier understandable for users and have less chances of containing errors [10]. If a process model is not block-structured, it can be transformed into a block-structured one in most practically relevant cases [11].

In this paper, we discuss the foundations of measuring distance and similarity between block-structured process models. In particular, our contribution is an approach

to evaluate the distance and similarity between two block-structured process models based on the alignment of process structure trees. Some useful properties of the process distance metric are proved so that it can be used for processes discovering or clustering correctly and feasibly.

The remainder of paper is organized as follows. Section 2 introduces preliminaries as the basis of this paper. Section 3 describes the concept of process structure trees and proposes the distance measure based on the alignment of process structure trees. Section 4 illustrates features and feasibility of the distance metric with case studies and experiment. Section 5 discusses related work to our approach, and Section 6 concludes and discusses the future work.

## 2 Preliminaries

A process model is represented as a directed graph, which comprises a set of nodes - either representing process steps (i.e., activities) or controlling connectors (e.g., And/Xor-Split) and a set of control edges between them. The latter specifies precedence as well as loop backward relations. Each process model contains a unique start and a unique end node. For control flows modeling the following patterns are available: Sequence, AND-split, AND-join, XOR-split, XOR-join, and Loop [12]. These patterns constitute the core of any process specification language and cover most of the process models that can be found in practice [13]. Further, they can be easily mapped into other process execution languages like WS-BPEL (Web Service Business Process Execution Language) as well as formal languages like Petri Nets. Based on these patterns we are also able to compose more complex process structures if required (e.g., in principle, an OR-split can be mapped to AND- and XOR-splits [14]. Finally, by only using these basic process patterns, we obtain better understandable and less erroneous models [15].

In this paper we focus only on the construction of block-structured processes [9-11]. A block in a block-structured process *S* can be a single activity, a sequence, a parallel branching, a conditional branching, or even *S* itself. And these blocks may be nested, but cannot be overlapped, that is, their nesting must be regular.

Fig. 1 shows an example of block-structured process associated with fulfilling orders for online stores designed to process orders according to user requests [16]. At the highest level, the process is a sequential construct that consists of order creation, stock payment check, and shipping arrangement. Stock payment check is a parallel construct that comprises payment check and stock check. Payment check is in turn an exclusive choice construct composed of Paypal check and credit card check, and its selection is determined by the user as indicated in the order. Stock check is a loop construct on get items, which is executed for each ordered item. Each item will be acquired from the warehouse (get from warehouse) if it is in stock and from vendor (get from vendor) otherwise.

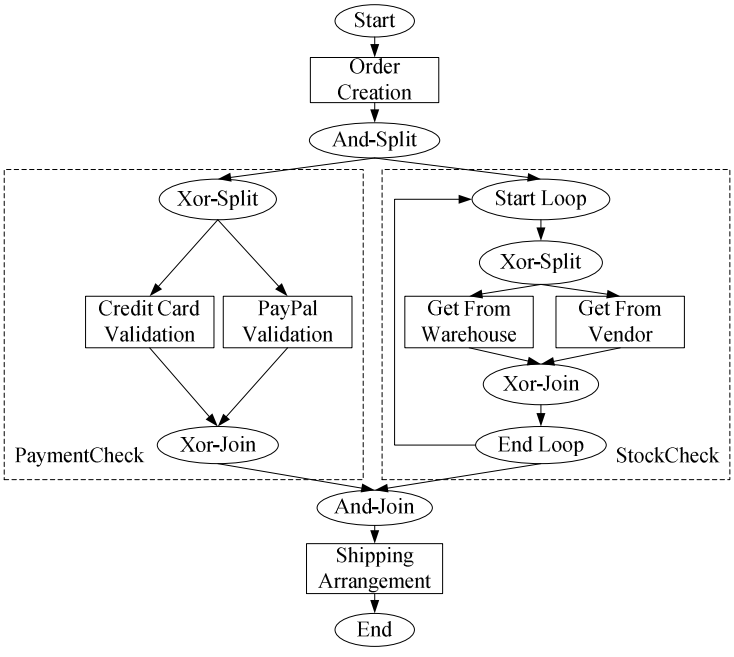



**Fig. 1.** An example of block-structured process designated order fulfillment

## 3 Processes Distance Metric Based on Alignment of Process Structure Trees

According to the definition of block-structured process, if a process model is block-structured, the process can be decomposed into more substructures using the various structured constructs (sequence, parallel, choice and loop). The decomposition process can be used on all substructure of a block-structured process recursively until each substructure is an atomic activity. Therefore, a block-structured process can be described as an equivalent tree. We define this kind of tree as a Process Structure Tree.

**Definition 1 (Process Structure Tree).** A process structure tree $T=<N;E;L>$ is a labeled tree that satisfies the following properties:

- $N=A \cup C$ is a set of nodes, where $A$ is a set of leaf nodes which represent atomic activities of a process, and $C$ is a set of non-leaf nodes which represent substructures of a process.
- $E \subseteq (A \times C) \cup (C \times C)$ is a set of edges.
- $L$ assigns a label $L(n)$ to each node $n \in N$ based on the node structure pattern. $L(n) \in \{Seq, AND, XOR, Loop\}$ for all $n \in C$, and $L(n) \in \{Act\}$ for all $n \in A$, where *Seq*, *AND*, *XOR*, *Loop* and *Act* represents Sequence, AND-block, XOR-block, Loop and atomic activity respectively.

For a process structure tree $T$, we denote the set of nodes and edges by $N(T)$ and $E(T)$ respectively. The size of $T$, denoted by $|T|$, is $|N(T)|$. Based on the approach presented in [11], a block-structured process can be transformed into a refined process structure tree in linear time. The refined process structure tree constitutes a unique representation of a block-structured process model. Fig. 2 shows the corresponding process structure tree of the block-structured process model illustrated in Fig.1. In such a tree, nodes (represented as rectangles) correspond to activities while connectors (represented as ellipses) represent their relations based on process patterns like Sequence, AND-block, XOR-block, and Loop. The precedence relations (expressed by connector *Seq* and *Loop*) are parsed from left to right, e.g., activity Order Creation precedes the *And* block (i.e., connector *And* and all its successors) in the corresponding process model since Order Creation is on the left side of connector *And*. In a process structure tree, activity nodes correspond to leaves, while substructures are non-leaves. Further, a process structure tree has a unique root node which represents the structure pattern of the whole process.

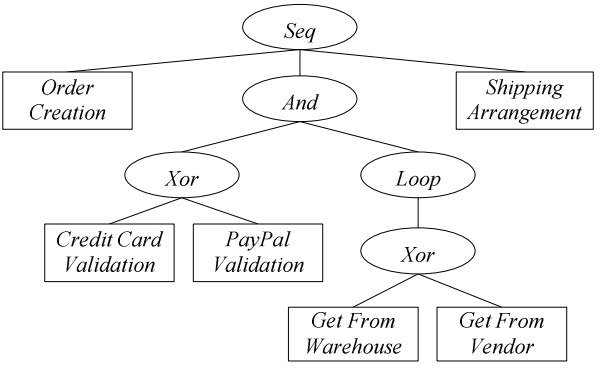



**Fig. 2.** Structure tree of the process in Fig. 1

The process structure tree provides a clear picture of the process model's structure and the relations between the activities. By analyzing structure trees, we can get the difference information between two processes and evaluate the distance between two processes. We first describe the concept of process tree alignment and then define the distance measure based on the alignment.

**Definition 2 (Adding Operation of Process Structure Tree).** Let $T$ be a process structure tree. Adding a node $u$ into $T$ means that for some node $v$ in $T$, we make $u$ the rightmost child of $v$ and there is no any other child node of $v$ that becomes the child node of $u$.

The adding operation differs from the insertion operation in tree edit [17] on that insertion operation can make some node in $T$ as child of the inserted node.

**Definition 3 (Alignment of Process Structure Trees).** Let $T_1$ and $T_2$ be two labeled process trees. An *alignment* $A(T_1,T_2)$ of $T_1$ and $T_2$ is obtained as follows. First we add nodes labeled with *spaces* (denoted by -) into $T_1$ and $T_2$ so that the two resulting trees

$T_1$' and $T_2$' have the same structure, i.e., they are identical if the labels are ignored. The resulting trees $T_1$' and $T_2$' are then overlaying on top of each other giving the alignment $A(T_1,T_2)$, which is a tree where each node $n$ is labeled by a pair of labels $(l_1,l_2)$, where $l_1 \in L_1$, $l_2 \in L_2$.

An example alignment of process structure tree is shown in Fig. 3.

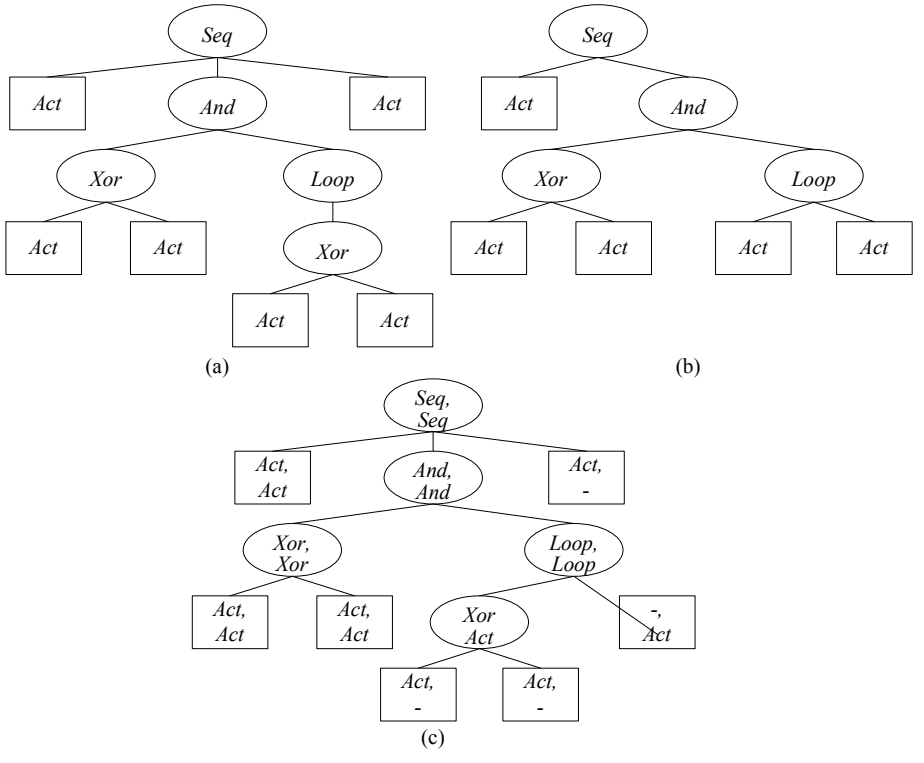



**Fig. 3.** (a) Tree $T_1$. (b) Tree $T_2$. (c) The alignment of $T_1$ and $T_2$.

In the alignment of two process structure trees, if the two components of a pair of labels $(l_1,l_2)$ for a node are different, then the corresponding substructures of two processes are different as well. By this way, we can intuitively find all structural differences between two processes. Therefore, we define the distance between process $p_1$ and $p_2$ as the number of different substructures between these two processes.

**Definition 4 (Process Distance).** If the process structure tree of process $p_1$ is $T_1$ and the process structure tree of process $p_2$ is $T_2$, then the *distance* $D(p_1, p_2)$ between $p_1$ and $p_2$ is the number of the pairs of opposing labels in alignment $A(T_1,T_2)$ of their structure tree, specified as follows:

$$D(p_1, p_2)=|\{n|n\in N(A(T_1,T_2))\wedge L(n)=(l_1,l_2)\wedge(l_1\neq l_2)\}| \quad (1)$$

As a distance function, process distance is required to satisfy some basic conditions include non-negativity, identity of indiscernible, symmetry, and triangle inequality [4]. These conditions of the distance metric express intuitive notions about the concept of distance. For example, that the distance between two processes is positive or zero (non-negativity) and the distance from a process to itself is zero (identity of indiscernible). On the other hand, the distance from process $p_1$ to $p_2$ is the same as the distance from process $p_2$ to $p_1$ (symmetry). Finally, the triangle inequality means that the distance from $p_1$ to $p_3$ via $p_2$ is at least as great as from $p_1$ to $p_3$ directly.

**Theorem 1.** Process distance measure satisfies distance metric properties, i.e., all following conditions are satisfied for all different processes $p_1$, $p_2$, $p_3 \in P$ ($P$ is the set of all block-structured processes):

1. Non-negativity, $D(p_1, p_2) \geq 0$.
2. Identity of indiscernible, $D(p_1, p_2)=0$ if and only if $p_1=p_2$.
3. Symmetry, $D(p_1, p_2)=D(p_2, p_1)$.
4. Triangle inequality, $D(p_1, p_3) \leq D(p_1, p_2)+D(p_2, p_3)$.

*Proof*

1. For any alignment $A(T_1,T_2)$, we know $|\{n|n \in N(A(T_1,T_2)) \wedge L(n)=(l_1,l_2) \wedge (l_1 \neq l_2)\}| \geq 0$, so $D(p_1, p_2) \geq 0$.
2. If $p_1=p_2$, then $|\{n|n \in N(A(T_1,T_2)) \wedge L(n)=(l_1,l_2) \wedge (l_1 \neq l_2)\}|=0$, so $D(p_1, p_2)=0$.
3. If $D(p_1, p_2) = 0$, then $|\{n|n \in N(A(T_1,T_2)) \wedge L(n)=(l_1,l_2) \wedge (l_1 \neq l_2)\}|=0$, which means $l_1=l_2$ for all node in $A(T_1,T_2)$. According to the definition of alignment, there must be $T_1=T_2$, and then $p_1=p_2$.
4. According to the definition of alignment and process distance, for any two processes $p_1$ and $p_2$, $A(T_1,T_2)=A(T_2,T_1)$, and then $D(p_1, p_2)=D(p_2, p_1)$.
5. $D(p_1, p_3)=|\{n|n \in N(A(T_1,T_3)) \wedge L(n)=(l_1,l_3) \wedge (l_1 \neq l_3)\}|$
$=|\{n|n \in N(A(A(T_1,T_2),T_3)) \wedge L(n)=(l_1,l_2,l_3) \wedge (l_1 \neq l_3)\}|$
$=|\{n|n \in N(A(A(T_1,T_2),T_3)) \wedge L(n)=(l_1,l_2,l_3) \wedge (l_1 \neq l_2) \wedge (l2 \neq l_3)\}|-$
$|\{n|n \in N(A(A(T_1,T_2),T_3)) \wedge L(n)=(l_1,l_2,l_3) \wedge (l_1 \neq l_2) \wedge (l_2 \neq l_3) \wedge (l_1=l_3)\}|$
$=|\{n|n \in N(A(T_1,T_2)) \wedge L(n)=(l_1,l_2) \wedge (l_1 \neq l_2)\}|+$
$|\{n|n \in N(A(T_2,T_3)) \wedge L(n)=(l_2,l_3) \wedge (l_2 \neq l_3)\}|-$
$|\{n|n \in N(A(A(T_1,T_2),T_3)) \wedge L(n)=(l_1,l_2,l_3) \wedge (l_1 \neq l_2) \wedge (l_2 \neq l_3) \wedge (l_1=l_3)\}|$
$= D(p_1, p_2)+D(p_2, p_3)-$
$|\{n|n \in N(A(A(T_1,T_2),T_3)) \wedge L(n)=(l_1,l_2,l_3) \wedge (l_1 \neq l_2) \wedge (l_2 \neq l_3) \wedge (l_1=l_3)\}|$
$\leq D(p_1, p_2)+D(p_2, p_3)$

Theorem 1 guarantees the process distance measure is, in fact, a distance metric.

The notion of distance is somewhat dual to similarity. A distance measure is a function that also associates a numeric value with a pair of processes, but with the idea that the larger the distance, the smaller the similarity, and vice versa. Obviously, the less different substructures between two processes mean the greater similarity of them, and vice versa. For this reason, we define the similarity $S(p_1, p_2)$ of processes $p_1$ and $p_2$ as follows:

$$S(p_1, p_2)=|\{n|n\in N(A(T_1,T_2))\wedge \mathrm{L}(n)=(l_1,l_2)\wedge(l_1=l_2)\}| \quad (2)$$

Traditionally, similarity measure is normalized so that similarity ranges from 0 to 1. So we can further define normalized similarity $S^n(p_1, p_2)$ as:

$$S^n(p_1, p_2)=S(p_1, p_2)/|A(T_1,T_2)| \quad (3)$$

Based on (1) and (3), it is easy to obtain:

$$S^n(p_1, p_2)=[|A(T_1,T_2)|-D(p_1, p_2)]/|A(T_1,T_2)| \quad (4)$$

Equation (4) denotes that the distance and the similarity of processes are interchangeable in the sense that a small distance means high similarity, and vice versa.

## 4 Example and Empirical Validation

To demonstrate the proposed process distance measure, an example of business process distance measurement is presented in this section. Figure 4 illustrates the example of four processes $p_1$, $p_2$, $p_3$ and $p_4$. These block-structured processes are derived from the process in Fig. 1 by means of a series of deletion or replacement operations [18]. Including the process illustrate in Fig. 1 (denoted as $p_5$), we have 5 processes in our case study.

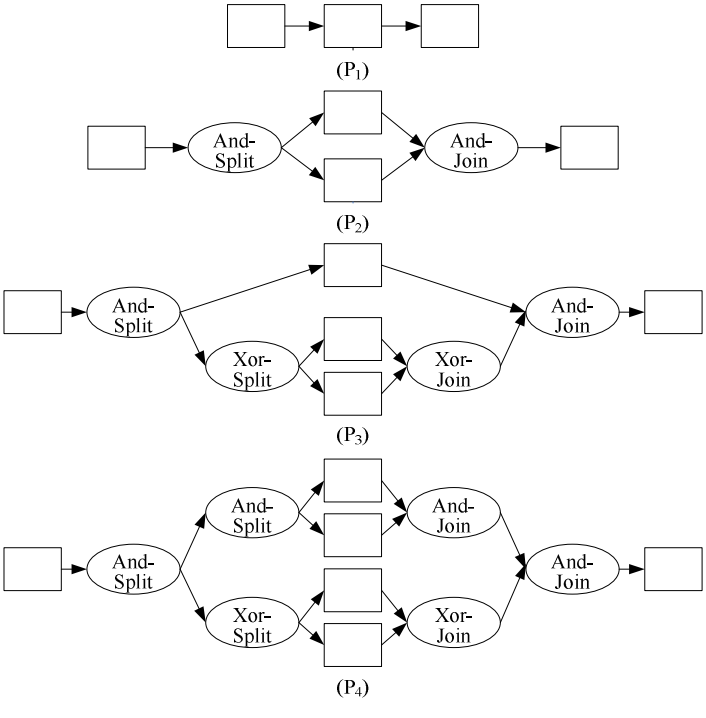



**Fig. 4.** Sample processes derived from the process in Fig. 1

All these processes are block-structure processes which can be transformed into process structure trees. Furthermore, the distance values between each pair of these 5 processes were calculated through the method explained in Sect. 3 (See Table 1). AS the distance between two processes is symmetrical (Theorem 1), we only fill in the upper triangle matrix in Table 1.

**Table 1.** Distances between illustrative processes

| | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|---|---|---|---|---|---|
| $p_1$ | 0 | 3 | 5 | 7 | 8 |
| $p_2$ | | 0 | 3 | 6 | 7 |
| $p_3$ | | | 0 | 5 | 8 |
| $p_4$ | | | | 0 | 7 |
| $p_5$ | | | | | 0 |

Accordingly, we can get the normalized similarity value between these 5 processes listed in Table 2.

**Table 2.** Similarities between illustrative processes

| | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ |
|---|---|---|---|---|---|
| $p_1$ | 1 | 0.5 | 0.38 | 0.3 | 0.27 |
| $p_2$ | | 1 | 0.63 | 0.4 | 0.36 |
| $p_3$ | | | 1 | 0.5 | 0.27 |
| $p_4$ | | | | 1 | 0.36 |
| $p_5$ | | | | | 1 |

Obviously, all the distances between any two processes in Table 1 satisfy the four distance measure properties, i.e., non-negativity, identity of indiscernible, symmetry and triangle inequality. The result also shows that the larger the distance, the smaller the similarity, and vice versa. These properties ensure the proposed distance measure is a distance metric.

In addition, we validated our approach to calculate the degree of similarity by computing its correlation with a similarity assessment of experiment participants. We obtained the similarity assessment using questionnaire that was distributed among trained process modelers. This questionnaire consisted of 50 pairs of process models. For each pair of models, we asked the participants whether they agreed or disagreed (on a 1 to 7 Likert scale) with the proposition: “These processes are similar.” To obtain a representative collection of model pairs, we selected the model pairs to be evenly distributed over the 0 to 1 similarity degree range.

Fig. 5 shows the correlation between the similarity computed by our proposed measure and the similarity assessment obtained from the questionnaire. Each point in Fig. 5 represents a pair of processes, with a similarity as indicated by its x-value and a human similarity assessment as indicated by its y-value. For this metric we got a high Spearman correlation coefficient of 0.96 with the human judgments. The correlation is represented as a straight line in Fig. 5. The result indicates that the similarity obtained with our measure highly coincides with the human judgment.

The properties indicated in the examples and the experiment mean the proposed similarity measure enables the feasible deployment of process classification, clustering, and retrieval problems—all desirable functionalities that are critical for fully supporting the effective process model management activities of an enterprise.

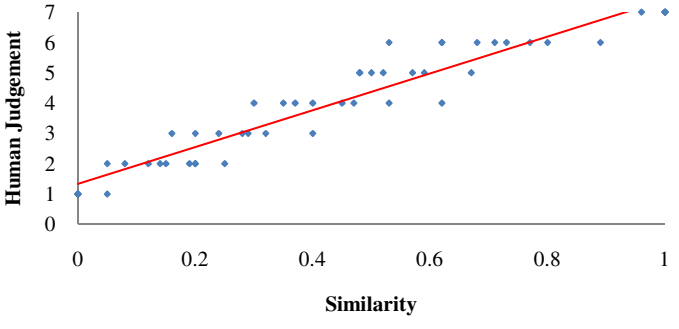



**Fig. 5.** Correlation between calculated similarity and human judgment

## 5 Related Work

Some papers have studied the process similarity problem and provided some useful results. By the measure in [6], the similarity between two process models can be obtained based on the observed behavior as captured in execution logs. This approach applies trace equivalence to decide whether two process models are similar or identical. Therefore, they can only be used in the context of processes have been executed.

There are few techniques measuring similarity between process models based on their structure. Regarding Petri Nets and state automata, similarity between process models can be measured based on behavioral inheritance rules [7]. Such techniques are similar to our distance measurement but can only provide a binary answer to whether two models are the same or whether one model is a sub-model of the other.

The approaches presented in [3] and [4] measure the distance between two process models by determining their difference with respect to dependencies among activities. However, it does not take the structural aspects of process models into consideration. A process model contains richer information than just nodes and edges (e.g., concerning split and join semantics). Similarly, work [5] provides an interesting approach to measure the similarity between two process models by using causal footprints, which describe a collection of the essential behavioral constraints imposed by a process

model. An approach was provided in [8] to evaluate the distance and similarity between two process models based on high-level change operations (e.g., to add, deletes or move activities). This work can measure similarity in terms of a unique number but it doesn't take the nesting control structures of process models into account.

Our approach focuses on the distance measure of block-structured processes. By transforming processes into structure tress, we can measures distance or similarity in terms of the control structure of processes. On the other hand, the differences of substructures between processes are considered in our approach.

## 6 Conclusion and Future Work

We provide a method to quantitatively measure the distance and similarity between two block-structured process models based on the alignment of process structure trees. Each process is transformed into a process structure tree. We further calculate the distance based on the alignment of two process structure trees. The proposed distance measure satisfies four distance measure properties, i.e., non-negativity, identity of indiscernible, symmetry and triangle inequality. So the proposed method enables the flexible deployment of process mining, discovery, or integration.

Some other research work is desired to enrich our knowledge on process distance. As a first step, we will extend our method so that the similarity between process models using additional constitutes (e.g., label of atomic activity) can be measured. The next step will incorporate data flow, temporal constraints, and resources, so that the distance measure can be further applied to realistic situations.

**Acknowledgements.** This work is partially supported by the National Natural Science Foundation of China (No. 71161015, 61063009), the Applied Fundamental Research Project of Yunnan Province (No. 2009CD040), and the Key Project of Science Foundation of Yunnan Provincial Department of Education (No. 2010Z009).

## References

1. Li, C., Reichert, M., Wombacher, A.: Discovering Reference Process Models by Mining Process Variants. In: 2008 IEEE International Conference on Web Services, pp. 45–53. IEEE Press, New York (2008)
2. Rosemann, M.: Potential Pitfalls of Process Modeling: part B. Business Process Management Journal 12(3), 377–384 (2006)
3. Jung, J.Y., Bae, J., Liu, L.: Hierarchical Business Process Clustering. In: 2008 IEEE International Conference on Services Computing, pp. 613–616. IEEE Press, New York (2008)
4. Bae, J., Liu, L., Caverlee, J., Zhang, L.J., Bae, H.: Development of Distance Measures for Process Mining, Discovery and Integration. International Journal of Web Services Research 4(4), 1–17 (2007)
5. van Dongen, B., Dijkman, R., Mendling, J.: Measuring Similarity Between Business Process Models. In: Bellahsène, Z., Léonard, M. (eds.) CAiSE 2008. LNCS, vol. 5074, pp. 450–464. Springer, Heidelberg (2008)

6. van der Aalst, W.M.P., Alves de Medeiros, A.K., Weijters, A.J.M.M.: Process Equivalence: Comparing Two Process Models Based on Observed Behavior. In: Dustdar, S., Fiadeiro, J.L., Sheth, A.P. (eds.) BPM 2006. LNCS, vol. 4102, pp. 129–144. Springer, Heidelberg (2006)
7. van der Aalst, W.M.P., Basten, T.: Inheritance of Workflows: an Approach to Tackling Problems Related to Change. Theoretical Computer Science 270(1-2), 125–203 (2002)
8. Li, C., Reichert, M., Wombacher, A.: On Measuring Process Model Similarity Based on High-Level Change Operations. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) ER 2008. LNCS, vol. 5231, pp. 248–264. Springer, Heidelberg (2008)
9. Kiepuszewski, B., ter Hofstede, A.H.M., Bussler, C.J.: On Structured Workflow Modelling. In: Wangler, B., Bergman, L.D. (eds.) CAiSE 2000. LNCS, vol. 1789, pp. 431–445. Springer, Heidelberg (2000)
10. Reijers, H., Mendling, J.: Modularity in Process Models: Review and Effects. In: Dumas, M., Reichert, M., Shan, M.-C. (eds.) BPM 2008. LNCS, vol. 5240, pp. 20–35. Springer, Heidelberg (2008)
11. Vanhatalo, J., Vzer, H., Koehler, J.: The Refined Process Structure Tree. Data & Knowledge Engineering 68(9), 793–818 (2009)
12. Van Der Aalst, W.M.P., Ter Hofstede, A.H.M., Kiepuszewski, B., Barros, A.P.: Workflow Patterns. Distributed and Parallel Databases 14(7), 5–51 (2003)
13. Muehlen, M., Recker, J.: How Much Language is Enough? Theoretical and Practical Use of the Business Process Modeling Notation. In: Bellahsène, Z., Léonard, M. (eds.) CAiSE 2008. LNCS, vol. 5074, pp. 465–479. Springer, Heidelberg (2008)
14. Mendling, J., van Dongen, B.F., van der Aalst, W.M.P.: Getting Rid of or-joins and Multiple Start Events in Business Process Models. Enterprise Information Systems 2(4), 403–419 (2008)
15. Mendling, J., Reijers, H.A., van der Aalst, W.M.P.: Seven Process Modeling Guidelines (7pmg). Information and Software Technology 52, 127–136 (2010)
16. Hwang, S.Y., Wang, H., Tang, J., Srivastava, J.: A Probabilistic Approach to Modeling and Estimating the QoS of Web-Services-Based Workflows. Information Sciences 177(23), 5484–5503 (2007)
17. Tai, K.C.: The Tree-to-Tree Correction Problem. Journal of the ACM 26, 422–433 (1979)
18. Reichert, M., Dadam, P.: ADEPTflex-Supporting Dynamic Changes of Workflows Without Losing Control. Journal of Intelligent Information Systems 10(2), 93–129 (1998)

# Adoption of Cloud Computing in Supply Chain Management Solutions: A SCOR-Aligned Assessment

Holger Schrödl

Magdeburg Research and Competence Cluster VLBA, Chair of Business Informatics,
Otto-von-Guericke University Magdeburg, Magdeburg, Germany
holger.schroedl@ovgu.de

**Abstract.** Efficient supply chains are a vital necessity for many companies. Supply chain management acts on operational processes, divergent and consolidated information flows and interaction processes with a variety of business partners. Efforts of recent years are usually facing this diversity by creating and organizing central information system solutions. Taking in account all the well-known problems of these central information systems, the question arises, whether cloud-based information systems represent a better alternative to establish an IT support for supply chain management.

Aim of this paper is to investigate this question. This is done by considering fundamental aspects of cloud-based solutions under the perspectives of the SCOR model. We present the SCOR model shortly and provide a current market perspective on cloud computing and supply chain from the position of logistics and from the perspective of the software industry. Along the five key processes of the SCOR model we evaluate the potential of cloud-based information system architectures, and we discuss successful implemented examples from practice, but also challenges in future implementation. The paper concludes with recommendations for design and implementation to cloud-based information system support for supply chain management.

**Keywords:** supply chain management, SCOR, cloud computing.

## 1 Motivation

The realization of an efficient supply chain is a challenging task for many companies. Supply chains are networks which organize manufactures, service providers and distribution sites that supply raw material, perform a transformation of raw material into both intermediate and finished products and distribute them to the customers [1]. Supply chain management (SCM) denotes all task related to manage the supply chain like planning and control, organizational structuring, product and information flow facility structure, management methods and risk and reward structure [2]. The importance of establishing and managing relationships among the participants of a supply chain is discussed widely [3–5]. With regards to a short term perspective, supply chain management is primarily to increase productivity,

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 233–244, 2012.

reduce inventory and inventory cycle times. On a long term perspective, supply chain management should lead to increasing market share, customer satisfaction and increasing profits for all participants in the supply chain [6]. It is essential to realize value-added processes that ensure optimum satisfaction of customer needs. In addition, this should be accompanied in accordance to an optimal design of customer service processes. Both are goals to conduct a thorough analysis of methods cause of supply chain management. This is called the design of logistics processes in which various actors are involved [7]. First, the design of supply chains means the specification of integrated business processes. Secondly, the question arises to the need to implement adequate information systems. Due to the aspect that in the inter-company supply chain management business process integration stays in the foreground, this leads to the question of appropriate information systems, often distributed in the area of information systems. First attempts have been made in terms of web-based service-oriented frameworks for establishing IT systems for supply chain management [8], which seems a promising way to face the challenges of the information flows in complex supply chain. In the wake of the emergence of cloud computing and SaaS (Software as a Service), the question can be made whether information systems for supply chains can be made more efficient through the use of these technologies.

To answer this question a structured approach in conducted to investigate the impact and opportunities of cloud computing to the establishment of supply chain management information systems.

The paper is structured as follows: in chapter 2, a brief introduction to the SCOR model is given as research framework. In chapter 3, a market-perspective on cloud computing is given to introduce the different aspects, factors and players of cloud computing. In chapter 4, an assessment is made on the different aspects of cloud computing structured by the perspectives of the SCOR model. In chapter 5, recommendations are given for enhancing information structures for supply chain management based on the results of the assessment. Chapter 6 closes with a summary and a research outlook.

## 2 Research Method

Aim of the paper is to provide a new insight into the intersection between supply chain management and cloud computing. Therefore, we have chosen to develop an exploratory conduction of the study. Exploratory studies are suitable for investigating contemporary phenomenon within its real-life context [9]. For structuring purposes, we have taken the most popular reference model in supply chain management – SCOR – as outline for the study. Data for the study came on the one hand from a comprehensive literature review and on the other hand from the study of market information and expert opinion. All the gathered information was classified into the five key processes of the SCOR model. From this classification recommendation were drawn.

## 3 Investigation Framework SCOR Model

There exist several business process frameworks in the literature to structure supply chain management processes. Hewitt gave a framework consisting of 14 business processes which are used by supply chain management executives in practice [10]. Cooper et al. identified eight supply chain processes: customer relationship management, customer service management, demand management, fulfillment, procurement, manufacturing flow management, product development and commercialisation and reverse logistics [2]. In order to achieve a holistic view on the opportunities of cloud computing a supply chain management processes we decide to take a more high-level framework. The framework chosen for such a holistic approach should is the SCOR model (Supply Chain Operations Reference-model) [11]. This model was designed by the Supply Chain Council as a reference model for describing business processes in the Supply Chain [12]. It draws on both corporate as well as enterprise-wide business processes described. SCOR has established itself as a model on the market, especially shown by the fact that more than 1000 companies worldwide have joined the Supply Chain Council. The active development of the model currently in Version 10.0 highlights the efforts to establish the SCOR model as a standard in a growing market. It is not only appropriate to look at complex supply chains, but it also offers the opportunity to improve basic requirements, which contributes significantly to the acceptance of the model.

The SCOR model includes five key supply chain operations Plan, Source, Make, Deliver and Return and is organized into four levels of observation (Figure 1):

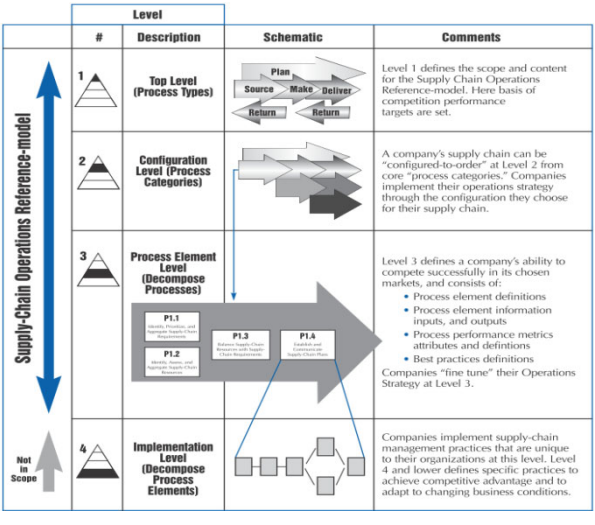



**Fig. 1.** SCOR reference model [12]

The level 1 is the process level and thus represents the highest level of the model defined here. The scope of this level is the company and the content of its supply chain. There are five processes are considered (see table 1):

**Table 1.** Processes and their content of the SCOR model

| Process | Content |
|---|---|
| Planning (plan) | the interplay of supply and demand |
| Sourcing (source) | procurement of products, components and services for service provision |
| Manufacturing (make) | the manufacture of products, intermediate products and services to different manufacturing |
| Delivery (deliver) | the supply of products and services to the customer with the appropriate accompanying |
| Return (return) | to receive a faulty product or return of primary products or raw materials to the supplier |

The second level of consideration is the configuration level. On this level of observation, the core processes are divided into process categories. A distinction is made in planning processes, implementation processes and support processes. Linking these two levels of observation produces a connection matrix. This matrix represents the set of all process combinations that should be used for the construction and development of a supply chain between the different business partners.

The detailing of these processes takes place in the level 3. On this level of consideration the specific process steps, the sequence and the input and output information are described. This level of consideration, referred to as a design level, completes the SCOR model. The viewing plane 4 is the implementation level, which is not included in the model. In this level, it is about company-specific considerations, not by general considerations concerning all types of companies.

## 4 Cloud Computing – A Market Perspective in the Context of Supply Chain Management

Cloud computing is an issue that is usually placed much attention from IT managers in companies meet. The term "Cloud Computing" goes back to a collaboration announcement between Google and IBM [13]. Before this time, several other technologies were discussed in the market which may be handles as predecessors of the term Cloud Computing like "Grid Computing" [14], "Computer in the Cloud" [15] or "Dreaming in the Cloud" [16]. An actual definition of "Cloud Computing" is given by Buyya: "A Cloud is a type of parallel and distributed system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified resource(s) based on service-level agreements established through negotiation between the service provider and customers." [17]. Breaking this rather complex definition down we can describe Cloud Computing as the delivering of infrastructure, platform and software in a service model based on a pay-per-use model provided to the customer [18]. In the market, this differentiation is denoted as Infrastructure as a Service (IaaS), Platform as a Service (Paas) and Software as a Service (SaaS).

Following a recent study by Sterling Commerce, a software provider from Duesseldorf, Germany, 87% of senior IT managers in Germany plan the move to cloud-based information systems in the B2B sector [19]. Main driver according to the survey is the reason of cost pressure: most companies promise a cost reduction through the use of cloud-based IT-structures due to a usage-based billing for services. Other aspects are also better use of their own IT staff, a reduction of manual processes and improve transparency of processes. Emphasis in the consideration of cloud-based systems lies in the areas of security and trust [20].

### 4.1 Cloud Computing in Logistics

In logistics, there exist highly branched communication structures between the different market participants. From this arises obviously the question whether cloud architectures are appropriate to optimize these communication structures. A survey by the Fraunhofer Innovation Cluster Cloud computing in logistics actually about 64% of all surveyed companies can imagine using logistics software that is not installed locally in the house, but on servers which are available on the Internet [21]. Therefore, willingness to engage such technologies constructively seems to be present. Larger companies (250 employees or 50 million € sales per year) have already been predominantly in the outsourcing experience of data and processes, whereas in small and medium enterprises this is rather less the case. In summary, it can be seen, therefore, that to depict the willingness to realize logistical processes and their corresponding data in a cloud architecture typically exist already in by large companies, but small and medium-sized enterprises rather not. This may have its origin, that large companies are able to build up effective information system structures and to operate them accordingly, small and medium-sized companies typically have neither the financial nor human resources to realize this. Exactly to these aspects, solutions based on a cloud architecture that offers both financial benefits as well by providing services, an independence of their IT knowledge achieved. It is expected that this will just be increased by the impact of financial crisis on the logistics market to expand solutions based on SaaS (Software as a Service) or Cloud architecture for small and medium-sized enterprises [22].

### 4.2 Cloud-Based Solution Providers for Logistics

There are already some cloud-based solutions that address the specific needs of supply chain management adapted. In this concise market view, three directions may be distinguished. One direction is the appropriate industry software provider to make their existing solution available in a cloud architecture on the market. They make use of the corresponding advantages of minimizing risk and often better affordability for the customer. On closer examination, it must be said but often that the manufacturers usually offer a SaaS model and use less of the specific benefits of a cloud architecture. One example is the British company OmPrompt (www.omprompt.com), one of the optimized logistics processes EDI solution in a SaaS architecture offering, thus enabling a complex integration of business partners, without requiring that a company invest in their own EDI infrastructure.

The second direction is the solution providers of platforms that are already being used to process the appropriate logistical operations. Here, one can see that taking place gradually increasing the supply portfolio. For example, the German company Transporeon (www.transporeon.com) provides a logistics platform that offers the elements of transport assignment, time slot management and tracking. In addition, a reporting and various additional functions are integrated. This platform has been extended by a special solution to the Book of charging slots. The U.S. company GT Nexus offers at international level to a platform that is complemented by a network of market participants and can be used not only for the processing of processes, but also to establish new business.

A third variant on the way to the cloud-based solution provider service that can be seen is getting the demand for such services from their own customers. These service providers are usually distinguished by their business content and use new technologies to market them accordingly into solutions. Example is Cargoclix (www.cargoclix.com), a provider of an electronic marketplace for transport and logistics. Besides the core business for the award of short-and long-term transport contracts, has now taken on the marketplace the opportunity to book time slots for loading and unloading of trucks. Due to the architecture of the solution offered in the form of a seamless market place, this was possible by using existing business content.

Especially the last example shows that cloud computing is not just a technical aspect. The technology allows for existing operators to expand its range, and the reorientation of existing business models. The result is first, a variety of providers for certain issues, which usually leads to customers to lower prices for the services. Second, to develop new pricing models, as the costs of cloud offerings have to be passed on to market participants. This leads to a shift of costs, the classification is still observed to date. Overall, it can be stated that the offer is growing cloud-based solutions for supply chain still on, but it offers because of the market situation in the logistics a high potential.

A study by IDC shows that the expected growth of cloud technology in logistics. Small and medium enterprises will significantly benefit from this technology, especially as regards the cash flow. IDC expects an increase of 19% by 2012 for the market of SaaS applications for Supply Chain Management [23].

### 4.3 Market Examples

There are already several examples existing on the market that describe the successful transition of enterprises towards a cloud-based architecture to their logistics systems. A current example is the decision of the Mexican brewer Grupo Modelo Group to adjust its export logistics to a cloud-based system [24]. Grupo Modelo sells 13 brands in 160 countries, including the highly successful beer Corona Extra and making it the leading Mexican beer manufacturer. The cloud-based logistics system allows for Grupo Modelo an optimized adaptation of products to demand a new level of transparency and control of all logistics processes. This example shows that (as already confirmed in the Fraunhofer-survey), large companies with a high innovation ratio are already heavily involved with the new opportunities and demonstrate first use of the potentials and improvements in practice.

# 5 A differentiated Assessment of the SCOR Model Involving Cloud Computing

The growing market acceptance and the first successful demonstration projects show that cloud architectures have a growing influence on the design of supply chains. Now companies are raising the question of how to approach this issue. For this purpose, the five main processes of the SCOR model in the context of cloud computing are considered separately.

## 5.1 Main Process: Plan

The process design of the process Plan includes the planning and management processes. In this case, a consultation between the existing resources and future needs is done and plans for the entire supply chain and procurement processes for exporting, manufacturing and delivery are created. It is a management of the business rules, evaluating the performance and various other aspects of the supply chain such as logistics management and risk management. Finally, an alignment of the plans of the supply chain has to be made to the financial plans of the company [12].

Overall, it is necessary to synchronize the data and information flow in a cloud-based application infrastructure with the flow of material and services in the real world. To achieve this, there has to be a mapping between elements of the internet data and information flow and the elements in the material flow. For this mapping, standards like the EPCglobal framework architecture may be used [25]. Another possibility is to use ID@URI, which describes a mapping between Internet addresses and unique product identifier to enrich the products which additional information retrieved from other sources in the Internet [26].

Subjects presented in example 2.3 of the Grupo Modelo can see that in this area a cloud-based architecture can achieve significant positive effects. Just mentioned for example the optimized adaptation of products to meet the demand of the customers was achieved through closer communication between business partners. Despite this successful example in the planning it is still lagging behind in terms of appropriate cloud-based solutions. It is observed that the sub-processes are modeled in company-specific implementation, which is more of a hindrance for the use of cloud-based solutions. Therefore, existing examples on the market are very much tailored to the needs of each project. It is expected that further efforts in the standardization process must take place in this area to establish an appropriate range of solutions.

## 5.2 Main Process: Source

The process Source includes the acquisition, receipt and inspection of incoming material. In addition, it includes the procurement processes and the identification and selection of appropriate suppliers. The management of business rules, supplier performance and the processing of specific aspects such as supplier contracts and risk management complete the purchasing process.

An interesting aspect in the context of cloud architectures is the identification and selection of suppliers [27]. In a growing global markets given the task to identify competent business partners, gets an ever-growing importance. Traditional search mechanisms such as an Internet search using appropriate search engines and a catalog-based search in relevant supplier catalogs come very quickly to the limit of their possibilities. Especially when it comes to building long-term business relationships play next to the question of whether the supplier can provide an appropriate material, other factors play a role. These other factors such as reliability and general business practices can often only be established beyond pure search engines. Often, here are recommendations or personal contacts relevant. Recent studies have shown that the proportion of partnerships that have a personal contact as a background take a significant proportion of all business collaborations [28].

One way to transfer such personal contacts is to provide a digital level like digital social networks. Examples such as XING, Facebook show how such digital networks are used to generate recommendations and to support new business development. The positive effect of digital social networks in the identification of possible business partnerships continue in subsequent phases of the selection and qualification. It is observed that these compounds are equipped with an increased trust in it [29], which influences the negotiations for the conclusion of cooperation and development in the course of a business relationship significantly positive. Thus, the integration of a component for digital social networking based on a cloud architecture makes an compelling contribution to optimize the supplier identification.

### 5.3 Main Process: Make

The process Make include processes such as production planning, production design, assembly, quality control and packing. Even in these areas there already exist software solutions, available on a cloud basis and thereby allowing a cloud-based production planning. On a closer examination of tenders, it can be seen, however, that all offers are substantially SaaS solutions that offer an Internet-based access to a production planning system based on an ERP system. A true cloud solution is not yet apparent.

The manufacturing processes have in common is that they can be highly individualized. Almost every manufacturing company has its own manufacturing processes that often make the value of the company through their individuality. Cloud solutions are living in standards: in the data formats, the information formats used in the processes. Standardized processes with the standard interfaces for integration. It is expected that the production area still needs some development in order to exploit the potential of cloud architectures. A promising way seems to be the establishment of cloud-based process management infrastructures which are suitable for a flexible composition of functional components [30]. These platforms are able to provide basic information and process infrastructure which may be delivered from a cloud provider. Based on this platform it would be suitable to define components for the functionality for supply chain management processes which will be provides from different suppliers.

### 5.4 Main Process: Deliver

The delivery processes include order processing and warehouse and transportation management. In these areas, currently the widest field of existing services can be recognized. In the area of order processing, there are corresponding CRM systems available on cloud-based systems or in a SaaS architecture. Order processing has the advantage to be more homogeneous in the processes in many areas, so that a variety of vendors can adopt them. A distributed control structure of employees with tasks in the order management favors the development of cloud-based systems. But also in transport management there is already a large number of solutions that can be included as part of a cloud-based overall architecture.

An outstanding example is the logistic tracking, which is now offered by virtually every logistics company for its business customers, but also for private clients. Here, there are small components that can be used to offer its customers the appropriate information service. An example is the UPS provided widget that evidence (widget.ups.com). This widget is a software component for the desktop of a computer. In this component, UPS displays relevant data to the particular user information about their deliveries available. It is expected that such components make a large contribution in the future motivated through the standardization of processes and information to realize significant potential for cloud-based solutions in the delivery processes.

### 5.5 Main Process Return

The return processes include the return and the withdrawal of all in the course of the supply chain resulting undesirable or not or no longer needed goods. The areas of the withdrawal process are currently largely not directly related to cloud-based solutions supported. Support refers indirectly to the offer to the CRM solutions and an offer to the logistics solutions instead. These sub-processes are integrated to optimize the return about a product or the award of an RMA number. Dedicated solutions for the return area are currently not identified yet on the market. Nevertheless, it is expected to be created solely on the basis of existing legislative provisions here relevant offers, which can then be re-integrated under a cloud architecture into an existing offer.

## 6 Approaches and Recommendations

In summary, it can be stated that cloud computing will act as an essential enabler for future supply chains. To take advantage of the benefits as a company, it is essential to keep for a rapprochement with the main processes which already have a corresponding range of solutions. Moreover, it is beneficial to have a focus on the following aspects:

- ***Cooperation***
  Cloud-based information system architectures are a key technology for an in-house, but just a single enterprise-wide collaboration. Intra-company cooperation in this case relates to issues such as seamless data availability and uniform

interfaces to different business areas as well as integration of mobile business processes. Company-wide collaboration focused on issues such as the sharing of business-related information among the participating partners as well as efforts on the standardization of business processes.

- ***Central Information and Process Models***
  In a cloud architecture, it is necessary that all partners use common data and process models to ensure interoperability. For example, a transfer order has to be described uniformly, regardless of the logistics company that created it. The challenge here is that a consensus is found on a common information structure and process among the partners. In these areas there are several standardization efforts (ecl@ss [31], BMECat [32], etc.) to establish a firm and generally accepted information modeling. It is assumed that precisely the key processes and information models enable the real growth lever for cloud-based solutions in the supply chain. The central idea of standardized items such as purchase orders, delivery notes or product numbers, held in one central place, accessible to all participants with relevant information flows transferred, changes the idea of supply chain management to a new level. Furthermore, seamless integration of mobile business processes will extend the influence of a well-organized decision base.
- ***Community Idea***
  Whether cloud technology is a nice-to-have technology or provides a real economic advantage for all concerned, lies in the possibility of forming a community. Community here means a set of electronic communications related merger of various market participants with the common goal of business to succeed. It has already started some providers of marketplaces in supply chain to establish functionality on their platform, enabling a closer personal collaboration. This enables market participants (and in this sense, the people who participate in this market) exchange messages and data, and set up user profiles that reflect their role and interests in the market. These solutions leverage the already connected networks of suppliers, customers and other partners with the aim of routine processes to improve the cooperation.

## 7 Summary and Outlook

Aim of this paper was the investigation of influence and opportunities of cloud computing to the establishment of supply chain management information systems. To answer this question, a structured assessment has been conducted by using the SCOR model as assessment perspectives. The SCOR reference model was taken to identify the potential of cloud computing in every single process step of the model. By means of a literature study, in addition to the discussion of existing examples in the market, it could be shown that in every main process of the process model, cloud computing may act as enabler for future requirements. It could be shown that in the Plan, Source, Deliver and Return processes the potential of cloud computing technologies is significant, whereas in the process Make we see that there is a lack of support. Reason for this might be the individuality of the manufacturing processes of every company.

From this structured assessment, three basic recommendation could be given to guide companies in the enhancement of supply chain management information systems through cloud computing.

Future research will take care of a more in-depth view of the subprocesses of the SCOR model to identify single process steps which may be enhanced significantly through cloud computing technologies. In addition to this, a broad market study will be conducted to evaluate existing offerings of cloud-based solution to the requirements of current and future supply chain management solutions.

## References


1. Lee, H.L., Billington, C.: Managing supply chain inventory: Pifalls and opportunities. MIT Sloan Management Review, 65–73 (1992)
2. Cooper, M.C., Lambert, D.M., Pagh, J.D.: Supply Chain Management: More Than a New Name for Logistics. The International Journal of Logistics Management 8, 1–14 (1997)
3. Stevens, G.C.: Integrating the Supply Chain. International Journal of Physical Distribution & Logistics Management 19, 3–8 (1989)
4. Ellram, L.M., Cooper, M.C.: Supply Chain Management, Partnership, and the Shipper - Third Party Relationship. The Int. J. of Logistics Management 1, 1–10 (1990)
5. Lambert, D.M., Emmelhainz, M.A., Gardner, J.T.: Developing and Implementing Supply Chain Partnerships. The International Journal of Logistics Management 7, 1–17 (1996)
6. Tan, K.C., Kannan, V.R., Handfield, R.B.: Supply Chain Management: Supplier Performance and Firm Performance. Int. J. of Purchasingand Materials Mgt. 34, 2–9 (1998)
7. Becker, J., Schütte, R.: Handelsinformationssysteme: Domänenorientierte Einführung in die Wirtschaftsinformatik Verlag Moderne Industrie, Landsberg (2004)
8. Zhang, M., Xu, J., Cheng, Z., Li, Y., Zang, B.: A Web Service-Based Framework for Supply Chain Management. In: Eighth IEEE ISORC 2005, pp. 316–319. IEEE (2005)
9. Yin, R.: Case Study Research: Design and Methods, 3rd edn. Applied Social Research Methods Series, vol. 5. Sage Publications, Inc. (2002)
10. Hewitt, F.: Supply Chain Redesign. The International Journal of Logistics Management 5, 1–10 (1994)
11. Poluha, R.G.: Anwendung des SCOR-Modells zur Analyse der Supply Chain. Diss. Univ. Köln (2005)
12. Supply-Chain Council: Supply-Chain Operations Reference-model: SCOR Overview Version 9.0, http://www.supply-chain.org/f/SCOR%2090%20Overview%20Booklet.pdf
13. New York Times: Google and I.B.M. Join in 'Cloud Computing' Research, http://www.nytimes.com/2007/10/08/technology/08cloud.html
14. Foster, I., Kesselman, C.: The grid. Blueprint for a new computing infrastructure. Morgan Kaufmann, Amsterdam (2004)
15. Naone, E.: Computer in the Cloud - Technology Review. Online desktop systems could bridge the digital divide, http://www.technologyreview.com/infotech/19397/?a=f
16. Reimer, J.: Dreaming in the "Cloud" with the XIOS web operating system, http://arstechnica.com/business/news/2007/04/dreaming-in-the-cloud-with-the-xios-web-operating-system.ars

17. Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I.: Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 25, 599–616 (2009)
18. Buyya, R., Pandey, S., Vecchiola, C.: Cloudbus Toolkit for Market-Oriented Cloud Computing, 24–44 (2009)
19. Sterling Commerce, Inc.: 87 Prozent deutscher Unternehmen planen Investitionen in Cloud-Services (2010)
20. Weinhardt, C., Anandasivam, A., Blau, B., Borissov, N., Meinl, T., Michalk, W., Stößer, J.: Cloud-Computing - Eine Abgrenzung, Geschäftsmodelle und Forschungsgebiete. Wirtschaftsinformatik 51, 453–462 (2009)
21. Fraunhofer-Innovationscluster Cloud Computing für die Logistik: Cloud Computing in der Logistics Mall - Akzeptanz der Logistics Mall bei den Stakeholdern (2009), http://www.ccl.fraunhofer.de/Schnelleinstieg/umfrage/index.jsp
22. Heistermann, F., Schmitt, H.: Finanzkrise wirdSaaS Einsatz in der Logistik beschleunigen (2009)
23. Solomon, S.E.: Cloud brings supply chain apps to SMBs (2009), http://www.zdnetasia.com/cloud-brings-supply-chain-apps-to-smbs-62053162.htm
24. GT Nexus: GrupoModelo Selects GT Nexus to Provide Global Logistics Operating Platform, http://www.gtnexus.com/press-room/press-release/162/
25. Hribernik, K.A., Warden, T., Thoben, K.-D., Herzog, O.: An Internet of Things for Transport Logistics. In: An Approach rocennecting the Information and Material Flows in Autonomous Cooperating Logistics Processes. MITIP (2010)
26. Främling, K., Ala-Risku, T., Kärkkäinen, M., Holmström, J.: Agent-based model for managing composite product information. Comput. Ind. 57, 72–81 (2006)
27. Schrödl, H.: Value enhancement of strategic supply networks for value bundles through digital social networks. In: Proc. of the 10th ICEB, Shanghai, China, pp. 35–41 (2010)
28. Thomé, U., Kortzfleisch, H.F.O., von Szyperski, N.: Kooperations-Engineering - Prinzipien, Methoden und Werkzeuge. Online-Kooperationen, 41–58 (2003)
29. Teten, D., Allen, S.: The virtual handshake. Opening doors and closing deals online AMACOM, New York (2005)
30. Anstett, T., Leymann, F., Mietzner, R., Strauch, S.: Towards BPEL in the Cloud: Exploiting Different Delivery Models for the Execution of Business Processes. In: Zhang, L.-J. (ed.) 2009 IEEE Congress on Services Proceedings, Los Angeles, CA, July 6-10. IEEE Computer Society, Los Alamitos (2009)
31. eCl@sse.V.: eCl@ss Classification and Product Description, http://www.eclass.de
32. Bundesverband Materialwirtschaft, E.u.L.e.V.: BMEC at, http://www.bmecat.org

# Exploiting Space-Time Status for Service Recommendation

Changjian Fang, Bo Mao, Jie Cao, and Zhiang Wu*

Jiangsu Provincial Key Laboratory of E-Business,
Nanjing University of Finance and Economics, Nanjing, P.R. China
zawuster@gmail.com

**Abstract.** The prevalence of smart devices allows people to record their space-time status. This paper focuses on exploiting user space-time status and the related semantic information for service recommendation. Firstly, event DAG is employed to organize the space-time information generated based on the service invocation history. Generation algorithm of the event DAG is then proposed. Secondly, a novel collaborative filtering based recommendation algorithm is designed. Potentially interesting services in the target node and its subsequent nodes can be recommended. In our implementation, the user space-time status is generated from the 4D city models (3D location + time) with semantic information. A prototype system is implemented to generate service invocation logs of different users. These simulative logs are utilized to evaluate the effectiveness and efficiency of our proposed method.

**Keywords:** 4D city model, space-time status, service recommendation.

## 1 Introduction

Along with the development of wireless networking, portable mobile devices and mobile broadband Internet access technologies [1,2], it becomes convenient to acquire space-time status of the mobile user. It is possible for the user to enjoy recommendation of services at any time in any place. The requirements of the people are related with their location and time. Meanwhile, the prevalence of Web 2.0 technologies leads to a great increscent of real time generated data [3]. It is getting more difficult to make a recommendation for a user because of not only huge amount data, but also lack of semantic information of the user.

This paper focuses on exploiting space-time status and related semantic information of users for service recommendation. The location of the user can be obtained by the GPS or mobile positioning method. The semantic information of location can be gained from the semantic city models. The OGC 3D city model standard [4], CityGML, defines the semantic representation of urban objects e.g. the usage of a building. Based on the position and semantic 3D city models, we can identify the semantic location of the user e.g. a shopping mall, a traffic station or a hospital. By combining

* Corresponding author.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 245–256, 2012.

3D location and time (or 4D city model), more semantic information will be revealed such as local holiday, weather condition, big events, and so on. In this paper, the concept of service is not limited to Web Service, but the activity from which people can reap the benefits e.g. alarm clock, news, traffic, discounting, and so on.

It is possible and necessary to recommend the services for the users according to their positions and the current time. In fact, the requirements are not limited to the 3D location and time but their semantic information behind. In this paper, we try to find out the relationship between user requirements and their space-time status information and make use of that for the service recommendation.

The rest of the paper is structured as follows. Related work is given in section 2. Section 3 describes the definition and generalization of user space-time status. Section 4 gives the service recommendation mechanism based on the proposed space-time status. A stimulated case study is shown in section 5. Section 6 concludes the whole paper.

## 2 Related Work

### 2.1 Recommender Systems

In recent years, recommender systems have emerged as one tool helping people look for items that they are interested in. The items include commodity products, movies, advertisements, CD, Web services, and so forth. The recommendation algorithm, the kernel of recommender systems, has been a hot research topic for a long time. The initial algorithm used by recommender systems is the content-based algorithm which suggests items that are similar to the ones the current users has shown a preference for in the past [5]. Content-based algorithm relies on rich content descriptions of the items (services for example) that are being recommended. For instance, if a recommender system wants to suggest a coat, the content-based algorithm will depend on information such as brand, style, color, price, etc. Therefore, the designer of the content-based algorithm should obtain abundant domain knowledge which may not be readily available or straightforward to maintain.

Collaborative filtering (CF) is different from the content-based methods. It collects opinions from users in the form of ratings on items. Its advantage over the content-based algorithms is that the CF algorithm does not need a representation of the items but only rely on historical information of users. CF algorithms can be divided into two categories: neighborhood-based approach and latent factor models [6]. User-based CF (UCF), a representative neighborhood-based CF algorithm, is adopted by a multitude of well-known recommender systems such as Tapestry, MovieLens, Amazon Book Store, YouTube, Facebook, and so forth. UCF utilizes the opinions from a user's $k$ nearest neighbors ($k$NN) to help the user to identify his/her interested content from an overwhelming set of potential choices. In this paper, we propose a new CF algorithm in order to incorporate semantic 4D (space-time status) information for improving the recommendation.

Adomavicius et, al. use a multi-dimensional approach to incorporate contextual information [7]. Zheng et, al. report on a personalized friend and location recommender

for the geographical information systems (GIS) on the Web [8,9]. GPS trajectories in [8,9] are collected by 75 volunteers over a period of one year. Limited by time and condition, the experimental data utilized in this paper is generated by a simulator. However, smart devices will be equipped on volunteers in the future in order to generate genuine dataset including space-time status of users and service invocation records.

### 2.2 Context Generation

Context aware computation is an interesting subject for many researchers [10-12]. Context is the information/data for characterizing the situation of an entity e.g. a person, place, or object that is considered relevant to the interaction between a user and an application, including location, time, activities, and the preferences of each entity [13,14]. The existing context information employed in the service recommendation is mainly about the user [15]. Along with the development of mobile techniques, the location based context information can be gathered for recommendation [16]. Quercia et al. [17] use location data from mobile phone to recommend the social events.

By reference to the 3D city models, not only location but also the semantic information of the location can be retrieved. Kolbe [18] suggests the semantic information should be contained in the 3D city models, and OGC issued the 3D city model representation standard CityGML [4]. Many cities such as Berlin, Stuttgart have delivered their official 3D city models online in CityGML. We can expect the semantic 3D city model will be available as the city information infrastructure. Through the semantic 3D city models, we can further get the semantic information of a position.

Based on the semantic information of the 3D city models and the time of user service invocation, we can get the space-time status about the service which is essential in the user service recommendation, because most of people have the experience that the need certain service when they are in certain place at certain time. Therefore, the space-time status of user will be generated and stored for the service recommendation in this paper.

## 3 Space-Time Status

### 3.1 Definition

We use directed acyclic graph (DAG) to organize the space-time information generated during the service invocation. The DAG is defined as follows:

**Definition 1** (Event DAG, G): A DAG $G=<N, E>$ consists of a set of nodes $N$ representing the events and a set of directed edges $E$ representing dependencies among events. Assume there are $n$ nodes and $m$ edges, namely $N=\{node_i|0<i<n\}$and $E=\{edge_j|0<j<m\}$. A event node $node_i=<Time, Location, Services>$ consists of time, location and invoked services. The edge set $E$ contains edges $<node_a, node_b, ber>\in E$ for each event $node_a$ (parent) that $node_b$ (child) depends on.

In fact, a node represents an event: the services invoked in a space-time coordinate (time, location). The edge is the number of the conversion from $node_a$ to $node_b$.

The time is an interval containing start and end. In this paper, the time is looped by the week that most people apply to. In the city area, the surrounding city objects affect the selection of services. For example, we trend to check in to the nearby hotel when the time is late; or find out the closest public traffic facilities. The location is defined as follows:

**Definition 2** (Location): The *Location={City_object*$_i$*|0<i<nc}*is composed by the surrounding city objects. A city object is denoted by City_object$_i$ = {Type, Class, Function, Usage} where Type = {Building | Transportation | WaterBody | Vegetation | City furniture} and *Class*, *Function, Usage* are specified in the CityGML standard.

The definitions of the *Class*, *Function* and *Usage* have been specified in the CityGML standard. Generally, each city object has the attributes *class*, function and *usage*, unless it is stated otherwise. The class attribute can occur only once, while the attributes usage and function can be used multiple times. The class attribute describes the classification of the objects, e.g. road, track, railway, or square. The attribute function contains the purpose of the object, like national highway or county road, while the attribute usage may define if an object is e.g. navigable or usable for pedestrians (OGC, 2008). For example *BuildingType* can be habitation, sanitation, sport, education, traffic; *BuildingFunction* can be office building, court, post office.

The services of a node are the user's invoked services during the time interval at a certain location. The semantic 4D (3-dimensional location + time) and the history of the services become the basis of the recommendation.

### 3.2 Generation

To generate the DAG of user status, we need the service invocation log and the city object database. Initially, the DAG only contains two nodes: the start node and the end node as shown in Fig. 1(a). We then read the service invocation log one by one and construct or modify the DAG accordingly as in Fig. 1(b).

In our implementation, the user log is formatted as: *L=<ID, TS, P, S>* in which *ID* is the user identity, *TS* is the time stamp, *P* is the coordinate of the user represented by (*x*, *y*, *z*) and *S* is the invoked services. The user log can be automated gathered from the "smart devices" with GPS function. In the specific recommend system, the *S* should be filtered to preserve the interested target services and remove these uninterested ones.

Another resource required for the user DAG creation is the city object database. It should provide the function of objective inquiring from coordinate. The stored city objective not only contains the general semantic information such as name, function, type etc. but also the personal semantic information. For example, one building may be the home for a person and also the work place for another people. Also the time can be decided the personal semantic, such as a same road in different time may be the way home or way work in different time. Therefore, the inquiring is processed in two steps: the first is getting city object (building, road etc) from the coordinate; the second is getting personal semantic information. The semantic information will be recorded in the nodes of user space-time status which are the many bases to implement the recommendation.

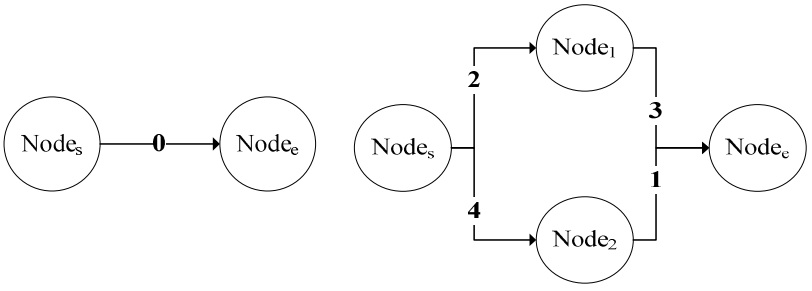



**Fig. 1.** An example of the space-time DAG

From the user log, DAG of user space-time status is generated as following three steps:

**Step 1: Initialization** Initially, the DAG of user space-time status just contains two controlled nodes start and end without real semantic information. We define the current node (CN) indicating the present status of the user. The current node is point to the start node in the beginning.

**Step 2: Updating** In the updating step, fetch one record in the user log as current record. Assume the current record $L_i$=<$ID_i$, $TS_i$, $P_i$, $S_i$>, and the current status is $node_j$=<$Time_j$, $Location_j$, $Service_j$>. Then the updating process can be implemented according to the pseudo-code in table 1. Since the time is cycled in the week basis, the current status will automatically go to the start point when the week is over.

**Step 3: Simplification** Usually the generated user space-time status DAG may contain many nodes and edges when the user log data is large. Therefore, it is necessary to simplify the DAG for faster matching. In the proposed framework, we detect the number of service invoking recorded in edge. If the number is smaller than a set value and one of the nodes in the edge is one degree node, the edge and its connected degree one node will be removed.

**Table 1.** Pseudo-code of the user space-time status DAG generation

| |
|---|
| Input: ***Lset***: user service invocation log set |
| Output: G=<*N*, *E*>: a DAG to describe the user space-time status |
| **1: Initiate the G=<*N*,*E*> in which *N*={*Node$_s$*, *Node$_e$*} *E*={<*Node$_s$*, *Node$_e$*, 0>};** |
| **2: Current Node *CN*=*Node$_s$*;** |
| **3: for** all $S_i \in$ Target Service Set |
| 4: *Location$_i$* = getLocation(*ID$_i$*, *TS$_i$*, *P$_i$*); |
| **5: if *location$_i$* is the same as the location of current node** |
| **6: extend time of *CN* to *TS$_i$*; add *S$_i$* into the services of *CN*** |

**7:** **continue;**
**8:** **else**
9: **for** every $node_l$ in $edge_l$ <*CN*, $node_l$, *n*>
10: **if** $location_i$ is the same with location in $node_l$
11: extend time of $node_l$ to $TS_i$; add $S_i$ into the services of $node_l$;
12: $Edge_l.n$+=1; set *CN* to $node_l$;
13: **continue**;
14: **end if**
15: **end for**
16: **end if**
17: $node_k$=<$TS_i$, $Location_i$, $S_i$>; $edge_k$=<$node_j$,$node_k$,*1*>;
18: add $node_k$ into *N*; add $edge_k$ into *E*;
19: **end for**
20: **remove the edge with its *n* value smaller than 5 and the degree of one of its node is 1.**

# 4 Recommendation Based on Space-Time Status

Service recommendation in this paper aims to provide potentially interesting services when the user is in a *target node*. Furthermore, potentially interesting services when the user is the next possible nodes will be also recommended. For instance, assuming the target node of a child is <(7:30, 8:00), home> of which the next possible nodes include <(8:30, 11:30), school>, <(9:00, 12:00), carnie>, or <(8:20, 9:20), bookstore>, all potentially interesting services in these four nodes will be recommended by the proposed approach. First, *k*NN are selected based on the distance computation. Second, the list of recommended services in the target node is generated. Third, the lists of recommended services in subsequent nodes are generated.

## 4.1 Distance Computation among Nodes and *k*NN Selection

The distance between two nodes in *G* is measured by the semantic 4D information including time, coordinate, semantic information of city object. The variable types of these data are heterogeneous. For instance, time and coordinate are interval-scale variables, and type, class and function are categorical variables. We propose to utilize Eq. (1) to compute the distance between $node_a$ and $node_b$.

$$distance(node_a, node_b) = \frac{\sum_{f=1}^{p} \delta_{ab}^{f} d_{ab}^{f}}{\sum_{f=1}^{p} \delta_{ab}^{f}} \quad (1)$$

Assume there are *p* variables of the semantic 4D information. In Eq. (1), if $node_a$ or $node_b$ do not assign a value to the *f*-th variable, $\delta^f_{ab}=0$; otherwise $\delta^f_{ab}=1$. The distance

of the *f*-th variable between $node_a$ and $node_b$ is written $d^f_{ab}$ which depends on the variable type. We divide the computation of $\delta^f_{ab}$ into two cases according to variable types.

- Case (1). the *f*-th variable is binary variable or categorical variable: if $x_{af} = x_{bf}$, $d^f ab = 0$; otherwise $d^f ab = 1$.
- Case (2). the *f*-th variable is interval-scale variable: time and coordinate belong to this type. *Manhattan* distance is employed to measure the distance of this type. Let $<x_i, y_i>$ and $<x_j, y_j>$ denote coordinates or time interval of $node_a$ and $node_b$ respectively, and $d^f ab$ can be computed as Eq. (2) shown.

$$d^f_{ab} = |x_i - x_j| + |y_i - y_j| \tag{2}$$

Selecting *k*NN of nodes is an important step for making accurate recommendation. Let $node_t$ denote the target node and *k*NN($node_t$) denote *k* nearest neighbors of $node_t$. Based on the distance calculated by Eq. (1), we select *k* nodes whose distance with the target node are as small as possible.

### 4.2 Recommendation in the Target Node

This phase predicts the potentially interesting services in the target node on the basis of its top-*k* neighbors. Since *Services* in $node_i$ is a binary vector where $S_i = 1$ denotes the user has invoked $S_i$, we predict the interesting degree according to the number of votes of *k*NN($node_t$). Let $S_{t,j}$ denote predicted interesting degree of $S_j$ in $node_t$, and $S_{t,j}$ is given by Eq. (3).

$$S_{t,j} = \sum_{node_k \in k\mathrm{NN}(node_t)} \frac{S_{k,j}}{distance(node_t, node_k)} \tag{3}$$

where $S_{k,j}$ is the binary value of $S_j$ in $node_k$, and $distance(node_t, node_k)$ represents the distance between $node_i$ and $node_k$ which is calculated as mentioned in Eq. (1). Then, top-*N* services having the most votes are recommended as recommended services in the target node. Let $RST_N(node_t)$ denote the list of recommended services in the target node, and $RST_N(node_t)$ is given by Definition 3.

**Definition 3** (Top-*N* recommended services, $RST_N(node_t)$). $RST_N(node_t)$ is a set of services with the size *N* and satisfies that $\forall S_j \in RST_N(node_t),\ \forall S_m \in \overline{RST_N(node_t)},\ S_{t,j} \geq S_{t,m}$, where $\overline{RST_N(node_t)}$ is the complementary set of $RST_N(node_t)$.

Once we have calculated predicted interesting degrees for all services in $node_t$, the list of recommended services with size *N* can be readily obtained.

### 4.3 Recommendation in Subsequent Nodes

The target node has several subsequent nodes. The edge between two nodes represents the number of the conversion from which the probability of this conversion can be calculated. Let $node_t^m$ be the *m*-step node from the $node_i$, and $< node_t^0, node_t^1, \ldots, node_t^m >$ be the path from $node_t$ ($node_t = node_t^0$) to $node_i^t$. Let

$P(node_t^q, node_t^{q+1})$ denote the probability between $node_t^q$ and $node_t^{q+1}$, and $S_{t,j}^t$ denote predicted interesting degree of $S_j$ in the *t*-step node from the $node_i$. $S_{i,j}^t$ can be computed by Eq. (4).

$$S_{t,j}^m = S_{m,j} \cdot \prod_{q=0}^{m} P\left(node_t^q, node_t^{q+1}\right) \tag{4}$$

where $S_{m,j}$ represents predicted interesting degree of $S_j$ in $node_t^m$, and $S_{m,j}$ can be calculated by Eq. (3). As the increase of $m$, $S_{t,j}^m$ verges on a minimum value and the strength of $S_{t,j}^m$ for service recommendation reduces sharply. It is in accordance with the fact that the longer distance between two nodes is, the lower accuracy of prediction is. From the Definition 3 and $S_{t,j}^m$, top-*N* services in $node_t^m$ can be obtained.

## 5 Case Study

Since there is no open dataset including GPS trajectories and service invocation histories, we design and implement a prototype system in order to generate service invocation logs of different types of users. We then generate event DAG based on these logs and recommendation in nodes of DAG. In this section, we first present the setup of our prototype system, including user types, events, services and places. Then, we report the experiments of event DAG generation and recommendation based on space-time status.

### 5.1 Setup

The prototype system simulates behaviors of different types of users on workday or weekend. The user behavior in one day consists of a series of random events in which service invocation histories are kept. Table 2 depicts user types, events, services and places of the prototype system.

**Table 2.** Attributes and possible values of our prototype system

| Attribute | Possible Values |
|---|---|
| user types | young man \| young woman \| old man \| old woman \| child |
| events | get up \| eat breakfast \| trip \| working \| eat lunch \| shopping \| entertainment |
| services | alarm_clock \| news \| traffic \| Internet \| print\| repast \|group_purchase |
| places | home \| office \| school \| road \| bazaar \| restaurant \| bookstore \| carnie \| resort |

Service invocation logs of users are generated in two scenarios that are workday and weekend respectively. For example, a young woman might prefer to read world news in the workday morning, and on weekends to read movie reviews and do shopping.

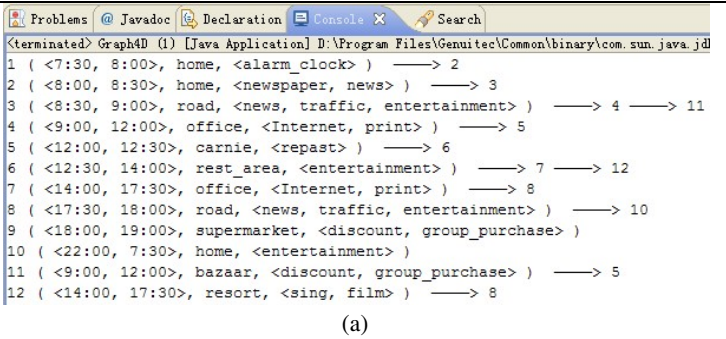
## 5.2 Experimental Results

### 5.2.1 Event DAG Generation.

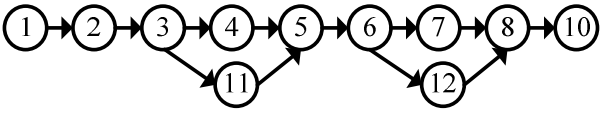
Event DAG generation algorithm shown in Table 1 is implemented in our prototype system. In this section, we utilize an example to reveal the process of the event DAG generation. Table 3 gives a representative log of a young woman on Wednesday and Saturday. *Adjacency list* is employed to store the event DAG in our implementation. Fig. 2(a) gives the output of the adjacency list obtained from the log in Table 3, and Fig. 2(b) shows the corresponding DAG of Fig. 2(a).

**Table 3.** An example of user log

| EventID | UID | UType | Day | Start-time | End-time | Location | Invoked Services |
|---|---|---|---|---|---|---|---|
| 1 | 6 | YWoman | Wednesday | 7:30 | 8:00 | home | alarm_clock |
| 2 | 6 | YWoman | Wednesday | 8:00 | 8:30 | home | newspaper, news |
| 3 | 6 | YWoman | Wednesday | 8:30 | 9:00 | road | news, traffic, entertainment |
| 4 | 6 | YWoman | Wednesday | 9:00 | 12:00 | office | Internet, print |
| 5 | 6 | YWoman | Wednesday | 12:00 | 12: 30 | carnie | repast |
| 6 | 6 | YWoman | Wednesday | 12:30 | 14: 00 | rest_area | entertainment |
| 7 | 6 | YWoman | Wednesday | 14:00 | 17:30 | office | Internet, print |
| 8 | 6 | YWoman | Wednesday | 17:30 | 18:00 | road | news, traffic, entertainment |
| 10 | 6 | YWoman | Wednesday | 22:00 | 7:30 | home | entertainment |
| 1 | 6 | YWoman | Saturday | 7:30 | 8:00 | home | alarm_clock |
| 2 | 6 | YWoman | Saturday | 8:00 | 8:30 | home | newspaper, news |
| 3 | 6 | YWoman | Saturday | 8:30 | 9:00 | road | news, traffic, entertainment |
| 11 | 6 | YWoman | Saturday | 9:00 | 12:00 | bazaar | discount, group_purchase |
| 5 | 6 | YWoman | Saturday | 12:00 | 12: 30 | carnie | repast |
| 6 | 6 | YWoman | Saturday | 12:30 | 14: 00 | rest_area | entertainment |
| 12 | 6 | YWoman | Saturday | 14:00 | 17:30 | resort | sing, film |
| 8 | 6 | YWoman | Saturday | 17:30 | 18:00 | road | news, traffic, entertainment |
| 10 | 6 | YWoman | Saturday | 22:00 | 22:30 | home | entertainment |




(a)




(b)

**Fig. 2.** The adjacency list and its corresponding DAG

### 5.2.2 Performance of the Recommendation

Service invocation logs of different types users in one week are generated firstly. An event DAG is constructed for one user in a week. We then select some target nodes and some of their subsequent nodes, and suppose the invoked services in these selected nodes are unknown. The list of recommended services are generated by the method mentioned in Section 4. Performance of the recommendation method is evaluated by comparing the list with the logs. The metric of *Precision* is defined as follows:

$$Precision = \frac{N_{rs}}{N_s} \tag{5}$$

In Eqs. (5), $N_{rs}$ is the number of services in both recommended list and the list of invoked services recorded in the log, and $N_s$ is the length of the recommended list.

The first experiment investigates how $m$ affect the accuracy. We range $m$ from 0 to 6 and set $k=8$ and $N=5$. We select 10 target nodes in the event DAG of five types of users on workday and weekend respectively. Fig. 3 shows how the precision of recommendation vary with the increase of $m$. As $m$ increases, the precision decreases firstly and then varies irregularly. The smaller distance from the target node is, the higher precision is. However, when $m$ reaches a bigger value, the stringency of recommendation reduces sharply, resulting in the irregularity of the precision variance.

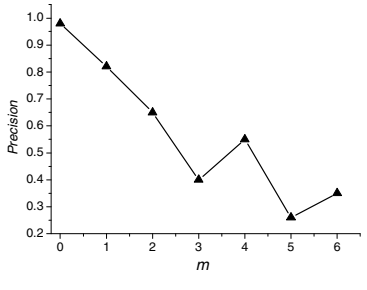



**Fig. 3.** Impact of $m$ on precision of recommendation

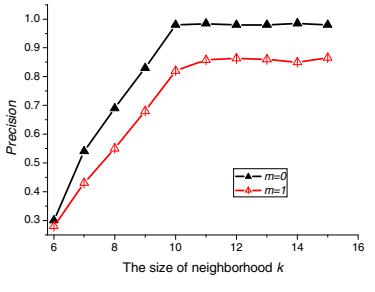



**Fig. 4.** Impact of $k$ on precision of recommendation

The second experiment investigates how $k$ affect the accuracy. We select 10 target nodes in the same way. And we range $k$ from 6 to 15 and set $N=5$. Recommendation results on the target nodes and their single-step nodes are recorded. The experimental

results shown in Fig. 4 indicate that as the increase of $k$, precision in two scenarios increases firstly and then tends towards stability. Therefore, it is not necessary to select nearest neighborhoods as much as possible.

## 6 Conclusion

This paper focuses on exploiting space-time status and its related semantic information of users for service recommendation. We propose to utilize event DAG to organize the space-time information generated during the service invocation. The generation algorithm of the event DAG is also proposed. Based on the event DAG, a novel collaborative filtering based recommendation algorithm is presented. Potentially interesting services in the target node and its subsequent nodes can be recommended. A prototype system is designed and implemented to generate service invocation logs of different users. We then utilize an example to elaborate the process of the event DAG generation. Finally, how parameters $m$ and $k$ affect the precision of recommendation is demonstrated by experimental results.

In the future, smart devices will be purchased and volunteers will be recruited so that abundant realistic service invocation logs with 4D city model can be obtained. We will propose more skillful recommendation algorithms and conduct more detailed experiments.

**Acknowledgement.** This research is supported by National Natural Science Foundation of China under Grants Nos.71072172 and 61103229, Industry Projects in the Jiangsu S&T Pillar Program under Grants No.BE2011198, Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No. BM2003201, Transformation Fund for Agricultural S&T Achievements under Grants No. 2011GB2C100024 and Innovation Fund for Agricultural S&T in Jiangsu under Grants No. CX(11)3039.

## References

1. Wang, S., Wu, C.: Application of Context-aware and Personalized Recommendation to Implement an Adaptive Ubiquitous Learning System. Expert Systems with Applications 38(9), 10831–10838 (2011)
2. Cao, J., Wu, Z.: An Improved Protocol for Deadlock and Livelock Avoidance Resource Co-allocation in Network Computing. World Wide Web: Internet and Web Information Systems 13(3), 373–388 (2010)
3. Wu, D., Ke, Y., Yu, J., Yu, P., Chen, L.: Leadership Discovery when Data Correlatively Evolve. World Wide Web: Internet and Web Information Systems 14(1), 1–25 (2011)
4. Gröger, G., Kolbe, T.H., Czerwinski, A., Nagel, C.: OpenGIS City Geography Markup Language (CityGML) Encoding Standard Copyright. Open Geospatial Consortium Inc. (2008)
5. Melville, P., Mooney, R., Nagarajan, R.: Content-boosted Collaborative Filtering for Improved Recommendations. In: 8th National Conference on Artificial intelligence (AAAI 2002), pp. 187–192 (2002)

6. Koren, Y.: Factor in the Neighbors: Scalable and Accurate Collaborative Filtering. ACM Transactions on Knowledge Discovery from Data 4(1), 1–24 (2009)
7. Adomavicius, G., Sankaranarayanan, R., Sen, S., Tuzhilin, A.: Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. ACM Trans. on Information Systems 23(1), 103–145 (2005)
8. Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.: Recommending Friends and Locations based on Individual Location History. ACM Trans. on the Web 5(1), Article 5, pages 44 (2011)
9. Zheng, Y., Xie, X.: Learning Travel Recommendation from User-generated GPS Traces. ACM Trans. on Intelligent Systems and Technology 2(1), Article 2, pages 29 (2011)
10. Brown, P.J.: The Stick-e Document: a Framework for Creating Context-aware Applications. In: Proceedings of EP 1996, Palo Alto, pp. 259–272 (1996)
11. Rosemann, M., Recker, J.: Context-aware Process Design: Exploring the Extrinsic Drivers for Process Flexibility. In: 18th International Conference on Advanced Information Systems Engineering, Proceedings of Workshops and Doctoral Consortium, pp. 149–158 (2006)
12. Schmidt, A., Beigl, M., Gellersen, H.: There is more to Context than Location. Computers & Graphics Journal 23(6), 893–902 (1999)
13. Abowd, G.D., Atkeson, C.G., Hong, J., Long, S., Kooper, R., Pinkerton, M.: Cyberguide: a Mobile Context-aware Tour Guide. Wireless Networks 3(5), 421–433 (1997)
14. Yau, S.S., Karim, F.: An Adaptive Middleware for Context-Sensitive Communications for Real-Time Applications in Ubiquitous Computing Environments. Real-Time Systems 26(1), 29–61 (2004)
15. Adomavicius, G., Tuzhilin, A.: Context-aware Recommender Systems. In: Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys 2008), pp. 335–336 (2008)
16. Park, M.H., Hong, J.H., Cho, S.B.: Location-Based Recommendation System Using Bayesian User's Preference Model in Mobile Devices. In: Indulska, J., Ma, J., Yang, L.T., Ungerer, T., Cao, J. (eds.) UIC 2007. LNCS, vol. 4611, pp. 1130–1139. Springer, Heidelberg (2007)
17. Quercia, D., Lathia, N., Calabrese, F., Lorenzo, G.D., Crowcroft, J.: Recommending Social Events from Mobile Phone Location Data. In: IEEE ICDM, Sydney, Australia (2010)
18. Kolbe, T.H.: CityGML–3D Geospatial and Semantic Modeling of Urban Structures. Presentation on the GITA/OGC Emerging Technologies Summit in Washington, http://www.citygml.org/fileadmin/citygml/docs/CityGML_ETS4_2007-03-21.pdf (accessed July 2, 2011)

# Invariant Analysis for Ordering Constraints of Multi-view Business Process Model*

Jidong Ge[1,2], Haiyang Hu[1,3], Hao Hu[1], and Xianglin Fei[1]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China, 210093
[2] School of Software, Nanjing University, China, 210093
[3] College of Computer, Hangzhou Dianzi University, China, 310018
gjdnju@163.com

**Abstract.** Workflow technology is applied in the mobile computing environment to improve the efficiency of modern business collaboration. Business process modeling language is a core element in WFMS (Workflow Management Systems). Business process model includes multiple views: activity view, artifact view and role view. Based on the similarity between multi-view business process modeling and object Petri nets, this paper proposes the MOPN-WF-net model which is a multi-view business process model based on multi-object Petri nets. To ensure the soundness of MOPN-WF-net, we propose an approach to check ordering constraints for necessary conditions of soundness property with invariant analysis.

## 1 Introduction

Workflow technology is applied in the mobile computing environment to improve the efficiency of modern business collaboration. A critical success factor for current enterprises is how to catch the details of business process nicely and timely. An important trend of modern business process management is that the details of business process become more and more complex. As we know, WFMS can manage and monitor business process. Process modeling language is a core element in WFMS [2]. A complete business process model includes multiple views: activity view, artifact view and role view. We should provide a paradigm to model multi-view business process so that each view of the business process can be checked and monitored in time. According to the principle of "Separation of Concerns", by investigating the similarity between multi-view business process modeling and object Petri nets, this chapter proposes MOPN-WF-net which is a multi-view business process model based on object Petri nets and enhances the reusability and the flexibility of business process modeling. The model includes two-level models: system net and object net. As an example, a multi-view paradigm including activity views and artifact views is

* This work was supported by NSFC (61100039, 61021062, 60973044, 61073030, 61003019, and 61073031), the Fundamental Research Funds for the Central Universities, and the Fund of State Key Laboratory for Novel Software Technology.

H. Wang et al. (Eds.): APWeb 2012 Workshops, LNCS 7234, pp. 257–268, 2012.

provided. Activity views are described by system nets and artifact views are described by object nets.

To ensure the business process model can be completely and correctly executed by the WFMS engine, the soundness property [1] of the process model must be taken into account. Towards the soundness of MOPN-WF-net, this paper presents some ordering constraints as necessary conditions based on invariant analysis.

The paper is structured as follows. Section 2 applies MOPN (Multi-Object Petri Nets) to model multi-view business process, and define the MOPN-WF-net model (MOPN-based Business process net). Section 3 provides some necessary conditions about the ordering constraints for the soundness of MOPN-WF-net. Section 4 presents the approach to calculate ordering relations with invariant analysis, which can be used to checking the ordering constraints for soundness of MOPN-WF-net. Section 5 provides some case study of invariant analysis for ordering constraints of multi-view business process model. Finally, we conclude this paper.

## 2 MOPN-WF-Net for Multi-view Business Process Modeling

As Figure 1 shows, a complete business process model includes several views: activity view, artifact view, role view [4]. In order to fully understand and monitor the business process, business process modeling should contain multiple views of business process. The multi-view paradigm separates the activity view modeling and the artifact view modeling clearly, and combines the interaction relations between activity view and artifact view.

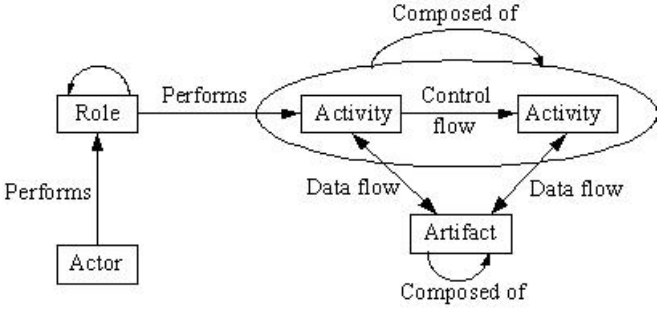



**Fig. 1.** Basic elements of business process model

According to the principle of "Separation of Concerns", the different views of business process should be modeled separately. In order to model multi-view of business process and describe the interaction relations between different views clearly, based on the similarity between multi-view business process modeling and object Petri nets, this paper proposes MOPN-WF-net which is a multi-view business process model based on object Petri nets. This paper applies the Valk's "Object Petri Nets" to model multi-view business process model [6, 7, 12]. From the structural

perspective, the object Petri net model includes two levels: system net and object net. There are interaction relation set between system net and object net.

As an example, combining the activity view and the artifact view, activity view is modeled by system net and the artifact view is modeled by object net. The tokens in the system net represent the behavior of some artifact views. The interaction set between the system net and the object net describes the interaction relations between the activity view and the artifact view. Based on the above idea, the activity view and the artifact view can modeled separately and clearly, in the meaning while, the interaction set can combine the activity view and the artifact view. The MOPN-WF-net includes multiple views of business process model, which is characterized with clearer hierarchy, simpler structure, and more extendibility. In general, based on the MOPN-WF-net model, the role view can be attached to activity view too. The basic concepts can be referred to [10, 11].

**Definition 1.** A Multi-Object Petri Net (MOPN) is a triple $MOPN = (SN, ON_S, \rho)$, $SN$ is the system net, $ON_S$ is the set of object nets, $\rho$ is the interaction set between $SN$ and $ON_S$.

(1) $SN = (P, T, F, M_0)$ is a Petri net;

(2) $ON_S = \{ON_1, ON_2, ..., ON_n\}$ is the finite set of object nets, $ON_k = (\hat{P}_k, \hat{T}_k, \hat{F}_k, \hat{M}_{0k})$ is a Petri net, which is the element of $ON_S$, on the other hand, it is the referred object of $SN$'s token labled with number $k$. $\hat{T} = \bigcup_{k=1}^{n} \hat{T}_k$, $\hat{P} = \bigcup_{k=1}^{n} \hat{P}_k$;

(3) $\rho_k \subseteq T \times \hat{T}_k$ is the interaction relation set between $SN$ and $ON_k$, $\rho = \bigcup_{k=1}^{n} \rho_k$. □

For applying the MOPN to model multi-view business process, we need to define MOPN-WF-net (Definition 2). As a paradigm (see Fig.2), combining the activity view and the artifact view, in the business process model, the activity view is on the core position, and several artifact views are attached to the activity view. This paradigm can improve the flexibility and the scalability of the business processes [9].

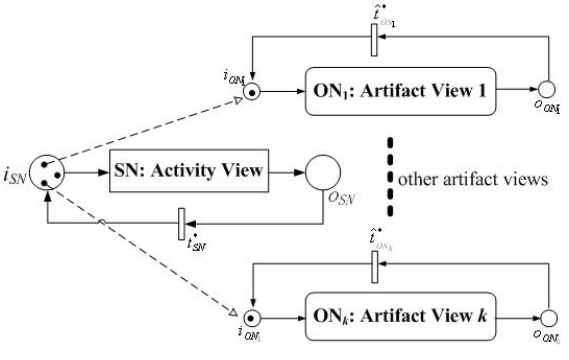



**Fig. 2.** A multi-view business process modeled by *MOPN-WF-net*

**Definition 2.** A $MOPN = (SN, ON_S, \rho)$ is *MOPN-WF-net* iff

(1) $SN = (P, T, F)$ is a WF-net, which describe the activity view of business process. *SN* has two special places: $i_{SN}$ and $o_{SN}$. $i_{SN}$ is the source place $\bullet i_{SN} = \phi$, and $o_{SN}$ is the sink place $o_{SN}\bullet = \phi$.

(2) $ON_S = \{ON_1, ON_2, ..., ON_n\}$ is the set of object nets. Here *n* is a finite natural number. $ON_k = (\hat{P}_k, \hat{T}_k, \hat{F}_k)$ is WF-net and describe the artifact view of business process. $ON_k$ has two special places: $i_{ON_k}$ and $o_{ON_k}$. $i_{ON_k}$ is the source place $\bullet i_{ON_k} = \phi$, and $o_{ON_k}$ is the sink place $o_{ON_k}\bullet = \phi$.

(3) $\rho_k \subseteq T \times \hat{T}_k$ describes the interaction relations between activity view *SN* and artifact view $ON_k$. The total set of interaction relations between *SN* and $ON_S$ is denoted by $\rho = \bigcup_{k=1}^{n} \rho_k$.

(4) $\rho_k(\hat{t}) = \{t \in T \mid (t, \hat{t}) \in \rho_k\}$ denotes the transition set in *SN* having interactions with $\hat{t}$ in $ON_k$. $\rho_k(t) = \{\hat{t} \in \hat{T}_k \mid (t, \hat{t}) \in \rho_k\}$ denotes the transition set in $ON_k$ having interactions with *t* in *SN*. □

The initial state of *MOPN-WF-net* is $\forall k \in \{1,...,n\} : ON_k.(M_0, \hat{M}_0) = (i_{SN}, i_{ON_k})$, so the total initial state of *MOPN* is $MOPN.Mark_0 = ((i_{SN}, i_{ON_1}), (i_{SN}, i_{ON_2}), ..., (i_{SN}, i_{ON_k}), ..., (i_{SN}, i_{ON_n}))$. The corresponding total final state of *MOPN* is $MOPN.final = ((o_{SN}, o_{ON_1}), (o_{SN}, o_{ON_2}), ..., (o_{SN}, o_{ON_k}), ..., (o_{SN}, o_{ON_n}))$.

In order to ensure that the business process model described by MOPN-WF-net can be completely and correctly executed by process engine in WFMS, soundness property of the process model should be considered. In WF-net, essentially, soundness property is the combination of two basic properties of Petri net: liveness and boundedness [1].

Because *MOPN-WF-net* is a multi-dimensional net and its firing rules are different from traditional Petri net, compared to the definition of soundness of WF-net and according to firing rules of object Petri nets [12], we define the soundness property of *MOPN-WF-net* as follows.

**Definition 3.** *MOPN-WF-net* is sound, if and only if:

(1) $\forall ON_k.(M, \hat{M}) : ON_k.(i_{SN}, i_{ON_k}) \xrightarrow{*} ON_k.(M, \hat{M}) \Rightarrow$
$ON_k.(M, \hat{M}) \xrightarrow{*} ON_k.(o_{SN}, o_{ON_k})$, $ON_k.(i_{SN}, i_{ON_k})$ is the initial state of $ON_k$, $ON_k.(o_{SN}, o_{ON_k})$ is the final state of $ON_k$.

(2) $\forall ON_k.(M, \hat{M}) : ON_k.(i_{SN}, i_{ON_k}) \xrightarrow{*} ON_k.(M, \hat{M}) \wedge$
$(ON_k.(M, \hat{M}) \geq ON_k.(o_{SN}, o_{ON_k})) \Rightarrow ON_k.(M, \hat{M}) = ON_k.(o_{SN}, o_{ON_k})$

(3) $\forall k \in \{1,...,n\}.\forall t \in T$,
a) if $\rho_k(t) = \phi$, then

$\exists ON_k.(M,\hat{M}), ON_k.(M',\hat{M}') : ON_k.(i_{SN}, i_{ON_k}) \xrightarrow{*}$

$ON_k.(M,\hat{M}) \xrightarrow{[t,\lambda]} ON_k.(M',\hat{M}')$;

b) if $\rho_k(t) \neq \phi$, then

$\forall \hat{t} \in \rho_k(t). \exists ON_k.(M,\hat{M}), ON_k.(M',\hat{M}') : ON_k.(i_{SN}, i_{ON_k}) \xrightarrow{*}$

$ON_k.(M,\hat{M}) \xrightarrow{[t,\hat{t}]} ON_k.(M',\hat{M}')$.

(4) $\forall k \in \{1,...,n\}. \forall \hat{t} \in \hat{T}_k$,

a) if $\rho_k(\hat{t}) = \phi$, then

$\exists ON_k.(M,\hat{M}), ON_k.(M',\hat{M}') : ON_k.(i_{SN}, i_{ON_k}) \xrightarrow{*}$

$ON_k.(M,\hat{M}) \xrightarrow{[\lambda,\hat{t}]} ON_k.(M',\hat{M}')$;

b) if $\rho_k(\hat{t}) \neq \phi$, then

$\forall t \in \rho_k(\hat{t}). \exists ON_k.(M,\hat{M}), ON_k.(M',\hat{M}') : ON_k.(i_{SN}, i_{ON_k}) \xrightarrow{*}$

$ON_k.(M,\hat{M}) \xrightarrow{[t,\hat{t}]} ON_k.(M',\hat{M}')$. □

## 3 Ordering Constraints for Soundness of MOPN-WF-Net

To ensure the soundness of MOPN-WF-net, we will provide some necessary conditions about ordering constraints for soundness of MOPN. We firstly introduce the ordering relations [3] between two transitions, and then provide some necessary conditions about ordering constraints for soundness of MOPN.

Let $PN = (P,T,F)$ be a workflow net, $T$ be transition set, $\sigma \in T^*$ be a process firing sequence and $W = \{\sigma \mid [i] \xrightarrow{\sigma} [o]\}$ be the set of process firing sequences. For example, $a_1a_2a_4a_5a_7a_8$ is a process firing sequence.

**Definition 4.** (Ordering Relations). Let $PN = (P,T,F)$ be a circuit-free Petri net, $a_1, a_2 \in T$ are two transitions in *PN*. $\sigma = \sigma_1 a_1 \sigma_2 a_2 \sigma_3$ and $\sigma' = \sigma_1 a_2 \sigma_2 a_1 \sigma_3$ are two possible process firing sequences in *PN* such that $[i] \xrightarrow{\sigma} [o]$. Here, $\sigma_1$, $\sigma_2$ and $\sigma_3$ are three subsequences and allowed to be empty sequence $\varepsilon$.

(1) $a_1 \parallel a_2$ iff there exist two sequences both $\sigma_1 a_1 \sigma_2 a_2 \sigma_3$ and $\sigma_1 a_2 \sigma_2 a_1 \sigma_3$. $a_1 \parallel a_2$ means that the occurring ordering relation between $a_1$ and $a_2$ is non-determined, i.e. the ordering relation between $a_1$ and $a_2$ is parallel. Obviously $a_1 \parallel a_2 \Leftrightarrow a_2 \parallel a_1$.

(2) $a_1 \mapsto a_2$ iff there exists a sequence $\sigma_1 a_1 \sigma_2 a_2 \sigma_3$ not $\sigma_1 a_2 \sigma_2 a_1 \sigma_3$. $a_1 \mapsto a_2$ means that $a_1$ must always occur before $a_2$ in any sequence containing both $a_1$ and $a_2$.

(3) $a_1 \# a_2$ iff $\neg(a_1 \mapsto a_2) \wedge \neg(a_2 \mapsto a_1) \wedge \neg(a_1 \parallel a_2)$. $a_1 \# a_2$ means that there exists no sequence containing $a_1$ and $a_2$ simultaneously. Obviously $a_1 \# a_2 \Leftrightarrow a_2 \# a_1$. □

$a_1 \parallel a_2$, $a_1 \mapsto a_2$ and $a_1 \# a_2$ are three basic ordering relations between two transitions $a_1$ and $a_2$. The ordering relations can be generated by special algorithms. For the soundness of MOPN, there are some ordering constraints between transitions under the interaction relations (in Theorem 2, 3 & 4).

**Definition 5.** Projection Sub-Sequence (PSS) in MOPN

Let $MOPN = (SN, ON_S, \rho)$, $ON = (P, T, F)$, $ON_S = \{ON_1, ON_2, ..., ON_n\}$, $ON_k = (\hat{P}_k, \hat{T}_k, \hat{F}_k)$. Suppose $\eta = v_1 v_2 ... v_i ... v_n$ is a sequence for $ON_k$ such that $ON_k.(M_1, \hat{M}_1) \xrightarrow{\eta} ON_k.(M_2, \hat{M}_2)$, there exist two subsequences $\delta = x_1 x_2 ... x_i ... x_n$ related to the system-level state changes, and $\hat{\delta} = y_1 y_2 ... y_i ... y_n$ related to the object-level state changes. These two subsequences are derived from $\eta = v_1 v_2 ... v_i ... v_n$ by selecting the actions and composing the projection subsequence over $ON_k$ or *SN*. It is according to the following projection rules.

$$x_i = \begin{cases} x_i = a & \text{if } v_i = [a, \hat{a}] \text{ or } v_i = [a, \lambda] \\ x_i = \varepsilon & \text{if } v_i = [\lambda, \hat{a}] \end{cases} \qquad y_i = \begin{cases} y_i = \hat{a} & \text{if } v_i = [\lambda, \hat{a}] \text{ or } v_i = [a, \hat{a}] \\ y_i = \varepsilon & \text{if } v_i = [a, \lambda] \end{cases}$$

The subsequence $\delta$ is called the system-level projection subsequence of $\eta$. It is denoted by $\eta|_{SN}$. The subsequence $\hat{\delta}$ is called the object-level projection subsequence of $\eta$. It is denoted by $\eta|_{ON_k}$. Obviously, $ON_k.M_1 \xrightarrow{\delta} ON_k.M_2$ and $ON_k.\hat{M}_1 \xrightarrow{\hat{\delta}} ON_k.\hat{M}_2$. For example, if $\eta = [a_1, \lambda][a_2, \hat{a}_2][a_3, \hat{a}_3][\lambda, \hat{a}_4][a_5, \hat{a}_5]$, $(a_1, a_2, a_3, a_5 \in T) \wedge (\hat{a}_2, \hat{a}_3, \hat{a}_4, \hat{a}_5 \in \hat{T}_k)$, then $\delta = \eta|_{SN} = a_1 a_2 a_3 a_5$ and $\hat{\delta} = \eta|_{ON_k} = \hat{a}_2 \hat{a}_3 \hat{a}_4 \hat{a}_5$. □

**Theorem 1.** Let $MOPN = (SN, ON_S, \rho)$, $ON_S = \{ON_1, ON_2, ..., ON_n\}$. *SN* is a circuit-free workflow net, and each $ON_k$ is also a circuit-free Petri net. If *MOPN-WF-net* is sound, then *SN* is sound alone, and each $ON_k$ is also sound alone.

**Proof.** (Omitted) □

**Theorem 2.** Let $MOPN = (SN, ON_S, \rho)$, $ON_S = \{ON_1, ON_2, ..., ON_n\}$ be an *MOPN-WF-net*. *SN* and $\forall k \in \{1, ...n\}.ON_k$ are all circuit-free Petri nets. $SN = (P, T, F)$, $ON_k = (\hat{P}_k, \hat{T}_k, \hat{F}_k)$, $\rho_k \subseteq T \times \hat{T}_k$, $a, b \in T$, and $\hat{a}, \hat{b} \in \hat{T}_k$.

(1) Let $(a, \hat{a}) \in \rho_k$, $(b, \hat{b}) \in \rho_k$. If *MOPN-WF-net* is sound and $a \mapsto b$ in *SN*, then $(\hat{a} \mapsto \hat{b}) \vee (\hat{a} \parallel \hat{b})$.

(2) Let $(a, \hat{a}) \in \rho_k$, $(b, \hat{b}) \in \rho_k$. If *MOPN-WF-net* is sound and $\hat{a} \mapsto \hat{b}$ in $ON_k$, then $(a \mapsto b) \vee (a \parallel b)$.

**Proof.** (Omitted) □

**Theorem 3.** Let $MOPN = (SN, ON_S, \rho)$, $ON_S = \{ON_1, ON_2, ..., ON_n\}$ be an *MOPN-WF-net*. *SN* and $\forall k \in \{1, ...n\}.ON_k$ are all circuit-free Petri nets. $SN = (P, T, F)$, $ON_k = (\hat{P}_k, \hat{T}_k, \hat{F}_k)$, $\rho_k \subseteq T \times \hat{T}_k$, $a, b \in T$, and $\hat{a}, \hat{b} \in \hat{T}_k$.

(1) Let $(a,\hat{a}) \in \rho_k$ , $(b,\hat{b}) \in \rho_k$ . If *MOPN-WF-net* is sound and $a \parallel b$ in *SN*, then $(\hat{a} \parallel \hat{b}) \vee (\hat{a} \mapsto \hat{b}) \vee (\hat{b} \mapsto \hat{a})$ .

(2) Let $(a,\hat{a}) \in \rho_k$ , $(b,\hat{b}) \in \rho_k$ . If *MOPN-WF-net* is sound and $\hat{a} \parallel \hat{b}$ in $ON_k$, then $(a \parallel b) \vee (a \mapsto b) \vee (b \mapsto a)$ .

**Proof.** (Omitted) □

**Theorem 4.** Let $MOPN = (SN, ON_S, \rho)$ , $ON_S = \{ON_1, ON_2, ..., ON_n\}$ be an *MOPN-WF-net*. *SN* and $\forall k \in \{1,...n\}.ON_k$ are all circuit-free Petri nets. $SN = (P,T,F)$ , $ON_k = (\hat{P}_k, \hat{T}_k, \hat{F}_k)$ , $\rho_k \subseteq T \times \hat{T}_k$ , $a,b \in T$ , and $\hat{a},\hat{b} \in \hat{T}_k$ .

(1) Let $(a,\hat{a}) \in \rho_k$ , $(b,\hat{b}) \in \rho_k$ . If *MOPN-WF-net* is sound and $a \# b$ in *SN*, then $\hat{a} \# \hat{b}$ . $(a \# b) \wedge (\hat{a} \mapsto \hat{b})$ or $(a \# b) \wedge (\hat{b} \mapsto \hat{a})$ or $(a \# b) \wedge (\hat{a} \parallel \hat{b})$ result in non-sound in *MOPN-WF-net*.

(2) Let $(a,\hat{a}) \in \rho_k$ , $(b,\hat{b}) \in \rho_k$ . If *MOPN-WF-net* is sound and $\hat{a} \# \hat{b}$ in $ON_k$, then $a \# b$ .

**Proof.** (Omitted) □

## 4 Calculating Ordering Relations with Invariant Method

Incidence matrix is an important approach to represent the Petri nets with formal mathematics. There are two kinds of invariants: P-invariant (Place-invariant) and T-invariant (Transition-invariant). In this paper, we apply T-invariants. For the formal model of Workflow net, invariant method as a basic approach can be used in soundness verification.

**Definition 6.** Incidence matrix, place invariants, transition invariants

(1) A Petri net $PN = (P,T,F)$ can be represented by an incidence matrix $PN : (P \times T) \to \{-1,0,1\}$ , which is defined by

$$PN(p,t) = \begin{cases} -1 & if \quad (p,t) \in F \\ 0 & if \quad (p,t) \notin F \wedge (t,p) \notin F \ or \ (p,t) \in F \wedge (t,p) \in F \\ 1 & if \quad (t,p) \in F \end{cases}$$

(2) A T-invariant of a net $PN = (P,T,F)$ is a rational-valued solution of the equation $PN \cdot Y = 0$ . The solution set is denoted by $J = \{J_1, J_2, ..., J_n\}$ . In essence, a T-invariant $J_k$ is a T-vector, as a mapping $J_k : T \to \mathrm{Z}$ . A T-invariant $J_k$ is called semi-positive if $J_k \geq 0$ and $J_k \neq 0$ . A T-invariant $J_k$ is called positive if $\forall t \in T : J_k(t) > 0$ .

(3) Minimal invariants: A semi-positive P-invariant $I_k$ is minimal if no semi-positive P-invariant $I_x$ satisfies $I_x \subset I_k$ . A semi-positive T-invariant $J_k$ is minimal if no semi-positive T-invariant $J_x$ satisfies $J_x \subset J_k$ . Every semi-positive invariant is the sum of minimal invariants [5]. If a net has a positive

invariant, then every invariant is a linear combination of minimal invariants.

(4) Fundamental property of T-invariant: Let $(PN, M_0)$ be a system, and let $J_k$ be a T-invariant of *PN*, then the Parikh vector $\vec{\sigma}$ is a T-invariant iff $M \xrightarrow{\sigma} M$ (i.e., iff the occurrence of $\sigma$ reproduces the marking *M*).

(5) A T-invariant $J_k$ of a $(PN, M_0)$ is realizable iff: there exists an $M_n \in RS(M_0)$ and a firing sequence $M_0 \xrightarrow{t_1} M_1 \xrightarrow{t_2} \ldots \xrightarrow{t_n} M_n$ such that $\forall t \in T : J_k(t) = |\{x \mid 1 \le x \le n \wedge t_x = t\}|$. □

Compared to ordinary Petri net, workflow net has special structure restriction. The invariants of workflow net have some special characteristics and special meanings.

**Definition 7.** T-Invariants of Workflow net [8]

Let *PN* =(*P*, *T*, *F*) be a workflow net. *t** is additional transition to connect source place *i* and sink place *o*. *PN** =(*P*, *T*∪{*t**}, *F*∪{(*o*, *t**), (*t**, *i*)} is the extended workflow net of *PN*. $J_k$ is called LMST-invariant (Legal Minimal Semi-positive T-invariant), if $J_k(t^*) = 1 \wedge J_k \ge 0$ and is minimal T-invariants of *PN**. A LMST-invariant $J_k$ of *PN** means an actually sound execution, and there exists a firing sequence $(\sigma = u_1 u_2 \ldots u_n t^*) \wedge (u_x \in T)$. Corresponding to $J_k$ such that $[i] \xrightarrow{u_1} M_1 \xrightarrow{u_2} M_2 \longrightarrow \ldots \xrightarrow{u_{n-1}} M_{n-1} \xrightarrow{u_n} [o] \xrightarrow{t^*} [i]$. Let $\pi(\sigma)$ be a function to record the occurrence times of each transitions over the sequence, then $\pi(\sigma) = J_k$. $\pi(\sigma, t) = J_k(t)$ denotes the times of transition *t* fired in the sequence $\sigma$. In MOPN, $ON_k.J$ denotes LMST-invariant set of $ON_k$. □

**Proposition 1.** A WF-net *PN* is sound iff $(PN, [i])$ is live and bounded [1]. □

**Proposition 2.** Every well-formed net has a positive T-invariant [5]. □

**Theorem 4.** A sound workflow net has a positive T-invariant.

Let *PN* =(*P*, *T*, *F*) be a workflow net. If *PN* is 1-sound, then *PN** =(*P*, *T*∪{*t**}, *F*∪{(*o*, *t**), (*t**, *i*)} has a positive T-invariant. *PN** is covered by LMST-invariants.

**Proof.** According to Proposition 1, if a WF-net *PN* =(*P*, *T*, *F*) is sound, then (*PN**, [*i*]) is live and bounded. So in the Petri net *PN** =(*P*, *T*∪{*t**}, *F*∪{(*o*, *t**), (*t**, *i*)}, there exists a marking $M_0$ = [*i*] such that (*PN**, $M_0$) is live and bounded system. Then according to the definition of well-formed Petri net, we conclude *PN** is well-formed. So, according to Proposition 2 *PN** has a positive T-invariant. According to the basic concepts about LMST-invariants in Definition 6, *PN** is covered by LMST-invariants.

**Proposition 3.** Necessary condition for liveness [5]

If (*PN**, $M_0$) is a live system, then every semi-positive P-invariant $I_k$ of *PN* satisfies $I_k \cdot M_0 > 0$.

**Definition 8.** The Decomposition Based on LMST-invariants

Let $PN = (P, T, F)$ be a 1-sound workflow net. $PN^*=(P, T\cup\{t^*\}, F\cup\{(o, t^*), (t^*, i)\}$ is its extended workflow net, and $J_k$ is an LMST-invariant of $PN^*$, then the subnet decomposed from $J_k$ is denoted $PN|_{J_k} = (P_{J_k}, T_{J_k}, F_{J_k})$, where:

(1) $T_{J_k} = \|J_k\| \setminus \{t^*\}$,

(2) $P_{J_k} = \{p \in \bullet T_{J_k} \mid p \in P\} \cup \{p \in T_{J_k} \bullet \mid p \in P\}$,

(3) $F_{J_k} = \{(p,t) \mid p \in P_{J_k} \wedge t \in T_{J_k} \wedge (p,t) \in F\} \cup \{(t,p) \mid p \in P_{J_k} \wedge t \in T_{J_k} \wedge (t,p) \in F\}$. □

According to the above discussion, if a workflow net $PN$ is 1-sound, then $PN^*$ can be decomposed by T-invariants and P-invariants. Fig.3 shows a workflow net. So, there are two subnets decomposed by LMST-invariants shown in Fig.4. From the decomposition by LMST-invariants, we can see that the LMST-invariant of the workflow net means a particular execution branch of the workflow process model.

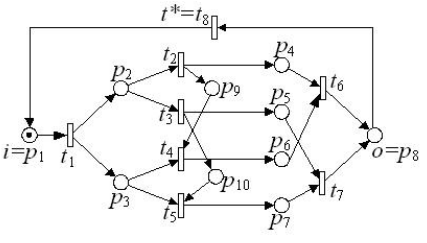



**Fig. 3.** An example: a workflow net

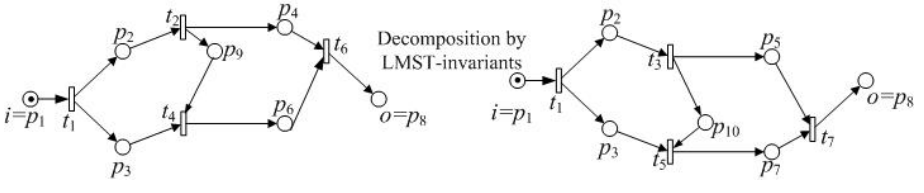



**Fig. 4.** The decomposition results by LMST-invariants from Fig. 3

We will propose the two theorems (Theorem 5 & 6) for calculating ordering relations with invariant method according to the above LMST-invariants definition and theory.

**Theorem 5**. Let $PN = (P, T, F)$ be a workflow net. $PN^*=(P, T\cup\{t^*\}, F\cup\{(o, t^*), (t^*, i)\}$ is its extended workflow net. Suppose that $J_k$ is an LMST-invariant of $PN^*$, there exist two transitions $a$ and $b$ such that $a, b \in \|J_k\|$.

(1) If there exists a directed path from transition $a$ to transition $b$, i.e., $(a, b) \in F^+$, then $a \mapsto b$ ;

(2) If there does not exist a directed path from transition $a$ to transition $b$, nor a directed path from transition $a$ to transition $b$, i.e., $(a, b) \notin F^+ \wedge (b, a) \notin F^+$, then $a \| b$.

**Proof.** Because the LMST-invariant of the workflow net means a particular execution branch of the workflow process model, the subnet decomposed from an LMST-invariant is a Marked Graph. When there are two transitions $a$ and $b$ belonging to the same LMST-invariant, there will be only two ordering relations: $a \mapsto b$ or $a \| b$.

(1) If there exists a directed path from transition $a$ to transition $b$, then according to the firing rules and the occurrence ordering in the firing sequence, for any firing sequence which includes both transition $a$ and transition $b$ together, transition $a$ must occurrence before transition $b$. So, $a \mapsto b$.

(2) If there exists no directed path from transition $a$ to transition $b$, i.e., it is not connected from transition $a$ to transition $b$. When transition $a$ and transition $b$ belong to the same concurrent branch for, the occurrence ordering between transition $a$ and transition $b$ is non-determined. So, $a \| b$. □

**Theorem 6.** Let $PN=(P, T, F)$ be a workflow net. $PN^*=(P, T \cup \{t^*\}, F \cup \{(o, t^*), (t^*, i)\})$ is its extended workflow net. Suppose that $J_k$ is an LMST-invariant of $PN^*$, there exist two transitions $a$ and $b$. If there exists no LMST-invariant including transition $a$ and transition $b$ together, i.e., $\neg(\exists J_k \in J : a \in \|J_k\| \wedge b \in \|J_k\|)$, then $a \# b$.

**Proof.** Because the LMST-invariant of the workflow net means a particular execution branch of the workflow process model, the subnet decomposed from an LMST-invariant is a Marked Graph. If in the LMST-invariant set of $PN^*$, there exists no LMST-invariant including both transition $a$ and transition $b$ together, i.e., $\neg(\exists J_k \in J : a \in \|J_k\| \wedge b \in \|J_k\|)$, then there does not exist a firing sequence including transition $a$ and transition $b$ together. So, according to the definition of ordering relations, we conclude $a \# b$. □

## 5 Case Study

In this section, we will provide several cases of checking ordering constraints with invariant analysis. In Fig.5(1), in $SN$, $|J_{S1}|=\{a_1, a_2, a_3\}$, in $ON_k$, $|J_{k1}|=\{\hat{b}_1, \hat{b}_2, \hat{b}_3\}$, $(a_1, \hat{b}_3), (a_3, \hat{b}_1) \in \rho$, $a_1 \mapsto a_3$, $\hat{b}_1 \# \hat{b}_3$, so, according to Theorem 2, $MOPN_1$ is not sound. In Fig.5(2), in $SN$, $|J_{S1}|=\{a_1, a_2, a_3\}$, in $ON_k$, $|J_{k1}|=\{\hat{b}_1, \hat{b}_2, \hat{b}_3\}$, $|J_{k2}|=\{\hat{b}_4, \hat{b}_5, \hat{b}_6\}$, $(a_1, \hat{b}_2), (a_3, \hat{b}_5) \in \rho$, $a_1 \mapsto a_3$, $\hat{b}_2 \# \hat{b}_5$, so, according to Theorem 2, $MOPN_2$ is not sound. In Fig.5(3), in $SN$, $|J_{S1}|=\{a_1, a_2, a_3, a_4\}$, in $ON_k$, $||J_{k1}|=\{\hat{b}_1, \hat{b}_2, \hat{b}_3\}$,

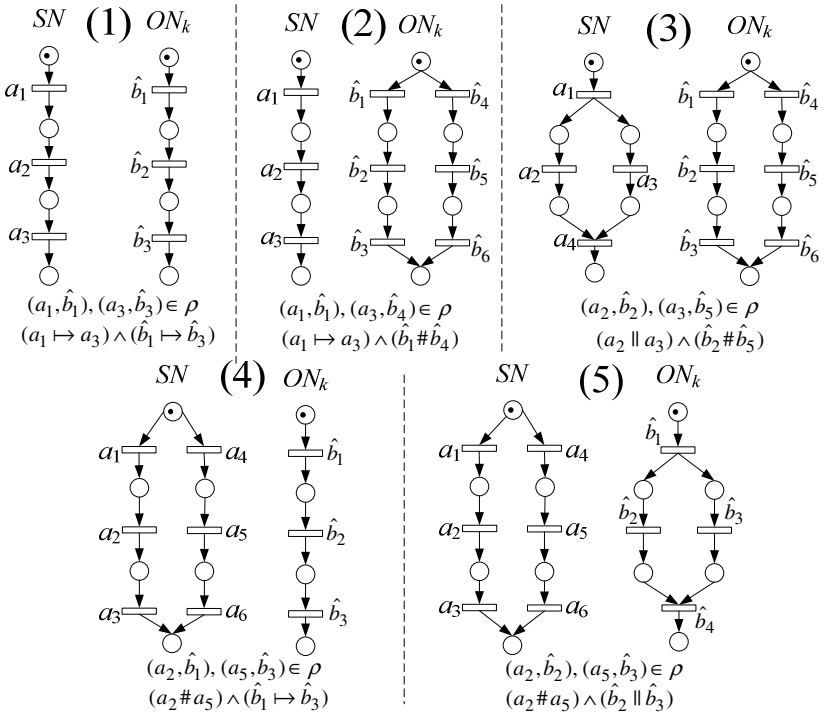



**Fig. 5.** Case study of checking ordering constraints in MOPN with invariant analysis

$\lfloor J_{k2}\rfloor=\{\hat{b}_4,\hat{b}_5,\hat{b}_6\}$, $(a_2,\hat{b}_2),(a_3,\hat{b}_5)\in\rho$, $a_2\|a_3$, $\hat{b}_2\#\hat{b}_5$, so, according to Theorem 3, $MOPN_3$ is not sound. In Fig.5(4), in $SN$, $\lfloor J_{S1}\rfloor=\{a_1, a_2, a_3\}$, $\lfloor J_{S2}\rfloor=\{a_4, a_5, a_6\}$, in $ON_k$, $\lfloor J_{k1}\rfloor=\{\hat{b}_1,\hat{b}_2,\hat{b}_3\}$, $(a_2,\hat{b}_1),(a_5,\hat{b}_3)\in\rho$, $(a_2\#a_5)\wedge(\hat{b}_1\mapsto\hat{b}_3)$, so, according to Theorem 4, $MOPN_4$ is not sound. In Fig.5(5), in $SN$, $\lfloor J_{S1}\rfloor=\{a_1, a_2, a_3\}$, $\lfloor J_{S2}\rfloor=\{a_4, a_5, a_6\}$, in $ON_k$, $\lfloor J_{k1}\rfloor=\{\hat{b}_1,\hat{b}_2,\hat{b}_3,\hat{b}_4\}$, $(a_2,\hat{b}_2),(a_5,\hat{b}_3)\in\rho$, $(a_2\#a_5)\wedge(\hat{b}_2\parallel\hat{b}_3)$, so, according to Theorem 4, $MOPN_5$ is not sound.

## 6 Conclusion

Workflow technology is applied in the mobile computing environment to improve the efficiency of modern business collaboration. WFMS can manage and monitor business process. Process modeling language is a core element in WFMS. A complete business process model includes multiple views: activity view, artifact view and role view, which should be considered in modeling business processes. In order to model multiple views of business process and describe the interaction relations between different views clearly, based on the similarity between multi-view business process modeling and object Petri nets, this paper proposes MOPN-WF-net which is a multi-view business process model based on object Petri nets and enhances the reusability

and the flexibility of business process modeling. The model includes two-level models: system net and object net. As an example, combining the activity view and the artifact view, the activity view is modeled by system net and the artifact view is modeled by object net. The tokens in the system net represent the behavior of certain artifact views. The interaction set between the system net and the object net describes the interaction between the activity view and the artifact view. MOPN-WF-net includes multiple views of business process model, which is characterized with clearer hierarchy, simpler structure, and more extendibility. Soundness property is important to ensure the process model can be executed correctly and completely. We provide the ordering constraints as necessary conditions for the soundness of MOPN-WF-net. What is more, we present the approach to calculate ordering relations with invariant analysis, which can be used to checking the ordering constraints for soundness of MOPN-WF-net.

## References

1. van der Aalst, W.M.P.: The Application of Petri Nets to Workflow Management. Journal of Circuits, Systems, and Computers 8(1), 21–66 (1998)
2. van der Aalst, W.M.P., van Hee, K.: Workflow Management – Models, Methods and Systems. MIT Press (2002)
3. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow mining: discovering process models from event logs. IEEE Transactions on Knowledge and Data Engineering 16(9), 1128–1142 (2004)
4. Acuna, S.T., Antonio, A.D., Ferre, X., Lopez, M., Mate, L.: The Software process: Modeling, Evaluation and Improvement. Handbook of Software Engineering and Knowledge Engineering, vol. 1, pp. 193–237. World Scientific Publishing Company (2001)
5. Desel, J., Esparza, J.: Free choice Petri nets. Cambridge University Press, Cambridge (1995)
6. Huang, L., Boehm, B., Hu, H., Ge, J., Lü, J., Qian, C.: Applying the Value/Petri Process to ERP Software Development in China. In: Osterweil, L.J., et al. (eds.) Proceeding of 28th International Conference of Software Engineering (ICSE 2006), pp. 502–511. ACM Press (2006)
7. Ge, J., Hu, H., Gu, Q., Lü, J.: Modeling Multi-View Business process with Object Petri Nets. In: Proceedings of International Conference on Software Engineering Advances (ICSEA 2006). IEEE Computer Society (2006)
8. Ge, J., Hu, H., Hu, H.: Calculating the Order Relations of Workflow Net with Invariant Method. In: Matsuo, T., Ishii, N., Lee, R. (eds.) Proceedings of the 9th International Conference on Computer and Information Science (ICIS 2010), pp. 714–719. IEEE Computer Society (2010)
9. Köhler, M., Moldt, D., Rölke, H.: Modelling Mobility and Mobile Agents Using Nets within Nets. In: van der Aalst, W.M.P., Best, E. (eds.) ICATPN 2003. LNCS, vol. 2679, pp. 121–139. Springer, Heidelberg (2003)
10. Murata, T.: Petri nets: properties, analysis and applications. Proceedings of the IEEE 77(4), 541–580 (1989)
11. Reisig, W.: An Introduction to Petri Nets. Springer, Heidelberg (1985)
12. Valk, R.: Petri Nets as Token Objects: An Introduction to Elementary Object Nets. In: Desel, J., Silva, M. (eds.) ICATPN 1998. LNCS, vol. 1420, pp. 1–25. Springer, Heidelberg (1998)

# Author Index